UNIVERSITY OF CALIFORNIA, SAN DIEGO

Expanding the toolbox of tandem mass spectrometry with algorithms to identify mass spectra from more than one peptide

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Jian Wang

Committee in charge:

Professor Philip E. Bourne, Chair
Professor Pieter C. Dorrestein, Co-Chair
Professor Ruben Abagyan
Professor Vineet Bafna
Professor Nuno Bandeira
Professor Pavel Pevzner

2013

The dissertation of Jian Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2013

DEDICATION

To my dearest grandmother Dr. Yuan Zhu Guo, thank you for all the sacrifices and dedications your made for your grandchildren. Without you, the next one hundred and thirty pages or so of this document will not have been possible. Although it saddens me greatly that you are not able to see this in person, I know you would have been very proud and will always be with us and watching us from above.

Your grandson

–Big

EPIGRAPH

"There are only two mistakes one can make along the road to truth;
not going all the way, and not starting."
—Buddha

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thanks to whoever deserves credit for Blacks Beach, Porters Pub, and all the coffee and tea shops in San Diego. Thanks also to hottubs. Thanks to my advisors Phil and Nuno for your guidances and mentoring all these years. Thanks to my experimental collaborators for performing some kick-ass experiments, so I can do what I do. Thanks to all my families for having a special place that I can always go to no matter what. Thanks to all my friends for making life more colorful in San Diego. Thanks to all my fellow bioinformaticians at UCSD or afar for accompanying me through grad school and all the stimulating discussions about science and life we had over the years.

Chapter 1, in part, is a reprint of the material as it appears in J. Wang, J. Perez-Santiago, J.E. Katz, P. Mallick, and N. Bandeira. "Peptide identification from mixture tandem mass spectra", Molecular & Cellular Proteomics, 9(7):14768, 201 and J. Wang, P.E. Bourne, and N. Bandeira. "Peptide identification by database search of mixture tandem mass spectra", Molecular & Cellular Proteomics, 10(12), 2011. It is also, in part, has been submitted for publication of the material as it appear in J. Wang, P.E. Bourne, and N. Bandeira. "Spectral probabilities for mixture tandem mass spectra of more than one peptides". The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, has been submitted for publication of the material as it appear in J. Wang, VG, Anania, J. Knott, J. Rush, J.R. Lill, P.E. Bourne, N. Bandeira. "A turn-key approach for large-scale identification of complex post-translational modifications", Journal of Proteome Research. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full is being prepared for submission for publication of the material. J. Wang, VG, Anania, J. Knott, J. Rush, J.R. Lill, P.E. Bourne, N. Bandeira. "Approach for large-scale identification of linked peptides from tandem mass spectrometry". The dissertation author was the primary investigator and author of this material.

VITA

2001-2005	Bachelor of Science in Bioengineering, University of California, Berkeley
2005-2006	Engineering Research Associate, Lawrence Berkeley National Laboratory
2006-2013	Doctor of Philosophy in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

J. Wang, VG, Anania, J. Knott, J. Rush, J.R. Lill, P.E. Bourne, N. Bandeira. Approach for large-scale identification of linked peptides from tandem mass spectrometry (in preparation)

J. Wang, VG, Anania, J. Knott, J. Rush, J.R. Lill, P.E. Bourne, N. Bandeira. A turn-key approach for large-scale identification of complex post-translational modifications (submitted)

J. Wang, P.E. Bourne, and N. Bandeira. Spectral probabilities for mixture tandem mass spectra of more than one peptides (submitted)

J. Wang, P.E. Bourne, and N. Bandeira. Peptide identification by database search of mixture tandem mass spectra. Molecular & Cellular Proteomics, 10(12), 2011.

J. Wang, J. Perez-Santiago, J.E. Katz, P. Mallick, and N. Bandeira. Peptide identification from mixture tandem mass spectra. Molecular & Cellular Proteomics, 9(7):14768, 201

L. Xie, J. Wang, and P. E. Bourne. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. PLoS Comput Biol, 3(11), Nov 2007.

N. Gupta, J. Benhamida, V. Bhargava, D. Goodman, E. Kain, I. Kerman, N. Nguyen, N. Ollikainen, J. Rodriguez, J. Wang, M. S. Lipton, M. Romine, V. Bafna, R. D. Smith, and P. A. Pevzner. Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. Genome Res, 18(7):1133 C 1142,Jul 200

ABSTRACT OF THE DISSERTATION

Expanding the toolbox of tandem mass spectrometry with algorithms to identify mass spectra from more than one peptide

by

Jian Wang

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2013

Professor Philip E. Bourne, Chair
Professor Pieter C. Dorrestein, Co-Chair

In high-throughput proteomics the development of computational methods and novel experimental strategies often rely on each other. In several areas, mass spectrometry methods for data acquisition are ahead of computational methods to interpret the resulting tandem mass (MS/MS) spectra. While there are numerous situations where two or more peptides are co-fragmented in the same MS/MS spectrum, nearly all mainstream computational approaches still make the ubiquitous assumption that each MS/MS spectrum comes from only one peptide. In this thesis we addressed problems in three emerging areas where computational tools that relax the above assumption are crucial for the success application of these

approaches on a large-scale. In the first chapter we describe algorithms for the identification of *mixture* spectra that are from more than one co-eluting peptide precursors. The ability to interpret mixture spectra not only improves peptide identification in traditional data-dependent-acquisition (DDA) workflows but is also crucial for the success application of emerging data-independent-acquisition (DIA) techniques that have the potential to greatly improve the throughput of peptide identification. In chapter two, we address the problem of identification of peptides with complex post-translational modification (PTM). Detection of PTMs is important to understand the functional dynamics of proteins. Complex PTMs resulted from the conjugation of another macromolecule onto the substrate protein. The resultant modified peptides not only generate spectrum that contains a mixture of fragment ions from both the PTM and the substrate peptide but they also display substantially different fragmentation patterns as compared to conventional, unmodified peptides. We describe a hybrid experimental and computational approach to build search tools that capture the specific fragmentation patterns of modified peptides. Finally in chapter three we address the problem of identification of linked peptides. Linked peptides are two peptides that are covalently linked together. The generation and identification of linked peptides has recently been demonstrated to be a versatile tool to study protein-protein interactions and protein structures, however the identification of linked peptides face many challenges. We integrate lessons learned in the previous chapters to build an efficient and sensitive tool to identify linked peptides from MS/MS spectra.

Chapter 1

Identification of mixture spectra from co-eluting peptides

Over the past several years there have been substantial advances in the sensitivity of protein identification thanks to technological developments in chromatography and tandem MS. In shotgun proteomics, researchers can routinely identify thousands of proteins from complex biological samples in a single experiment [1, 2, 3]. But, despite this rapid progress, there are still challenging issues that remain unsolved [4, 5]. One such challenge is that in any high throughput MS/MS experiment only a fraction of MS/MS spectra can be identified by current computational methods. While there are many factors contributing to this low spectrum identification rate, recent studies suggest that one reason is the occurrence of co-eluting peptides. As instruments with high mass accuracy enable us to better distinguish peptides with close precursor masses, it was recently shown that as many as 50% of the MS/MS spectra collected in typical proteomics experiments comes from more than one peptide precursor [6, 7, 8, 9], giving rise to *multiplexed* or *mixture* MS/MS spectra. These spectra can confuse current computational methods because most mainstream approaches make the assumption that each MS/MS spectrum comes from a single peptide. Houel et. al. estimated that identification rates for mixture spectra can be as low as only half of those spectra from a single peptide [9]. Thus computational methods that can handle mixture spectra can readily improve our current ability to analyze MS/MS spectral data

in traditional experimental workflows.

More importantly, in recent years there have been numerous development in data-independent acquisition (DIA) technologies where multiple peptide precursors are intentionally selected for co-fragmentation in each MS/MS spectrum [10, 11, 12, 13, 14]. Recent improvements in instrumentation have made it possible to perform DIA with high speed and high resolution in both MS and MS/MS modes. With continuing improvements in sensitivity and dynamic ranges, these emerging technologies can addresses some of the major disadvantages of traditional data-dependent acquisition methods such as reproducibility of data and can potentially increase the throughput of peptide identification by 10–20-fold [13, 15].

However despite the growing importance and the enormous potential of mixture spectra, there is still a shortage of computational methods that can analyze mixture spectra. In their pioneering work Zhang et. al. described the first database search method for mixture spectra, ProbIDtree [16] and showed that it is possible to identify co-eluting peptides from single MS/MS spectra. However in that study not all peptides could be confidently identified and in most cases ProbIDtree only identified the most prominent peptide in each mixture spectrum. In addition, ProbIDtree's accuracy for identifying mixture spectra is relatively low compared to the accuracy of current database search methods in identifying single-peptide spectra[17]. Other methods approach the mixture spectra identification problem by reporting spectra with more than one significant single-peptide match and do not explicitly attempt to model the occurrence of fragment ions from more than one peptide in the same spectrum. False Discovery Rates (FDR) are also left unadjusted [16, 18, 19] and may result in higher than expected FDR for mixture spectra (e.g., when co-eluting peptides share a substantial number of fragment masses). Finally while much attention has been dedicated to computing the statistical significance of a PSMs for single-peptide spectra [20, 21], this fundamental question still remain essentially unanswered for mixture spectra. Below we discuss three approaches that address the challenges of identifying mixture spectra from two peptides. This relatively simple case accounts for a large fraction of mixture spectra present in traditional DDA workflows [8] and thus allows us to test and de-

velop algorithmic concepts using DDA data since data from DIA workflows is still not widely available in public repositories. We demonstrated even for this relative simple cases we are able improve data analysis in the traditional DDA workflow by identifying significantly more peptides compared to existing approaches.

1.1 M-SPLIT: spectral library search of mixture spectra

Spectral library search has been recognized as a more sensitive method (as compared to database search methods) for peptide identification from MS/MS spectra due to the fact that spectral libraries have more information about the actual fragmentation pattern of a particular peptide [22]. Given the inherent difficulties in identifying peptides from mixture spectra, it is expected that extending spectral library search to mixture spectra will significantly improve peptide identification rates. In order to alleviate the algorithmic bottlenecks, we describe a new approach, M-SPLIT (Mixture-Spectrum Partitioning using Library of Identified Tandem mass spectra), that is able to reliably and efficiently identify peptides from *mixture* spectra - spectra that are generated from a pair of peptides. In brief, a mixture spectrum is modeled as linear combination of two single-peptide spectra and peptide identification is done by searching against a spectral library. We show that efficient filtration and accurate branch-and-bound strategies can be used to avoid the huge computational cost of searching all possible pairs. Thus equipped, our approach is able to identify the correct matches by considering only a minuscule fraction of all possible matches.

Beyond potentially enhancing the identification capabilities of current tandem MS acquisition protocols, we argue that the availability of methods to reliably identify MS/MS spectra from mixtures of peptides could enable the collection of MS/MS data using accelerated chromatography setups to obtain the same or better peptide identification results in a fraction of the experimental time currently required for exhaustive peptide separation.

1.1.1 Problem formulation

A mixture spectrum is defined as an MS/MS spectrum from two different peptides and a spectral library is a collection of identified MS/MS spectra. Analogous to the identification of MS/MS spectra by comparison against a database of known protein sequences, our goal is to identify mixture spectra by comparison against a spectral library. More formally, we modeled a mixture spectrum M as $M = A + \alpha B$, where A and B are MS/MS spectra from two different peptides and α, the mixture coefficient, indicates their relative abundance. Without loss of generality, we assume that A and B are scaled to Euclidean norm 1 and that $0 \leq \alpha \leq 1$ (i.e., A always corresponds to the higher abundance peptide). We can now formulate the following computational problem:

Mixture Spectrum Identification Problem (MSIP)
Input A putative mixture spectrum M and a spectral library $\mathcal{L}$.
Output A constant $0 \leq \alpha \leq 1$ and pair of spectra $A, B \in \mathcal{L}$, maximizing $similarity(M, A + \alpha B)$

While there are several ways to define similarity between two peptide spectra [23, 24, 25, 26], the *normalized dot product (ndp)* or cosine[1] measure of spectral similarity is widely accepted to be robust and makes no special assumptions concerning peptide mass spectra [26]. Moreover, as we show below, cosine similarity has a number of useful mathematical properties that allow us to derive theoretical bounds to guide our approach.

1.1.2 Simulation of mixture spectra

Since there is currently no publicly available data with validated identifications of mixture MS/MS spectra, we created a dataset of simulated mixture spectra to develop and benchmark our approach. To this end, we used the human MS/MS spectral library from the National Institute of Standards and Technology (*ver. 6/06*) and grouped the spectra according to their identified peptide. This

[1]Since all of spectra were scaled down to norm 1, *ndp* simply reduces to estimating the cosine between two unit vectors. Also, we reduce the disproportionate influence of high intensity peaks by first applying the square-root transform to all peak intensities [27]

resulted in 27,966 groups in the library, each containing two or more spectra belonging to the same peptide. The spectral library was then divided into two sets: i) a set $\mathcal{X}$, which has exactly one spectrum per peptide, used to create the simulated mixture spectra and ii) a spectral library $\mathcal{L}$ containing all the remaining spectra, used for searching. All spectra in the library are first scaled to norm 1 and since in a mixture the two peptides will most likely be present at different abundances, mixture spectra were created by randomly selecting two spectra $A_{\mathcal{X}}$ and $B_{\mathcal{X}}$ from $\mathcal{X}$ and linearly combining them using a predefined mixture coefficient α. In other words, a mixture spectrum is of the form $M = A_{\mathcal{X}} + \alpha B_{\mathcal{X}}$, where M represents a simulated mixture spectrum and $A_{\mathcal{X}}$ and $B_{\mathcal{X}}$ represent two single-peptide spectra, $0 \leq \alpha \leq 1$. Below we benchmark our approach for $\alpha \in \{0.1, 0.2, 0.5, 1\}$.

1.1.3 Overall algorithm

While the MSIP formulation is simple, the rapidly growing size of target spectral libraries (already on the order of 10^5-10^6 spectra) makes searching all possible *pairs* of spectra a prohibitive approach (10^{11} comparisons per query spectrum). We note that while one can pre-filter the target spectral library to consider only combinations of spectra with the same precursor mass as the query spectrum , such an approach would currently not provide a realistic estimate of performance on quickly growing proteome-scale spectral libraries. By not enforcing any parent mass filters on our performance estimates, we argue that the approach proposed here should seamlessly scale to much larger spectral libraries and be directly applicable to complex searches (e.g., metaproteomics studies). We propose two ways to avoid the quadratic penalty of searching all pairs: first we use an efficient *projected-cosine* filter to eliminate a large fraction of spectra in the library. After filtering, we use a branch-and-bound search strategy to find the best-matching *pairs* by considering only a subset of all possible pairs. The overall strategy is detailed in pseudocode below.

Input : Mixture Spectrum M, Spectral library $\mathcal{L}$
Output: A pair of spectra: $A^*,\ B^* \in \mathcal{L}$ and α^* such that `cosine` $(M, A^* + \alpha^* B^*)$ is maximized

Filter the library $\mathcal{L}$ by retaining top K candidate spectra with highest projected-cosine to M and create a filtered library $\mathcal{L}^{'}$
Sort the filtered library according to `cosine(`M, S`)`, $S \in \mathcal{L}^{'}$
$BestScore = 0$
for i =1 to `Size(`$\mathcal{L}^{'}$`)` **do**
  $A = i^{th}$ *spectrum in* $\mathcal{L}^{'}$
  for j = i+1 *to* `Size(`$\mathcal{L}^{'}$`)` **do**
    $B = j^{th}$ *spectrum in* $\mathcal{L}^{'}$
    if `upperBound(`M, A, B`)` $< BestScore$ **then**
      break
    else
      $\alpha =$ `estimateAlpha(`M, A, B`)`
      $score =$ `cosine`$(M, A + \alpha B)$
      if $score \geq BestScore$ **then**
        $BestScore = score,\ A^* = A,\ B^* = B,\ \alpha^* = \alpha$
      end
    end
  end
end

1.1.4 Filtering with projected-cosine

While cosine is generally a good measure of spectrum similarity, a mixture spectrum M derived from peptides A and B may have limited similarity to the corresponding single-peptide spectra - e.g, the presence of B in the mixture results in many unmatched peaks between M and A. We address this with a *projected-cosine* similarity, a modified cosine function that only considers coordinates in M if the corresponding coordinate in A is not zero. More precisely for two vectors A and M, the projection of M on A ($M_{p(A)}$) is defined as:

$$M_{p(A)}[i] = \begin{cases} M[i] & \text{if } A[i] > 0 \\ 0 & \text{otherwise} \end{cases}$$

The projected cosine between M and A is then simply the cosine of the $M_{p(A)}$ and A:

$$\cos_p(M, A) = \frac{M_{p(A)} \cdot A}{\left\|M_{p(A)}\right\| \|A\|}$$

Given a spectrum M, the filtering step consists of computing the projected-cosine similarity between M and all spectra in $\mathcal{L}$ and retaining the top most similar matches. The filtering efficiency of projected-cosine similarity is determined by the highest (i.e., worst) rank of a correct match of M to the library $\mathcal{L}$. Note that a correct match in $\mathcal{L}$ has the same peptide as M - single-peptide spectra have one correct match, mixture spectra have two correct matches. As shown in Figure 1.1a, the resulting ranks of correct matches indicate that projected-cosine is an efficient filter that, in most cases, retains the correct matches at ranks less than 500 in a library of about 27,966 spectra. In fact, for 95% of cases the correct pair of peptides in a mixture spectrum M can be identified by considering only the top 100 library spectra with highest projected cosine similarity to M (for $\alpha \geq 0.2$).

1.1.5 Searching with Branch-and-Bound

To better describe the concepts behind the branch and bound search strategy, let us assume for the moment that a mixture spectrum M is obtained from two single-peptide spectra with same abundance(i.e. $\alpha = 1$; see Supplementary

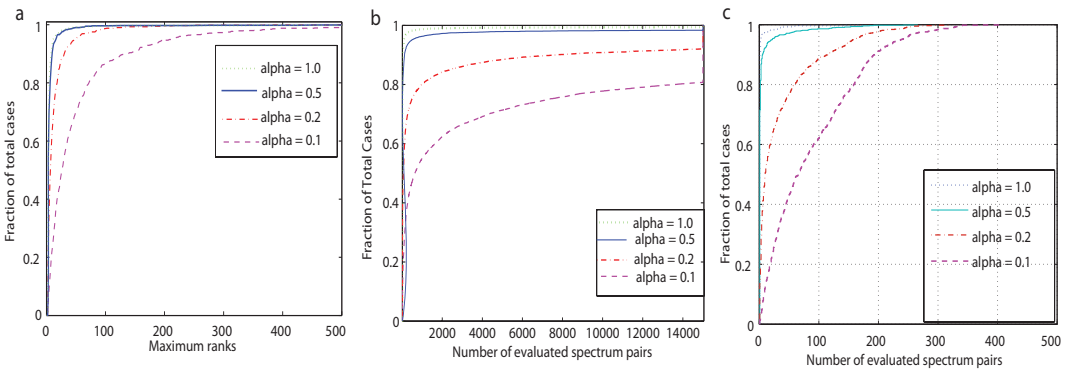

Figure 1.1: Effectiveness of filtering and branch-and-bound strategy: **a)** Cumulative distributions of maximum rank for correct matches to the spectral library. Spectra in the library are first sorted according to decreasing projected-cosine similarity to the mixture spectrum (library containing 27,966 spectra). The rank of correct matches are then determined. Correct matches are spectra identified as one of the peptides in the mixture. Since each mixture spectrum has two correct matches (it is generated from two peptides) we take the maximum (i.e., worst) rank of the two matches. **b)** Effectiveness of the branch and bound strategy. To avoid considering all *pairs* of spectra in the library, we derive a branch-and-bound search strategy to eliminate a large fraction of all possible pairs. The number of evaluated pairs of spectra is shown. Since the total number of possible pairs is 3.9×10^8 and our approach never evaluates more than 15,000 pairs, this self-adjusting strategy achieves speedups of at least 2×10^4. **c)** Combining the projected-cosine filter (part a) with branch-and-bound search (part b): we first filter the spectral library with projected-cosine and retain only the top 500 candidates; the branch-and-bound search strategy is then applied to further reduce number of pairs of spectra that needed to be evaluated. The curves for $\alpha = 0.1$ clearly shows that projected-cosine is a effective pruning filter - note that pre-filtering the library with cosine results in the evaluation of more pairs of spectra. The combined filters typically achieve speedups of approximately six orders of magnitude ($\approx \frac{3.9 \times 10^8}{500} = 7.8 \times 10^5$ speedups

materials for analysis when $\alpha < 1$). Therefore, for any pair of spectra (A, B) we

have the following relation for our objective function:

$$\begin{aligned}
\cos(M, A+B) &= \frac{M \cdot (A+B)}{\|M\| \, \|A+B\|} \\
&= \frac{M \cdot A + M \cdot B}{\sqrt{A \cdot A + B \cdot B + 2A \cdot B}} \\
&= \frac{M \cdot A + M \cdot B}{\sqrt{2 + 2A \cdot B}} \\
&\leq \frac{M \cdot A + M \cdot B}{\sqrt{2}} \\
\text{thus we define:} \quad upperBound(M, A, B) &= \frac{M \cdot A + M \cdot B}{\sqrt{2}}
\end{aligned}$$

Assume that at certain a stage of our search the best solution we've seen so far is: $A^* + B^*$, and without loss of generality let us also assume $cos(M, A^*) \geq cos(M, B^*)$. By the above equations, we do not need to pair A^* with any spectrum C such that $upperBound(M, A^*, C) < cos(M, A^* + B^*)$. This is because $upperBound(M, A^*, C)$ is never less than $cos(M, A^* + C)$. Moreover, a spectrum D with $cos(M, D) \leq cos(M, C)$ necessarily implies that $upperBound(M, A^*+D) \leq upperBound(M, A^*+C)$ - thus implying that the pair (A, D) can be excluded from consideration. This leads to the following search strategy: 1) sort spectra in the library according to their cosine similarity to the query spectrum M; 2) set A to the spectrum with highest $cos(M, A)$ in the library; 3) pair A with remaining spectra $C \in \mathcal{L}$ until we find a spectrum that has $upperBound(M, A, C) < cos(M, A^*+B^*)$; 4) delete A from the library, and repeat from step 2.

We determine the efficiency of this method by counting the number of *pairs* that are evaluated before the algorithm terminates with the optimal answer. As shown in Figure 1.1b, in most cases we only consider hundreds to thousands of combinations, approximately five order of magnitude less than the total number of possible pairs ($\approx 3.9 \times 10^8$). To take advantage of both the projected-cosine filter and the branch-and-bound strategy, we first filter the library with projected-cosine to retain only the top five hundred candidates and then apply the branch-and-bound strategy to limit the number of evaluated pairs. As shown in Figure 1.1c, only a few hundred *pairs* of spectra need to considered before M-SPLIT finds the

optimal answer. We also note that projected-cosine is a better filter than cosine - as shown in Figure 1.1c for $\alpha = 0.1$, pre-filtering the library with cosine results in more pairs of spectra being matched to each query spectrum(yellow line).

1.1.6 Estimating the mixture coefficient α

When trying to identify a mixture spectrum $M = A + \alpha B$, the mixture coefficient α is generally not known in advance. Since an incorrect α will distort the cosine similarity between M and its correct library matches, it is important to estimate it correctly. To distinguish the true and estimated values of α, we denote the estimated mixture coefficient as $\hat{\alpha}$ and compare two methods to compute $\hat{\alpha}$. In the *residual-spectrum* approach, we first identify the dominant component in the mixture (A) and construct a residual spectrum R by removing from the mixture spectrum all common peaks between A and M. It can be shown that $\hat{\alpha}$ is directly related to the magnitude of the residual spectrum ($\|R\|$) and can be estimated by solving the following equation:

$$\hat{\alpha} \quad = \frac{\|R\|^2}{1 - \|R\|^2}$$

In the *optimal-cosine* approach $\hat{\alpha}$ is chosen to maximize the cosine similarity between M and $A + \hat{\alpha}B$. By taking the derivative of the cosine similarity function with respect to $\hat{\alpha}$, setting it to zero and solving for $\hat{\alpha}$ (

$$\hat{\alpha} = \frac{M \cdot B - (M \cdot A)(A \cdot B)}{M \cdot A - (A \cdot B)(M \cdot B)}$$

The performance of both methods is shown in Table 1.1. While the performance of the residual-spectrum method is reasonable when α is large, the error becomes quite substantial when α is small. By contrast, the optimal-cosine method is robust in the presence of noise and delivers comparable performance across different values of α.

Table 1.1: Mean and standard deviation of the log-2 ratios of estimated ($\hat{\alpha}$) and true (α) mixture coefficients. While both approaches are roughly equivalent when $\alpha \geq 0.5$, optimal-cosine estimation performs substantially better on the more difficult cases of smaller mixture coefficients.

	Residual-spectrum approach		Optimal-cosine approach	
True α	mean	standard deviation	mean	standard deviation
1.0	-0.1312	0.4278	-0.0393	0.4646
0.5	0.051	0.4449	-0.0103	0.4770
0.2	0.4592	0.6767	0.0816	0.5264
0.1	1.0139	0.9021	-0.0014	0.5317

1.1.7 Classification of spectral library matches

As with regular database search of MS/MS spectra from isolated peptides, a spectral library search will always identify some top-scoring pair for any given query. To assess whether a match is significant we consider three possible outcomes when searching a given query spectrum S:

- No-match: S does not match any spectrum in the library
- Single-peptide match: S matches one peptide in the library.
- Mixture match: S is identified as a pair of peptides in the library.

Let $A^* + \hat{\alpha}\ B^*$ be the best pair of spectra in the library returned by M-SPLIT; we distinguish between the possible outcomes using P and Δ defined as follows:

$$\begin{aligned} P &= Max(cos(S, A^*),\ cos(S, B^*)) \\ \Delta &= cos(S, A^* + \hat{\alpha}B^*) - P \end{aligned}$$

Intuitively, if S is from a peptide not present in the library, both A^* and B^* should have low cosine similarity to S. It follows that P should be low in the No-match case but relatively high in the other two cases. Also, in Mixture matches the term B^* should increase the similarity to S by a significant amount, as determined by Δ. We thus determine the outcome of a particular match by a simple two-step process: a) a match is classified as No-match if P is below a certain threshold; b) distinguish Single-peptide and Mixture matches by checking whether Δ is below or above a chosen threshold, respectively.

To determine the actual threshold used in this process, we constructed two negative control datasets. One consists of five thousand mixture spectra (with $\alpha = 1.0$) where the peptides used to create the mixture spectra are deleted from the library. The second dataset consists of five thousand single-peptide spectra. These two datasets were combined with another mixture dataset and searched against the library for the best pairs of matches. As shown in Figure 1.2, when the peptides are not present in the library (No-match case) P has relatively low values (green dots) and can thus distinguish these from Single or Mixture-match cases by placing a threshold on P (see Figure 1.3a for Precision/Recall curves).; In distinguishing Single-peptide from Mixture matches, Figure 1.2 shows that Δ is higher for Mixture matches than for Single-peptide matches. However, Figure 1.2 also shows that this threshold depends on α. To build a general model, we first choose the threshold for cases where $\alpha \in \{0.1, 0.2, 0.5, 1.0\}$ and use linear regression to obtain the relationship between Δ and α. During our experiments we also found low-complexity spectra (i.e. spectra dominate by only a few peaks) can lead to artificially high P or Δ, we computed a measure similar to dot-bias used in [28] and use this to filter out any significant matches that may due to low-complexity spectra.

1.1.8 Identification of simulated mixture spectra

Our running hypothesis is that a mixture spectrum can be identified by matching it to a linear combination of single-peptide spectra. To test this hypothesis, we simulated a series of different mixture spectra (as described in Methods) and verified if the resulting Mixture matches correctly identified the peptides used to construct each simulated mixture. As shown in Table 1.2, the performance of our approach varies with α but is able to select the correct peptides in $90 - 99\%$ of all cases. As expected, as α decreases it becomes more difficult to identify *both* peptides in the mixture spectra because the signal-to-noise ratio substantially decreases for the low-abundance peptide. Also, the accuracy decreases faster at a ratio of 1:0.1, suggesting this may be the lowest α that can be handled without substantially decreasing sensitivity. Of course, high MS/MS mass accuracy should

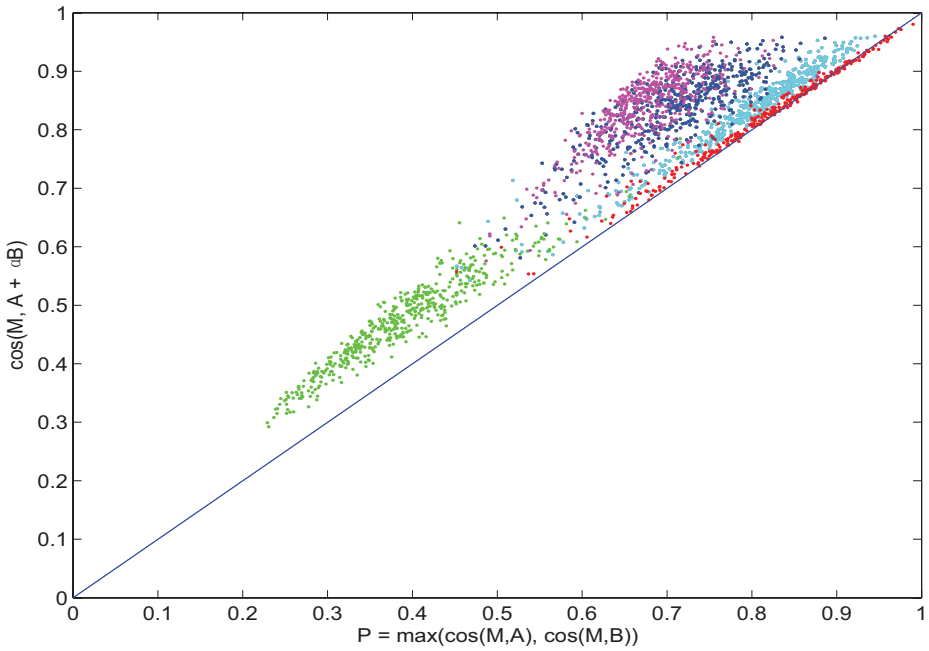

Figure 1.2: Comparison of spectral library search outcomes: Searching a query spectrum S against a spectral library has three possible outcomes: 1) No-match when S does not match any spectrum in the library (green dots); 2) Single-peptide match when S matches only one peptide in the library (red dots); 3) Mixture match when S is identified as a pair of peptides in the library (pink, blue, cyan dots represent Mixture matches when $\alpha = 1.0, 0.5, 0.1$, respectively). As illustrated by the colored sets, M-SPLIT can distinguish No-match from the rest by thresholding $P = max(cos(M, A), cos(M, B))$, shown on the x-axis. Similarly, Single-peptide and Mixture matches can be distinguished by thresholding $\Delta = cos(M, A+B)-P$, (shown on the y-axis as the distance from the main diagonal line

seamlessly elevate the performance of this approach to lower values of α.

Due to multiple factors in MS/MS data acquisition, it is possible that not all peaks in a single-peptide spectrum will appear in a mixture spectrum containing the same peptide. Intuitively, it is reasonable to assume that high-intensity peaks in the single-peptide spectrum will be detectable while low-intensity peaks may not be observed. We simulate this scenario by applying a window filter where a peak is kept if it has rank less than or equal to N in a window of W Daltons around its mass. We show that our method is robust against missing peaks using different values of W and N. This is consistent with previous studies showing that one

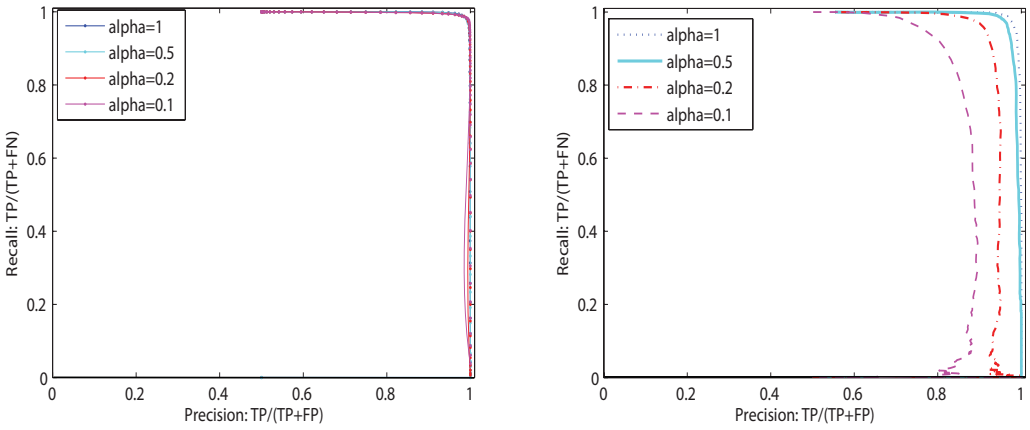

Figure 1.3: Classification of spectral library matches: **Left:** Precision/Recall curves when distinguishing No-match from Single-peptide and Mixture matches; decisions are made by checking whether $P = max(cos(M, A), cos(M, B))$ is above a predetermined threshold. **Right:** Precision/Recall curves when distinguishing Single-peptide from Mixture matches; decisions are made by checking whether $\Delta = cos(M, A+B) - P$ is below or above a predetermined threshold (respectively).

Table 1.2: Selecting the correct pair of peptides from the spectral library; each row indicates the percentage of cases when the top ranking pair is correct. M-SPLIT is compared with an iterative approach where one first identifies the spectrum with the top-scoring projected-cosine, removes shared peaks between the top-scoring spectrum and the query spectrum and finally searches the library a second time to identify the second peptide in the mixture. As shown here, the iterative approach is generally worse than M-SPLIT and especially error-prone for low values of mixture coefficients, consistent with our observations on estimation of mixture coefficients. For smaller values of α, M-SPLIT gains an advantage by simultaneously considering both peptides in the mixture.

Mixture coefficient (α)	M-SPLIT	Iterative approach
1:1	99.4	98.4
1:0.5	98.7	98.3
1:0.3	96.8	96.4
1:0.1	89.6	77.1

does not need all peaks in a spectrum for single-peptide identification purposes - in XHunter [25], the authors speed up the computation by showing that it is generally enough to retain only the top 20 peaks per spectrum.

Having observed that the highest-abundance peptide in a mixture can be identified as the top ranking match using projected-cosine, one could reason that

if the peaks from this peptide are removed from the mixture spectrum, we are left with a non-mixture spectrum. This leads to an *iterative* strategy to identify peptides in mixture spectra: first identify the spectrum with top-scoring projected-cosine, remove shared peaks between the top-scoring spectrum and the mixture spectrum, and search the library a second time to identify the second peptide in the mixture. The accuracy of the iterative method is compared with that of M-SPLIT in Table 1.2 and observed to be worse. Note that this is consistent with our results on estimation of α: as α gets smaller, it is important to consider both components in the mixture for accurate identification and quantification of *both* peptides .

1.1.9 Peptide identification with compressed chromatography

While the simulation experiments demonstrate the ability of M-SPLIT to reliably identify mixture spectra against large spectral libraries, we further validated our method on experimental data. The dataset consists of six bovine proteins (Apotransferrin, Carbonic Anhydrase, Catalase, Glutamate Dehydrogenase, Lactoperoxidase, Serum Albumin) from Michrom. 500 pmol of each protein were mixed in an equimolar ratio in an 50/50 mix of acentonitrite and water, reduced, alkylated and trypsinized. This same sample was analyzed under two different chromatographic time scales: one dataset was obtained with an 80-minute chromatography (Long dataset) while the other dataset was obtained with a short 3-minute chromatography (Short dataset). Mass Spectrometry data were acquired on a Thermo LTQ-Orbitrap XL operating on an acquisition cycle of two consecutive survey scans (first in the linear ion trap, second in the Orbitrap at 60K resolution) followed by MS/MS scans at unit resolution (linear ion trap, centroid mode, AGC on). We note that while the high-resolution survey scans readily provide accurate precursor masses, these particular settings assign MS/MS precursor masses based on the low-resolution survey scans, thus allowing us to verifiably test the performance of our approach as if operating in the (still) most common data acquisition mode. Peak lists in RAW files were converted to mzXML using

ReAdW. Excluding the initial load and final wash periods, we obtain 251 MS/MS spectra in the Short dataset that could possibly be mapped to spectra in the Long dataset. Under these chromatographic conditions, we assume that each spectrum in the Long dataset comes from only one peptide and use these as our library of single-peptide spectra. Conversely, since the Short dataset was obtained from the same sample with much less chromatography time, we assumed that some spectra might contain pairs of peptides that had been separated in the Long run; the Short dataset was thus used as our set of query spectra against the spectral library defined by the Long dataset.

The Long dataset was annotated using InsPecT[29] to search SwissProt (*ver.15.9*) with parent mass tolerance 2 Da and fragment mass tolerance 0.5 Da; a 5% false discovery rate was enforced using a standard target/decoy strategy [30] and no modifications were allowed. We note that while 5% FDR is generally too high for peptide identification purposes, our main utilization of search results was in grouping repeated spectra from the same peptide. To further increase the coverage of peptide identification we grouped spectra in the library by assigning two spectra to the same group if their parent masses are within tolerance (2 Da; 0.05 Da if precursor masses are corrected using the high accuracy survey scans) and their cosine similarity is high. Then if any spectrum in a group is annotated by InsPecT, annotations are transferred to every member in the group. To reduce potential errors, annotations are transferred only when coherent across all identified spectra in the same group; otherwise all spectra in that group are considered unidentified. Spectra in the Short dataset were annotated in two different ways: a) M-SPLIT with parameters determined from the simulation experiments and b) by InsPecT using the same search parameters used for the Long dataset. Out of 251 MS/MS spectra, M-SPLIT returned a total of 187 matches (see Figure 1.4) and InsPecT returned 22 IDs. As a first level of validation, we ran M-SPLIT without a parent mass filter and used parent mass as an *a posteriori* independent test to approximate the accuracy and sensitivity of our approach. The lack of a parent mass filter also allows us to estimate the performance of M-SPLIT on a much larger spectral library (e.g., proteome-scale spectral library) where searches would be conducted

only against spectra with matching parent masses, thus resulting in a comparable number of candidate matches. We manually compared the MS1 isotopic profile of the query spectra to the MS1 isotopic profile of the top match(es) returned by M-SPLIT and verified whether these were the same. Two isotopic profiles are considered the same if both indicate the same peptide charge and if isotopic peaks have m/z difference less than 0.05 Da. We also manually visualize both the MS1 and MS/MS spectra of Mixture match cases to verify that the matches are valid. The estimated accuracy for both Single-peptide and Mixture matches are shown in Table 1.3a.

Out of all 251 MS/MS spectra in the Short dataset, 64 did not match any spectra in the Long dataset (Table 1.3b). After manual investigation of the Long dataset, it turned out that for most cases (54/64) M-SPLIT did not find a match because either the corresponding peptide is missing (i.e., no corresponding MS1 isotopic profile was found) or it was not selected for MS/MS (MS2 not found). Hence, these unannotated spectra are likely a limitation in the library derived from the Long dataset and not a shortcoming of M-SPLIT that correctly classifies them as No-match cases. Considering the remaining ten cases as false negatives leads to an estimate of M-SPLIT's sensitivity of $\approx$ 94%(186/196). Note that these numbers, although not identical, are close to those seen in our simulated dataset and thus indicate that our simulation is able to capture important aspects of mixture spectra in real chromatographic settings.

To quantify the peptide identification gain in M-SPLIT we further compared the number of spectra and unique peptides identified in the Short dataset to those obtained by InsPecT. While this comparison violates the common assumption of database search methods that each spectrum comes from a single peptide, it nevertheless mimics the typical setup in MS/MS experiments and allows us to estimate the expected gains from using M-SPLIT. The 187 matches from the Short dataset to the Long dataset are divided into four groups in Table 1.3b according to their InsPecT annotations. While InsPecT is only able to annotate 22 spectra in the Short dataset, M-SPLIT is able to successfully annotate about four times as many spectra in the same dataset. When comparing the number of unique pep-

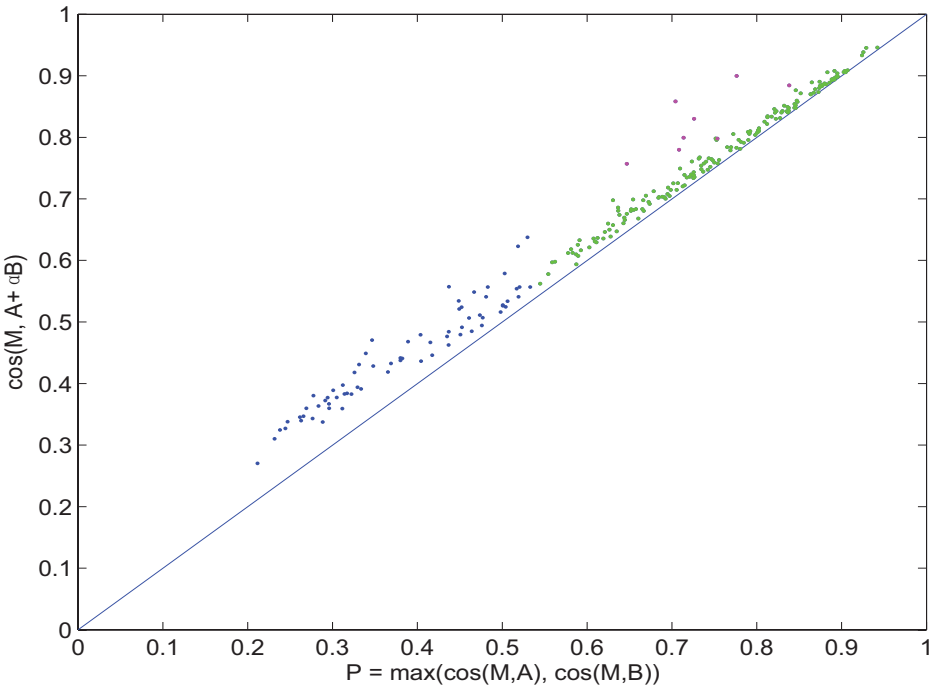

Figure 1.4: Classification of spectral library matches between Short (3-min) and Long (80-min) chromatography runs of the same sample: We assume that each MS/MS spectrum in the Long dataset comes from only one peptide and use these as our library of single-peptide spectra. On the other hand, since the Short dataset was obtained from the same sample with compressed chromatography, we would expect that some MS/MS spectra might contain pairs of peptides that were separated in the Long run and thus use this as our set of query spectra. Each spectrum in the Short dataset was searched against the Long dataset for the best pair and labeled as Mixture, Single-peptide and No-match, shown here as purple, green and blue dots, respectively.

tides identified in the Short dataset, InsPecT identified only $\approx 6\%$ of the peptides identified in the Long run, while M-SPLIT matches recover about $\approx 20\%$ of all identifications in the Long dataset, including IDs from mixture spectra.

1.1.10 Peptide identification in Yeast

To illustrate the utility of our method in a typical scenario, we further tested M-SPLIT on a larger experimental *Yeast dataset* [31], generously made publicly available in Tranche/ProteomeCommons [32] by researchers at the University of

Table 1.3: M-SPLIT results on the compressed-chromatography (Short) dataset. Out of 251 spectra, 186 have a match to the spectral library obtained from an 80-minute run of the same sample (Long dataset). **a)** Precision was estimated by comparing the MS1 isotopic profile of each query spectrum and the top matches returned by M-SPLIT in the Long dataset. Two isotopic profiles are considered matched if they indicate the same peptide charge, have correlated intensities and isotopic peaks have m/z difference $\leq$ 0.05 Daltons. **b)** M-SPLIT matches are divided into four categories according to whether the spectra were identified by InsPecT. **c)** The 64 spectra that did not match to the Long dataset were further investigated manually. For most cases (54 out of 64) this was due to missing data in the Long dataset - either there was no MS/MS spectra for the corresponding MS1 precursor or no matching MS1 precursor was found. **d)** Number of unique peptides identified by M-SPLIT and InsPecT.

a) All M-SPLIT matches

Category	Precision
Single-peptide matches	97% (174/179)
Mixture matches	87% (7/8)

b) Identified M-SPLIT matches

Identified by InsPecT		
Long dataset	Short dataset	Counts
No	No	95
Yes	No	73
No	Yes	8
Yes	Yes	11

c) Spectra in the Short dataset not matched to the Long dataset

Category	Counts
MS1 not found	17
MS2 not found	37
MS1 and MS2 found	10

d) Unique peptide identifications

	Number of peptides identified	
Method	Long dataset	Short dataset
InsPecT	211	14
M-SPLIT	n/a	43

Vanderbilt. In brief, a tryptic digest of a Saccharomyces cerevisiae was analyzed on an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific) and MS/MS

spectra were acquired using a data-dependent scanning mode in which one full MS scan (m/z 3002000) was acquired on the Orbitrap at a resolution of 60 000, followed by 8 MS/MS scans collected on the LTQ (see [31] for full details). To retain the utility of accurate precursor masses for a posteriori validation of search results, InsPecT was ran with 2.5 Da parent mass tolerance and 0.5 Da fragment mass tolerance on SGD yeast protein database (*ver.5/8/2009*); a 1% false discovery rate was enforced using a target/decoy strategy and no modifications were allowed. M-SPLIT was ran with default parameters against the Yeast spectral library from NIST (*ver.5/4/2009*); a 3 Da parent mass filter was used to pre-filter the library before the search. The results are summarized in table 1.4. In short, InsPecT is able to identify a total of 19,297 spectra and 4,486 unique peptides. On the other hand, M-SPLIT is able to identify 28,993 single-peptide spectra, 1,505 mixture spectra and a total of 6,089 unique peptides. Since the Yeast dataset was acquired with high-accuracy survey scans, this information was further used to validate our annotations by comparing the theoretical m/z value of the peptides returned by InsPecT/M-SPLIT and the observed precursor m/z in the corresponding survey scans. An annotation is considered correct if the theoretical precursor m/z is within 5ppm of the observed m/z; the estimated accuracies are summarized in table 1.4. The comparison between M-SPLIT and InsPecT further reveals that their annotations are same in $\approx 99\%$ of the cases where both make an annotation, thus demonstrating the coherence of these two independent methods.

M-SPLIT identifications indicate that mixture spectra consist of about 5% of all identifiable spectra in the Yeast dataset, suggesting that these constitute a modest but significant fraction of identifiable spectra in typical proteomics experiments. It should be emphasized that even though the number of mixture spectra is not large, these result in more than one peptide identification per spectrum and thus carry more information than single-peptide spectra. In the Yeast dataset there are a total of 28,993 single-peptide spectra identified by M-SPLIT as 5,873 unique peptides. In addition, M-SPLIT further identifies 1,505 mixture spectra as 1,627 unique peptides, 239 of which are only identified in mixture spectra, a summary of the overlap between the two methods is shown in figure 1.5.

Table 1.4: M-SPLIT and InsPecT search results on the Yeast dataset [31]. a) Numbers of identified spectra (single-peptide and mixture) and unique peptides. b) The precision of peptide identifications was estimated by comparing the theoretical precursor m/z of peptides returned by M-SPLIT or InsPecT and the observed precursor m/z values in the corresponding MS1 scan (isotopic profile). An identification is considered correct if the difference between theoretical and observed precursor m/z values is less than 5ppm. For mixture spectra the overall precision is computed by dividing the number of correct peptide identifications by the total number of identifications (i.e., twice the number of mixture spectra). The precision for the second-peptide identifications is also shown (in parenthesis); this precision is lower because the second peptide in the mixture is usually of low-abundance (average $\alpha = 0.3$) and thus harder to identify.

a) Spectrum and peptide identifications in the Yeast dataset

	Spectrum Identifications			Unique peptides
	Single-peptide	Mixture	Total	
InsPecT	19,297	n/a	19,297	4,486
M-SPLIT	28,993	1,505	30,498	6,089

b) Estimated precision in the Yeast dataset

Method	Single-peptide matches	Mixture matches
InsPecT	98%	n/a
M-SPLIT	98%	95.7% (91.4%)

1.1.11 Discussion

Despite the success of mainstream software for peptide identification from MS/MS spectra, the ubiquitous assumption that each spectrum arises from only one peptide is often not valid, making the interpretation difficult in such scenarios. To address this computational bottleneck, we propose the first spectral library-based approach (M-SPLIT) to the identification of mixture spectra generated from pairs of peptides. Theoretical bounds were derived to prune the search space using branch-and-bound techniques and further improved using a new projected-cosine metric. Thus, M-SPLIT dramatically reduces the search space by six orders of magnitude and is able to deliver results at an average of 2 seconds/spectrum (on a regular laptop with a Pentium Core2Duo, 1.6Ghz, 2Gb RAM), even when searching against proteome-scale spectral libraries. Despite considering only a tiny fraction of the whole search space, our benchmarks on both simulated and experimental data consistently show that M-SPLIT has both high sensitivity ($\approx 94\%$) and high

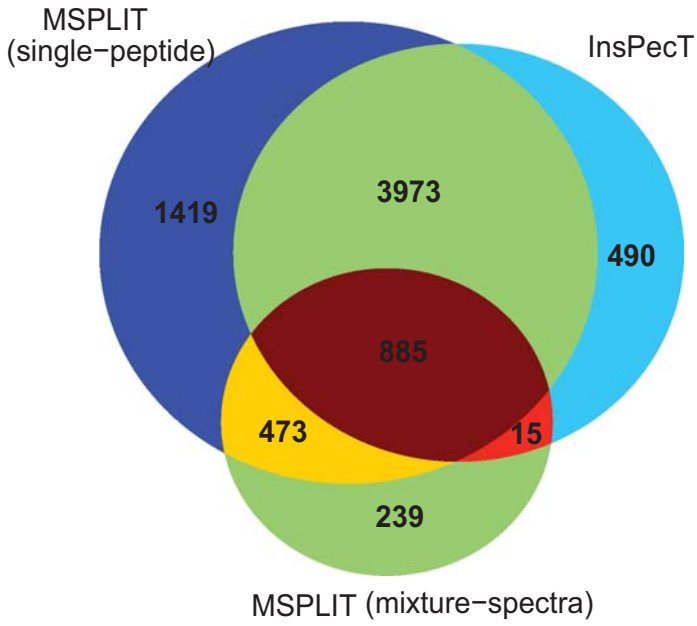

Figure 1.5: Peptide identifications with M-SPLIT and InsPecT in yeast dataset:Peptides identified by M-SPLIT and InsPecT are compared in a Venn diagram indicating the numbers of unique peptides in each category.

accuracy (up to 98%).

In addition to accurate peptide identification, M-SPLIT robustly quantifies the relative abundances of co-eluting peptides at the time of MS/MS acquisition, as determined by the fraction of MS/MS ion current assigned to each peptide. In principle, extending this approach to relative peptide abundances per run (e.g., in Data Independent Acquisition setups [33]) could be as simple as adding the estimated intensities over consecutive MS/MS scans followed by a posteriori computation of per-run relative abundances. It should be noted that, as in other label-free MS-based quantification approaches [34], there are MS-specific confounding factors that may result in distortion of the observed relative abundances (e.g., peptide-specific ionization efficiencies) and thus require follow-up experiments to validate the observed relative abundances.

We further note that M-SPLIT makes no assumptions about the type of query or library spectra. While M-SPLIT was developed and tested on peptide MS/MS spectra, the current implementation is readily applicable to any type of

spectra. In particular, it would be straightforward to extend any target spectral library to include spectra of common peptide and chemical contaminants and thus reduce their negative effect on peptide identifications (by matching experimental contaminant spectra to library contaminant spectra). By blindly identifying the best pairs in a given spectral library, M-SPLIT automatically classifies each query spectrum as a Mixture-match, Single-match, or No-match and thus it is a general self-adjusting tool that can be used on experimental setups promoting the acquisition of either/both single-peptide or/and mixture spectra.

The development of mass spectrometry algorithms typically requires large datasets with validated identified spectra that are difficult to obtain. The unavailability of datasets with validated identifications of mixture spectra was a limiting factor that we addressed in two different ways: by generating large datasets of simulated mixture spectra and by acquiring MS/MS spectra from the same sample using different chromatographic time-scales. The level of control afforded by the generation of simulated mixture spectra was instrumental in determining spectrum identifiability over a range of relative abundances of co-eluted peptides. These results were then corroborated using an experimental dataset where it was possible to provide exhaustive manual validation. As such, we were able to determine both the accuracy *and sensitivity* of our approach - a commonly difficult task since the set of true positives (and its complementary false negatives) is typically not known in advance. After our validation, we estimate that M-SPLIT delivers a false *negative* rate of only 5% at accuracy levels of up to 98%.

Focusing M-SPLIT on the identification of mixture spectra from pairs of peptides allowed us to derive theoretical bounds and filtration techniques that can be extended for spectra from more complex mixtures. In particular, the utility of the projected cosine metric is likely to increase as mixture spectra become more complex. Also, while M-SPLIT is already able to reliably annotate mixture spectra with inaccurate fragment masses (still the dominant MS/MS acquisition mode), its performance is very likely to further improve for high accuracy MS/MS data. Such data could seamlessly enable the identification of co-eluted peptides at more disparate relative abundance ratios and would likely greatly simplify the extension

to mixture spectra from more than two peptides.

1.2 MixDB: database search of mixture spectra

Despite the rapid growth of spectral libraries, methods based on spectral matching suffer from the limitation that peptides cannot be identified if they have not been observed before. Moreover, while spectral library from CID fragmentation is quite comprehensive, spectral libraries from other fragmentation methods (e.g. ETD, HCD) are still limited. Hence database search methods are still the mainstream approach for peptide identification. Some database search tools approach the mixture spectra identification problem by reporting spectra with more than one significant single-peptide match and do not explicitly attempt to models the occurrence of fragment ions from two peptides in the same spectrum. False Discovery Rates (FDR) are also left unadjusted [16, 18, 19] and may result in higher than expected FDR for second IDs (e.g., when co-eluting peptides share a substantial number of fragment masses). Different from previous approaches, our new database search tool, MixDB, uses a scoring model specifically designed for matching spectra against more than one peptides and determines separate FDRs for identification of single-peptide spectra and mixture spectra. Below we described our approach for the case when a mixture spectra is from two peptides.

Similar to M-SPLIT a mixture spectrum M is modeled as an MS/MS spectrum from two different peptides: $M = A + \alpha B$ and our goal is to identify mixture spectra by comparison against all possible *pairs* of peptides in a given protein sequence database. More formally, we define a peptide sequence as a vector $P = p_1, p_2, ...p_n$, where p_i is non-zero if there is at least one theoretical ion mass in the corresponding mass bin. A database D is simply a set of peptides $D = \{P^1, P^2, ...P^n\}$. We can now formulate the following computational problem:

Mixture Spectrum Identification Problem (MSIP)

Input A putative mixture spectrum M and a sequence database D.

Output A pair of peptides $P^i, P^j \in D$,
maximizing $PPSM(M, P^i, P^j)$
where $PPSM$ is a given Peptide/Peptide Spectrum Match scoring function that describes how well a pair of peptides (P^i, P^j) matches the spectrum M.

1.2.1 Filtration strategy

As in the case for spectral library search, the large number of possible peptide candidates in a sequence database (10^4-10^5 peptides in the yeast database after precursor mass filtering with 3Da tolerance) makes searching all possible *pairs* of peptides a prohibitive approach: the resulting $\approx 10^{10}$ comparisons per query spectrum would mean that a typical dataset of 15,000 spectra would take ≈50 days to process (based on average InsPecT runtimes, previously shown to be ≈100 times faster than SEQUEST [29]), even without considering the additional computational burden of scoring spectra against pairs of peptides. The explosion in the number of candidate matches per spectrum also dramatically increases the chances for false-positive matches. We used a filtration strategy that is similar to one introduced in M-SPLIT using the concept of project-spectrum. While in M-SPLIT we used the normalize dot-product to evaluate how good a match between the projected-spectrum and a candidate spectrum in spectral library, here we used a probabilistic scoring function, which will be described below, to evaluated the match between the projected-spectrum and a candidate *peptide* in the sequence database.

The efficiency of the projected-spectrum filter is determined by the highest (i.e., worst) rank of a correct peptide match of a mixture spectrum to the database D. Note that a correct match in D is a match to one of the peptides generating M – single-peptide spectra have one correct match, mixture spectra have two correct matches and thus two correct-match ranks. As shown in Figure 1.6, the resulting ranks of correct matches indicate that the projected-spectrum filter is an efficient filter that, for over 96% of cases, retains *both* correct matches (i.e., maximum ranks) at ranks less than 500 from about 10,000 candidate peptides (yeast database, 3 Da

precursor mass filtering). The left panel in Figure 1.1 also shows that one of the correct matches, presumably the higher-abundance peptide in the mixture, almost always has rank less than 10 (i.e., minimum ranks). This means that for almost all cases one only needs to pair the top 10 candidates with the top 500 candidates to find the correct match. Using this strategy at most $10 \times 500 = 5000$ candidate pairs need to be considered, thus conferring a $\approx \frac{10{,}000 \times 10{,}000}{2 \times 5000} = 10,000\times$ speedup compared to considering all $\approx 5 \times 10^7 = \frac{10{,}000 \times 10{,}000}{2}$ possible candidate peptide pairs.

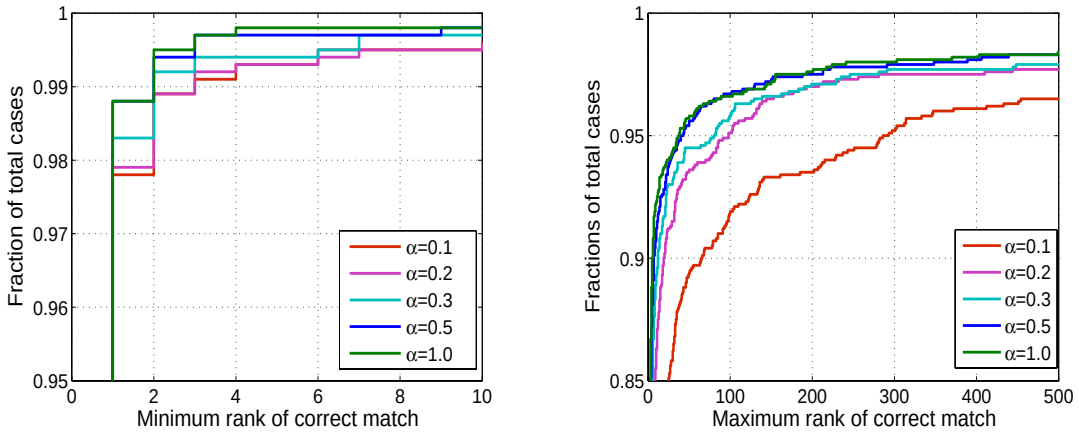

Figure 1.6: Filtration efficiency: Cumulative distributions of minimum rank (left) and maximum rank (right) for correct matches of simulated mixture spectra to the yeast database. Candidate peptides in the database are first sorted according to decreasing score against the corresponding projected mixture spectra (typical number of peptide candidates after precursor mass filtering is ≈10,000). The ranks of the correct matches are then determined. Correct matches are peptides in the database that correspond to one of the peptides used to generate the simulated mixture spectrum. Since each mixture spectrum has two correct matches we report both the minimum (i.e., best) and maximum (i.e., worst) rank of the two matches in the left/right panels, respectively. As shown, in more than 96% of cases, one of the correct matches has rank ≤ 10 (left) while the other correct match ranks ≤ 500 (right). Thus it is sufficient to pair the top 10 candidates with the top 500 candidates to find the correct pair. Using this strategy at most 10×500 peptide *pairs* need to be considered, resulting in a speedup of four orders of magnitude compared with considering all $\approx 5 \times 10^7$ possible peptide pairs.

1.2.2 Scoring function for Peptide/Peptide Spectrum Match

While scoring a peptide against an MS/MS spectrum is a well studied problem in proteomics, few scoring functions have been designed to handle more than one peptide [16]. Here we describe a general probabilistic model for scoring Peptide/Peptide Spectrum Matches (PPSMs). First we briefly review the model for single-peptide matches (PSMs) [35] and show how to extend the approach for PPSMs.

As described above, an MS/MS spectrum is represented as a vector of n bins, each representing a mass interval of width δ. A bin has value zero if there is no peak in the corresponding δ-Dalton interval otherwise it is non-zero. For experimental MS/MS spectra the raw intensity in each bin is first transformed into peak intensity rank (ranked from most to least intense), whereas for a theoretical spectrum bin values indicate the ion type (e.g., b-ion or y-ion) that generates the peak. Hence we define an experimental spectrum $S = s_1, s_2, ...s_n$ as a vector where $s_i \in R$ (peak ranks, always positive integers) and a peptide $P = p_1, p_2, ...p_n$ as a vector where $p_i \in I$ (ion types). When multiple ion types fall into the same bin, we keep track of all the ion types associated with that particular bin. The probability of a peptide P generating a spectrum S is defined as $Prob(S|P) = \prod_{i=1}^{n} Prob(s_i|p_i)$, where $Prob(x|y)$ is an arbitrary $|R| \times |I|$ matrix representing the probability that an ion type y in the peptide generates a peak with rank x in the spectrum. When there are multiple ions associated with a particular bin in peptide P we choose the ion that maximizes $Prob(s_i|p_i)$. Formally, if we denote p_{ij} as each of the ion types that associate with the i^{th} bin in P then we have $Prob(s_i|p_i) = max_j Prob(s_i|p_{ij})$. Finally, the score of a peptide P against a spectrum S is defined as the ratio of the probability that S is generated by the peptide P versus the probability that S is generated by a peptide string of all zeros (i.e., all peaks interpreted as noise). We express this score as the sum of a log odds ratio:

$$Score(S, P) = log\left(\frac{Prob(S|P)}{Prob(S|0)}\right) = \sum_{i=1}^{n} log\frac{Prob(s_i|p_i)}{Prob(s_i|0)} = \sum_{i=1}^{n} score(s_i, p_i)$$

The values of $Prob(s_i|p_i)$ can be learned from a training dataset of annotated

single-peptide spectra; similarly the noise model $Prob(s_i|0)$ can be trained using the rank distribution of unassigned peaks in the same annotated single-peptide spectra. The learning is done separately for peptides of different precursor charge and length to account for their different fragmentation statistics (see [35] for full details of this model).

In order to score mixture spectrum matches, we extend this model for pairs of peptides (P, Q). We first define a score vector as:
$\overrightarrow{Score}(P, S) = [Score(s_1, p_1), Score(s_2, p_2)\ ...\ Score(s_n, p_n)]$
where each element is the value of scoring the i^{th} element in P against the i^{th} element in S. Without loss of generality we refer to the highest-abundance peptide in the pair as P. To score a pair of peptide against the observed spectrum we first generate a score vector for each peptide:
$\overrightarrow{Score}(P, S, Hi) = [Score(s_1, p_1, Hi), Score(s_2, p_2, Hi)\ ...\ Score(s_n, p_n, Hi)]$
$\overrightarrow{Score}(Q, S, Lo) = [Score(s_1, q_1, Lo), Score(s_2, q_2, Lo)\ ...\ Score(s_n, q_n, Lo)]$
where Hi and Lo indicates the relative abundance of P and Q. Then we combine the two score vectors into the final score for the $PPSM$:
$Score(M, (P, Q)) = \sum_{i=1}^{n} max(Score(s_i, p_i, Hi), Score(s_i, q_i, Lo))$.
The max operation handles the dependency among the two peptides matched to the spectrum. This way score for a particular peak will not be double-counted when theoretical fragment ions from both peptides matched to the same observed peaks in the mixture spectrum.

Also noted that different scoring models are used for the high-abundance peptide P and low-abundance peptide Q respectively to capture their difference in fragmentation pattern. To show this we generated a dataset of 100,000 simulated mixture spectra and compute their fragmentation statistic for each peptide. As shown in Figure 1.7, fragment ions from high-abundance and low-abundance peptides have quite different peak rank distributions. Scoring models for single-peptide spectra do not capture these characteristics of mixture spectra. Therefore, a scoring model that explicitly models fragment ions from pairs of peptides is needed.

Because peptides with different charge states and length have different frag-

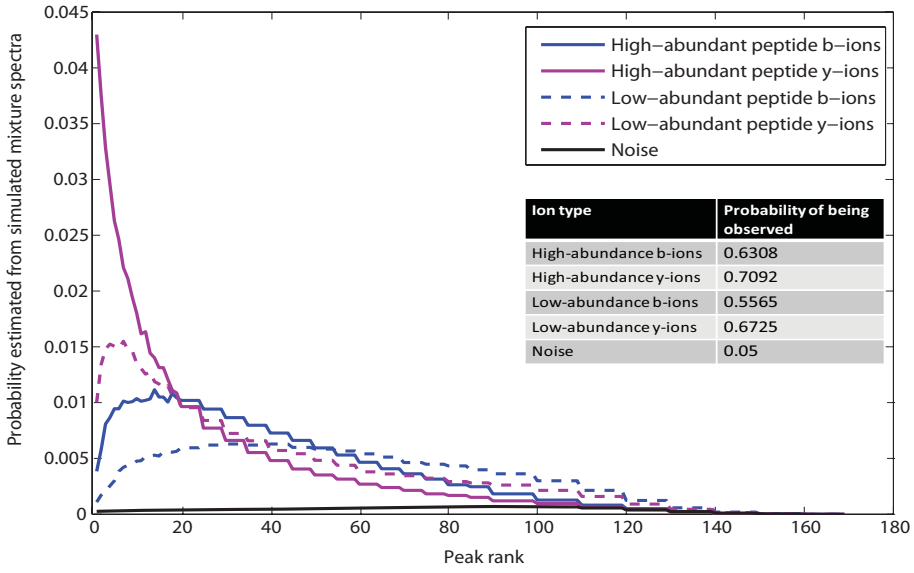

Figure 1.7: Fragmentation statistics for mixture spectra: Statistics are computed from a set of simulated mixture spectra. As shown in figure, fragment ions from high-abundance and low-abundance peptides in mixture spectra have different distribution of peak ranks (peak ranked from most intense to least intense). Scoring function that design only for single-peptide spectra do not capture this characteristic of mixture spectra. Thus scoring function that distinguish and explicitly model fragment ions from high/low-abundance peptides in mixture spectra are used in MixDB. Probability shown in the table is the total probability that a particular ion type is being observed in the training data, they correspond to the area under each curve.

mentation patterns [35], we divided library spectra into four categories according to their identified peptides as shown on the left below:

a) Number of spectra per peptide category

	Peptide Length		
	Long	Short	
Precursor charge, $z = 2$	17016	15197	Short: $\leq 13aa$
Precursor charge, $z = 3$	6819	6949	Short: $\leq 19aa$

b) Mixture spectrum categories

$$\left\{ \begin{array}{c} Long, z = 2 \\ Long, z = 3 \\ Short, z = 2 \\ Short, z = 3 \end{array} \right\} \times \left\{ \begin{array}{c} Long, z = 2 \\ Long, z = 3 \\ Short, z = 2 \\ Short, z = 3 \end{array} \right\}$$

This separation results in a total of sixteen categories of mixture spectra by pairing spectra from each category with spectra from another category (see

above right); the 100,000 simulated spectra were divided into 16 sets using the above categorization. A separate scoring model was then trained for each different type of mixture spectra. The considered ion types were: $b, b(iso), b - H_2O, b - NH_3, y, y(iso), y - H_2O, y - NH_3$, where b(iso) and y(iso) indicate the first isotopic peak of b/y ions, respectively. We consider doubly-charged peaks for spectra from charge two precursors and both doubly- and triply- charged peaks for spectra from charge three precursors. Peak ranks are divided into bins as follows: 1) one rank per bin for rank 1-20; 2) five ranks per bin for ranks 20-60; 3) ten ranks per bin for ranks 60-150 and one last bin for all peaks ranks 150 or higher. Because our scoring function distinguishes between high-abundance and low abundance peptides in mixture spectra and during searching we do not know which candidate peptide is of higher abundance, we score a query spectrum against a candidate pair (P, Q) by comparing the observed spectrum against two theoretical spectra, one with P and another with Q as the higher abundance peptide; the higher score is the final PPSM score.

The performance of this scoring model is first evaluated using a set of simulated mixture spectra generated from a different dataset than that used to generate the training dataset. Using single-peptide spectra identified by both InsPecT and M-SPLIT in a previous study [36], we simulated mixture spectra with $\alpha = 1.0, 0.5, 0.3, 0.2, 0.1$ as described above. The percentage of cases where the top peptide pair returned by MixDB is correct is shown in Table 1.5. As expected, as α decreases the performance worsens because it becomes harder to identify the lower abundance peptide. We also compared the performance of MixDB with M-SPLIT, our spectral library search tool previously shown to be efficient and robust in identifying mixture spectra. On average, M-SPLIT correctly identifies 15% more mixture spectra than MixDB. In general, spectral library search methods have two main advantages over database search methods: the relative intensities of different fragment ions are known in advance and the number of peptide candidates is smaller. In order to understand the relative importance of these two factors, we also evaluated MixDB when searching only against peptides with spectra in the NIST spectral library. As can be seen from Table 1.5, with a reduced search

space, MixDB has similar performance to M-SPLIT. We also compare MixDB with an iterative search strategy where one first identifies the highest-scoring peptide, removes all annotated peaks from the spectrum and then uses the "residual" spectrum to search against the database a second time to identify the second peptide. As we can see in Table 1.5 when α is high, the performance of the iterative method is comparable with that of the combined scoring function. However, as α becomes smaller, it is better to consider both peptides at the same time.

Table 1.5: Sensitivity of selecting the correct pair of peptides from the spectral library/protein sequence database. MixDB and M-SPLIT were evaluated on a set of 5000 simulated mixture spectra. Each row indicates the percentage of cases when the top ranking pair is correct; numbers in parenthesis indicate fraction of cases when one of the top ten peptide pairs is correct. Two types of database search were performed. In the first (MixdB), each spectrum was searched against all Yeast tryptic peptides. In the second (MixDB*), each spectrum was searched only against peptides with spectra in the NIST Yeast spectral library. In general M-SPLIT spectral library search was found to have better performance than MixDB database search. However, as shown in the fourth column, the main advantage of spectral library search was the reduction in the number of possible peptide candidates. We also compared the MixDB's performance with an iterative strategy where: 1) the dominant peptide is identified; 2) peaks explained by the top peptide match are removed from the spectrum and 3) the residual spectrum is again searched against the database to identify the second peptide. When the mixture coefficient α is high the iterative strategy has similar performance to MixDB; however, as α becomes smaller, MixDB's scoring of both peptides at the same time results in considerably better performance than the iterative strategy.

Mixture coefficient (α)	M-SPLIT	MixDB	MixDB*	Iterative method
1:1	97	87 (97)	95 (98)	81
1:0.5	92	79 (92)	90 (98)	74
1:0.3	80	66 (86)	79 (92)	57
1:0.2	63	50 (77)	69 (87)	30
1:0.1	34	19 (43)	34 (70)	6

1.2.3 Classification of database search matches

A database search of MS/MS spectra will always identify some top-scoring peptide or peptide pair for any given query spectrum, even if the true match is not in the database. To assess whether the top match is significant we use a two-stage classifier to distinguish true matches from false positive matches. Since our

goal is to build a general search tool that can identify both single-peptide and mixture spectra we consider three possible outcomes when searching a given query spectrum S:

- No-match: S does not match any peptide in the database
- Single-peptide match: S matches one peptide in the database.
- Mixture match: S matches a pair of peptides in the database.

Classification of the top matches is done using two Support Vector Machines(SVM) [37]. The first SVM distinguishes No-match cases from Single-peptide / Mixture matches and the second SVM distinguishes Single-peptide matches from Mixture matches (see Figure 1.9). To build the SVMs we consider the PPSM score described above and several other features that have been found useful in distinguishing true matches from false positives in single-peptide spectra, namely:

- 1) likelihood score for one peptide match: likelihood score while considering only matched peaks from one peptide.
- 2) likelihood score divide by peptide length: score from 1) divided by the number of amino acids in the top candidate peptide.
- 3) explained intensity: total intensity of annotated peaks divided by total intensity of the spectrum.
- 4) fraction of b and y ions present (2 features): number of b and y ions present in the spectrum divided by the number of b/y ions possible from the peptide.
- 5) longest consecutive series of b and y ions (2 features).
- 6) average mass error between theoretical and observed masses.

Features 2-6 are computed separately for each peptide in the top pair, resulting in a total of 16 feature inputs to the SVMs To train the SVM models, we constructed two negative control datasets. The No-match dataset consists of 5000 mixture

spectra where the peptides used to create the mixture spectra are deleted from the database. The Single-match dataset consists of 2500 single-peptide spectra and 2500 mixture spectra where one peptide in the mixture is removed from the database. These two datasets were combined with another dataset of 5000 mixture spectra (Mixture-match dataset) and searched against the database. For all simulated mixture spectra, the mixture coefficient α was selected uniformly from 0.1 to 1. The top matches from each dataset were used as training data for the SVM models. The training is carried out in a two-step fashion. In the first step, top matches from the No-match dataset were treated as negative cases and correct top matches from the Single-match and Mixture-match dataset were used as positive cases. In the second step correct top matches from the Mixture-match dataset were used as positive cases while all top matches from the Single-match dataset and No-match dataset were used as negative cases. The performance of the SVM models were assessed using 10-fold cross-validation (shown in figure 1.8).

1.2.4 Estimation of False Discovery Rates

False Discovery Rates (FDRs) were estimated by extending the standard Target-Decoy strategy for database search [30]. Depending on whether each peptide in the top peptide pair comes from the Target or Decoy databases, matches are divided into the following possibilities: TT – both peptides matched Target, TD – the most abundant peptide matched Target and the least abundant peptide matched Decoy, DT – the most abundant peptide matched Decoy and the least abundant peptide matched Target and DD – both peptides matched Decoy. When searching mixture spectra there are two possible outcomes for identification: single-peptide matches and mixture matches, each with a separate corresponding FDR. The FDR for single-peptide matches is defined as: $FDR_{match} = \frac{DT+DD}{TT+TD}$ and the FDR for mixture-spectrum identification is defined as: $FDR_{mixture} = \frac{TD}{TT}$ (see figure 1.9). Note that DT and DD peptide pairs will never advance to the second SVM because they are rejected as false positives after FDR_{match}; only TD and TT matches are considered by the second SVM as candidate mixture matches.

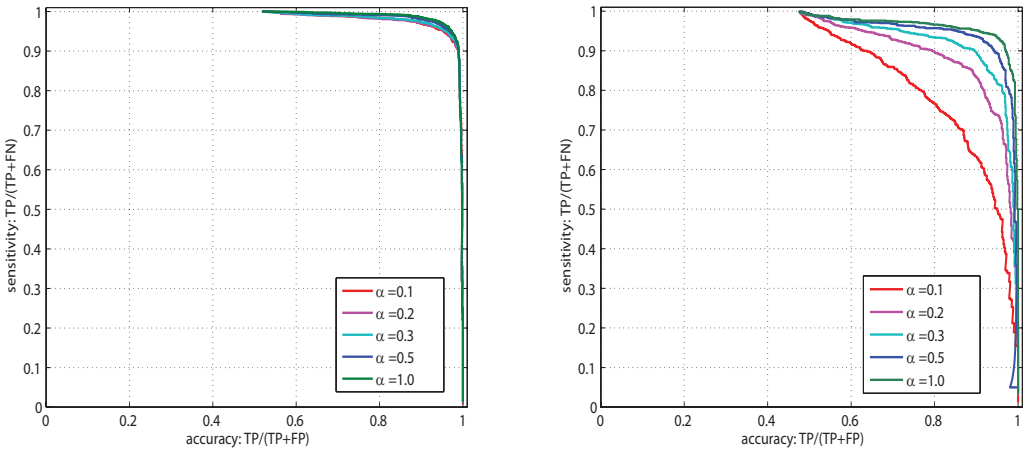

Figure 1.8: Classification of database search matches: Three sets of simulated mixture spectra were constructed: 1) spectra where both peptides match the Yeast protein database (Mixture match), 2) spectra where only one peptide matches the Yeast protein database (Single-peptide match) and 3) spectra where neither peptide matches the Yeast protein database (No match). Each spectrum was searched against the Yeast protein database and the top matches were used as training data to build two SVM models: one distinguishing No-match from Mixture and Single-peptide matches and a second SVM model distinguishing Mixture matches from the other two types of matches. The performance of the SVMs was assessed using cross-validation. **Left:** Precision/Recall curves when distinguishing No-matches from Single-peptide and Mixture matches and **Right:** Precision/Recall curves when distinguishing Single-peptide matches from Mixture matches.

1.2.5 Identification of mixture spectra in Yeast data

To illustrate the utility of our method in a typical scenario, we tested our database search method on an experimental *Yeast dataset* [31], generously made publicly available in Tranche/ProteomeCommons [32] by researchers at Vanderbilt University. In brief, a tryptic digest of Saccharomyces cerevisiae was analyzed on an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific) and MS/MS spectra were acquired using a data-dependent scanning mode in which each full MS scan (m/z 300–2000) was acquired on the Orbitrap at resolution 60,000, followed by 8 MS/MS scans collected on the LTQ (see [31] for full details). The data were analyzed using InsPecT [29], MixDB and ProbIDtree [16] with 3 Da parent mass tolerance and 0.5 Da fragment mass tolerance against the SGD yeast

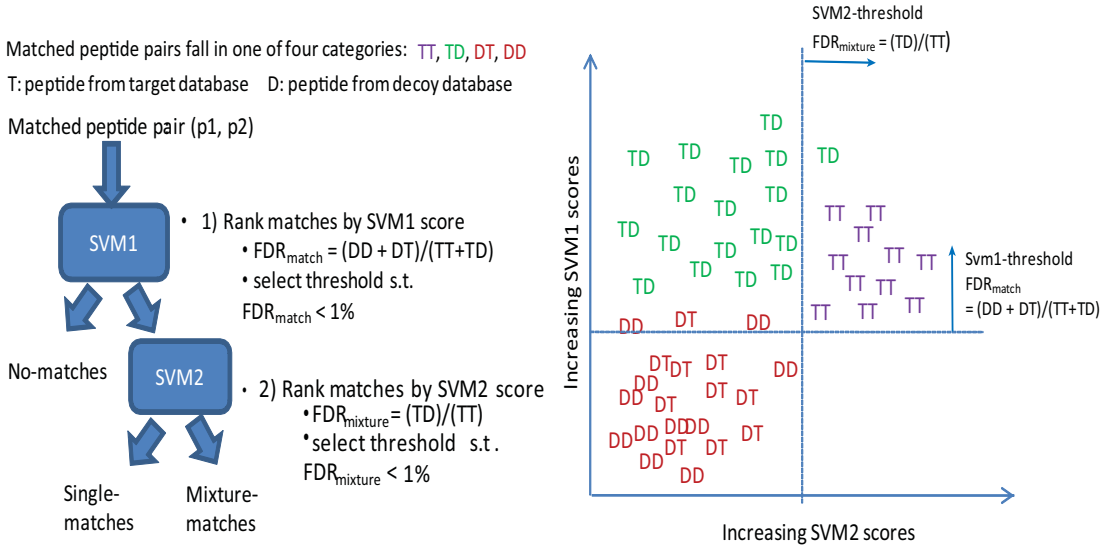

Figure 1.9: Workflow of SVM classification and FDR estimation: Every query spectrum searched against the database becomes assigned to some peptide pair (P_A, P_B) that best matches the spectrum. Depending on whether the matched peptides come from the Target or Decoy databases, each match falls in one of four categories: Target/Target(TT), Target/Decoy(TD), Decoy/Target(DT), Decoy/Decoy(DD). All matches are ranked according to their SVM1 score to assess whether the most abundant peptide (P_A) in each paired match (P_A, P_B) is significant. As with the standard Target/Decoy strategy, matches with P_A from Target are considered positive matches and matches with P_A Decoy are considered negative matches. Therefore, the FDR for single-peptide matches is computed as: $FDR_{match} = \frac{DT+DD}{TD+TT}$ and the SVM1 score is thresholded to yield $FDR_{match} < 0.01$. Matches with SMV1 score above the threshold are then ranked by their second SVM scores to evaluate whether the second peptide (P_B) in the top match (P_A, P_B) is also significant. As before, matches P_B from Target database are considered positive matches and those with P_B from Decoy are negative matches. Thus the FDR for mixture matches is computed as: $FDR_{mixture} = \frac{TD}{TT}$ and SVM2 scores are also thresholded such that $FDR_{mixture} < 0.01$.

protein database (*ver.5/8/2009*); a 1% false discovery rate was enforced using a target/decoy strategy and the only modification allowed was carboxamidomethylation on cysteine. Although data collected with these survey scan setting typically features very high mass accuracy, we allowed a large precursor mass tolerance in the searches to find possible MS/MS spectra from co-eluting peptides. In short, InsPecT is able to identify a total of 22,658 single-peptide spectra, MixDB is able to identify 23,930 single-peptide spectra plus 978 mixture spectra and ProbIDtree

is able to identify 19,840 single-peptide spectra plus 821 mixture spectra. Since the Yeast dataset was acquired with high-accuracy survey scans, this information was further used to validate the annotations by requiring the presence of the theoretical monoisotopic m/z value of the identified peptides in the corresponding survey scans. Since for low-abundance peptides the monoisotopic m/z may not be visible in the corresponding survey scan (observed in only $\approx 0.7\%$ of all cases and only for the lower abundance peptide in mixture spectra), in such cases we also check one preceding and one subsequent survey scan for the monoisotopic m/z. An annotation is considered correct if the theoretical precursor m/z is within 5ppm of the observed m/z; the results are summarized in Table 1.6. We first focus our attention on single-peptide cases. As seen in Table 1.6b) all three database search methods achieve similar precision of $\approx 97\%$. A more detailed comparison in Figure 1.10 shows that 78.4% and 68.2% of MixDB's annotations overlap with those of InsPecT and ProbIDtree, respectively. For all spectra for which both MixDB and InspecT/ProbIDtree make an annotation, more than 96% of them have the same peptide ID, indicating that these independent methods are consistent. Among those spectra that are only identified by MixDB or InsPecT/ProbIDtree, we further divide them into two categories – those where two methods identify the same peptide as the top hit and those where the two methods have a different top hit. Those in the "same-top-hit" category are more likely to be correct, since different scoring functions rank them as the same top peptide candidate. If we consider these cases, it increases MixDB's overlap with InsPecT and ProbIDtree to 90% and 85%, respectively, further indicating very good agreements between these different methods. Overall, for identification of single-peptide spectra, all three database search methods have comparable accuracy while MixDB identifies approximately 6% more spectra than InsPecT and 21% more spectra than ProbIDtree. This shows that MixDB's performance in identifying single-peptide spectra was not diminished by trying to identify more than one peptide per spectrum.

For mixture spectra, MixDB identifies 978 spectra while ProbIDtree identifies a total of 821 spectra (see Table 1.6a) . MixDB identified two peptides in each mixture spectrum, while ProbIDtree found 32 mixture spectra with more

Table 1.6: Search results on the Yeast dataset for MixDB, M-SPLIT, InsPecT and ProbIDtree. a) Numbers of identified spectra (single-peptide and mixture) and unique peptides are compared. b) To allow for identification of co-eluting peptides, all searches were run using 3 Da precursor mass tolerance. The accurate precursor mass information was then used a posteriori to estimate the precision of peptide identification by comparing the theoretical precursor m/z of peptides returned by each method and the observed precursor m/z values in the corresponding MS1 scan (isotopic profile). An identification is considered correct if the difference between theoretical and experimental precursor m/z values is less than 5 ppm. For mixture spectra the precision is slightly lower because the second peptide in the mixture is usually of low-abundance (average $\alpha = 0.3$) and thus harder to identify.

a) Spectrum and peptide identifications in the Yeast dataset

	Spectrum Identifications			Unique Peptides		
	Single-pep	Mixture	Total	Single-pep	Mixture	Total
MixDB	23930	978	24908	5476	1128	5802
ProbIDtree	19840	821	20660	4420	820	4739
InsPecT	22658	n/a	22658	5272	n/a	5272
M-SPLIT	28417	2567	30984	5997	2394	6684

b) Estimated precision in the Yeast dataset

Method	Single-peptide matches	Mixture matches
MixDB	98.3%	95.9%
ProbIDtree	97.8%	90.1%
InsPecT	97.1%	n/a
M-SPLIT	97.3%	95.4%

than two identified peptides. This indicates that even though more than two identifiable co-eluting peptides may appear in one MS/MS spectrum this is relatively rare. Thus by limiting mixture spectra to two peptides per spectra MixDB does not lose much sensitivity in peptide identification. Furthermore, MixDB compensates by identifying about 20% more mixture spectra than ProbIDtree (978 versus 821 mixture spectra). In addition, by limiting its search space MixDB is more accurate. As shown in Table 1.6b, MixDB achieves a precision of almost 96% while ProbIDtree has only ≈90% precision for mixture spectra. Two main reasons contribute to MixDB's higher precision. First, MixDB only searches up to two peptides per spectrum (ProbIDtree attempts to look for up to eight peptides per spectrum), thus theoretically it has a smaller search space than ProbIDtree. More

importantly, MixDB applies different scores and FDRs for the first and second identified peptides in each MS/MS spectrum (FDR_{match} and $FDR_{mixture}$, respectively). This is crucial because high-abundance peptides are likely to have very different match statistics (e.g., % explained intensity, % b/y ions presented) than low-abundance peptides in mixture spectra. Because there are many more single-peptide spectra than mixture spectra in this dataset, applying a single global FDR for the combined identification of both single-peptide and mixture spectra can lead to an underestimation of FDR for mixture spectra. Just as in the case of ProbIDtree, it has an estimated precision of $\frac{90.1\%\times 821+97.8\%\times 19840}{19840+821} = 97.94\%$ or an overall FDR of 2.06%. However, the precision for mixture spectra is only 90.1%, bringing its FDR on mixture spectra to 9.9% which is much higher than its FDR for single-peptide matches. In summary, we show that MixDB has both higher sensitivity and precision than ProbIDtree in identifying mixture spectra while having comparable performance to InsPecT in identifying single-peptide spectra.

Since spectral library methods are in general considered to be both more sensitive and more accurate in peptide identification than database search methods [22], we use identifications from spectral library searches to further validate the different database search methods. The Yeast dataset was analyzed using an extension of the M-SPLIT algorithm [36] (see Supplementary Materials) and SpectraST [26] with default parameters against the Yeast spectral library from NIST (*ver.5/4/2009*) with a precursor mass tolerance of 3 Da; a 1% false discovery rate was enforced using the decoy library strategy described in [38]. Since we do not allow PTMs in the database searches we also removed all the entries in the spectral library that contain PTMs. To evaluate the performance of M-SPLIT, we first compare it to SpectraST, the most popular publicly available method for peptide spectral library search. In short we showed that both methods identified similar number of single-peptide spectra and are consistent with each other. Only in approximately 6% of the spectra do the two methods identify different peptides as top hits (see Supplementary Materials). Therefore, from here on we use results from M-SPLIT as a reference to evaluate results from different database search methods. As shown in Figure 1.10, M-SPLIT misses about $\approx$ 3600 single-peptide

spectra identified by either MixDB or InsPecT. However, in 85%–90% of these spectra, the identified peptides are not in the spectral library, indicating that M-SPLIT has very high sensitivity and therefore we can use spectra identified by M-SPLIT as a reference and compare the relative sensitivity of database search methods. Figure 1.10 and Supplementary Figure 2 show that MixDB, InsPecT and ProbIDtree identify 68%, 61.3% and 51.8% of the spectra identified by M-SPLIT, respectively. Among these shared annotations more than 97% of them have the same peptide annotation as M-SPLIT, again indicating that these database search methods have high precision for single-peptide spectra. If we look further into those spectra that are only identified by M-SPLIT but not by database search methods, we find that for 60%, 30% and 11% of the cases MixDB, InsPecT and ProbIDtree ,respectively, identify the same top hit as M-SPLIT. Altogether these indicate that MixDB, InsPecT and ProbIDtree, have a sensitivity of 89%, 75% and 58% for ranking the correct peptide as the top candidate. This implies that these database search methods are able to correctly identify these spectra, but the current scoring functions/SVM models do not have enough discriminative power to distinguish experiment-wide true matches from false matches. Thus to keep the FDR low, database search methods have to discard some possibly valid but low-scoring matches.

Next, we turn our attention to mixture spectra. Similar to the single-peptide case, most mixture spectra identified by database search methods but missed by M-SPLIT corresponded to peptides that did not have spectra in the spectral library (see Figure 1.10). Assuming the union of mixture spectra returned by M-SPLIT and MixDB as the total number of mixture spectra in this dataset we get that M-SPLIT has a false negative rate of $\frac{78}{2567+95+78} = 0.0357$ at ranking the correct peptide pair as the top candidate and $\frac{95}{2567+95} = 0.0274$ at the classification stage. Similar calculations for ProbIDtree results estimate the false negative rate of M-SPLIT as 0.056 and 0.0403 at ranking and classifying, respectively. This agrees with what we observed before [36] and again shows that M-SPLIT has high sensitivity and can also be used as a reference to compare database search methods for the identification of mixture spectra.

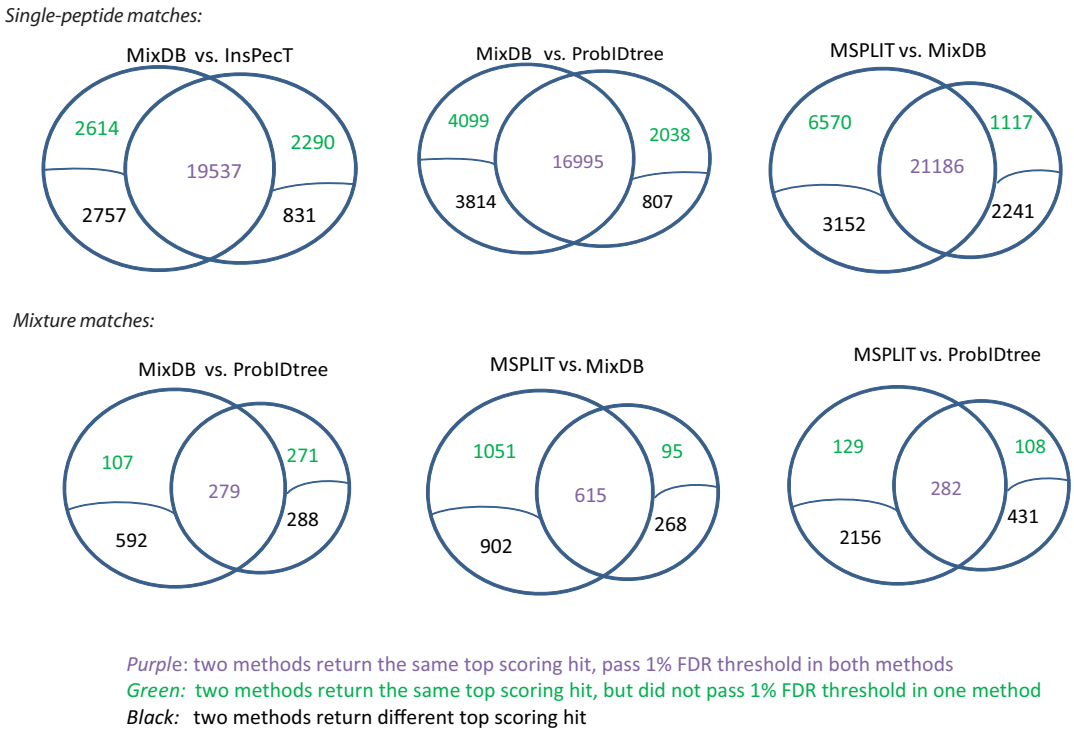

Figure 1.10: Comparison of identifications from MixDB, M-SPLIT, InsPecT, and ProbIDtree: All pairwise comparisons between MixDB, M-SPLIT, InsPecT and ProbIDtree compare identification results at 1% FDR as determined by the Target/Decoy strategy. In each pairwise comparison, spectra identified by both methods (in the intersection, shown in *purple*) were assigned to the same peptide in $96-97\%$ of cases, indicating that the methods are consistent and the precision is in good agreement with our estimates. Spectra identified by one method but not the other are subdivided into two categories: cases where the two methods return the same peptide (peptide pair in the case of mixture matches) as the top hit but it was below the FDR threshold for one of the methods (shown in *green*) and cases where the two methods do not return the same peptide as the top hit (shown in *black*). In general MixDB has high overlap with other database search methods. For single-peptide spectra MixDB finds the same top peptide match as other methods in 85% – 90% of cases. When using spectra identified by spectral library search as a reference set, MixDB is able to identify 6% – 16% more single-peptide spectra and 38% more mixture spectra than current database search methods. Taken together, these show that MixDB has better/comparable sensitivity and accuracy in identifying single-peptide spectra as well as significantly higher sensitivity and accuracy in the identification of mixture spectra.

Out of the 2567 mixture spectra identified by M-SPLIT, MixDB is able to identify 615 while ProbIDtree is able to identify 282. If we look at these cases

closely, we find that about 2460 mixture spectra identified by M-SPLIT come from peptides with charge 2 and 3, the only possibilities considered by MixDB. This means MixDB has a sensitivity of 25% while probIDtree has a sensitivity of 12%. If the two peptides in mixture spectra are considered independent and observing that MixDB has a sensitivity of 60% for single-peptide spectra, we expect MixDB to identify $60\% * 60\% = 36\%$ of all mixture spectra identified by M-SPLIT. This calculation makes the assumption that the two peptides are present at similar abundance, which, as shown in previous section, is the easiest scenario for identification of both peptides in mixture spectra. In practice, one peptide is present at lower abundance in most mixture spectra and in the yeast dataset, the average mixture coefficient α estimated by M-SPLIT is only 0.3. Thus having a sensitivity of 25% is a reasonable performance for MixDB. In addition, for a large fraction of mixture spectra (1051/1953) that MixDB did not classify as mixture matches, it identified the same top peptide pair per spectrum as M-SPLIT, but again the current SVM model does not have enough discriminative power to separate true matches from false positives. The extension of more sophisticated statistical models that have been proposed for single-peptide spectra [21] to mixture spectra is likely to increase the sensitivity of database search methods.

In order to estimate how the presence of more than one peptide affects current computational techniques in peptide identification, we performed a more detailed comparison between MixDB and InsPecT on the 2567 mixture spectra identified by M-SPLIT. As described above, the problem of peptide identification consists of two sub-tasks: 1) ranking the correct peptide as the top scoring candidate per spectrum among all the other peptide candidates in the database and 2) experiment-wide discrimination of correct peptide matches from false, but top scoring matches. We evaluated how the presence of more than one peptide affected each of these tasks. First we compared whether InsPecT and MixDB are able to correctly rank one of the peptides in the mixture as the correct top match. As shown in Figure 1.11, InsPecT is able to rank one of the correct peptides at the top for 75% of cases. On the other hand, MixDB is able to achieve this goal for more than 95% of the cases and is further able to rank *both* correct peptides as

the top for 65% of cases. This shows that by relaxing the assumption that each MS/MS spectrum comes from only one peptide, we gain a sensitivity of $\approx 20\%$ in ranking the correct peptides in mixture spectra as top candidates. At a 1% false discovery rate, InsPecT is able to classify 53% of all mixture spectra as single-peptide matches, while MixDB is able to classify 68% of cases as matches, with 24% of these classified as mixture matches. Since each mixture spectrum contains information for two peptides and InsPecT is able to identify one peptide for 53% of the cases, this means InsPecT is able to recover about $53\% \times 0.5 = 26.5\%$ of all peptide information in mixture spectra. In contrast, MixDB is able to identify both peptides in 24% of cases while identifying one peptide in another 43% of cases, thus resulting in a recovery rate of $24\% + 43\% \times 0.5 = 46\%$ of all identification contained in mixture spectra. Recall that MixDB and InsPecT have a sensitivity of 68% and 61.3% for identifying single-peptide spectra. This means that the presence of co-eluting peptides does not interfere significantly with the ability of current database search methods to identify the most dominant peptides (only a 9% drop in sensitivity). As for MixDB the sensitivity for the presence of co-eluting peptides does not affect its ability to identify the most dominant peptide in the spectra, since for mixture spectra it also has a sensitivity of approximately 68%.

1.2.6 Discussion

As increasingly more complex samples are analyzed in high-throughput proteomic experiments [7] and new data acquisition protocols evolve [10, 11, 12, 39], the occurrence and detectability of co-eluting peptides per MS/MS spectrum is likely to increase. The almost ubiquitous assumption that most mainstream computational methods make, namely that every MS/MS spectrum comes from one peptide affects their ability to identify spectra from co-eluting peptides [9]. While in this study the effect is only moderate for InsPecT, a $\approx 10\%$ decrease in sensitivity was observed, it is also worth noting that mixture spectra contain more information than single-peptide spectra. In our analysis of a yeast dataset, M-SPLIT, MixDB and ProbIDtree identified 5997, 5476 and 4420 unique peptides

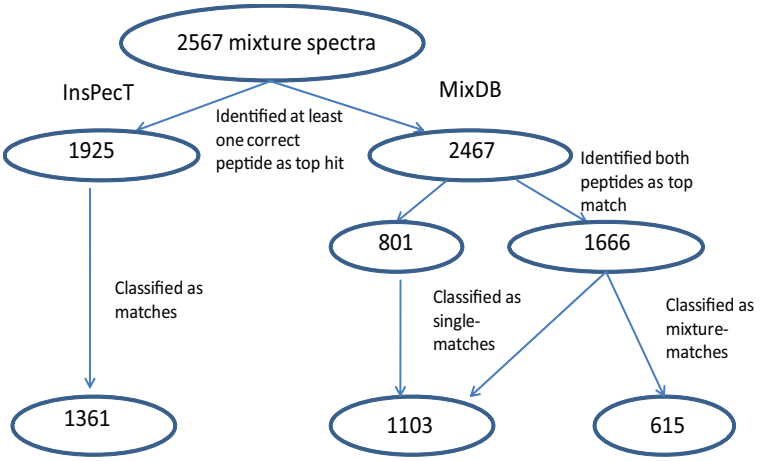

Figure 1.11: Comparison of InsPecT and MixDB result on mixture spectra: To study how the presence of co-eluting peptides affects database search results, mixture spectra identified by M-SPLIT were searched against the Yeast database using MixDB and InsPecT. InsPecT was able to rank one of the correct peptides as the top match in 75% of cases, while MixDB was able to rank one of the peptides as the top match in 95% of cases and ranked both correct peptides as the top match in 65% of cases, showing that MixDB gains 20% higher sensitivity in spectrum identification by relaxing the assumption that each spectrum comes from only one peptide. After imposing score thresholds at 1% FDR, InsPecT was able to classify 53% of cases as single-peptide matches, while MixDB was able to classify 24% of cases as mixture matches and an additional 43% of cases as single-peptide matches. Since each mixture spectrum contains information for two peptides, InsPecT recovers 27% = 53%×0.5 and MixDB recovers 46%=24%+43%×0.5 of the peptide information contained in mixture spectra.

from 28417, 23930 and 19840 single-peptide spectra, respectively. However, they were able to identify 2394, 1128, and 820 unique peptides from 2567, 978, 821 mixture spectra, thus revealing MixDB's rate of 1128/978=115% peptide IDs per mixture spectrum versus 5476/23930=23% peptide IDs per single-peptide spectrum. These results show that if dynamic range challenges can be addressed (e.g., by selectively isolating precursors of comparable abundance [39]) then protocols designed to generate mixture spectra have the potential to substantially improve peptide identifications while considerably decreasing the number of MS/MS scans required to obtain these identifications. In addition, due to the lower-abundance nature of most of second IDs from mixture spectra, they are less likely to be sampled again by the instrument and may not even be detectable without a co-eluting

peptide. Altogether, methods that consider more than one peptide per MS/MS spectrum can potentially double the sensitivity in recovering peptide information from mixture MS/MS spectra.

However, this higher information content comes at a cost in that the combinatorial explosion of searching for best peptide pairs dramatically increases the search space for mixture spectra. Thus, effective filtration strategies and special consideration for controlling the FDR become crucial in achieving high precision in the identification of mixture spectra. Similar to the development of any mass spectrometry algorithm, a large dataset of reliably identified spectra is crucial and is often hard to come by. In this study we addressed this issue with comprehensive simulations and by taking advantage of the public availability of the yeast spectral library [40] and experimental data [31]. Spectral library search methods are in general considered to be more sensitive and accurate for peptide identification [22] and in this case the yeast spectral library is also comprehensive (i.e., only a small fraction of peptides identified by database search method are not in the library). Thus we can use identifications from spectral library searches as the reference "truth" and comprehensively benchmark various aspects of different database search methods. We showed that MixDB has good sensitivity and high precision in identifying both single-peptide spectra and mixture spectra.

The development of computational methods and novel experimental strategies often rely on each other. For example, because mainstream computational approaches assume each MS/MS spectra comes from one peptide, development of chromatography and mass spectrometry protocols has focused on making this assumption valid for most cases. However, with a new generation of algorithms that are able to identify mixture spectra, experimental protocols designed to generate mixture spectra may lead to many interesting applications. There are already several alternative data acquisition approaches that rely on mixture spectra to overcome the limitations of current instrument scan rates [10, 11, 12]. Mixture spectra also arise from peptides that are covalently linked in the sample, examples include disulfide bridges [41], SUMOlyated peptides [42] and peptides from cross-linking experiments [43, 44]. The identification of these cross-linked peptides can

provide valuable information on protein structures and interactions. Thus solving the problem of peptide identification from mixture spectra represents an important step toward addressing related and emerging problems in proteomics.

1.3 MixGF: computing statistical significance for mixture spectra matches

Previous sections shows that MixDB can accurately and efficiently identify mixture spectra and comparison with other database search methods shows that MixDB can identify significantly more mixture spectra [17]. However it still only identifies about half as many mixture spectra as our mixture spectral library search tool M-SPLIT [36]. Comparison of the results between the two methods reveals that current version of MixDB suffers from relative low sensitivity due to its limited ability to separate true matches from false positives - still an ongoing research problem even for the case of single-peptide spectra [18, 21, 20, 21]. Capitalizing on recent advances using a generating function approach to rigorously compute the statistical significance of PSM, we propose to extend the MS-GF [21] approach. Below we first show how to rigorously compute statistical significance for a PPSM in mixture spectra. Applying this approach (MixGF) on both simulated and real datasets, we show that MixGF improves the sensitivity of identification of mixture spectra by 26-76% over MixDB [17] and by 110-162% over ProbIDtree [16].

1.3.1 Scoring function for mixture spectrum

We represent a tandem mass (MS/MS) spectrum as a real-value vector with N bins where each bin represents a mass interval of width δ Da (δ depends on instrument resolution). For an experimental spectrum: $V = v_1...v_N$, each v_i is the sum of the intensity of all peaks fall into the ith bin. We represent a peptide P as a set of prefix residue masses (PRMs) which is defined as the sum of amino acid mass for each peptide prefix. For P with PRMs: $p_1....p_n$, we define p_n as the parentmass and note that for each p_i, we can compute a set of theoretical fragment ions, T_{pi}, that can be generated from the peptide. For example, if we consider only b and y ions, a singly charged b-ion will have mass

$p_i + 1$ and a singly-charged y-ion will have mass $p_n - p_i + 1$. Therefore, the set of all theoretical fragment ions from a peptide, T, can be considered as the union of all T_{p_i}: $T = T_{p_1} \cup T_{p_2} ... \cup T_{p_n}$. A probabilistic scoring model such as that described in [35] defines the score for a Peptide-spectrum-match (PSM) as the sum of scores of matching each theoretical fragment ion against the observed peak in the spectrum. Using such additive scoring model we can then compute a combined score for a set of theoretical fragment ions, T_{p_i}, and associate this score with the corresponding PRM, p_i. Since the above operation can be done for any PRM, we can construct a PRM spectrum as a spectrum: $S = s_1 ... s_N$ where each mass bin i has a score s_i that represents the log-likelihood that the peptide generated the observed spectrum has a prefix mass i(see [45] for details). For a peptide P with prefix masses $p_1 .... p_n$ the score of matching it against a spectrum is the sum of all the scores at its prefix residue masses in the PRM spectrum: SCORE(P, S) = $s_{p_1} + s_{p_2} ... + s_{p_n}$.

We define a mixture spectrum as a spectrum from two different peptides. When interpreting an MS/MS spectrum as a mixture spectrum M, we construct two PRM spectra, M^H and M^L, to represent the two scoring models for the high and low-abundance peptides present in the mixture spectrum, respectively. As we showed in MixDB [17], different scoring models are needed for high and low-abundance peptides because they have quite distinct fragmentation statistics in mixture spectra. Without loss of generality, when matching a mixture spectrum (M) against a pair of peptides (P, Q) we assume the first peptide is the high-abundance peptide. Thus the score of a pair of peptides (P, Q) against a mixture spectrum M will be the sum of scoring P with M^H and scoring Q with M^L: SCORE(P, Q, M) = $M^H_{p_1} + ... M^H_{p_n} + M^L_{q_1} + ... M^L_{q_n}$. To avoid double counting, when a prefix mass of P is the same as a prefix mass of Q, only the bin with the higher score is considered and the other peptide gets a score of zero for that particular mass position: $when\ p_i = q_j : if(M^H_{p_i} > M^L_{q_j})\{M^L_{q_j} = 0\} else \{M^H_{p_i} = 0\}$.

1.3.2 Spectral probability for a mixture spectrum

The statistical significance of a particular peptide P matched to a spectrum S with score T is determined by the probability that a random peptide R (out of all possible peptides) when matched to S has a score greater or equal to T: $\Pr(SCORE(R, S) > T)$. From here on we will refer to this as the *Single-peptide probability* in order to better distinguish it from the other definitions introduced below. Analogously, to compute the statistical significance of a particular peptide pair (P,Q) matched to a mixture spectrum (M) with a score of T, we are interested in two statistical questions: 1)*Joint probability* $\equiv \Pr(SCORE(R_1, R_2, M) > T)$: the probability that a random peptide pair (R_1, R_2) (out of all possible peptide pairs) when matched to M yields a score greater or equal to T and 2) *Conditional probability* $\equiv \Pr(SCORE(R_1, R_2, M) > T \mid R_1 = P)$: given a peptide P, the probability that a random peptide R_2 (out of all possible peptides) together with P when matched to M yields a score greater or equal to T. Intuitively a Peptide/peptide spectrum match (PPSM) can fall into three categories: 1) *Correct-matches*: both peptides are correct matches; 2) *Half-correct matches*: one peptide is correct and the other peptide is an incorrect match; 3) *Incorrect-matches*: both peptides are incorrect matches. We are interested in separating the correct matches from incorrect and half-correct matches. The formulation above addresses this question in two steps. The joint probability assesses the chance that two random peptides can have the same or higher score than the current best match. When this probability is very low, this means that at least one peptide is a statistically significant match to the spectrum (i.e. correct or half-correct match). Once we assume that at least one peptide is a true match, the conditional probability helps us evaluate whether the second peptide is also a statistically significant match (i.e. correct matches). In summary, one is looking for PPSMs with both low joint probability and conditional probability.

In order to compute the probabilities mentioned above we need to know the score distribution for all possible peptides and peptide pairs. The original MS-GF [21] approach uses dynamic programming to compute the single-peptide probability efficiently without explicitly consider the scores for all peptides. Here

we extend this generating function approach to compute the probability for joint and conditional probability. Let J_M be a three-dimensional dynamic programming matrix where each element $J_M(p_i, q_j, T)$ stands for the joint probability that a pair of peptides P, Q with parent mass p_i and q_j match to M with score higher than or equal to T. When there are no shared peaks between P and Q this means P matches to M^H up to the p_i^{th} bin and Q matches to M^L up to q_j^{th} bin. We can define the following recurrence relationship for computing joint probability:

$$J_M(p_i, q_j, T) = \left\{ \begin{array}{l} if\ p_i < q_j: \\ \quad \sum_{all\ amino\ acids\ a} J_M(p_i, q_j - mass(a), T - M^L_{q_j}) \times prob(a) \\ if\ p_i > q_j: \\ \quad \sum_{all\ amino\ acids\ a} J_M(p_i - mass(a), q_j, T - M^H_{p_i}) \times prob(a) \\ if\ p_i == q_j: \\ \quad \sum_{a_1} \sum_{a_2} J_M(p_i - mass(a_1), q_j - mass(a_2), T - max \left\{ \begin{array}{c} M^H_{p_i} \\ M^L_{q_j} \end{array} \right\}) \\ \qquad \times prob(a_1) \times prob(a_2) \end{array} \right\}$$

In the equation above a, a_1, and a_2 denote amino acids; $mass(a)$ denotes the mass of an amino acid and $prob(a)$ denotes the probability that a particular amino acid occurs in a peptide. When considering all possible peptide sequences this probability is uniform and has a value of $\frac{1}{20}$ for each of the 20 standard amino acids. To better reflect the amino acid composition observed in real protein sequences we can also obtain this probability by computing the frequency of each amino acid in the protein sequence database against which the spectra are searched. To start the computation of the recurrence, we initialize $J_M(0, 0, 0) = 1$.

The computation of the conditional probability is very similar to that of single-peptide probability, except that it is conditioned on the first peptide being accepted as a match. Specifically, for a peptide pair (P, Q) matched to a spectrum M with score T, we define that peptide P and Q contribute T_P and T_Q to the total score, respectively. Assuming that peptide P was matched to M, we define a two-dimensional dynamic programming matrix C_M where each element, $C_M(q_j, T|P)$,

represents the conditional probability that a peptide with parent mass q_j together with P match M with a score greater than or equal to T. To compute this probability, we first modify M^L by setting all the bins corresponding to a prefix mass of P to zero if M^H has a higher score at the same location. Then Conditional probability can be computed using the following recursion:

$$C_M(q_j, T|P) = \sum_{all\ amino\ acids\ a} C_M(q_j - mass(a), T - M^L(q_j)|P) \times prob(a)$$

We initialize the recurrence with the base case: $C_M(0, T_P|P) = 1$. The base case starts at score T_P rather than zero because the first peptide P already contributes T_P to the total score.

We note that even though the joint probability assesses whether at least one peptide is a significant match to the spectrum, it does not determine which peptide is the significant match in the case only one peptide is significant match. More importantly, when calculating the conditional probability one assumes that the first peptide is a true match but it is unclear which peptide is the first peptide from the joint probability assessment. In order to resolve this ambiguity, for a candidate peptide pair (P,Q) matched to a spectrum M, we compute their respective single-peptide probabilities and the peptide with lower (i.e. statistically more significant) single-peptide probability is designated as the first peptide.

1.3.3 Approximating joint probability

The dynamic programming approach above enables us to compute the Joint probability without explicitly computing the scores for all peptide pairs. However the computational complexity still scales exponentially with the number of peptides that possibly generated the observed spectrum (e.g., quadratic for two peptides). Thus it is desirable to find a way to approximate this probability efficiently. To derive this approximation we borrow an intuition from the definition of conditional probability where the joint probability of two random events (R_1, R_2), is equal to the probability of one event times the conditional probability of the second event given the first event: $Prob(R_1 \wedge R_2) = Prob(R_1) \times Prob(R_2|R_1)$. Analogously we

can decompose the joint-probability question into two simpler questions: 1) what is the probability of finding a random peptide that matches to M with a score equal or better than T_P? and 2) once we find a first peptide P what is the probability of finding a random peptide that together with P scores equal or higher than T when matched to M? Note that the first question is just the single-peptide probability and the second question can be answered with the conditional probability. Therefore we can define the following approximation: $\Pr(SCORE(R_1, R_2, M) > T) \approx \Pr(SCORE(R_1, M) > T_P) \times \Pr(SCORE(R_1, R_2, M) > T | R_1 = P)$. From here on we refer to this approximation as the *Product-probability*. While this formulation is not exactly equivalent to the definition of joint probability (because it does not explicitly consider the dependencies between all possible *pairs* of peptides that can be matched to the mixture spectrum), both single-peptide probability and conditional probability can be computed efficiently in linear time and we show in the next section that this approximation is sufficiently accurate for our main use of joint-probability – to separate correct from incorrect matches to mixture spectra.

1.3.4 Classification of matches

Since a typical proteomic dataset contain both single-peptide and mixture spectra we consider three possible outcomes when searching a given query spectrum M: 1) *No-match:* M does not match any peptide in the database; 2) *Single-peptide match:* M matches one peptide in the database and 3) *Mixture match:* M matches a pair of peptides in the database. We start out by assuming that each query spectrum is a putative mixture spectrum and considered its top-scoring PPSM. Then a two-step procedure is used to separate true mixture matches from false mixture matches. At the first stage, all PPSMs with joint probability greater than a threshold are filtered out and considered as incorrect mixture matches. Then PPSMs with conditional probability greater than a threshold are also filtered out and are considered as half-correct mixture matches. The probability thresholds are determined such that it enforces a particular false discovery rate (FDR, see next section). All PPSMs that pass both the joint and conditional probability threshold are considered as Mixture-matches. Next all the remaining spectra that

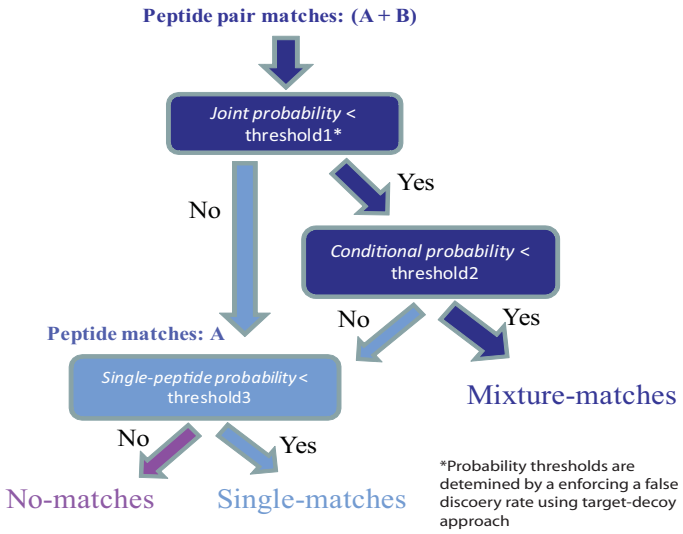

Figure 1.12: **Classification of matches**: All query spectra are first assumed to be putative mixture spectra and their top-scoring PPSMs are considered. PPSMs with joint and conditional probability passing a particular threshold are classified as *Mixture-matches* – spectra that match to two peptides in the database. The probability threshold for each joint and conditional probability is determined by enforcing a particular FDR using a target/decoy approach. Next, spectra that do not pass either probability threshold are treated as single-peptide spectra by considering the first peptide in each PPSM as the peptide match to the spectrum. Then single-peptide probabilities are calculated and those with probability passing a FDR-determined threshold are considered as *Single-matches* – spectra that match to only one peptide in the database.

do not pass either probability threshold are re-considered as single-peptide spectra. Each PPSM is converted into a PSM by considering the first peptide as the match to the spectrum. Single-peptide probabilities are computed for all PSMs and a probability threshold is determined to enforce a FDR for single-peptide spectra. The workflow of this classification procedure is illustrated in Figure 1.12.

1.3.5 Estimation of False Discovery Rates

In the classification step of PPSMs, the probability thresholds are determined such that they enforce a certain FDR. For the joint probability we are interest in the FDR that an incorrect mixture match is accepted as half-correct or correct matches: $FDR_{Joint} = \frac{\#incorrect}{\#correct+\#half-correct}$. For the condi-

tional probability we only want to accept correct matches thus we are interest in the FDR that half-correct or incorrect matches are accepted as correct matches: $FDR_{Conditional} = \frac{\#incorrect+\#half-correct}{\#correct}$. Each of the above FDR can be estimated by extending the Target-Decoy Approach (TDA) for single-peptide spectra [30]. However, the assumptions used in TDA first need to be generalized to the case of PPSMs and their validity also need to be tested. The detailed derivation of the TDA approach for PPSM will be described in next section, we first summarize the main results below. In a set of PPSMs if we define TT to be the number of PPSMs where both peptide matches are from the target database; TD or DT to be the number of cases one peptide is from the target while the other peptide is from the decoy database and DD to be the cases where both peptides are from the decoy database, the two FDRs mentioned above can be computed using the following formula:
$FDR_{Joint} = \frac{DD}{TT}$
$FDR_{Conditional} = \frac{1/2(TD+DT)}{TT}$.

1.3.6 Derivation of TDA for mixture spectra

False Discovery Rate (FDR) can be estimated by extending the Target-Decoy Approach (TDA) for database search [30]. Each top scoring peptide-peptide-spectrum match (PPSM) can be one of the following type: TT – both peptide matches are from the target database, TD or DT – one peptide is from the target while the other peptide is from the decoy database and DD – both peptides are from the decoy database. A peptide from the target database can be either a correct (C) or incorrect match (I) and a peptide from a decoy database is by definition an incorrect match. Therefore matches in each type can be further divided into subtypes: for example, TT can be divided into TT^{CC}, TT^{CI}, TT^{IC}, and TT^{II} where the superscript indicates whether the peptide match is correct or not. We can write the number of PPSMs belonging to each type as a sum of

PPSMs belonging to its subtypes:

$$TT = TT^{CC} + TT^{CI} + TT^{IC} + TT^{II} \quad (1.1)$$

$$TD = TD^{CI} + TD^{II} \quad (1.2)$$

$$DT = DT^{IC} + DT^{II} \quad (1.3)$$

$$DD = DD^{II} \quad (1.4)$$

TDA assumes that an incorrect peptide match has equal chance of coming from the target or the decoy database. Therefore matches of II type has equal chance of being TT, TD, DT and DD, making the number of II matches in equation 1-4 approximately the same: $DD^{II} = DT^{II} = TD^{II} = TT^{II}$. By similar argument, the number of matches of type CI in equation 2 and the number of matches of type IC in equation 3 should roughly be same as those in equation 1. Hence to extend the TDA to mixture spectra we made the following assumptions:

$$DD = DD^{II} = DT^{II} = TD^{II} = TT^{II} \quad (1.5)$$

$$TT^{CI} = TD^{CI} \quad (1.6)$$

$$TT^{IC} = DT^{IC} \quad (1.7)$$

To test whether these TDA assumptions hold true, we constructed a set of simulated mixture spectra (see next section) and extracted the top-scoring matches of type II, CI and CI returned by MixDB. Then we computed the relative frequency of each peptide match being from the target or decoy database. As shown in Figure 1.13a, matches of type II has $\sim 25\%$ chance of being TT, TD, DT, and DD. Figure 1.13b shows that within range of random variation, matches of type CI has equal probability of being TT and TD while Figure 1.13c shows that matches of types IC, has equal chance of being TT and DT. Taken together, these results show that the TDA assumption can be generalized to mixture spectra.

By substitution and rearranging terms, we can redefine the CI and IC term in equation 1 to be:

$$TT^{CI} = TD - DD \quad (1.8)$$

$$TT^{IC} = DT - DD \quad (1.9)$$

As described in the Method section, for MixGF, two different FDRs needed to be computed, one is used to determine the probability threshold for joint probability and the other for determining the threshold for conditional probability. For joint probability we want to accept PPSMs that are of the type TT^{CC}, TT^{CI}, TT^{IC} and reject matches of the type TT^{II}. Thus we are interest in controlling the following FDR: $FDR_{Joint} = \frac{TT^{II}}{TT^{CI}+TTIC+TT^{CC}}$. For conditional probability one wants to accept matches of the type TT^{CC} and reject matches of the other types. The FDR of the conditional probability can then be defined as: $FDR_{Cond} = \frac{1/2(TT^{IC}+TT^{CI})+TT^{II}}{TT}$. The $1/2$ in the equation accounts for the fact that matches of IC and CI type contribute one correct match and one incorrect match. Substituting the terms defined above, we get:

$$FDR_{Joint} = \frac{DD}{TT} \tag{1.10}$$

$$FDR_{Conditional} = \frac{1/2((TD - DD) + (DT - DD)) + DD}{TT} \tag{1.11}$$

$$= \frac{1/2(TD + DT)}{TT} \tag{1.12}$$

1.3.7 Testing the TDA assumption for mixture spectra

In order to test whether the TDA assumption can be extended to the case of mixture spectra, we generated a set of simulated spectra of type CI, IC and II. We started with two NIST spectral libraries [40], one from Yeast and one from E. Coli. Then we removed from the E. Coli spectral library any entries that have a peptide matched to protein sequences in Yeast. This is to ensure that any peptide matches to a spectrum from the E. Coli library is an incorrect match when searching a Yeast protein sequence database. Mixture spectra of type II are simulated by linearly combining two spectra from the E.Coli library while mixture spectra of type CI and IC are simulated by linearly combining one spectrum from the Yeast library and one spectrum from the E. Coli library. All the simulated mixture spectra were searched against a Yeast protein sequence database using MixDB [17]. The top-scoring match of type II. CI and IC are extracted. Then we compute the frequency that each match is from TT, TD, DT and DD respectively. Each experiment was performed on a set of three thousand simulated

mixture spectra (one thousand for each type: II, CI, CI). The experiment was repeated twenty times to estimate the random fluctuations in the data.

1.3.8 Datasets and Data Processing

The performance of MixGF was first evaluated on a set of simulated mixture spectra. As described before [36], mixture spectra were created by linearly combining two single-peptide spectra with mixture coefficients selected from 0.3 to 1.0. Mixture coefficient is a parameter that reflects the relative abundance of the two peptides present in the mixture spectrum. In addition, MixGF was tested on two experimental datasets. In brief, the *Yeast dataset* [31] is from a tryptic digest of Saccharomyces cerevisiae that was analyzed on an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific) and MS/MS spectra were acquired using a data-dependent scanning mode in which each full MS scan (m/z 300–2000) was acquired on the Orbitrap at resolution 60,000, followed by 8 MS/MS scans collected on the LTQ (see [31] for full details). The *Human dataset* is from a tryptic digest of HEK293 cell lysate that was fractioned using Strong Cation Exchange(SCX) and each fraction was analyzed by a LTQ Orbitrap XL ETD mass spectrometer (Thermo Fisher Scientific) in data dependent mode. MS full scan were acquired from m/z 350–1500 with a resolution of 60,000. The two most intense ions were fragmented in the linear ion trap using CID and ETD (see [46] for details). For the Human dataset only CID spectra were considered. All database searches were performed with precursor mass tolerance of 3.0 Da and fragment mass tolerance of 0.5 Da. For MixDB we used all the default parameters as described in [17], except that the FDR for mixture spectra was computed using the formula described in the previous section. For ProbIDtree, we separated all the spectra into two sets depending on whether ProbIDtree identified only one or multiple peptide matches to the spectrum, then an FDR was enforced using the standard TDA method [30] for each subset. The rationale for separate FDR determination is that we've shown previously that a FDR computation that combine both mixture and single-peptide spectra will leads to an underestimation of the FDR for mixture spectra [17]. The protein sequence databases used are the SGD yeast pro-

tein database (*ver.5/8/2009*) and the Human protein database (downloaded from NCBI refseq, *ver.10/29/2010*).

1.3.9 Separating true and false mixture spectrum matches

As described above, our main goal of computing the statistical significance of PPSMs is to separate correct mixture matches from the half-correct and incorrect ones. To test MixGF's ability in these tasks, we constructed a set of simulated mixture spectra by linearly combining pairs of single-peptide spectra. Since we know a priori the peptides that generated each simulated mixture spectrum, we can extract the top-scoring correct, half-correct and incorrect matches returned by MixDB and compute their joint and conditional probabilities. As shown in Figure 1.14a, joint probability does a very good job at separating correct matches from incorrect matches. However there is considerable overlap between the joint-probability of correct-matches and that of half-correct matches (see Figure 1.14b). Further investigation of cases in the overlap region shows that for correct-matches usually both peptides contribute moderate scores to the final combined score. On the other hand for the half-correct matches the correct peptide often contributes a very high score and thus even when paired with an incorrect match, the resulting combined high score still yields a low joint-probability. Intuitively in order to separate half-correct matches from correct matches we need to look for cases that have high combined score as well as both peptides contributing significantly to the total score. The concept of conditional probability defined above aims to address exactly this question - is the score of the peptide pair (P, Q) significantly higher than that of the single peptide P? As illustrated in Figure 1.14c, conditional probability indeed is a better feature to separate correct matches from half-correct matches. Therefore a two-step procedure is used to separate correct matches from false matches: at the first stage of MixGF, joint-probability is used to filter out incorrect matches and then conditional probability is used to filter out half-correct matches.

1.3.10 Joint-probability improves the detection of mixture spectra.

For mixture spectra, we expect joint probability to perform better at separating correct matches from incorrect matches by explicitly considering two peptides. Intuitively we expect that single-peptide probabilities for correct peptide matches to mixture spectra to be higher (i.e. worse) than those for correct matches to single-peptide spectra. This is because the presence of a second peptide in mixture spectra which will allow more peptides to match to the spectrum with high score. However for false matches, the single-peptide probability distribution remains roughly the same for both single-peptide and mixture spectra because they are random matches in either case. Therefore the distribution of single-peptide probabilities between correct and incorrect matches should be less well-separated for mixture spectra than for single-peptide spectra. To show this we generated a series of simulated mixture spectra where the first peptide is mixed with a second peptide at 100%, 50%, 30% of the first peptide's total intensity and then computed the single-peptide probability, joint probability and product probability for the correct matches as well as the top-scoring incorrect matches. The performance of each probability function in separating correct from incorrect matches is shown in Table 1.7. As expected when the second peptide is at relatively low abundance (i.e. 30%), the performance of single-peptide probability and joint probability is similar as the mixture spectrum is more similar to a single-peptide spectrum. However as we increase the relative abundance of the second peptide, joint probability performs considerably better at separating correct matches from incorrect matches. Thus we expect that as mixture spectra with more peptides become more common in experiments, joint probability and its approximation will substantially improve our ability to identify mixture spectra.

1.3.11 Product probabilities accurately estimate joint probabilities

Since it is computationally expensive to compute the joint probability exactly, we seek to approximate it by a product of single-peptide probability and

Table 1.7: A set of simulated paired-peptide mixture spectra were constructed with mixture coefficients $\alpha = 1.0, 0.5, 0.3$. Single-peptide probability, joint-probability and product-probability were computed for the correct matches as well as the top-scoring incorrect matches returned by MixDB. Then each probability was used to separate correct from incorrect matches. The sensitivity of accepting correct matches at different FDR levels is shown.

	Probability	False discovery rate (FDR)		
		1%	2%	5%
$\alpha = 1.0$	Single-probability	71.9	74.7	81.8
	Joint-probability	93.4	94.7	96.6
	Product-probability	93.6	94.0	96.7
$\alpha = 0.5$	Single-probability	85.7	87.5	92.0
	Joint-probability	93.5	94.3	96.0
	Product-probability	92.6	94.1	96.1
$\alpha = 0.3$	Single-probability	89.4	90.8	92.8
	Joint-probability	90.2	91.8	93.8
	Product-probability	90.4	91.7	93.8

conditional probability, both of which can be computed efficiently. Using our simulated mixture spectra dataset, we compare the joint probability and its approximation. As shown in Figure 1.15, in most cases the joint probability is accurately approximated by product probability as most data points clustered tightly along the main diagonal. For correct matches, the product probability is sometimes lower than the true joint probability. This can be attributed to the fact that the approximation does not explicitly consider all pairs of peptides - since P is fixed there are less opportunities for false positive matches to achieve high scores and thus the resulting spectral probability can be smaller in such cases. However, the range of probabilities where this underestimation occurs is well below the range where incorrect matches tend to occur. Therefore for the purpose of separating correct matches from incorrect matches, using the approximation is nearly equivalent to computing the exact joint probability. As shown in Figure 1.15b, correct matches and incorrect matches remain very well-separated whether using the product or joint probability.

1.3.12 MixGF increase the sensitivity of identification of mixture spectra

To illustrate MixGF's ability to identify mixture spectra in practical scenario, we tested it on a yeast and a human dataset [31] that represent typical proteomics analysis of cell lysate. We compared the performance of MixGF with two current state-of-the-art database search methods for identification of mixture spectra: MixDB and ProbIDtree. As shown in Figure 1.16 and Table 1.8 MixGF is able to outperform MixDB and ProbIDtree by identifying 26-76% and 110-492% more mixture spectra, respectively. We also compared different variants of MixGF in which a different probability function was used at stage one to separate correct matches from incorrect matches. They are all followed by using conditional proability at the second stage to separate correct matches from half-correct matches. Note that the performance is very similar when using the joint probability or its approximation (product-probability), further indicating that our approximation of the joint probability is sufficiently accurate as a score for FDR-controlled identifications. It is perhaps a little surprising that using single-peptide probability has comparable performance to using joint probability. As shown in previous section, for mixture spectra, joint probability perform better at separating correct matches from incorrect matches by explicitly considering two peptides. Further analysis shows that this can be explained by the fact that in typical experimental datasets, most mixture spectra have the second peptide at relatively low abundance (we estimated that on average, the low-abundance peptides are at 1/3 of the intensity of the high-abundance peptides [36]). As shown in the Method section when the second peptide in the mixture is at relatively low abundance, the mixture spectrum tends to be more similar to a single-peptide spectrum and thus the ability to separate correct and incorrect matches using either Joint or Single-peptide probability is similar in these cases.

1.3.13 Discussion

It is undoubtful that *mixture* MS/MS spectra from more than one peptide will become increasingly common and important as advancement in technology

Table 1.8: Total numbers of spectra and unique peptides identified by ProbIDtree, MixDB and MixGF at 1% FDR are summarized. 'Single' indicates spectra from which only one peptide is identifed and 'Mixture' indicates spectra from which more than one peptides are identified. For MixGF, IDs are shown for the variant where joint probability is used in the first stage to separate correct mixture matches from incorrect mixture matches.

Dataset	Method	Identified Spectra			Identified peptides		
		Single	Mixture	Total	Single	Mixture	Total
Yeast	ProbIDtree	21807	504	21807	4826	495	4936
	MixDB	25033	748	25778	5702	895	5924
	MixGF	28022	1320	29342	6315	1398	6637
Human	ProbIDtree	28614	1433	30036	8479	1675	9153
	MixDB	38855	5420	44275	13021	5735	15298
	MixGF	39701	7052	46783	13027	6982	16080

and instrument allow us to analyze increasingly large numbers of molecules presented in complex biological samples. These new methods will certainly provide us with a more complete and quantitative view of the proteome. However this will also require the development of accurate computational tools to identify multiple peptides contain inside each MS/MS spectrum. Two fundamental questions that are pre-requisites to build accurate computational tool: 1) To separate correct multiple-peptide-spectrum matches (mPSMs) from false positive ones and 2) To estimate the false positive or discovery rate in a set of MPSMS. Here we try to address these questions by computing the statistical significance of mPSMs. Given a MS/MS spectrum database search tools can always return a top-scoring peptide or multiple peptides matched to the query spectrum. By random chances it is always possible to have some false peptide matches to obtain a high score. This is especially true in the case for mixture spectra because the explosion of the search space will dramatically increase the occurrence of high-scoring false matches. Thus it is crucial to be able to compute the statistical significance of mPSMs accurately. Here we show that for two-peptide cases, it is possible to compute the statistical significance rigorously using the generating function approach and show that the joint and conditional probability are very good features that can separate true PPSMs from false positive ones. In addition we further show that the

computationally expensive joint probability can be approximate accurately using a product of conditional probabilities, which can be computed in linear time. In order to estimate the false discovery rate (FDR) for mixture spectra, we extends the traditional target-decoy approach (TDA). We noted that it is important to perform the database search using a concatenated target-decoy sequence database as this will allow us to estimate the occurrence of half-correct matches where one peptide in the PPSM is correct and one peptide is incorrect. This is important because just by random chance, they constitute a large fraction of false positive matches in mixture spectra as compare to cases where both peptides are incorrect and as we shown in the paper these cases are more difficult to separate from correct matches.

Benchmarking MixGF performance on two datasets that represent typical proteomic experiments, we further showed that the proposed approach is able to identify 25%–415% more mixture spectra and 5%-70% more unique peptides as compared to MixDB and ProbIDtree. It is worthwhile to point out that in the yeast dataset, the number of mixture spectra is only about 4.71% of the single-peptide spectra that were identified. However, in the human dataset where the sample is much more complex, the number of identified mixture spectra increase to be 17.76% of the single-peptide spectra that were identified by MixGF. This shows that as the complexity of the sample increase, mixture spectra indeed constitute a significant fraction of MS/MS spectra observed in the data. Such observation is not new, several previous studies have reported that multiple precursors with similar mass were observed within in precursor isolation window of many MS/MS spectra. However, it remains unclear that whether these co-fragmented peptides are identifiable and here we provide evidence that these peptides can be identified. Since mixture spectra contain two peptides per spectrum, the comparison between mixture spectra and single-peptide spectra is even more striking when we look at the numbers of unique peptides identified. In the yeast dataset even though mixture spectra is only 5% of single-peptide spectra, the number of unique peptides identified in mixture spectra is 22.1% of the total number of peptides identified in single-peptide spectra. For the human dataset the number of unqiue peptides

identified in mixture spectra is even more than half of (53.6%) the number of peptides identified in single-peptide spectra. These observation highlight the importance of moving away from the *one-peptide-one-spectrum* assumption for the next generation of computational tools for identifying MS/MS spectra as instruments advances and more complex samples are being analyzed in proteomic experiments. This also show us the potential of the emerging data-independent-acquisition protocols (DIA) [] where multiple peptide precursors are purposely co-fragmented in the same MS/MS spectra. Imagine that if all of our spectra in the human dataset are mixture spectra, one could have potentially identified twice as many peptides in the same experimental time.

Finally even though our focus in this paper is for mixture spectra that come from two peptides, we did not make any special assumption in developing our approach and thus MixGF should be readily extensible to cases for more than two peptides. Mixture spectra from two peptide are perhaps the simplest mixture spectra, but the concepts developed here to compute statistical significance of mPSMs are the same for cases with more than two peptides and this relative simple settings allow us empirically assess various aspects such as the performance of separating true from false matches and computational efficiency of our approach in computing the statistical significance. Therefore we believed solving the problem for the case of two peptide represents an important step toward addressing the more general scenario of mixture spectra from any number of peptides.

1.4 Acknowledegments

The author would like to thank NIST, ProteomeCommons and the University of Vanderbilt for the publicly availability of the mass spectrometry data used in this research and the reviewers for insightful suggestions. Chapter 1, in part, is a reprint of the material as it appears in J. Wang, J. Perez-Santiago, J.E. Katz, P. Mallick, and N. Bandeira. “Peptide identification from mixture tandem mass spectra”, Molecular & Cellular Proteomics, 9(7):14768, 201 and J. Wang, P.E. Bourne, and N. Bandeira. “Peptide identification by database search of mixture tandem mass spectra”, Molecular & Cellular Proteomics, 10(12), 2011. It is also, in part,

has been submitted for publication of the material as it appear in J. Wang, P.E. Bourne, and N. Bandeira. “Spectral probabilities for mixture tandem mass spectra of more than one peptides”. The dissertation author was the primary investigator and author of this material.

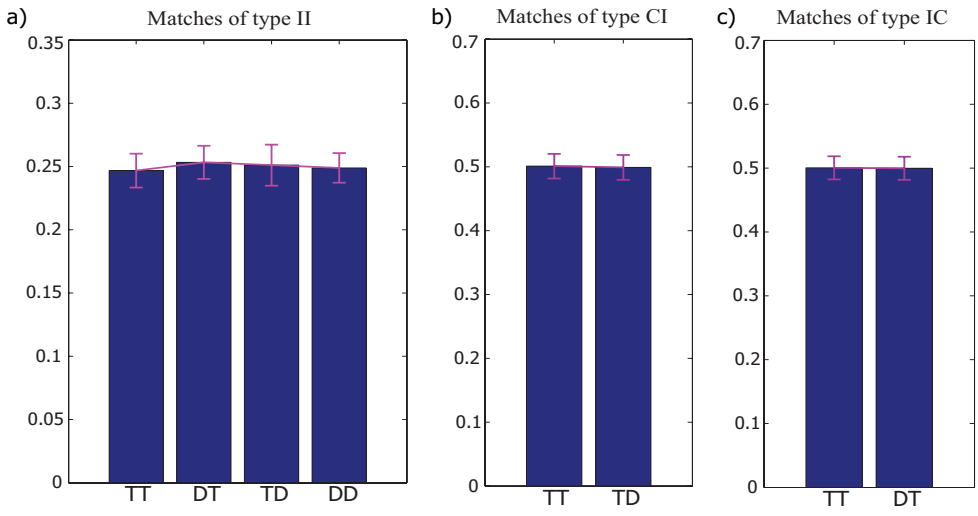

Figure 1.13: Target/Decoy Approach (TDA) for mixture spectra: TDA assumes that for an incorrect peptide-spectrum-match (PSM), the peptide has equal chance of coming from the target (T) or the decoy (D) database. For a peptide-peptide-spectrum match (PPSM) there are three kinds of incorrect matches: II–where both peptides are incorrect and CI/IC – where one peptide is a correct match and the other peptide is an incorrect match. Extending the TDA assumption to mixture spectra will imply that a match of type II will have equal chance of being TT, TD, DT or DD. Similarly a match of type CI will have equal chance of being TT and TD while a match of type IC will have equal chance of being DT and TT. In order to test these assumptions, we constructed a set of simulated mixture spectra and extracted the top-scoring matches of type II, CI, and IC returned by MixDB. Then we computed the relative frequency of each peptide being from the target (T) or the decoy (D) database. As shown in the figure, within the range of random variations, each incorrect peptide match (I) has approximately 50% chance of being from the target or the decoy database, confirming that the TDA assumption can be generalized to mixture spectra.

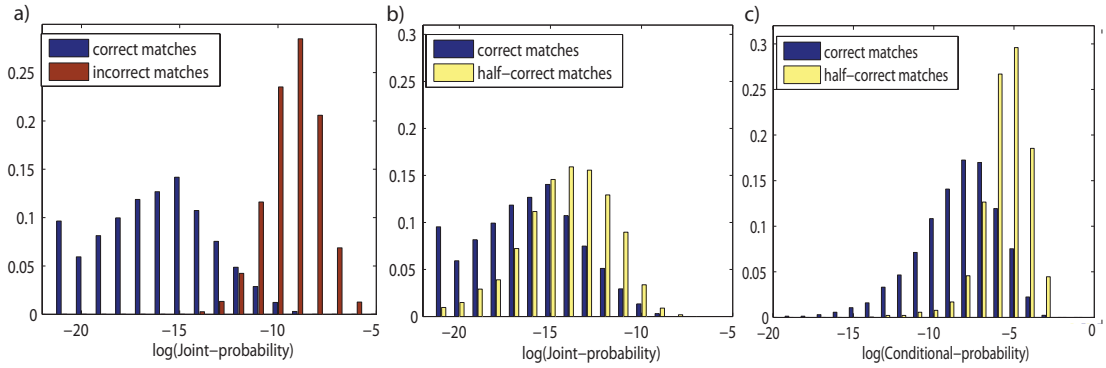

Figure 1.14: Separating true matches from false matches: Mixture spectra were simulated by a linear combination of two single-peptide spectra. Correct matches are cases where both peptides in a PPSM are correct; incorrect matches are cases where both peptides are incorrect and half-correct matches are cases where one peptide is correct and one peptide is incorrect. The distribution of Joint-probability and conditional probability for correct matches (blue bars) incorrect matches (red bars) and half-correct matches (yellow bars) are shown. As shown in a), the distributions of Joint-probability are well-separated between correct and incorrect matches. However, there is considerable overlap between the Joint-probability distribution of correct matches and half-correct matches (see b). On the other hand, conditional probability is a better approach for separating correct matches from half-correct matches as shown in c).

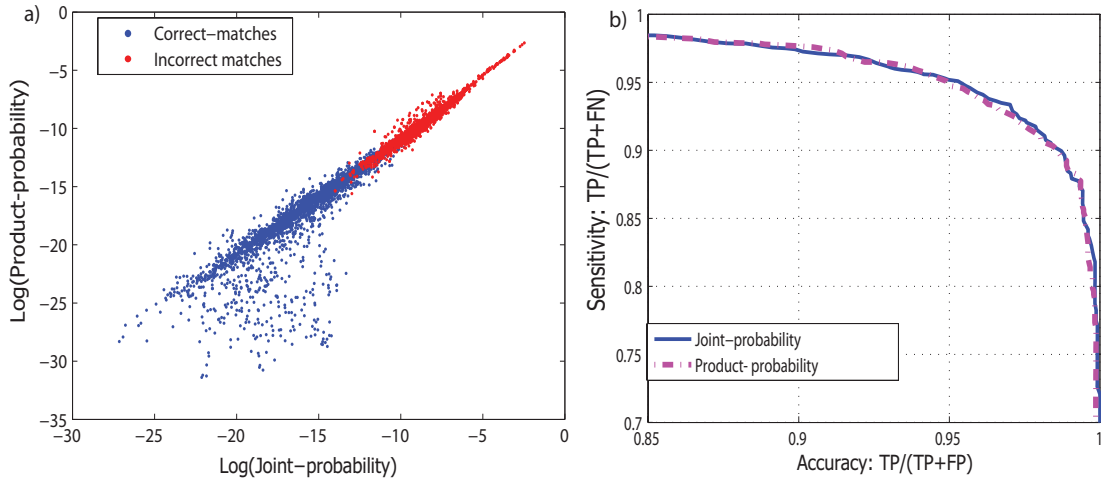

Figure 1.15: Approximation of joint probability: Since it is computationally expensive to compute the exact joint probability for a mixture spectrum (scales exponentially with the number of peptides), we approximate it using joint probability defined as the product of Single-peptide and Conditional probabilities, which can be computed in linear time. As shown in a): for most cases the Product-probability accurately approximate the joint probability (most points cluster tightly around the main diagonal). For correct-matches, there are some cases falling below the main diagonal. This shows that the approximated probability is sometimes lower than the true joint probability for correct matches. However, for the practical purpose of distinguishing correct-matches and incorrect-matches, the two distributions remain very well separated using either the exact joint probability or its approximation as shown in the Precision/Recall curve on the right.

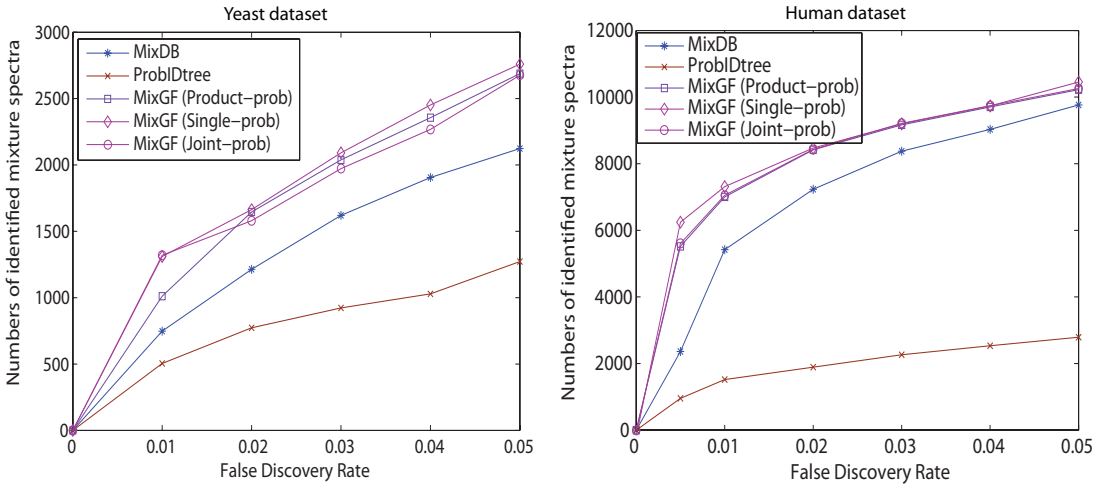

Figure 1.16: Identification of mixture spectra in yeast and human dataset: Numbers of mixture spectra identified by ProbIDtree, MixDB and MixGF are compared. The different variants of MixGF differ by the probability (indicated in parenthesis) that is used in the first stage to separate correct matches from incorrect matches. It is always followed by using conditional probability in the second stage to separate correct matches from half-correct matches.

Chapter 2

Identification of peptides with complex posttranslational modification

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 [3]. 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 [47, 48]. 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 [49, 50, 51, 52, 53]. 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 [54], Small Ubiquitin-like Modification (SUMOylation) [55] PUPylation [56] and AD-Pribosylation [57], 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, 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 PT-Modified peptides identification tool), to derive new algorithms for any type of modified peptide fragmentation. 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 attached to substrate proteins to modify their functions. SUMOylation have been shown to be involved in many cellular pathways such as cellular trafficking, cell cycle, DNA repair and replication [58]. It is also implicated in several neurodegenerative diseases such as Alzheimer's disease and Huntington disease [59, 60]. 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)] [61] 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 [62]. It has also been observed in many cases that SUMOylation can occur on lysine residues not located within any pre-determined 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-shape-like' linked peptides (see Figure 2.1a). 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 which are unusual in unconjugated tryptic peptides. In addition, the SUMO tag resulting from tryptic digestion is relatively large, ranging from 20–30 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/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[63, 64, 65, 66, 67]. 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 [68], 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 trained on 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 [69, 70]. 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.

2.1 Method overview and results

2.1.1 Fragmentation pattern of SUMOylated peptides

In order to obtain a large MS/MS dataset with identified SUMOylated peptides we designed and synthesized three combinatorial peptide libraries, each with a SUMO C-terminus tag (QQQTGG) attached via a lysine residue at different position along the peptide (see Figure 2.2). 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 2.2). 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 dataset 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 Supplementary Figure 2.3, SUMO fragment ions can contribute a large fraction of the total intensity in MS/MS spectra, ranging from 10-60% 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 2.1b). 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 2.4a).

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 2.1c). Linked fragment ions are from peptide fragments which 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 resulted in fragmentation patterns similar to those of common, unlinked peptides (see Supplementary Figure 2.4b) while linked-fragments resulted in fragmentation patterns substantially different from those of unlinked peptides (see Supplementary Figure 2.4c). 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 which 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.

2.1.2 Identifying SUMOylated peptides in combinatorial peptide library

To benchmark our new database search method, 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). Because the training and testing data were the same for Specialize in this case, we split the data evenly into two subsets and trained Specialize on one subset and tested it on the other subset to avoid overfitting (i.e., 2-fold cross validation).

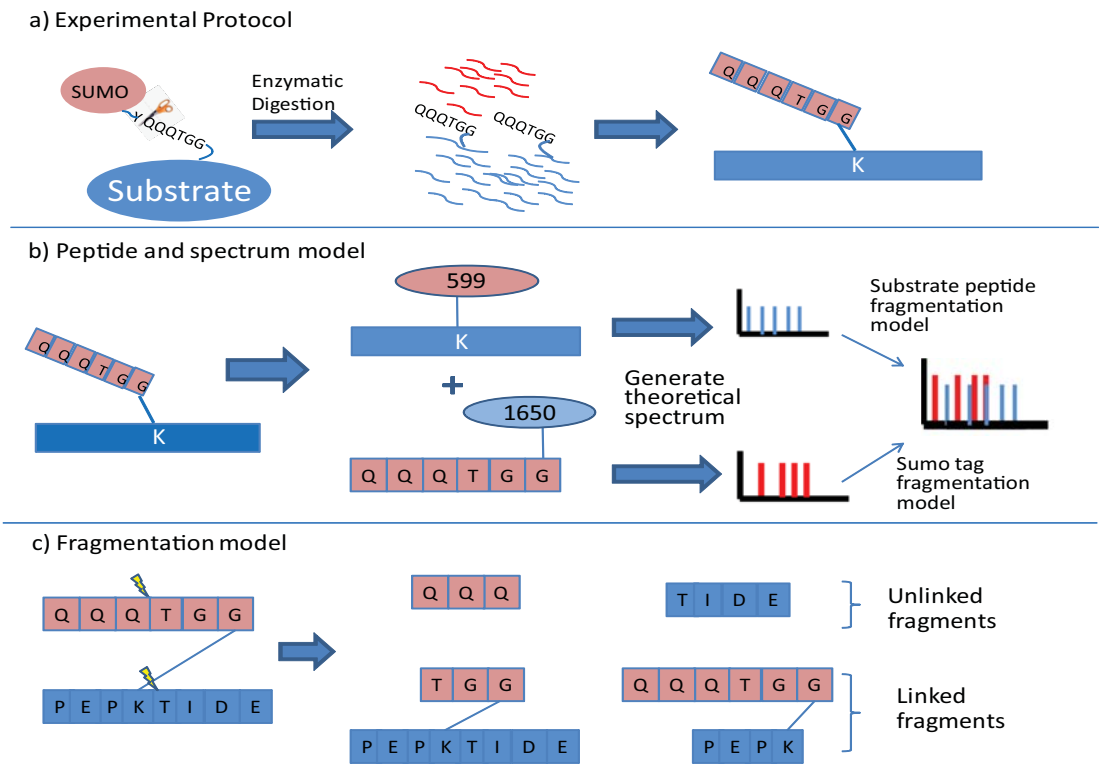

Figure 2.1: Conceptual model of SUMOylated peptides a) Small Ubiquitin-like Modifiers (SUMO) are small proteins that reversibly attach to substrate proteins to regulate their functions. Upon enzymatic digestion of SUMO-conjugated proteins, peptides that contain the SUMO conjugation site in the substrate 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.

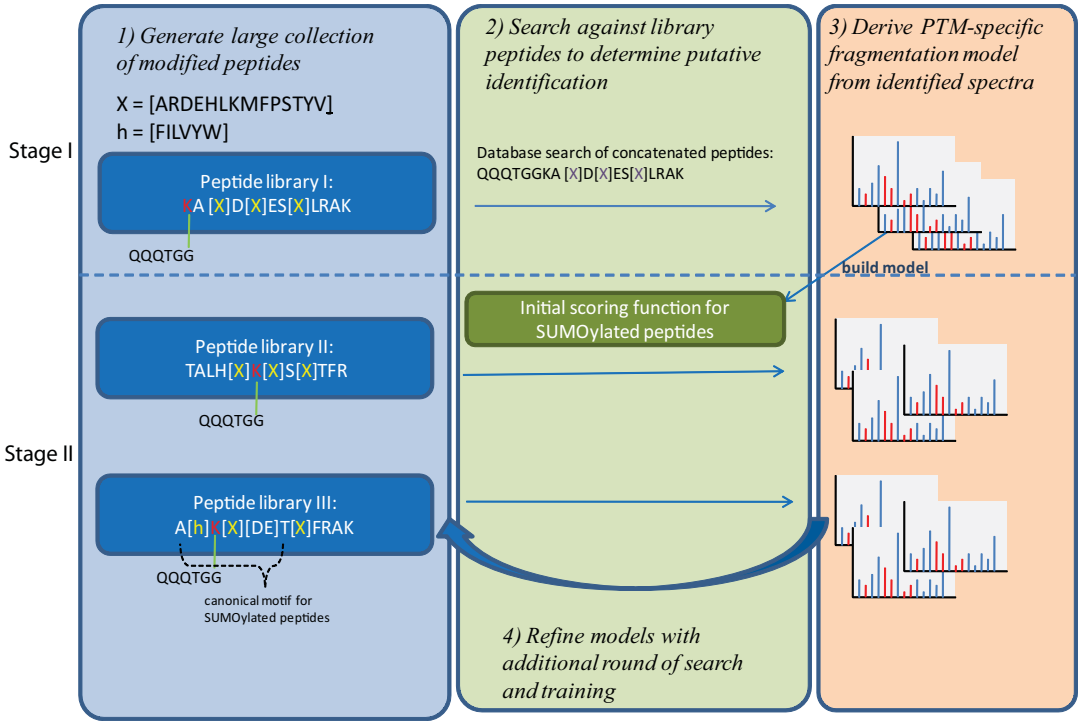

Figure 2.2: Generating training data using combinatorial synthetic peptide libraries We designed and synthesized three combinatorial peptide libraries, each with a SUMO tag (QQQTGG) attached via a lysine residue at a different position along the library peptide. The sequence pattern for each library 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.

As shown in Table 2.1, Specialize identified three to seven times more MS/MS spectra from SUMOylated peptides than InsPecT. To provide some perspective, we also ran InsPecT on a Yeast dataset [31] representing a typical proteomics ex-

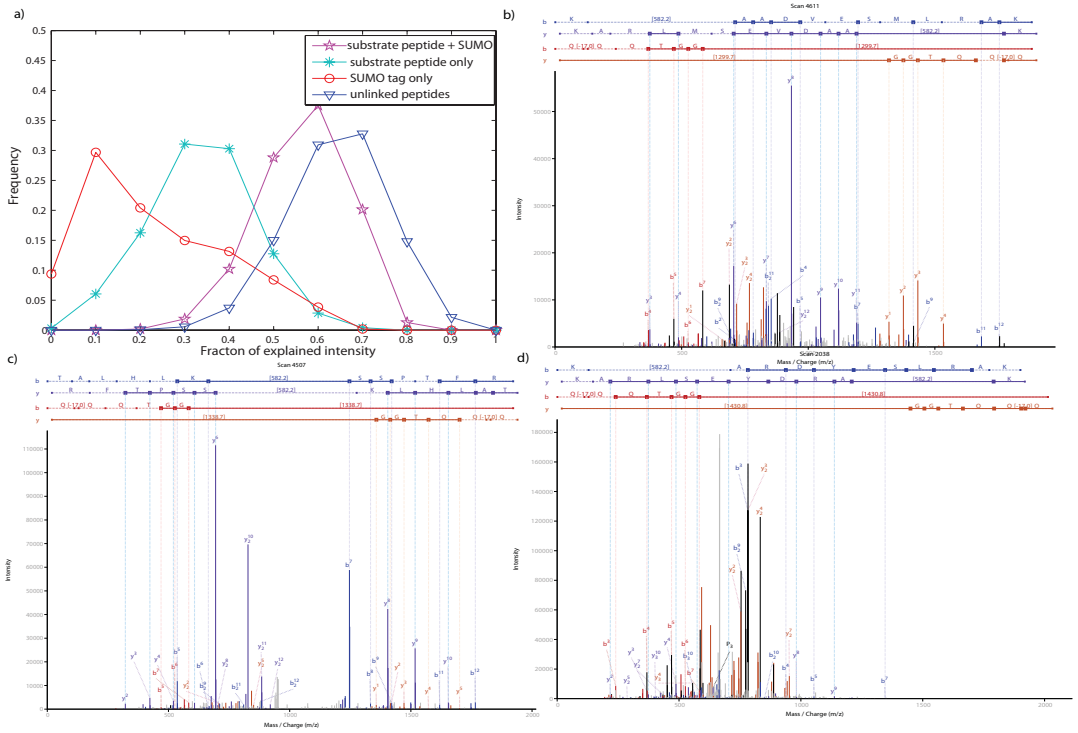

Figure 2.3: Contribution of ion intensity from SUMO tag We computed the fraction of ion intensity corresponding to fragment ions from the SUMO tag in our training data. As shown in a) the fragment ions from the SUMO tag contribute a significant fraction (10-60%) of total intensity in the MS/MS spectra (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) show 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 in red. Peaks correspond to neutral-losses or explained by both substrate peptide and SUMO tag are colored in black. Figure d) shows an example in which peaks from the SUMO tag dominates the observed MS/MS spectra.

periment designed for linear, unlinked peptides. Out of the 76,177 MS/MS spectra in the Yeast dataset, 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 2.1), substantially lower than that of the Yeast dataset even though the combinatorial peptide libraries are much less complex samples than the Yeast lysate. This sup-

ports 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 19-44.7%, which is comparable to InsPecT's identification rate for linear, unlinked peptides.

Table 2.1: 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 are shown. Numbers inside the parenthesis indicate the identification rate which is the percentage of total number of spectra that are identified. The Yeast dataset 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.

Numbers of 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%)
MXDB	2320 (19.0%)	4967 (44.7%)	2929 (24.2%)	n/a
# of spectra	12202	11113	12177	76177

	Numbers of unique SUMOylated peptide identified		
	Library I	Library II	Library III
InsPecT	543	941	385
MXDB	1018	1404	1070

2.1.3 Identifying SUMOylated peptides from cell lysate

In order to demonstrate Specialize's ability to process biological samples, we synthesized twenty 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

eighteen out of the twenty SUMOylated peptides synthesized. The remaining two peptides which 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 dataset). From this dataset Specialize was able to identify thirteen unique MCL1 SUMOylated peptides while InsPecT was able to identify three unique SUMOylated peptides. To estimate the sensitivity of identifying SUMOylated peptides, identified MS/MS spectra from the pure MCL-1 dataset 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. Spectral library search identified a total of 16 unique MCL1 SUMOylated peptides in the Jurkat lysate sample, thus this indicates that Specialize has a sensitivity of approximately $13/16 \approx 80\%$.

Specialize was also evaluated on two large-scale proteomic experiments from two previous studies on SUMOylation in Human [62] and Arabidopsis [71]. Most SUMOylation studies to date have focused on identifying potential substrate proteins. While these studies often identify 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 [63, 68, 64, 65, 72, 66]. As shown in Figure 2.5, in both SUMO datasets Specialize was able to increase the number of identified SUMOylated peptides by 83–325% over what was identified by Mascot [73]. A detailed comparison shows that Specialize was able to identify all the SUMOylated peptides found by Mascot in the Arabidopsis SUMO dataset while identifying 43 out of 64 SUMOylated peptides found by Mascot in the Human SUMO dataset. Further investigation into these twenty one SUMOylated peptides missed by Specialize showed that half of them either contain a phosphorylation

that was not considered by Specialize or the substrate peptides are very long (e.g. $\geq 27 a.a.$, see Supplementary Figure 2.6a). For the remaining cases, a common feature is that the fragment ions from SUMO tag displays 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 2.6b). It is perhaps not surprising that Specialize did not identify this sub-group 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 twelve 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 2.3). 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 datasets.

2.2 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 dataset 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 datasets 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. 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 datasets 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 but were not identified by Specialize. In principle, one could train a specific 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). This would be readily addressable in our framework as one can synthesize peptide libraries with more diversity in sequence composition and length. Finally, the core concepts of the proposed approach for developing a PTM-specific search method are not specific to SUMOylation and can thus be used to develop new tools to identify peptides with a wide range of complex PTMs.

2.3 Detailed Methods

2.3.1 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 position 1, 6 and 3 along the substrate peptides respectively (see Figure 2.2). 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 realis-

tic 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 [61].

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 Symmetry C18 column (1.7 m BEH-130, 0.1 x 100 mm, Waters, Dublin, CA) with a flow rate of $1\mu L$ 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% FA/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.

2.3.2 Synthetic MCL1 dataset

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 twenty SUMOylated peptides, including variants of the same peptide with SUMO attached to different lysine positions. This set of synthesized peptides serves as a benchmark dataset 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 $1\mu g$ whole cell lysate of the human Jurkat cell. The samples were then analyzed by LC-MS/MS as described for the combinatorial peptide libraries.

2.3.3 Identification of SUMOylated peptides from combinatorial peptide libraries

As illustrated in Figure 2.2, 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 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 which 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 while library III contains peptides whose sequence pattern conforms to the canonical sequence motif [XK(D/E)] [61] for SUMOylated peptides. After spectra from SUMOylated peptides were identified from libraries II and III, they were incorporated our training data and used to build a better scoring model for SUMOylated peptides. Finally, this improved method was used to re-search the spectra from all three libraries to get a final list of MS/MS spectra from synthetic SUMOylated peptides. 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 50ppm 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 50ppm 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.

2.3.4 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 fragments 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 2.1b). 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 [17]) we introduced a probabilistic model that describes how well a pair of peptides matches to a mixture MS/MS spectrum from co-eluting 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 unan-

notated 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: $Score(S,P) = \sum Score(s_i, p_i) = \sum log(\frac{Prob(s_i|p_i)}{Prob(s_i|0)})$. 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 ($Score(s, b_{substrate})$ and the SUMO tag ($Score(s, b_{tag})$). 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 towards 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 2.1). 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: $b, b(iso), b-H20, b-NH3, y, y(iso), y-H20, y-NH3$, where $b(iso)$ indicates the isotopic peak of b or y ions. Specialize further adds the ion types $b_X, b(iso)_X, b-H2O_X, b-NH3_X, y_X, y(iso)_X, y-H2O_X, y-NH3_X$

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 non-linked fragment ions when matching a SUMOylated peptide against an MS/MS spectrum.

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. 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 separated from unlinked PSMs and scored using a Support Vector Machine (SVM) [37] to distinguish true matches from false positive ones. The features used in SVM were: 1) likelihood score as described above; 2) likelihood score divided by peptide length - 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 sixteen features. Together with the combine likelihood score that consider fragments from both the substrate peptide and SUMO tag (as described above) this define the final list of seventeen 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 dataset, the correct PSMs were used as positive training data while top-scoring PSMs from the decoy database were used as negative training data.

Finally all PSMs from SUMOylated peptides were sorted by decreasing SVM score and FDR was determined using the standard target/decoy approach (TDA). The rationale for separate FDR determination for SUMOylated and non-SUMOylated peptides is that the match statistics for SUMOylated and unlinked peptides are different, and thus their score distributions will also be substantially different. Furthermore in typical large-scale experiments the number of SUMOylated PSMs is expected to be small and thus we usually observed that an FDR calculation that combines PSMs from SUMOylated and unlinked peptides together tends to result in the underestimation of FDR for SUMOylated peptides.

2.3.5 Identification of SUMOylated peptides in biological datasets

For the synthetic MCL1 SUMOylated peptides (*pure MCL1 dataset*), InsPecT search was run with 3.0Da parent mass tolerance and 0.5Da fragment mass tolerance allowing for +599 (SUMO) and +582 (SUMO with pyro-glutamate) on Lysine as variable modifications. Specialize was run with same parameters while allowing the following two SUMO tags: QQQTGG and Q(-17.0265)QQTGG where Q(-17.0265) indicates pyro-glutamate 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, then a 5% FDR was enforced using the standard target/decoy strategy (TDA). We used a slightly higher FDR threshold 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 dataset). 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 dataset with spiked-in MCL1 peptides (*Jurkat dataset*), 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 sensitivity of identifying SUMOylated peptides in the Jurkat dataset, identified spectra from SUMOylated peptides in the pure MCL1 dataset were compiled into a spectral library of MCL1 SUMOylated peptides. Then we searched this spectral library against the Jurkat dataset using M-SPLIT [36] to identify a list of potential SUMOylated peptides that are present in the sample and were selected for MS/MS analysis.

The Arabidopsis SUMO dataset and its Mascot search results were obtained from the original publication [71]. Because 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_091609q.mzXML, Vierstra_Sumo_062209d.mzXML and Vierstra_sumo_070109d.mzXML. The Specialize search was done using 50ppm precursor mass tolerance and 0.5Da fragment mass tolerance against an Arabidopsis Thaliana protein sequence database (downloaded from UniProt, ver. 5/13/2012). The SUMO tag considered was QTGG and Q(-17.0265)TGG. N-terminal acetylation (Nterm+42.011) and methionine oxidation (M+15.995) were allowed as variable modifications on the substrate peptides. All PSMs representing possible SUMOylated peptides were extracted and filtered to have a precursor mass error less than 15ppm and a 5% FDR was enforced using the TDA approach. For the Human SUMO dataset [62], we only considered MS/MS data that were generated in the Collision-induced dissociation (CID) mode since our training data was generated in CID only. Mascot search results were ob-

tained directly from the original publication [62]. The Specialize search was done the same parameters as the Arabidopsis dataset except the search was against a Human protein sequence database and the SUMO tags considered were QQTGG and Q(-17.0265)QTGG.

2.4 Acknowledgements

Chapter 2, in full, has been submitted for publication of the material as it appear in J. Wang, VG, Anania, J. Knott, J. Rush, J.R. Lill, P.E. Bourne, N. Bandeira. "A turn-key approach for large-scale identification of complex post-translational modifications", Journal of Proteome Research. The dissertation author was the primary investigator and author of this material.

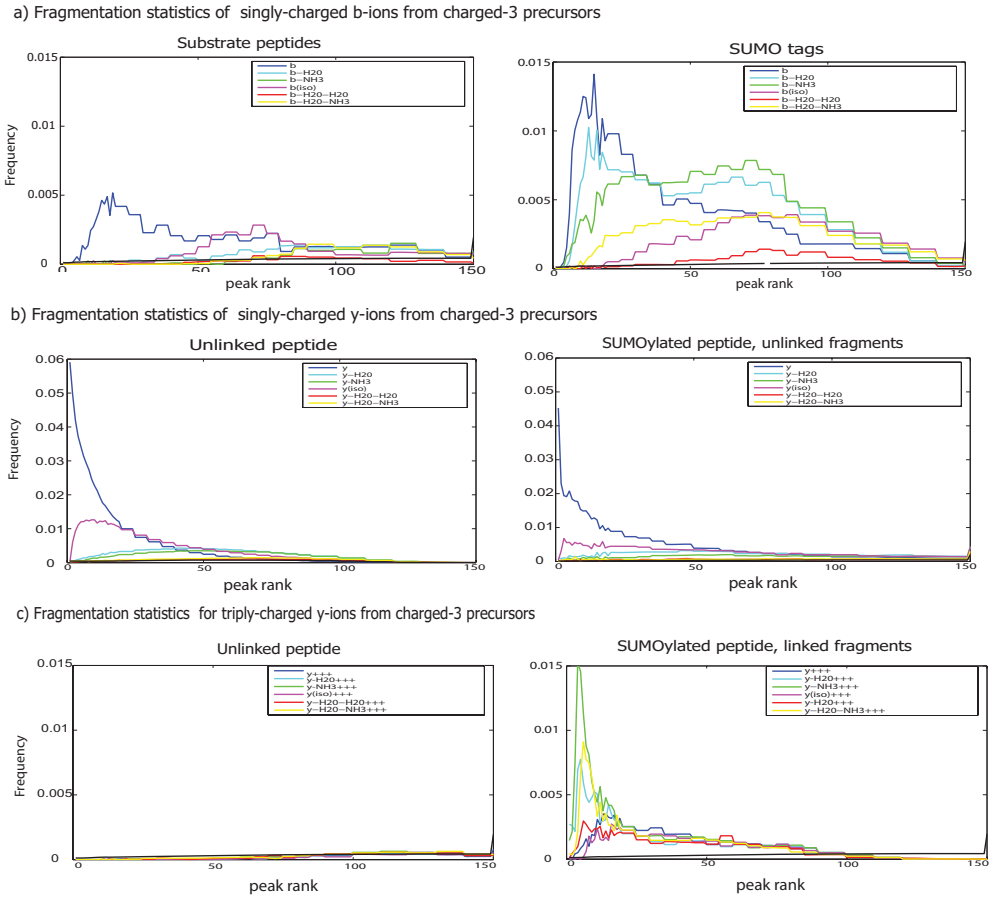

Figure 2.4: Comparison of fragmentation patterns of unlinked and SUMOylated peptides: Fragmentation patterns are represented as the distributions of peak intensity ranks for all peaks matched to a particular type of fragment ions; peaks are ranked from most intense to least intense. The fragmentation pattern of substrate peptides and SUMO tag are observed to have different statistics. For example, as shown in a), H_2O and NH_3 losses from b-ions are more frequently observed in SUMO tag peptides than in substrate peptides. To capture the fragmentation pattern of SUMOylated peptides, fragment ions are divided into two categories: linked-fragments and unlinked fragments (see Figure 2b in the main text). Linked fragments are from peptide fragment ions that are covalently linked to a second peptide. In general, unlinked fragments have fragmentation patterns similar to that of unlinked peptides. As an example, the fragmentation patterns of y-ion from unlinked peptide and SUMOylated peptide are compared in b). On the other hand, linked-fragments have fragmentation patterns different from those of unlinked peptides. In particular, it was observed that multiply-charged fragments are more prominent as compared to unlinked peptides. The fragmentation pattern of triply charged y-ion from unlinked peptides and SUMOylated peptides are illustrated in c).

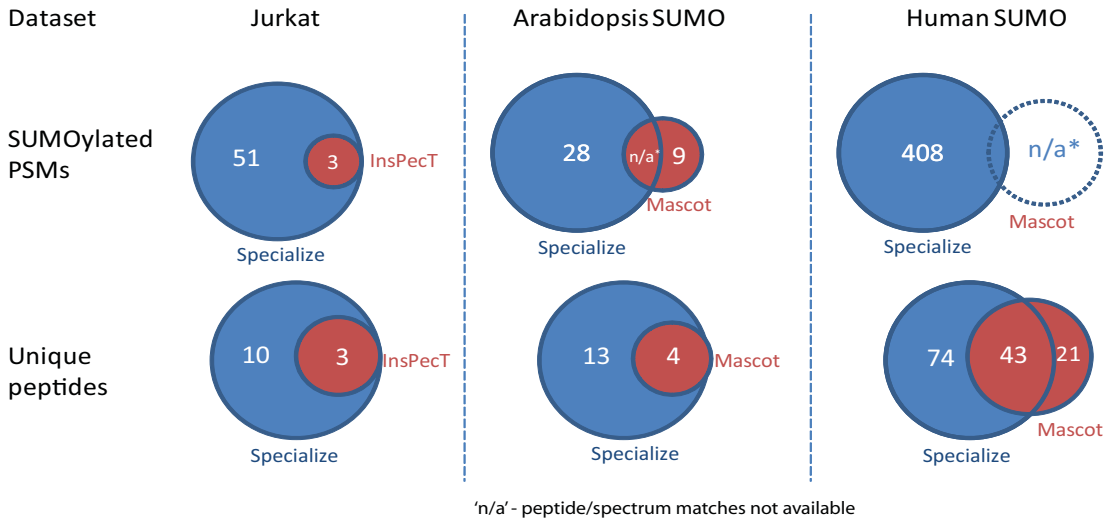

Figure 2.5: Comparison of identification of SUMOylated peptides between Specialize, InsPecT, and Mascot The ability of Specialize to identify SUMOylated peptides was tested on three datasets. The Jurkat dataset contains a set of 20 synthetic SUMOylated peptides from the human MCL1 protein spiked into a background of Jurkat human cell lysate. The Arabidopsis and Human SUMO datasets 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 by identified InsPecT and Mascot. As shown in the Figure, Specialize is able to improve the identification of SUMOylated peptides by 82.8%–325%.

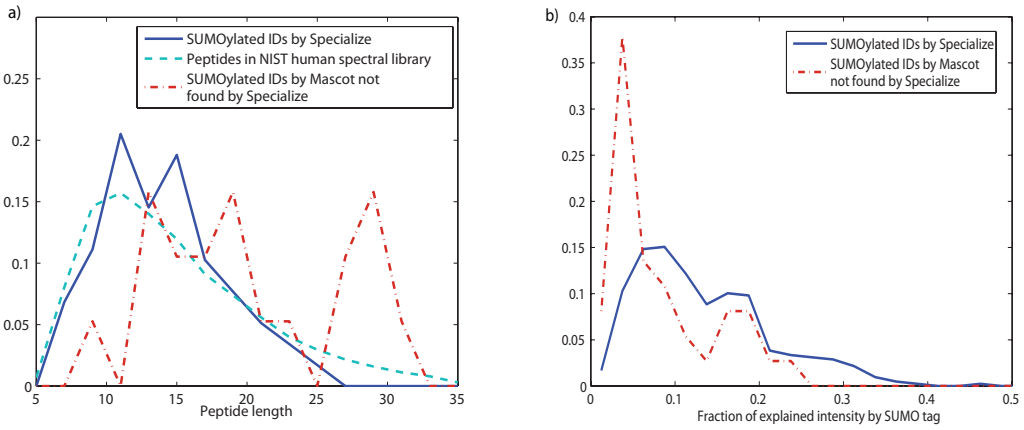

Figure 2.6: Features of SUMOylated peptides not identified by Specialize In the Human SUMO dataset, Specialize identified 43 out of the 64 SUMOylated peptides found by Mascot. SUMOylated peptides not identified by Specialize were found to have two common features: a) the substrate peptides are of very long length ($\geq 27 a.a.$) or b) the SUMO tag does not fragment very well and thus results in peaks with relatively low intensity in the MS/MS spectrum. As shown in a), SUMOylated peptides identified by Specialize (blue line) and peptides in the human NIST spectral library (cyan line) have similar length distributions indicating that Specialize can identify peptides with a wide range of lengths. However, a large fraction of SUMOylated peptides not identified by Specialize tend to be long ($\geq 25 a.a.$), indicating that the current model in Specialize may not generalize well to the class of very long peptides. b) Similarly, in the synthetic peptide libraries, the fragments from the SUMO tag usually contribute 10-20% of the total intensity in the resulting MS/MS spectra (see Figure 2 in the main text). Such trends were also observed for SUMOylated peptides identified by Specialize in the Human SUMO dataset (blue line). However for the SUMOylated peptides not identified by Specialize were observed to have much lower explained intensity from SUMO tags, on average contributing only 5% of the total intensity (red line).

Chapter 3

Identification of linked peptides from tandem mass spectra

3.1 Introduction

The study of protein-protein interactions (PPIs) is crucial to understand how cellular system functions as a whole because proteins do not act in isolation. Most cellular processes are carried out by large macromolecular assemblies and regulated through complex cascades of transient protein-protein interactions [74]. For the past several years we've seen numerous high-throughput studies pioneering the systematic characterizations of PPIs in model organisms [75, 76, 77]. These studies mainly utilize two techniques: the yeast two-hybrid system, which aims at identifying binary interactions, and affinity purifications followed by tandem mass spectrometry analysis (TAP-MS), for the identification of multi-protein assemblies. Together this has leaded to a rapid growth of known PPIs in human and other model organisms. Patche and Aloy recently estimated that there are more than one million interactions known to date [78].

Despite rapid progress, current methods are not without limitations [79]. For example, while the yeast-two hybrid method is plagued by poor reproducibility and high error rates, the more reliable TAP-MS method is time and labor intensive. Thus it is not easy to extend these methods to routinely study the dynamic nature of protein interactions, such as performing the same experiment in different cell types or perform time-point experiments to deduce the temporal

dynamics in interaction networks. More importantly, these techniques can only identify whether proteins interact and at best identify the particular domains that are involved in the interactions. This only represents the first step toward the understanding of how proteins interact. A fuller understanding will come from the three-dimensional structures of protein complexes as they provide mechanistic insights that govern how interactions occur and the high specificity observed inside the cell. Traditionally the gold-standard method in structural biology is x-ray crystallography and there have been several efforts similar to structural genomics [80] that aims to solve all protein complexes [81]. However, while we see the accelerated growth of structure for protein monomers in the Protein Data Bank (PDB) [82], the growth of structures for protein complexes remain relatively steady over the years [78]. Many factors, including the large size and the transient and dynamic nature the interactions has prevent many complexes from being solved by traditional approaches in structural biology. Thus, the developments of complementary analytic techniques to probe the structure of large protein complexes have been underway [83, 84, 85, 86, 87, 88].

Recently, one emerging and promising approach is to analyze protein structures and interactions by generating tandem mass (MS/MS) spectra of cross-linked peptides [87]. The fundamental idea behind this technique is to generate and detect pairs of amino acid residues that are close spatially close to each other. When these linked pairs of residues are from the same protein (intra-protein crosslinks), they provide distance constraints to infer the possible conformations of protein structures. On the other hand, when the pairs of residues come from different proteins (inter-protein crosslinks) they provide information about how the proteins interact with each other. Cross-linking strategy dated back almost a decade ago, but due to the difficulty in analyzing the convoluted MS/MS spectrum generated from linked peptides it is not widely applied. With the recent advances in instrumentation there has been a renewal of interest in using this strategy help determine protein structure and protein-protein interactions. However most studies so-far has been focus on purified protein complexes. With today's mass spectrometers capable of analyzing ten of thousand of spectra in a single experiment, it is valuable

to extend this approach to study complex biological samples. One of the main bottlenecks preventing this is lack of software that is able to search linked-peptide spectra against large sequence database. Accordingly there has been several recent efforts to develop computational method for the automatic identification of linked peptides from MS/MS spectra [44, 89, 90, 91, 92, 93, 94]. However due to the lack of large annotated training data, most current approach either use fragmentation model borrowed from unlinked, linear peptides or learn the fragmentation statistics from training data of limited size [44, 41] which may not generalized well across different data. Here we used disulfide-bridged peptides as an example to describe a generic procedure to a) efficiently generate a *large* mass spectral reference data for linked peptides and b) use this data to automatically train an algorithm that can efficiently and accurately identify linked peptides from MS/MS spectra.

3.2 Method overview and results

3.2.1 Building linked-peptide specific search method for disulfide-bridged peptides

There are three major computational challenges in identification of linked peptides. First, the covalent linkage of two peptides changes the physicochemical properties of the peptides and generates new types of fragment ions which display substantially different fragmentation statistics than those captured by existing models for linear, unlinked peptides. Second, while spectra from linked peptides contain fragment ions from *two* peptides, almost all MS/MS database search tools assume that each spectrum comes from a single peptide. The presence of two peptides in the same spectrum also creates a quadratic search space for peptide candidates. Efficient techniques are thus needed to efficiently search this vast search space. Finally, there is a small number of reliably identified publicly available spectra to learn the fragmentation models for linked peptides and benchmark search strategy, thus making the development of these tools quite difficult. In order to address these challenges we used disulfide-bridged peptides as an example, designed and synthesized three combinatorial peptide libraries, each with

a cysteine residue at different position along the library peptides. The peptide libraries were then promoted to form random disulfide-bridged dimers and analyzed with an LTQ-Orbitrap mass spectrometer. MS/MS spectra were identified using a two-step search strategy (see Figure 3.1). A total of 5952 MS/MS spectra from disulfide-bridged peptides, corresponding to 2976 unique linked-peptide pairs, were identified.

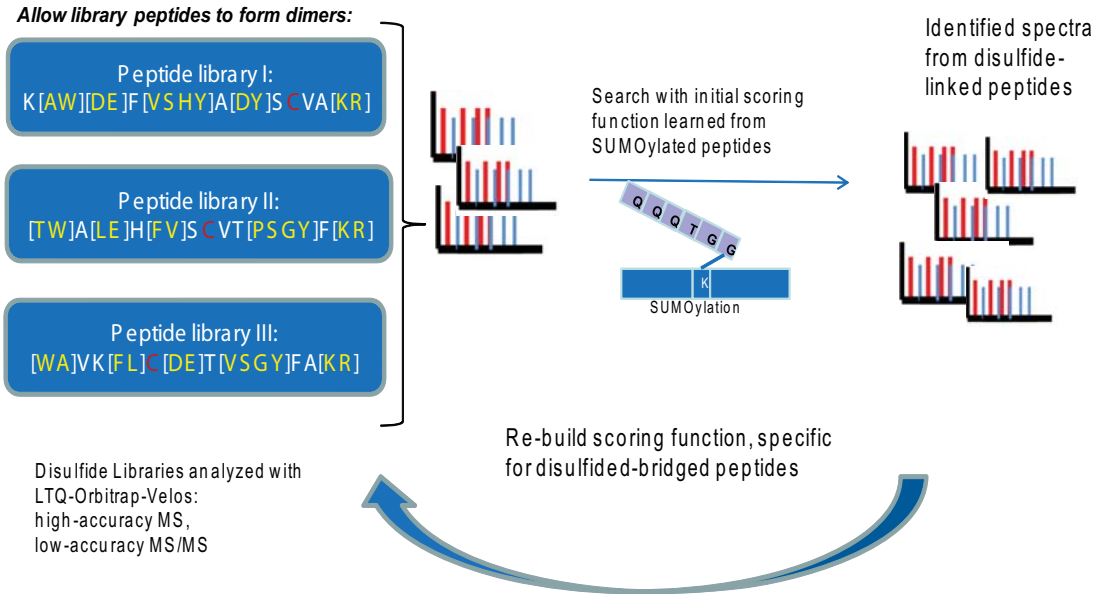

Figure 3.1: Generating training data using combinatorial synthetic peptide library: In order to generate a sufficiently large training data for linked peptides we designed and synthesized three combinatorial peptide libraries, each with a cysteine residue at different position along the peptide. The letters in the square bracket indicates that multiple residues are possible at that position. The peptides libraries were promoted to form disulfide-bridged dimers and analyzed using LTQ-Orbitrap-Velos mass spectrometers. MS/MS spectra from the disulfide-bridged peptide libraries were identified using a two-step strategy. An initial set of MS/MS spectra from disulfided-peptides were identified using scoring models learned from SUMOylated peptides. Using this initial set of spectra as training data, we built a scoring models specific for disulfide-peptides and used the improved scoring models to search the data again to get a final list of spectra from disulfided peptides.

From this training data, we studied the fragmentation patterns of disulfide-bridged peptides. We divided fragment ions from linked peptides into linked and unlinked fragments. Linked fragments are fragment ions that remain covalently linked to a second peptide and they are observed to have different fragmentation

patterns as compared to unlinked fragments. For example while a triply charged y-ions are quite common in linked fragments from a charge three precursors (see Figure 3.3), they are hardly observed for unlinked fragments. Furthermore we observed that both linked and unlinked fragments from disulfide peptides have different fragmentation patterns when compared to conventional, unlinked peptides. In general unlinked fragments display less intensity in the MS/MS spectrum as compared to the fragment ions of the same type from unlinked peptides (see Figure 3.3a). On the other hand, for linked fragments, highly-charged fragments tend to be more prominent as compared to those in unlinked peptides since linked fragments are covalently attached to a second peptide that has an additional N and/or C-termini able to capture charges(see Figure 3.3b). Finally we observed that different types of linked peptides tends to have different fragmentation patterns. In Figure 3.3c, we compared the fragmentation patterns of disulfide-bridged peptides and those of SUMOylated peptides, which are a special type of linked peptides where the c-terminus of the peptide QQQTGG is linked to the lysine residue of a second peptide (see Figure 3.1). Even though triply-charged y-ions are prominent in both disulfide-bridged peptides and SUMOylated peptides, they are twice as prominent in disulfide-bridged peptides as compare to those of SUMOylated peptides. Thus ultimately we would want to build a specific probabilistic model for each type of linked peptides in order to maximize the sensitivity of identifying linked peptides from MS/MS spectra.

In order to account for these fragmentation characteristics of linked peptides, our database search method, MXDB, used separate ion types for linked and unlinked fragments (e.g. linked b fragments vs. normal b fragments, see Method section). Therefore, the fragmentation statistics for linked and unlinked fragment ions are captured separately and different probability/weights are assigned to linked and unlinked fragments when scoring a candidate linked peptide against a MS/MS spectrum. To address the fact that a MS/MS spectrum from linked peptides contain fragments from two peptides, MXDB uses a mixture fragmentation models similar to those used in [17] that explicitly account for the co-fragmentations of two peptides (see Figure 3.2b). With this new probabilistic

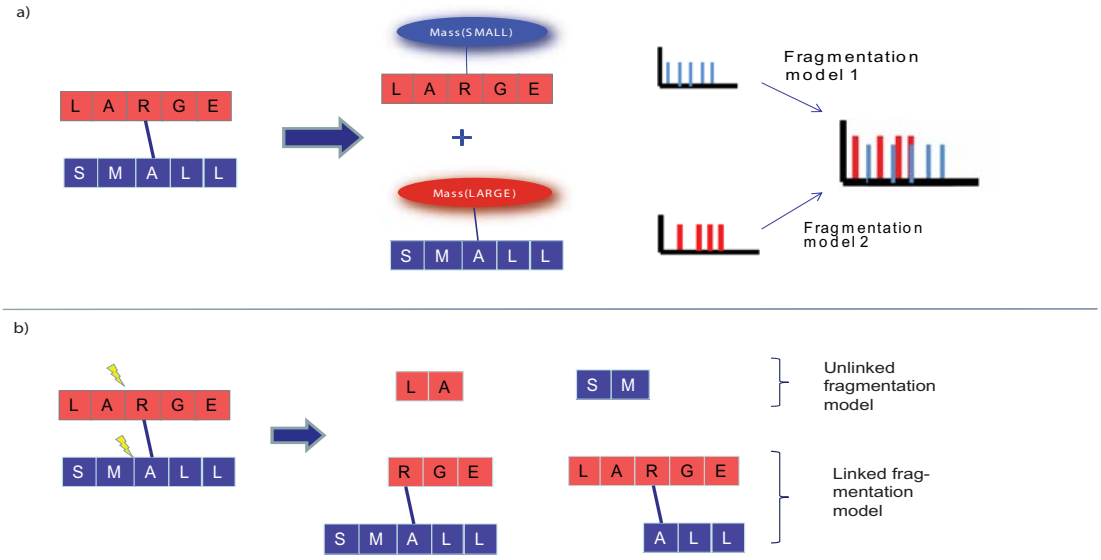

Figure 3.2: Fragmentation model of linked peptides: a) To account for fragment ions from both of the cross-linked peptides, we model a linked peptide as two peptides that carry a modification with mass equal to the other peptide. Therefore to generate theoretical spectrum for a linked peptide we generate theoretical spectrum for each peptide separately and combined them into a theoretical spectrum for the linked peptide. This allows us to build separate fragmentation models for each of the linked peptide separately, accounting for their difference in fragmentation patterns. b) To capture the fragmentation characteristics of linked peptides, we divide fragment ions into two types: linked-fragments and unlinked fragments. Linked fragments are fragment ions that are covalently linked to a second peptide. Linked fragments and unlinked fragments have different fragmentation pattern, our scoring models use separate probabilistic models for each type of fragments.

scoring model MXDB also uses a two-stage search strategy where it focus on identifying only one of the linked peptide during first stage and then identify the other linked peptide in subsequent stage (see Figure 3.4). This strategy helps us filter the search space of all possible peptide pairs by several order of magnitudes, and identify the correct linked peptide efficiently.

3.2.2 Identification of disulfide-bridged peptides from combinatorial peptide library

In order test MXDB's ability at identifying disulfide-bridged peptides, we searched MS/MS spectra from the three disulfide-bridged peptide libraries against

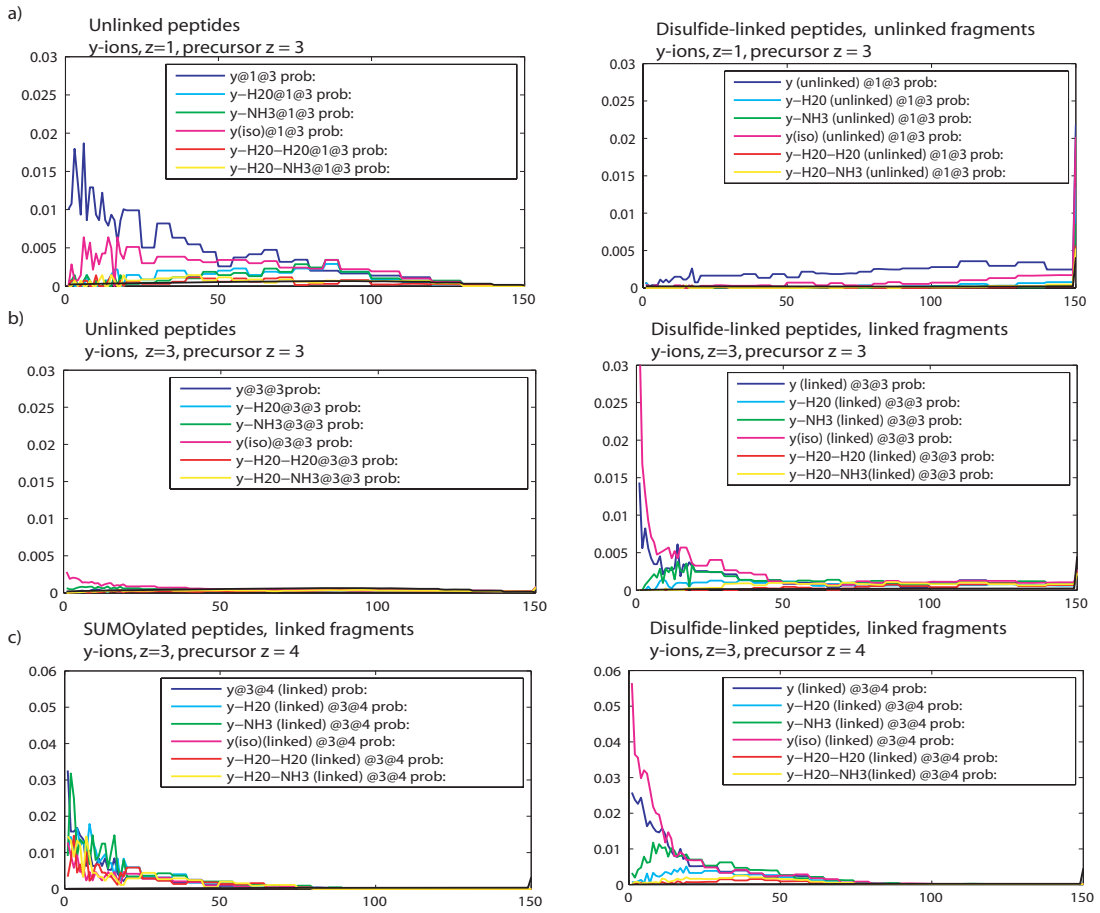

Figure 3.3: In order to capture the specific fragmentation patterns of linked peptides, we analyzed the fragmentation statistics of identified MS/MS spectra from disulfide-bridged peptides in our reference dataset. Fragmentation pattern is represented as the peak rank distribution of a particular type of ion observed in the MS/MS spectra. Peaks are ranked from most intense to least intense. We divide fragment ions from linked peptides into linked and unlinked fragments (see Figure 3.2b). As shown in a), in general, unlinked fragments are less prominent as compared to the same type of fragment ions from linear, unlinked peptides. On the other hand, linked-fragments have highly charged fragments that are much more prominent as compared to those from unlinked peptides. This make sense since linked fragments are covalently attached to a second peptide and thus contain an extra N and C-termi to acquire extra charges. Finally different types of linked peptides also tend to have different fragmentation patterns. For example, as shown in c), triply charged y-ions from disulfide-bridged peptides are twice as prominent in the MS/MS spectra as compared to SUMOylated peptides.

all possible library peptide sequences. Starting with an initial scoring model

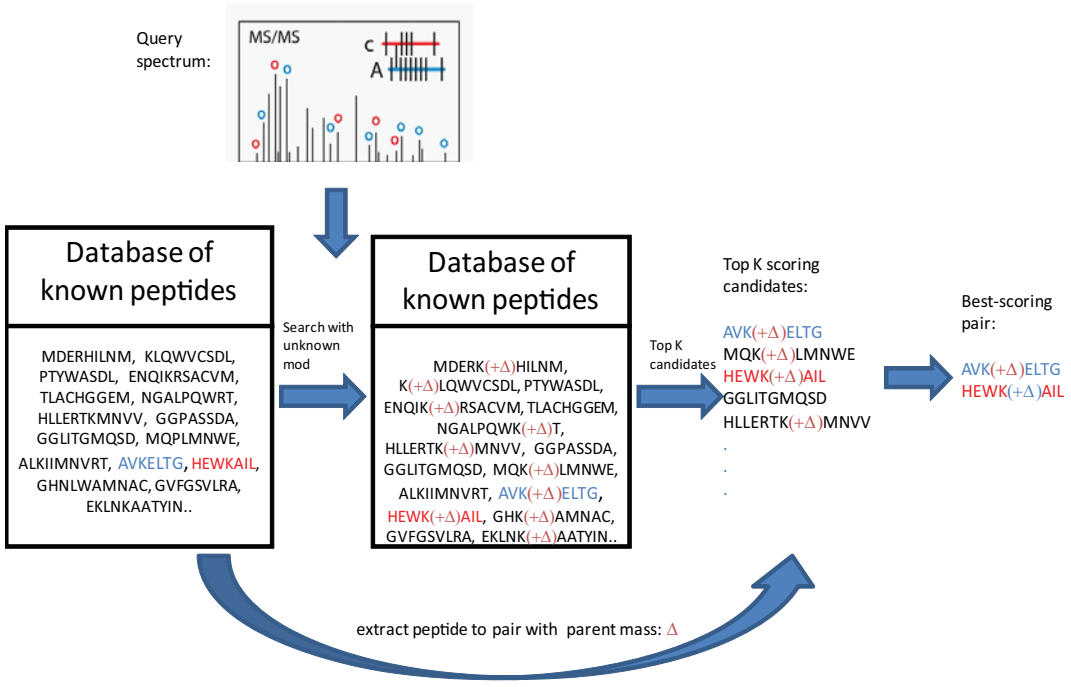

Figure 3.4: MXDB search strategy: In order to avoid the quadratic search space of all possible peptide pairs in the database we adopted a two-steps search strategy. During the first pass of the search, all candidate peptides in the database are treated as peptides with modification at the allowable linking residues and with a mass-offset equal to the difference between the parentmass of the query spectrum and the parentmass of the candidate peptide. We score all peptide candidates against the query spectrum and only retain the K top-scoring candidates for the next stage where each candidate is paired with the remaining peptides in the database that together with a combined parentmass added up to that of the query spectrum. Finally the pair of peptides with the best score is returned.

learned from SUMOylated peptides, we identified an initial set of 4232 MS/MS spectra from disulfide-bridged peptide at a 5% false discovery rate (FDR). From this initial training dataset, we built an improved scoring models specific for disulfide-bridged peptides and use it to identified an additional 31-62% more MS/MS spectra from each peptide library (see Figure 3.5a). This support our hypothesis that different types of linked peptides have different fragmentation patterns and properly capture these fragmentation patterns can improve the sensitivity of identifying linked peptides from mass spectra.

An important goal of building better tools for the identifications of linked

peptides is to enable the application of the crosslinking method in more complex biological samples. Thus, we tested whether MXDB search can scale up to large sequence databases. We appended the whole E. coli, Yeast and Human proteins database as decoys to all possible library peptides respectively and search the MS/MS spectra from the peptide libraries against these concatenated databases. The effect of database size on MXDB's sensitivity at identifying disulfide-bridged peptides is shown in Figure3.5b). The general trend indicates that as we increase the size of the database by a factor of 3.13, which corresponds roughly to a 9.8 times increase in the search space of linked peptides, there is a 20%-30% percent drop in sensitivity. In general, it is expected as we increase the size of our search space, the sensitivity of database search method will decreases. As a comparison, we also searched a typical trypsin-digested yeast cell lysate dataset[31] using MS-GFDB [46], a state-of-the-art database search tool for linear, unlinked peptides. MSGFDB identified $27,500$ and $23,300$ spectra with 50 and 500ppm precursor mass tolerance respectively. This means there is a approximately 18% drop in sensitivity when we increase the size of search space ten times. Even though MXDB's drop in sensitivity is slightly worse than that observed in traditional database search tool for unlinked peptides, the total search space for linked peptides is much larger than those of unlinked peptides. Thus, this demonstrates MXDB's ability to identify disulfide-bridges peptides against proteome-scale sequence database.

3.2.3 Identification of cross-linked peptides from protein complexes

Next we set out to test whether the fragmentation models we learned from disulfide-bridged peptides can help us identify other type of linked-peptides. We benchmarked MXDB on two MS/MS datasets from cross-linking experiments on the S. Pombe 26S. proteasome and the Rabbit 20S proteasome complexes respectively [95]. As we can see in Figure 3.6, in both datasets MXDB is able to identify significantly more cross-linked peptides as compared to two current state-of-the art database search tools for crosslinked peptides, xQuest [44] and pLink [94]. This demonstrates that the fragmentation models learned from disulfide-bridged pep-

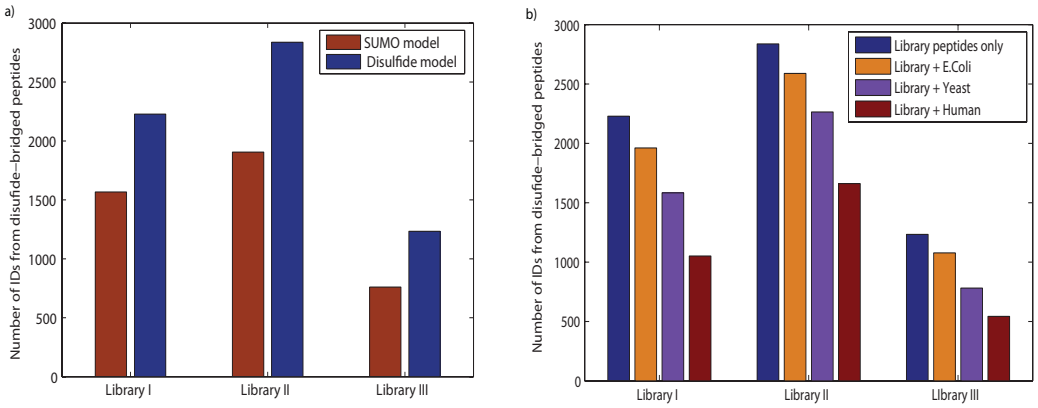

Figure 3.5: Identification of disulfide-bridged peptides from combinatorial peptide library: a) We compared the initial scoring model learned from SUMOylated peptides to the specific scoring models we built for disulfide-bridged peptides. The latter models improve the identification of disulfide-bridged peptides by 30-50% in the peptide libraries, underscoring the fact that different types of linked-peptides has different fragmentation patterns. b) To test MXDB's ability to identify disulfide-bridged peptides against whole-proteome sequence databases, we concatenated the library peptide sequences to all E. coli, Yeast and Human protein sequences respectively and search the spectra from peptide libraries against the concatenated database. MXDB was able to identify thousands of spectra from disulfided peptides these large databases and the general trend shows that as the search space of cross-linked peptides increase by approximately tenfold, the sensitivity of identifying disulfide-bridged peptides decreases by 20-30%.

tides can serve as a starting scoring model for the identification of linked peptides in general. We expect the sensitivity of identifying linked peptides will be further improved, once we have enough training data to build a specific scoring models for a specific type of cross-linked peptides.

To validate our search results, we mapped the identified cross-linked peptides to homologous proteins with available crystal structures found in the Protein Data Bank (PDB) [82]. We then computed the distance between the identified cross-linked residue pairs. As shown in Figure 3.7, most of the identified cross-links have distance within the range of the cross-linker used. Approximately 5.7% of the identified cross-linked peptides have distance greater than $30\mathring{A}$, which is considered to exceed the maximum distance range of the cross-linker. This is also consistent with the FDR we estimated in the search resutlt using the Target-

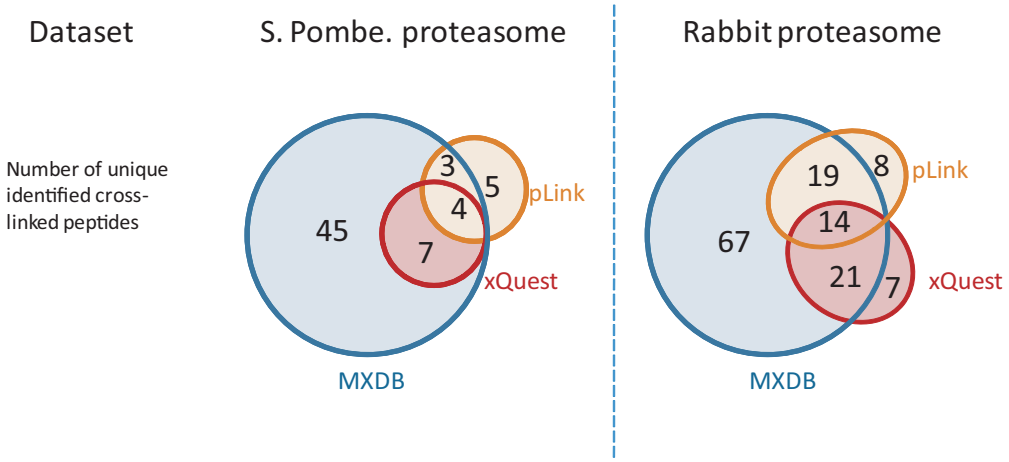

Figure 3.6: Identification of cross-linked peptides in proteasome complexes: We compared the identification of cross-linked peptides between MXDB, pLink, xQuest on two datasets from the crosslinking studies of proteasome complexes from S. Pombe and Rabbit respectively. MXDB is able to identify significantly more crosslinked peptides as compared to pLink or xQuest.

Decoy Approach (TDA). To evaluate whether MXDB can identify linked peptides against large sequence databases we appended the whole E. Coli and Yeast protein sequence database to all Rabbit proteasome proteins and searched the Rabbit proteasome dataset against the concatenated databases. As shown in Figure 3.8, while MXDB is still able to identify a large numbers of the cross-linked peptides, there is also a noticeable decrease in sensitivity. Thus, ideally in order to maximize the sensitivity of the identification of linked peptides, one would like to use a smaller database if possible. However in contrast to the studies on the proteasome complexes, which are relatively well-studied complexes, in many studies that involve protein complexes we would not know ahead of time the proteins subunits that constitute the complexes. To address this scenario, we noted that in cross-linking experiments there are usually many unlinked peptides presented in the samples that can inform us about the possible proteins presented in the complex. We evaluated a two-pass search strategy where we first identified a list of candidate proteins that are presented in the sample by searching for unlinked peptides and peptides with dead-end linkers using MSGFDB. Then we tried to identify linked peptides using MXDB and this reduced list of candidate proteins identified in the

first pass. As shown in Figure 3.8, this search strategy is able to recover 83% of the cross-linked peptides that are presented in the Rabbit proteasome sample while searching the whole Rabbit protein database. Hence it represents a balance between the assumptions we have to made and the sensitivity of identifying of linked peptides. This type of search strategy is readily applicable to many of the proteomic studies where protein complexes are extracted from cell lysate background using affinity purification methods.

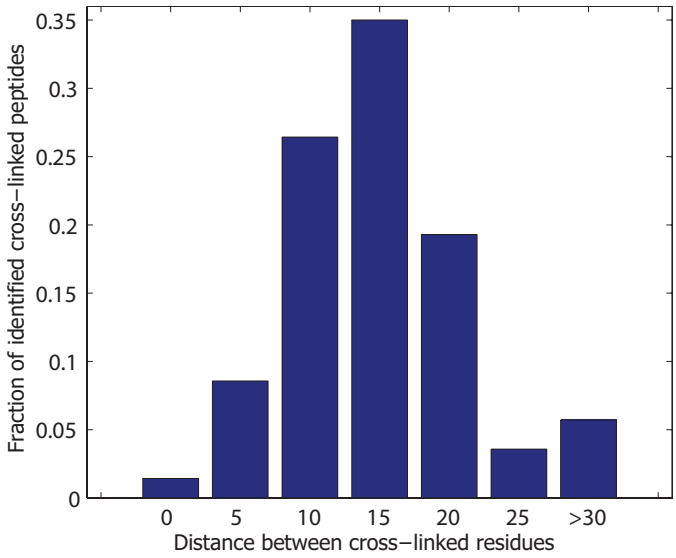

Figure 3.7: Structural validation of identified crosslinked peptides: To validate the cross-linked peptides identified by MXDB, we mapped the cross-linked peptides identified from the S. Pombe data to the crystal structure of a Yeast proteasome (PDBID:1FNT) and cross-linked peptides from the Rabbit data to a crystal structure of the Mouse proteasome respectively (PDBID:3UNE). The distances between the identified cross-linked residues were computed. When the distance is greater than 30Å, we considered it to exceed the possible distance the cross-linker can span. As shown in the Figure, out of all cross-linked peptides that can be mapped to the crystal structures, approximately 5.8% of them has distance exceed the maxmimum threshold, which is also in accord with the 5% false discovery rate we estimated during our search.

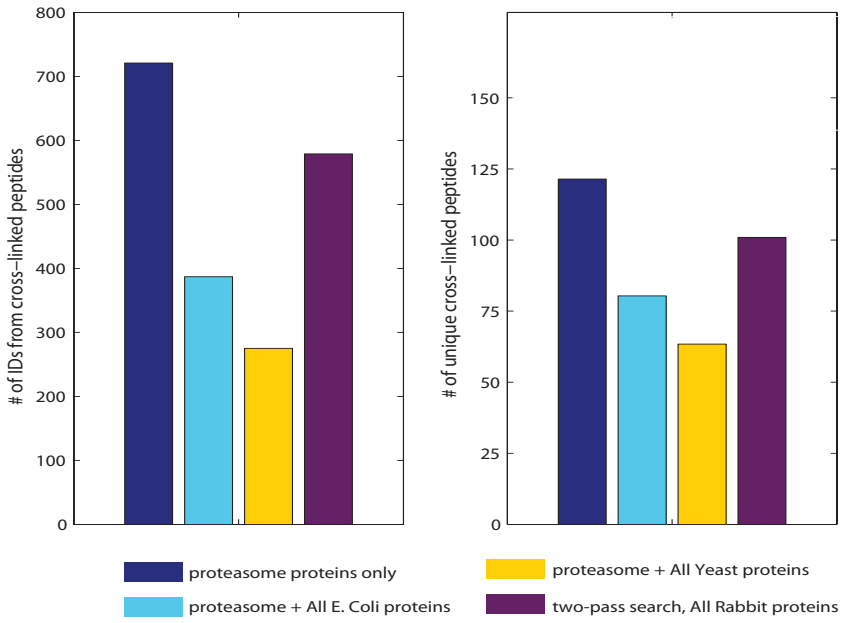

Figure 3.8: Identification of cross-linked peptides against proteome-scale databases: To evaluate MXDB's ability to identify cross-linked peptides against proteome-scale database, we appended whole E. coli and Yeast protein sequence databases to all Rabbit proteasome proteins and searched the Rabbit proteasome data against the concatenated databases. Even though MXDB can identify a large fraction of cross-linked peptides when searching against this large database there is a considerable drop in sensitivity. To achieve a good balance between sensitivity of identifying linked peptide and the complexity of database MXDB can handle, we tested a two-pass search strategy where a list of candidate proteins that are presented in the sample were constructed by identifying unlinked peptides against all proteins in the database. Then we tried to identify cross-linked peptides by searching against only candidate proteins identified in the first stage. As shown this strategy allowed us to recover most of the cross-linked peptides while searching against proteome-scale database.

3.3 Discussion

Chemical crosslinking and tandem mass spectrometry is a versatile and high-throughput method to study protein structures and protein-protein interactions. However, there are several challenges that needed to be addressed before we can routinely apply this method on a large scale. In recent years we've already seen many efforts in the development of novel cross-linkers [] and enrichment strategies [] to improve our ability to separate linked peptides from a large background of

other analytes in the sample and analyze them using mass spectrometry. Here we focus on the identification part of the problem and show that having appropriate computational methods can greatly improve our ability to identify linked peptides from MS/MS spectra. It was recently showed that even for linear, unlinked peptides, if they are products of different enzymatic digestion or analyzed by different type of mass spectrometers, they display rather different fragmentation patterns and properly model these fragmentation characteristics can greatly improve our ability to identify peptides from MS/MS spectra []. For linked peptide, because of their different physico-chemical properties the development of an appropriate fragmentation models is even more needed for their identification and the task is challenging because there is no sufficiently large public available reference dataset for us to learn the fragmentation pattern. By using a novel strategy with combinatorial peptide synthesis, we show that it is possible to efficiently generate large reference dataset. From this reference dataset we've show that linked-peptides indeed have quite different fragmentation statistics than unlinked peptides, making most current tools which mostly learn from unlinked peptide not suitable for identifying linked peptides. By incorporating these linked-peptide specific fragmentation statistics into our new database search tool MXDB, we showed that we are able to identify disulfide-bridged peptides against proteasome-scale sequence databases. This scoring model also allow us to develop an efficient filtration strategy that dramatically reduce the search space of all possible crosslinked peptides pairs by several order of magnitude. Aside from addressing fundamental challenges in developing sensitive and accurate method to identify linked peptides, we introduce a basic framework that can be adapted to any type of linked peptides, which will simplify and expedite the development of computational tools for the identification of other types of linked peptides.

3.4 Detailed methods

3.4.1 Builidng training MS/MS data for linked peptides

We synthesized three combinatorial peptide libraries with the following sequence pattern:

- I) K[AW][DE]F[VSHY]A[DY]SCVA[KR];
- II) [TW]A[LE]H[FV]SCVT[PSGY]F[KR];
- III) [WA]VK[FL]C[DE]T[VSGY]FA[KR];

The letters in the square bracket indicates multiple residues are possible at those positions. For example in library I, both Alanine (A) and Tryptophan (W) are possible residues at the second position. The sequence pattern are designed to have several desirable properties to facilitate identification:
1) Each library contain only one cysteine for disulfide-bond formation, this is to make sure there is no ambiguity for assigning the linking site. The incorporation of residues such as Proline that is known to produce MS/MS spectra with relatively poor fragmentation [96] is kept low (theoretically 1/4 of peptides in library II contain Proline).
2) We computed the theoretical parentmass of all possible disulfide-bridged peptide pairs that can be formed from the library peptides and try to find sequence pattern that can generate disulfided peptides with as many unique parentmass as possible.
3) In each variable position (i.e square bracket) we choose residues with different physicochemical properties (e.g. hydrophobicity, polarity, size etc.) to potentially maximize the separation of the possible library peptides with chromatography.

Each peptide library is designed to generate $2^6 = 64$ unique peptides. Theoretically each library can generated $64 * 63/2 + 1 = 2017$ unique disulfide-bridged peptide pairs. This ensures that we have a sufficiently large training dataset to learn the fragmentation pattern of disulfide-bridged peptides while at the same time have a manageable search space so we can identify an initial set of of linked peptides simply using a brute-force search strategy where all possible peptide pairs are considered.

After synthesis, the peptides libraries were put in condition that promotes the formation of disulfide-bridged dimers and were analyzed using tandem mass spectrometry. Samples from each library were injected via an auto-sampler for separation by reverse phase chromatography on a NanoAcquity UPLC system (Waters, Dublin, CA). Peptides were loaded onto Symmetry ? C18 column (1.7 m BEH-130, 0.1 x 100 mm, Waters, Dublin, CA) with a flow rate of 1 L 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% FA/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 60,000 resolution. 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.

MS/MS spectra from the disulfide-bridged peptide libraries were identified using a two-step search strategy. As illustrated in Figure 3.1, an initial search was done using scoring model learned from SUMOylated peptides against a database containing the all possible library peptides and decoy peptides. Decoy peptides were generated by randomly shuffling the amino acid in the library peptides while retaining the position of K or R to keep the enzymatic termini of tryptic peptides. SUMOylated peptides are a special type of linked peptides where one peptide is always fixed to the sequence QQQTGG. Thus scoring model learned from SUMOylated peptides can serve as a starting model to identify disulfide-linked peptides. The details about the identification of SUMOylated peptides was discussed in previous chapter. An initial set of MS/MS spectra from disulfide-peptides were identified at a 5% FDR. From this initial training dataset, a scoring model specific to disulfide-bridged peptides were built and was used to search the spectra from all three libraries again to get a final list of MS/MS spectra from disulfide-bridged peptides. Unless otherwise noted, all searches with MXDB were performed with 50ppm parent mass tolerance and 0.5 Da fragment mass tolerance. Then the

results were filtered to have a parentmass error of less than 10ppm and a False-Discovery-Rate(FDR) of 5% was enforced using a TDA [30] (see next sections)

3.4.2 Scoring models for linked peptides

In order to evaluate the match between a cross-linked peptide pair and a observed spectrum we conceptually think of it as two peptides, each carrying a modification at the cross-linked residue with mass equals to the parentmass of the other peptide(see Figure 3.2). In regular database searches, we try to evaluate how well a *single* candidate peptide matches to a MS/MS spectrum. For cross-linked peptides we evaluated how well does *a pair* of peptides matches to a MS/MS spectrum. In a previous work, MixDB [17], we introduced a probabilistic model that describes how well a pair of peptides matches to a mixture MS/MS spectrum from co-eluting 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 $Prob(s|0)$ model can be learned using unmatched peaks in the spectrum (where the symbol 0 represents noise). The scoring function for a Peptide Spectrum Match (PSM) is thus defined as 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: $Score(S, P) = \sum Score(s_i, p_i) = \sum log(\frac{Prob(s_i|p_i)}{Prob(s_i|0)})$. Since linked peptides are represented as pairs of peptides, we can represent a linked

peptide as two vectors (P, Q). The vector $P = p_1, p_2, ...p_n$ contains all possible fragment ions from the first peptide while the vector $Q = q_1, q_2, ...q_n$ contains all possible fragments ions from the second peptide. Without loss of generality, we define the first peptide to be the dominant peptide that accounts for more ion intensity in the observed MS/MS spectrum. This way we can account for possible differences in fragmentation patterns between the first and second peptides. The score that a spectrum S is matched to a pair of peptides (P, Q) is thus: $Score(S, (P, Q)) = \sum_i \max(Score(s_i, p_i), Score(s_i, q_i))$, where max is used to model the dependency between the two peptides. When theoretical fragment ions from both P, Q match to the same observed spectrum peak, the model only uses the fragment ion with higher probability, thus avoiding using the same peak twice to support the identification of linked peptides. If not explicitly prevented, such double-counting will incorrectly bias unusually high scores towards pairs of peptide candidates sharing many of their theoretical fragment ions.

In order to further capture the fragmentation statistics of crosslinked peptides we further divide the set of fragment ion types into linked and non-linked fragments (Figure 3.2). Linked fragments are fragment ions that are covalently linked to a second peptide. Thus for every ion type that is used to describe linear peptides we introduce its corresponding linked ion type in our probabilistic models. For example in our current implementation, we considered the ion types: $b, b(iso), b - H2O, b - NH3, y, y(iso), y - H20, y - NH3$ for linear, unlinked peptides, where $b(iso)$ indicates the first ^{13}C isotopic peak of a b-ion. Then we add the ion types $b_X, b(iso)_X, b - H2O_X, b - NH3_X, y_X, y(iso)_X, y - H2O_X, y - NH3_X$ to represent the corresponding linked-fragment ions that can be generated from linked peptides. For each ion type we consider charge states from one to the precursor charge of the observed MS/MS spectrum. With these new ion types, the fragmentation statistics specific to linked peptide fragments can be learned during training and different probability/weights are assigned to linked and non-linked fragment ions.

3.4.3 Efficient database search for linked peptides

With a scoring function that properly models the fragmentation characteristics of linked peptides, we can evaluate how well a pair of cross-linked peptides match to a observed MS/MS spectrum. However, we still need to evaluate all possible cross-linked peptide pairs in the sequence database to find the correct match. When the size of the protein sequence database is large, it is not practical to consider all possible peptide pairs in the database. Here we described a two-step search strategy to allow us to find the correct cross-linked peptide matches without considering all possible pairs. Since we model linked peptides as a pair of modified peptides, we argue that for a linked peptide pair generating the spectrum, at least one peptide should score reasonably well when matched to the spectrum alone. Thus in the first stage of our search we will match every peptide candidate against the query spectrum and sort them by their match score. In the second stage only the top scoring peptides are then paired with the remaining candidates to find the best-scoring linked peptide pairs.

Specifically, let S be a query spectrum with parent mass M_S, P be a peptide with parent mass M_P and $P_1, P_2, ...P_n$ be a database containing n peptides. A modified peptide $P(\Delta, t$ is peptide P with a mass-offset of Δ Da at the t-th amino acid residue. For each peptide P_i in the database we consider all of its modified variants $P^m(\Delta, t)$, where Δ is the mass difference between the mass of the spectrum and the mass of the candidate peptide: $\Delta = M_S - M_{P_i}$ $s.t. \Delta > 0$ and t is all valid linking sites for the candidate peptide P_i. For example, in the case of disulfide-bridged peptides, all cysteine residue positions are considered. We score all of these modified candidate peptides against the query spectrum S and sort all the peptide candidates according to their match score. As shown in Figure 3.9, when we search our training data against a concatenated database of all library peptide sequences and the whole E. Coli database which contain about 200,000 tryptic peptides, we see that one of the correct peptide always ranks top 50 scoring peptides and the other peptide always rank top 1000 scoring candidates. This mean that rather than consider $200,000 \times 200,000 = 2 \times 10^{10}$ peptide pairs, we can consider $50 \times 1000 = 50,000$ peptide pairs and still find the correct matches in more than

96% of the cases. In practice we can reduce the search space further by only consider peptide pairs such that their combined masses match the precursor mass of the MS/MS spectrum.

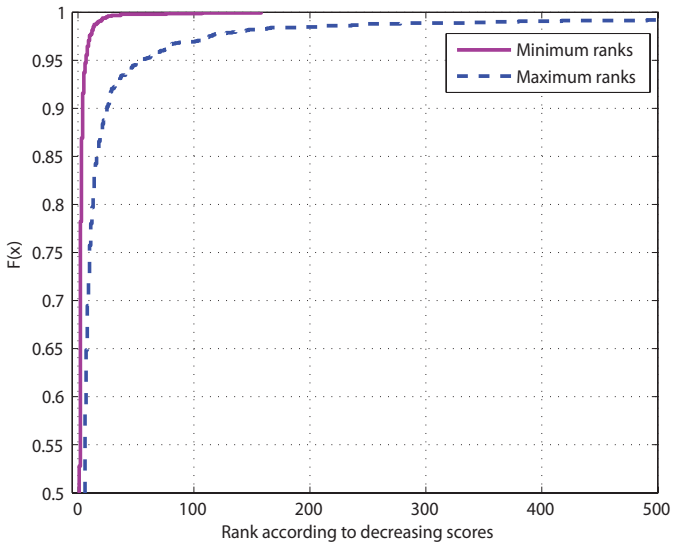

Figure 3.9: MXDB filtration efficiency: To evaluate the filtration efficiency of the search strategy in MXDB, we searched the MS/MS spectra from disulfide-bridged peptides against a concatenated database containing all library peptides and a E. coli protein sequence database as decoy. The number of tryptic peptides in this database is about $200,000$. We scored every candidate in the database and sort them according to decreasing score. Then we evaluated the ranks of the correct peptide matches in the list. A correct match is a library peptide that form a disulfide-bridged peptides, thus each spectrum has two correct matches. The distribution of the minimum and maximum rank of the correct peptide matches are shown. One peptide (the higher scoring one) has rank less than 50 while the other peptide always has a rank less than 1000. This means that we need only pair the top 50 candidate peptides with the top 1000 candidate peptides to find the correct cross-linked peptide pairs. This allow us to consider at most $50 \times 1000 = 5 \times 10^4$ peptide pairs rather than consider all possible pairs in the database which can amount to $1/2 \times 200,000 \times 200,000 = 2 \times 10^{10}$ candidates to evaluate.

3.4.4 Separation of linked-peptide matches from false positives

After database searching, the top scoring linked peptide spectrum matches (LPSMs) scored using a Support Vector Machine (SVM) to separate true matches from false positive ones. Features used in SVM are as followed:

- 1) likelihood score
- 2) likelihood score divide by peptide length: score from 1) divided by the number of amino acids in the top candidate peptide.
- 3) explained 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).
- 8) average mass error between theoretical and observed masses.

Noted we can compute the above features for each of the cross-linked peptide separately resulting in a total of sixteen features. These together with a combined likelihood score that consider both matched peaks from both cross-linked peptides constitute the final list of seventeen features used in the SVM. To train the SVM we used the identified MS/MS spectra from the combinatorial library of disulfide peptides as described in previous section. For each training dataset, the correct LPSMs are used as positive training data while top-scoring incorrect LPSMs from decoy database are used as negative training data.

All LPSMs are sorted according to their SVM scores and they are considered a match if their score pass a certain threshold. The SVM score threshold is chosen to enforce a particular false discovery rate (FDR). For a set of LPSMs the FDR is estimated using a TDA strategy as described in [97]. Briefly because each LPSM has two peptide match, it can fall into one of the following categories: TT–both

peptides matches are from target database; TD/DT – one peptide from target and another peptide from decoy database; DD–both peptides from decoy database. If we define N^{type} as the number of LPSMs fall in to particular category (i.e. N^{TT} is the number of matches of type TT), we can then define FDR for LPSMs as: $FDR_{linked} = \frac{N^{TD}+N^{DT}-N^{DD}}{N^{TT}}$.

3.4.5 Analysis of spectra from cross-linked samples

We analyzed the data from cross-linked sample of Schizosaccharomyces pombe (S. Pombe) and Rabbit proteasome from a previous study [95]. The search result from xQuest was obtained from the original publication [95]. The result from pLink was obtained by running the search tool with 50ppm precursor mass tolerance and default (0.5Da) fragment mass tolerance for CID spectra. The search results were filtered with 10ppm precursor mass tolerance and 5% FDR. Protein sequences for the proteasome complex are downloaded from UniProt [98](www.uniprot.org) by extracting all proteins from the corresponding species that contain the keyboard "'Proteasome"' using the advanced search function of UniProt. To validate the identified cross-linked peptides we obtained crystal structures of the proteasome complexes of a related species from the Protein Data Bank (www.pdb.org) [99]. We mapped the S. Pombe proteasome sequences to the crystal structure of Saccharomyces cerevisiae proteasome (PDBID: 1FNT) [100] and the Rabbit proteasome sequences to the crystal structure of Mouse proteasome (PDBID: 3UNE) [101]. Then we compute the distance between the C_α atoms of the two cross-linked residues in each identified crosslink peptide. If the distance is below $30\mathring{A}$ we consider the identified cross-linked peptide is within the distance constraints of the cross-linker. For the two-pass search we first identified unlinked peptides by searching the data against all Rabbit protein sequences using MSGFDB with the following variable modifications: +42 on n-terminus, +16 on methionine and +138, +156 and +168 on lysine (for identification of peptides with dead-end and intra-peptide crosslinkers). Search results were filtered with a 1% FDR. Then all Rabbit proteins that contain at least one identified peptides were extracted as candidate proteins presented in the sample. For degenerate peptides

that are shared among multiple proteins all proteins are considered as candidates. Next we searched for cross-linked peptides using MXDB against only on this list of candidate proteins identified in the first stage.

3.5 Acknowledgements

Chapter 3, in full is being prepared for submission for publication of the material. J. Wang, VG, Anania, J. Knott, J. Rush, J.R. Lill, P.E. Bourne, N. Bandeira. "Approach for large-scale identification of linked peptides from tandem mass spectrometry". The dissertation author was the primary investigator and author of this material.

Bibliography

[1] M P Washburn, D Wolters, and J R Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, 19(3):242–247, 2001.

[2] E Brunner, C H Ahrens, S Mohanty, H Baetschmann, S Loevenich, F Potthast, E W Deutsch, C Panse, U de Lichtenberg, O Rinner, and Others. A high-quality catalog of the Drosophila melanogaster proteome. *Nature biotechnology*, 25(5):576–583, 2007.

[3] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[4] R J Chalkley, P R Baker, K C Hansen, K F Medzihradszky, N P Allen, M Rexach, A L Burlingame, L Huang, K C Hansen, N P Allen, M Rexach, A L Burlingame, K F Medzihradszky, N P Allen, M Rexach, and A L Burlingame. Comprehensive Analysis of a Multidimensional Liquid Chromatography Mass Spectrometry Dataset Acquired on a Quadrupole Selecting, Quadrupole Collision Cell, Time-of-flight Mass Spectrometer. *Mol Cell Proteomics*, 4(8):1189–1193, 2005.

[5] B R Wenner and B C Lynn. Factors that affect ion trap data-dependent MS/MS in proteomics. *Journal of the American Society for Mass Spectrometry*, 15(2):150–157, 2004.

[6] G Alves, A Y Ogurtsov, S Kwok, W W Wu, G Wang, R F Shen, and Y K Yu. Detection of co-eluted peptides using database search methods. *Biology direct*, 3(1):27, 2008.

[7] Annette Michalski, Juergen Cox, and Matthias Mann. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC?MS/MS. *Journal of Proteome Research*, 10(4):1785–1793, 2011.

[8] Roland Luethy, Darren E Kessner, Jonathan E Katz, Brendan MacLean, Robert Grothe, Kian Kani, Vitor Faca, Sharon Pitteri, Samir Hanash,

David B Agus, and Parag Mallick. Precursor-Ion Mass Re-Estimation Improves Peptide Identification on Hybrid Instruments. *Journal of Proteome Research*, 7(9):4031–4039, 2008.

[9] Stephane Houel, Robert Abernathy, Kutralanathan Renganathan, Karen Meyer-Arendt, Natalie G Ahn, and William M Old. Quantifying the Impact of Chimera MS/MS Spectra on Peptide Identification in Large-Scale Proteomics Studies. *Journal of Proteome Research*, 9(8):4152–4160, 2010.

[10] J D Venable, M Q Dong, J Wohlschlegel, A Dillin, and J R Yates III. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*, 1:39–45, 2004.

[11] C Masselon, L Pasa-Tolic, S W Lee, L Li, G A Anderson, R Harkewicz, and R D Smith. Identification of tryptic peptides from large databases using multiplexed tandem mass spectrometry: simulations and experimental results. *Proteomics*, 3(7), 2003.

[12] A B Chakraborty, S J Berger, and J C Gebler. Use of an integrated MS-multiplexed MS/MS data acquisition strategy for high-coverage peptide mapping studies. *Rapid Communications in Mass Spectrometry*, 21(5), 2007.

[13] T Geiger, J Cox, and M Mann. Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation. *Molecular & Cellular Proteomics*, 9(10):2252, 2010.

[14] L C Gillet, P Navarro, S Tate, H Röst, N Selevsek, L Reiter, R Bonner, and R Aebersold. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics*, 11(6), 2012.

[15] K. Blackburn, F. Mbeunkui, S.K. Mitra, T. Mentzel, and M.B. Goshe. Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation. *Journal of proteome research*, 9(7):3621–3637, 2010.

[16] N Zhang, X Li, M Ye, S Pan, B Schwikowski, and R Aebersold. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics*, 5(16):4096–4106, 2005.

[17] J. Wang, P.E. Bourne, and N. Bandeira. Peptide identification by database search of mixture tandem mass spectra. *Molecular & Cellular Proteomics*, 10(12), 2011.

[18] L Käll, J D Canterbury, J Weston, W S Noble, M J MacCoss, L Kall, J D Canterbury, J Weston, W S Noble, and M J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, 2007.

[19] M Bern, Y Cai, and D Goldberg. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem*, 79(4):1393–1400, 2007.

[20] A I Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics*, 73(11):2092–2123, 2010.

[21] S Kim, N Gupta, and P A Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res*, 7(8):3354–3363, 2008.

[22] Xin Zhang, Yunzi Li, Wenguang Shao, and Henry Lam. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *PROTEOMICS*, 11(6):1075–1085, 2011.

[23] B L Atwater, D B Stauffer, F W McLafferty, and D W Peterson. Reliability ranking and scaling improvements to the probability based matching system for unknown mass spectra. *Analytical Chemistry*, 57(4):899–903, 1985.

[24] N Bandeira, H Tang, V Bafna, and P Pevzner. Shotgun protein sequencing by tandem mass spectra assembly. *Analytical Chemistry*, 76:7221–7233, 2004.

[25] R Craig, J C Cortens, D Fenyo, and R C Beavis. Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res*, 5:1843–1849, 2006.

[26] H Lam, E W Deutsch, J S Eddes, J K Eng, S E Stein, and R Aebersold. Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods*, 5:873–875, 2008.

[27] B E Frewen, G E Merrihew, C C Wu, W S Noble, and M J MacCoss. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal Chem*, 78:5678–5684, 2006.

[28] H Lam, E W Deutsch, J S Eddes, J K Eng, N King, S E Stein, and R Aebersold. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, 7(5):655–667, 2007.

[29] S Tanner, H Shu, A Frank, L C Wang, E Zandi, M Mumby, P A Pevzner, and V Bafna. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem*, 77:4626–4639, 2005.

[30] J E Elias and S P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–214, March 2007.

[31] J Li, L J Zimmerman, B H Park, D L Tabb, D C Liebler, and B Zhang. Network-assisted protein identification and data interpretation in shotgun proteomics. *Molecular Systems Biology*, 5(1), 2009.

[32] J A Falkner, J W Falkner, and P C Andrews. ProteomeCommons. org IO Framework: reading and writing multiple proteomics data formats. *Bioinformatics*, 23(2):262, 2007.

[33] J D Venable and J R Yates. Impact of ion trap tandem mass spectra variability on the identification of peptides. *Anal Chem*, 76:2928–2937, 2004.

[34] L N Mueller, M Y Brusniak, D R Mani, and R Aebersold. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res*, 7(1):51–61, 2008.

[35] S Kim, N Gupta, N Bandeira, and P A Pevzner. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Molecular & Cellular Proteomics*, 8(1):53, 2009.

[36] J Wang, J Perez-Santiago, J E Katz, P Mallick, and N Bandeira. Peptide identification from mixture tandem mass spectra. *Molecular & Cellular Proteomics*, 9(7):1476–1485, 2010.

[37] C Cortes and V Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[38] Henry Lam, Eric W Deutsch, and Ruedi Aebersold. Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics. *Journal of Proteome Research*, 9(1):605–610, 2010.

[39] A Michalski, E Damoc, J P Hauschild, O Lange, A Wieghaus, A Makarov, N Nagaraj, J Cox, M Mann, and S Horning. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Molecular & Cellular Proteomics*, 2011.

[40] S E Stein and P A Rudnick, editors. *NIST Peptide Tandem Mass Spectral Libraries.* National Institute of Standards and Technology, 2010.

[41] S Choi, J Jeong, S Na, H S Lee, H Y Kim, K J Lee, and E Paek. New Algorithm for the Identification of Intact Disulfide Linkages Based on Fragmentation Characteristics in Tandem Mass Spectra. *Journal of Proteome Research*, 9(1):626–635, 2009.

[42] P G A Pedrioli, B Raught, X D Zhang, R Rogers, J Aitchison, M Matunis, and R Aebersold. Automated identification of SUMOylation sites using mass spectrometry and SUMmOn pattern recognition software. *Nature Methods*, 3(7):533–539, 2006.

[43] B Schilling, R H Row, B W Gibson, X Guo, and M M Young. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *Journal of the American Society for Mass Spectrometry*, 14(8):834–850, 2003.

[44] O Rinner, J Seebacher, T Walzthoeni, L Mueller, M Beck, A Schmidt, M Mueller, and R Aebersold. Identification of cross-linked peptides from large sequence databases. *Nature methods*, 5(4):315–318, 2008.

[45] V Danc\'ik, T A Addona, K R Clauser, J E Vath, and P A Pevzner. De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol*, 6:327–342, 1999.

[46] Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert JR Heck, and Pavel A Pevzner. The generating function of cid, etd, and cid/etd pairs of tandem mass spectra: applications to database search. *Molecular & Cellular Proteomics*, 9(12):2840–2852, 2010.

[47] R G Krishna and F Wold. Post-Translational Modification of Proteins. *Advances in enzymology and related areas of molecular biology*, pages 265–298, 1993.

[48] X J Yang. Multisite protein modification and intramolecular signaling. *Oncogene*, 24(10):1653–1662, 2004.

[49] M Mann and O N Jensen. Proteomic analysis of post-translational modifications. *Nature biotechnology*, 21(3):255–261, 2003.

[50] E S Witze, W M Old, K A Resing, and N G Ahn. Mapping protein post-translational modifications with mass spectrometry. *Nature methods*, 4(10):798–806, 2007.

[51] O N Jensen. Interpreting the protein language using proteomics. *Nature Reviews Molecular Cell Biology*, 7(6):391–403, 2006.

[52] K Rikova, A Guo, Q Zeng, A Possemato, J Yu, H Haack, J Nardone, K Lee, C Reeves, Y Li, and Others. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell*, 131(6):1190–1203, 2007.

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

[54] R G Spiro. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology*, 12(4):43R—-56R, 2002.

[55] R. Geiss-Friedlander and F. Melchior. Concepts in sumoylation: a decade on. *Nature reviews Molecular cell biology*, 8(12):947–956, 2007.

[56] M.J. Pearce, J. Mintseris, J. Ferreyra, S.P. Gygi, and K.H. Darwin. Ubiquitin-like protein involved in the proteasome pathway of mycobacterium tuberculosis. *Science Signalling*, 322(5904):1104, 2008.

[57] K. Ueda and O. Hayaishi. Adp-ribosylation. *Annual review of biochemistry*, 54(1):73–100, 1985.

[58] E. Meulmeester and F. Melchior. Cell biology: Sumo. *Nature*, 452(7188):709–711, 2008.

[59] Y.Q. Zhang and K.D. Sarge. Sumoylation of amyloid precursor protein negatively regulates a [beta] aggregate levels. *Biochemical and biophysical research communications*, 374(4):673–678, 2008.

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

[61] M.S. Rodriguez, C. Dargemont, and R.T. Hay. Sumo-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. *Journal of Biological Chemistry*, 276(16):12654–12659, 2001.

[62] I Matic, J Schimmel, I A Hendriks, M A van Santen, F van de Rijke, H van Dam, F Gnad, M Mann, and A C O Vertegaal. Site-specific identification of SUMO-2 targets in cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif. *Molecular cell*, 39(4):641–652, 2010.

[63] M. Knuesel, H.T. Cheung, M. Hamady, K.K.B. Barthel, and X. Liu. A method of mapping protein sumoylation sites by mass spectrometry using

a modified small ubiquitin-like modifier 1 (sumo-1) and a computational program. *Molecular & Cellular Proteomics*, 4(10):1626–1636, 2005.

[64] I. Matic, M. van Hagen, J. Schimmel, B. Macek, S.C. Ogg, M.H. Tatham, R.T. Hay, A.I. Lamond, M. Mann, and A.C.O. Vertegaal. In vivo identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an in vitro to in vivo strategy. *Molecular & cellular proteomics*, 7(1):132–144, 2008.

[65] J. Schimmel, K.M. Larsen, I. Matic, M. van Hagen, J. Cox, M. Mann, J.S. Andersen, and A.C.O. Vertegaal. The ubiquitin-proteasome system is a key component of the sumo-2/3 cycle. *Molecular & Cellular Proteomics*, 7(11):2107–2122, 2008.

[66] H.A. Blomster, S.Y. Imanishi, J. Siimes, J. Kastu, N.A. Morrice, J.E. Eriksson, and L. Sistonen. In vivo identification of sumoylation sites by a signature tag and cysteine-targeted affinity purification. *Journal of Biological Chemistry*, 285(25):19324–19329, 2010.

[67] F. Galisson, L. Mahrouche, M. Courcelles, E. Bonneil, S. Meloche, M.K. Chelbi-Alix, and P. Thibault. A novel proteomics approach to identify sumoylated proteins and their modification sites in human cells. *Molecular & Cellular Proteomics*, 10(2), 2011.

[68] J.A. Wohlschlegel, E.S. Johnson, S.I. Reed, and J.R. Yates III. Improved identification of sumo attachment sites using c-terminal sumo mutants and tailored protease digestion strategies. *Journal of proteome research*, 5(4):761–770, 2006.

[69] P.G.A. Pedrioli, B. Raught, X.D. Zhang, R. Rogers, J. Aitchison, M. Matunis, and R. Aebersold. Automated identification of sumoylation sites using mass spectrometry and summon pattern recognition software. *Nature methods*, 3(7):533–539, 2006.

[70] H H Hsiao, E Meulmeester, B T C Frank, F Melchior, and H Urlaub. ChopNspice,a mass-spectrometric approach that allows identification of endogenous SUMO-conjugated peptides. *Molecular & Cellular Proteomics*, 2009.

[71] M.J. Miller, G.A. Barrett-Wilt, Z. Hua, and R.D. Vierstra. Proteomic analyses identify a diverse array of nuclear processes affected by small ubiquitin-like modifier conjugation in arabidopsis. *Proceedings of the National Academy of Sciences*, 107(38):16512–16517, 2010.

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

[73] J.S. Cottrell and U. London. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[74] Carol V Robinson, Andrej Sali, and Wolfgang Baumeister. The molecular sociology of the cell. *Nature*, 450(7172):973–82, 2007.

[75] P Uetz, L Giot, G Cagney, T A Mansfield, R S Judson, J R Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, A Qureshi-Emili, Y Li, B Godwin, D Conover, T Kalbfleisch, G Vijayadamodar, M Yang, M Johnston, S Fields, and J M Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[76] L Giot, J S Bader, C Brouwer, A Chaudhuri, B Kuang, Y Li, Y L Hao, C E Ooi, B Godwin, E Vitols, G Vijayadamodar, P Pochart, H Machineni, M Welsh, Y Kong, B Zerhusen, R Malcolm, Z Varrone, A Collis, M Minto, S Burgess, L McDaniel, E Stimpson, F Spriggs, J Williams, K Neurath, N Ioime, M Agee, E Voss, K Furtak, R Renzulli, N Aanensen, S Carrolla, E Bickelhaupt, Y Lazovatsky, A DaSilva, J Zhong, C A Stanyon, R L Finley, K P White, M Braverman, T Jarvie, S Gold, M Leach, J Knight, R A Shimkets, M P McKenna, J Chant, and J M Rothberg. A protein interaction map of Drosophila melanogaster. *Science*, 302(5651):1727–1736, 2003.

[77] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[78] Roland A Pache and Patrick Aloy. Incorporating high-throughput proteomics experiments into structural biology pipelines: identification of the low-hanging fruits. *Proteomics*, 8(10):1959–1964, 2008.

[79] Sylvie Lalonde, David W Ehrhardt, Dominique Loqué, Jin Chen, Seung Y Rhee, and Wolf B Frommer. Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. *The Plant Journal*, 53(4):610–635, 2008.

[80] Raymond C Stevens, Shigeyuki Yokoyama, and Ian A Wilson. Global efforts in structural genomics. *Science Signalling*, 294(5540):89, 2001.

[81]

[82] Helen Berman, Kim Henrick, Haruki Nakamura, and John L Markley. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, 35(Database issue):D301–D303, 2007.

[83] Haydyn D T Mertens and Dmitri I Svergun. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *Journal of Structural Biology*, 172(1):128–141, 2010.

[84] Chirlmin Joo, Hamza Balci, Yuji Ishitsuka, Chittanon Buranachai, and Taekjip Ha. Advances in single-molecule fluorescence methods for molecular biology. *Annual Review of Biochemistry*, 77(1):51–76, 2008.

[85] Keiji Takamoto and Mark R Chance. Radiolytic protein footprinting with mass spectrometry to probe the structure of macromolecular complexes. *Annual Review of Biophysics and Biomolecular Structure*, 35(January):251–276, 2006.

[86] Henning Stahlberg and Thomas Walz. Molecular Electron Microscopy: State of the Art and Current Challenges. *ACS Chemical Biology*, 3(5):268–281, 2008.

[87] A Sinz. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom Rev*, 25:663–682, 2006.

[88] Tord Berggård, Sara Linse, and Peter James. Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7(16):2833–2842, 2007.

[89] Feixia Chu, Peter R Baker, Alma L Burlingame, and Robert J Chalkley. Finding chimeras: a bioinformatics strategy for identification of cross-linked peptides. *Molecular & Cellular Proteomics*, 9(1):25–31, 2010.

[90] Leo J Koning, Piotr T Kasper, Jaap Willem Back, Merel A Nessen, Frank Vanrobaeys, Jozef Beeumen, Ermanno Gherardi, Chris G Koster, and Luitzen Jong. Computer-assisted mass spectrometric analysis of naturally occurring and artificially introduced cross-links in proteins and protein complexes. *FEBS Journal*, 273(2):281–291, 2005.

[91] Hua Xu, Liwen Zhang, and Michael A Freitas. Identification and characterization of disulfide bonds in proteins and peptides from tandem ms data by use of the massmatrix ms/ms search engine. *Journal of proteome research*, 7(01):138–144, 2007.

[92] Sean McIlwain, Paul Draghicescu, Pragya Singh, David R Goodlett, and William Stafford Noble. Detecting cross-linked peptides by searching against a database of cross-linked peptide pairs. *Journal of proteome research*, 9(5):2488–2495, 2010.

[93] Michael Götze, Jens Pettelkau, Sabine Schaks, Konstanze Bosse, Christian H Ihling, Fabian Krauth, Romy Fritzsche, Uwe Kühn, and Andrea Sinz. Stavroxa software for analyzing crosslinked products in protein interaction studies. *Journal of the American Society for Mass Spectrometry*, pages 1–12, 2011.

[94] Bing Yang, Yan-Jie Wu, Ming Zhu, Sheng-Bo Fan, Jinzhong Lin, Kun Zhang, Shuang Li, Hao Chi, Yu-Xin Li, Hai-Feng Chen, et al. Identification of cross-linked peptides from complex samples. *Nature Methods*, 9(9):904–906, 2012.

[95] Alexander Leitner, Roland Reischl, Thomas Walzthoeni, Franz Herzog, Stefan Bohn, Friedrich Förster, and Ruedi Aebersold. Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. *Molecular cellular proteomics MCP*, 11(3):M111.014126, 2012.

[96] Schutz F, Kapp EA, Simpson RJ, Speed TP, F Schutz, E A Kapp, R J Simpson, and T P Speed. Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem Soc Trans*, 31:1479–1483, 2003.

[97] Thomas Walzthoeni, Manfred Claassen, Alexander Leitner, Franz Herzog, Stefan Bohn, Friedrich Förster, Martin Beck, and Ruedi Aebersold. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nature Methods*, 9(9):901–903, 2012.

[98] A Bairoch, R Apweiler, C H Wu, W C Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, and Others. The universal protein resource (UniProt). *Nucleic acids research*, 33(suppl 1):D154—-D159, 2005.

[99] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, TN Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[100] F.G. Whitby, E.I. Masters, L. Kramer, J.R. Knowlton, Y. Yao, C.C. Wang, and C.P. Hill. Structural basis for the activation of 20s proteasomes by 11s regulators. *Nature*, 408(6808):115–120, 2000.

[101] E.M. Huber, M. Basler, R. Schwab, W. Heinemeyer, C.J. Kirk, M. Groettrup, and M. Groll. Immuno-and constitutive proteasome crystal structures reveal differences in substrate and inhibitor specificity. *Cell*, 148(4):727–738, 2012.