Identification of Post-translational Modifications via Blind Search of Mass-Spectra

Dekel Tsur* Stephen Tanner† Ebrahim Zandi‡ Vineet Bafna*
Pavel A. Pevzner*

Abstract

While identifying post-translational modifications (PTMs) is undoubtedly the next big step for proteomics, most MS/MS database search algorithms perform a *restrictive* search that can only take into account a few types of PTMs and ignore all others. We describe an *unrestrictive* PTM search algorithm (MS-Alignment) that searches for *all* types of PTMs at once in a *blind* mode, i.e., without knowing which PTMs exist in nature. The blind PTM identification opens a possibility to study the extent and frequency of different types of PTMs, still an open problem in proteomics. We use MS-Alignment to construct a two-dimensional *PTM frequency matrix* that reflects the number of MS/MS spectra in a sample for each putative PTM type and each amino acid. Application of this approach to lens proteins resulted in a largest set of PTMs reported in human crystallins so far. Our analysis of various MS/MS datasets implies that the biological phenomenon of modification is much more widespread than previously thought. We further argue that MS-Alignment reveals some still unknown types of modifications that warrant further experimental validation.

1 Introduction

Fueled by recent improvements in instrumentation and software, tandem mass spectrometry has become the tool of choice for protein identification. However, while popular algorithms like Sequest [11], and Mascot [21] are used extensively for peptide identification, many spectra remain unidentified by these searches. This can be attributed to several factors including poor quality of fragmentation/ionization, and the presence of post-translational modifications *(PTMs)* and mutations.

Post translational modifications greatly increase the complexity of the proteome and identifying PTMs is undoubtedly the next big step for proteomics [5, 7, 28]. However, reliable computational identification of PTMs remains a formidable challenge. The first approach to PTM identification was proposed by Yates et al., 1995 [32], who advocated the enumeration and scoring of all possible modifications for each peptide from the database. This exhaustive search approach has serious limitations since it can only take into account a few modifications and is prohibitively slow for mutation detection. In other words, a researcher has to "guess" in advance which PTMs are present in the sample, an unrealistic expectation. As a result, the current practice is to perform a *restrictive*

*Computer Science and Engineering, UC San Diego, {dtsur,vbafna,ppevzner}@cs.ucsd.edu
†Bioinformatics, UC San Diego, stanner@ucsd.edu
‡Molecular Microbiology and Immunology, School of Medicine, Univ. Southern California, zandi@usc.edu

search for a small set of PTMs (such as phosphorylation) and ignore all other PTMs. The question arises whether one can design an *unrestrictive* PTM search algorithm that searches for *all* types of PTMs at once in a *blind* mode, i.e., without knowing which PTMs exist in a sample. Another, more ambitious question is whether one can predict PTMs that are not known yet by mining large MS/MS datasets, something that was never done before. Additionally, the blind PTM identification approach opens a possibility to study the extent and frequency of different types of PTMs, still an open problem in proteomics.

The first blind approach to PTM identification (*spectral alignment*) was proposed by Pevzner et al., 2000, 2001, [22, 23]. Recently, Searle et al., 2004 (OpenSea) [27] and Han et al., 2004 [13] (SPIDER) proposed yet another approach to blind PTM identification. In contrast to spectral alignment, these approaches rely on de novo interpretation of MS/MS spectra. For example, Han et al. [13] formulate the problem as the identification of a modified peptide that best matches both the de novo interpretation and the database peptide. While this elegant formulation accommodates some de novo sequencing errors, the approach depends critically on a good de novo interpretation. We emphasize the important difference between the spectral alignment approach [22, 23] and approaches in [27] and [13]. Searle et al., 2004 [27] essentially use the heuristic branch and bound technique that, in contrast with spectral alignment, (i) does not guarantee the optimal solution and (ii) crucially depends on the quality of de novo reconstruction. Han et al., 2004 [13] use a rigorous dynamic programming algorithm (that is similar to spectral alignment in case there is no sequencing errors) but only compare a database peptide against a *single* de novo interpretation of an experimental spectrum. The spectral alignment, on the other hand, compares a database peptide against *every* possible interpretation of an experimental spectrum. In this paper, we describe a new blind approach to PTM identification that directly aligns the spectra against the database thus eliminating dependence on accurate de novo interpretations. Our MS-Alignment algorithm (code available at http://peptide.ucsd.edu) resolves a number of open problems outlined in Pevzner et al., [22] and leads to fast and reliable PTM identification.

Identification of all types of PTMs present in a large collection of MS/MS spectra is a difficult task. An even more difficult task is to distinguish between real PTMs and computational artifacts. Using our approach we were able, for the first time, to construct a *PTM frequency matrix* PTM(Δ, a) that reflects the number of MS/MS spectra in a sample with predicted PTM Δ on amino acid a for all possible shifts Δ and all amino acids a. For example, Table 2 represents a fragment of the PTM frequency matrix of IKKb dataset (below). Some entries in this table represent interpretation artifacts rather than real PTMs. However, one can notice that while most entries in this table are small, some entries are very large (highlighted in gray). We argue that since the noise in the PTM frequency matrix is random, the large values are likely to represent real PTMs rather than artifacts. For example, the largest entries in Table 2 (PTM(16,M)=613, PTM(71,C)=297, PTM(32, M) = 224 and PTM(1,N)=191) match common PTMs (singly or doubly oxidized methionine, PAM-cysteine, deamidated arginine) [1]. These four PTMs alone increase the number of interpreted MS/MS spectra in the sample by $\approx 15\%$, a significant increase. Known chemical modifications dominate the large entries in Table 2 and therefore suggest that PTM frequency matrix is a gateway to the study of modified peptides and mining for still unknown PTM types.

All previous approaches to blind PTM search required a manual validation of the results, thereby

making it difficult to mine large data-sets for interesting modifications. Our PTM frequency matrix approach bypasses this problem by introducing a new approach to validation that compares $PTM(\Delta, a)$ with shift Δ at amino acid a with $PTM(\sigma, y)$ for a "random" shift σ at a "random" amino acid y.

Here, we demonstrate an excellent correlation between high values in PTM frequency matrix and known PTMs thus validating our approach. High values in PTM frequency matrix may point to some still unknown modifications and we provide multiple supporting evidence that they indeed may correspond to previously unknown PTMs rather than artifacts.

2 MS-Alignment

The key idea of the spectral alignment [22, 23] is to represent spectrum of a peptide with parent mass M as a boolean 0-1 sequence of length M (the positions of peaks in the spectrum correspond to positions of 1s in the sequence). In this representation, a PTM with positive shift $\Delta > 0$ is simply an insertion of Δ zeroes in the sequence, while a PTM with negative shift $\Delta < 0$ is simply a deletion of Δ zeroes from the sequence. With this model in mind, comparison of two spectra is turned into comparison of two strings in 0-1 alphabet under insertion and deletion operation. Since this is the classical *edit distance* problem in bioinformatics, the spectral alignment approach essentially transfers the power of dynamic programming and sequence alignment algorithms from genomics to mass-spectrometry.

The above paragraph is an over-simplified description of spectral alignment. While it works for comparing spectra in a case-by-case fashion, it is rather slow, does not adequately model some specifics of MS/MS spectra, and therefore has difficulties in analyzing large datasets of MS/MS spectra in an automatic fashion. Pevzner et al. [22] themselves formulated a number of open problems that need to be resolved to make spectral alignment practical. Here, we describe *MS-Alignment* , a spectral alignment algorithm that resolves those problems, using a number of improvements, as described below:

- MS-Alignment is a local or *fitting* version of spectral alignment instead of global alignment as in [22, 23]. While the global spectral alignment compares a spectrum against *every* possible peptide from a protein, the local spectral alignment compares the spectrum against entire protein at once by a single swap of the dynamic programming matrix (like in Smith-Waterman algorithm for sequence alignment). This improvement leads to roughly an order of magnitude speed-up by reducing the time spent evaluating overlapping database peptides.
- Pevzner et al. [22] reduce the spectral alignment problem to finding longest paths in the *spectral product* graph with unit-cost vertices. A more adequate scoring requires assignment of (i) intensity-based vertex scores and (ii) PTM-based edge scores in the spectral product. While Pevzner et al. [22] approach can be easily modified to address (i), it is not clear how to address (ii) without an order of magnitude increase in the running time. MS-Alignment achieves this goal with a rapid running time. Figure 1 illustrates the spectrum graph for search against a small database.
- MS-Alignment employs a PTM-dependent p-value scoring instead of the rather naive scoring in [22].

Recently, Hansen et al., 2005 [14] and Tang et al., 2005 [30] studied a problem of interpreting

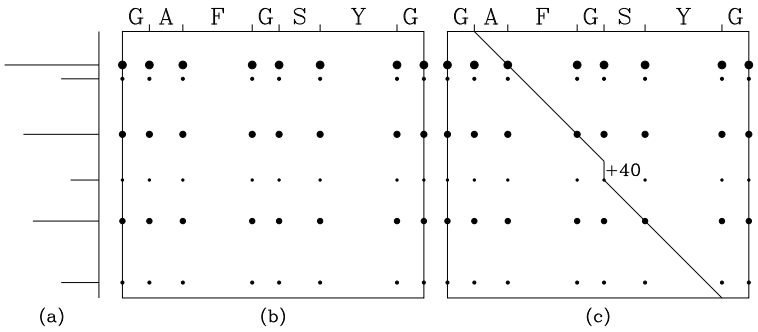

Figure 1: An example showing the spectral product graph for the spectrum $\{71.0, 100.5, 218.1, 315.1, 402.2, 532.6\}$ (shown in Figure a) and the database GAFGSYG. The spectral product graph (Figure b) contains a dot for every mass in the spectrum and every mass of a prefix of the database. In the figure, the size of the dot is proportional to the intensity of the corresponding peak in the spectrum. An interpretation of the spectrum, such as AFG#SYG (Figure c), corresponds to a path from top to bottom in the product graph. A path consists of 45^0 segments, vertical segments (corresponding to modifications with positive mass offsets), and horizontal segments (corresponding to modifications with negative mass offsets).

peptides with a single modification. In this case, the mass shift is known in advance (semi-blind search) and the edit distance is 1 thus allowing one to substitute the dynamic programming with an exhaustive search that analyzes every possible PTM position. We remark that exhaustive search for a single modification leads to a quadratic algorithm (experimental and theoretical spectra have to be compared n times where n is the peptide length) while the spectral alignment can be implemented in linear time (see Supplementary material). As a result, applicability of these approaches is mainly limited to tryptic peptides and small databases. Although the exhaustive search approaches use advanced scoring functions that are more accurate than the scoring function in MS-Alignment , we emphasize that we use MS-Alignment only for candidate generation and re-score all found hits afterwards with the rigorous InsPecT scoring [31]. This leads to significant speed-up since candidate generation is the most time-consuming step in blind PTM search.

Our method has the following steps:

Database Pruning: Note that a typical protein database is very large. We make the reasonable assumption that for every protein present in the spectral data-set, at least one unmodified peptide is present with an identifiable spectrum ([9]). We use this method to rapidly identify a much smaller subset of expressed proteins.

Candidate generation: We use MS-Alignment to align all spectra against the smaller data-set, and produce a list of 'hits' (peptides with one or more mass-offsets).

Re-scoring: We use a rigorous scoring function (see below) to re-score all the hits, and discard ones with a high P-value.

Candidate Validation: Scoring for modifications is still in early stages due to the lack of annotated spectra. We use a number of 'principles' to validate, and further cull from the list of

candidates. The remaining candidates are output.

Re-scoring and Candidate Validation

We limit our discussion to re-scoring module, and candidate validation. A complete description of the MS-Alignment algorithm can be found in the supplementary documents. There has been extensive research of late in improving the scoring of mass spectra of unmodified peptides [4, 10, 12, 16, 18–21, 24, 25, 29, 32, 33]. Unfortunately, these algorithms do not carry over to the identification of modified peptides, and most algorithms for identifying PTMs still consider simpler score functions. The issue of scoring was addressed recently by Tanner et al., 2005 [31]. They revisited restricted PTM search and developed the InsPecT algorithm which (i) incorporates recent advances in scoring into PTM analysis and (ii) is two orders of magnitude faster than other restricted PTM searches. While Tanner et al., 2005 [31] report success in revealing many previously unidentified PTMs, their algorithm also has to "guess" the types of PTMs in advance and cannot be run in a blind mode.

The common approach to assigning statistical significance [3, 16, 19, 20, 31] is to combine a number of features, including the percentage of b and y fragments found, the percentage of spectral peaks annotated, the percentage of total ion current in annotated peaks, etc. We use an SVM based approach, similar to Anderson et., 2003 [3] to optimally classify the correctly and incorrectly assigned peptides using the features described above. The SVM score for a match is compared to a histogram of SVM scores over incorrect peptide assignments to produce a p-value for the candidate (See Supplementary data).

Modifications on closely located amino acids often produce similar theoretical spectra, thus making it difficult to identify the exact position of modification. For example, consider a peptide with two consecutive Methionines, one of which is oxidized. Unless the spectrum is of particularly high quality, the two candidate peptides that place the oxidation on either residue will receive very similar scores. However, if these candidates greatly out-score other peptides, we can confidently assign a peptide annotation (with some uncertainty on the oxidation position). Therefore, we categorize search results as being either *incorrect*, Δ-*correct*, or *correct*. A Δ-*correct* result recovers the correct peptide (possibly with misplaced modifications), while a *correct* result recovers the original peptide sequence and position of modification exactly. In the case with two consecutive Methionines, and one oxidation, we would have high confidence that the match was Δ-correct, but lower confidence that it was correct.

Correspondingly, for the SVM training, we compute two δ-scores (similar to Δ_{cn} feature used by Sequest). The first is difference from the runner-up (the same peptide with different position of modification), and the second is difference in scores between the top scoring peptide and the highest scoring distinct peptide. The first δ-score is most useful for classifying matches as correct, the second for classifying them as Δ-correct.

2.1 Candidate validation via PTM frequency matrix

Even with improved scoring, reliable identifications of modifications is challenging. Many modifications, such as phosphorylation, introduce novel fragmentation patterns in the spectra. To incorporate these into the score function, one would have to train on extensive data-sets of annotated spectra of modified peptides, which do not yet exist.

On the other hand, a manual validation of the good hits is not feasible. Our search of the IKKb dataset revealed $3,905$ spectra of modified peptides with p-value lower than 0.05 ($3,571$ with a single modification and 334 with two modifications). To automate the validation, we construct a *PTM frequency matrix*.

For every peptide with predicted modification Δ on amino acid a, we incremented the count PTM(Δ, a) in the PTM frequency matrix (See Table 2). We varied Δ from -100 to 160 resulting in a 261×20 matrix. If all identified $3,571 + 334 \cdot 2 = 4239$ PTMs were incorrect, one could assume that PTM frequency matrix represents "noise" with mean value $PTM(\Delta, a) = \frac{4239}{261 \cdot 20} \approx 0.8$. In reality, while most $PTM(\Delta, a) = 0$ (90 percent of all entries), a few others represent very high values, a highly unlikely outcome for a "random" matrix. Assuming a Poisson process with mean 0.8, the probability of an entry of size 20 or higher is less than 1×10^{-20}.

To check that spurious hits are randomly distributed through the PTM frequency matrix, we performed a search of annotated spectra (from the ISB data-set) against a database containing no true matches (the human Lens database). The entropy of the resulting 2,312 matrix entries was 7.0, 90% of the maximum possible ($ln(2,312)$).

One may be tempted to output the list of large entries in PTM frequency matrix as the set of identified PTMs. With this approach, we will certainly miss the rare, yet correct modifications. Surprisingly, even for the common modifications, this approach has pitfalls. For example, $PTM(16, I) = 39$ may simply be a "shadow" of very common $PTM(16, M) = 612$ caused by Δ-correct interpretations with misplaced PTM. As another example of the complexity in interpreting this matrix, note the surprisingly many large entries $PTM(1, *)$ in the row 1 of PTM frequency matrix. The modifications with mass difference 1 should be taken with caution since they may represent an artifact caused by errors in parent mass and isotopic peaks artifacts rather than real PTMs. However, one of the entries in the first row $PTM(1, N) = 164$ is particularly large and we suggest that it represents deamidation of Asparagine rather than a parent mass artifact.

To address the issue of shadows we use a "strength in numbers" approach and *rank* all identified PTMs based on the values in PTM frequency matrix after "subtracting" of shadows as described below. The ranking procedure allows us to deal with the cases when the correct candidate was out-scored by a Δ-correct candidate containing a misplaced PTM. Below we list several ways this can happen:

- The PTM is shifted to an adjacent residue. A shift of one or two residues affects only one or two b and y peaks in the theoretical spectrum, so PTM sites may be difficult to pinpoint (particularly near peptide start/end, where peak information is scarce).
- The reported PTM is 1 Da too large. Note that a proposed PTM mass is much more likely to be too large than too small, since a candidate with a too-large PTM can still explain a run of secondary isotopic peaks.
- The match may be one residue too short (or too long), and include a PTM which compensates for the missing (or additional) residue. For example, peptide $QA[111]EVAHMSQTQEEK$ and slightly longer peptide $Q[-17]QAEVAHMSQTQEEK$ (from the IKKb data-set) produce near-identical spectra, but the latter has a simple explanation (ammonia loss) and the former does not (mass(Q)-17 $\approx$ 111).

On some spectra, our search may fall victim to these pitfalls. Therefore, we adopt an iterative

approach that begins with enumerating all spectral annotations whose match score has a p-value of .05 or less. Some spectra will have several such annotations, generally all variants of a single peptide.

Initially, we consider each spectrum to be unlabeled. At each stage, select the (Δ, a) entry which can annotate the greatest number of unlabeled spectra. Label all the unlabeled spectra which can be annotated using this new PTM, and keep a list L of the new peptide annotations. Then, label any unlabeled spectrum carrying an annotation which is a shadow of an annotation in L. A spectrum is a shadow of another if it has the same amino acid sequence, but (a) one PTM differs by 1 Da, or (b) a PTM attachment site has been shifted by one residue. This procedure allows true PTMs "explain away" their shadows. Note that distinct PTMs of similar masses (e.g. +42 and +43) can still be identified, as long as they do not co-occur at the same position. We stop when the next PTM can annotate few new spectra.

This process produces a ranked list of PTMs which are present in the sample with a high degree of confidence. The resulting list of PTMs is less noisy than the raw PTM frequency matrix. Table 3 illustrates the utility of this ranking procedure when applied to the PTM frequency matrix in Table 2. We remark that although PTM frequency matrix shows 612+83=695 spectra with $PTM(16, M)$ and $PTM(16, W)$ modifications, the ranking procedure reveals 803 spectra with these modifications. This is due to the fact that most annotations in the row $PTM(16, *)$ are shadows of $PTM(16, M)$ (and, to a smaller extent, $PTM(16, W)$) that were corrected by the ranking procedure.

Once the list of PTMs have been selected, we can further validate particular modification sites. The same PTM site supported by multiple spectra is more reliable than a PTM site supported by a single spectrum. In addition, two overlapping peptides validating the same PTM are more robust evidence than two identical spectral annotations for the same PTM. Such "overlapping witness" evidence is most abundant for low-complexity samples, and samples which have been subjected to non-specific digest. We also report, for each modified peptide, the number of times the unmodified peptide was observed. This can provide additional positive evidence for the modified annotation (since the peptide is known to be present), as well as a rough measure of the frequency with which the particular residue is modified.

3 Results

3.1 Datasets

We use the following MS/MS datasets in our computational experiments:

- **IKKb dataset** 45,500 spectra acquired from a digestion of the inhibitor of nuclear factor kappa B kinase beta subunit (IKKb protein) by multiple proteases (E. Zandi and T. Higashimoto, unpublished data). Spectra were acquired on a Thermo Finigan LTQ mass spectrometer. These spectra were previously used in a restrictive search [31]. A database pruning step on Swissprot (54Mb) resulted in a database containing the fusion protein and proteases (10 proteins, 5kb). The pruning step took 6 hours.
- **Lens dataset** 43,518 spectra acquired from human lens proteins [26]. Of these, 12,800 were acquired on a Micromass QTOF-2. The remainder were acquired on a ThermoFinnigan LCQ Classic ion trap mass spectrometer. A major component of the lens proteome comprises of

(a)

	0 mutations	1 mutation	2	3+	Overall
Correct	97.1%	54.8%	8.1%	0.0%	42.5%
Δ-correct	2.9%	35.5%	70.3%	45.5%	38.3%
Incorrect	0.00%	9.7%	21.6%	54.6%	19.2%

(b)

	1 PTM	2 PTMs	Overall
Correct	57.3%	15.6%	36.4%
Δ-correct	35.4%	67.2%	51.3%
Incorrect	7.3%	17.2%	12.3%

Table 1: (a) MS-Alignment performance on MutDB (a) and ModSpec (b) data-sets.

crystallins, which have very little turnover, and acquire modifications with age. When a person ages, the crystallins become insoluble, and the tissue increasingly opaque often leading to cataract. Post-translational modifications are known to play a major role in the process [19, 26]. A database pruning step resulted in a database of crystallin proteins that represent the majority of proteins in human lens tissue (20 proteins, 5kb).

- **ISB dataset** 37,000 spectra from a public collection of MS/MS spectra acquired on a ThermoFinnigan ESI-ITMS [17]. This data-set was chosen as it has been queried extensively, but many spectra remain unannotated. The ISB spectra were searched against a database of 37 proteins (25kb).

In addition we apply our analysis to two simulated data-sets. As large data-sets of annotated spectra of modified peptides are not currently available, these data-sets are valuable in testing the accuracy of our approach.

ModSpec We "modify" annotated spectra (by shifting some peaks) from the ISB data-set to simulate modifications and match them against the database. We constructed a data-set of modified spectra by adding feasible modifications to each peptide as described in Tanner et al., 2005 [31].

MutDb We match previously identified ISB spectra against a "mutated" database. The database was mutated to a sequence identity level of 90% (1 or 2 mutations for most tryptic peptides).

3.2 Results on ModSpec and MutDb

Table 1 presents simulation results for ModSpec and MutDB. Overall, 81% (88%) of interpretations were correct and Δ-correct for MutDb (ModSpec) simulations.[1] Certainly, at the first glance the rate of incorrect identifications seems high (12-19%). However, the PTM frequency matrix approach provides a way to ignore incorrect annotations since they are somewhat randomly distributed over PTM frequency matrix and thus get discarded during our ranking procedure.

3.3 Results on IKKb

We analyzed spectra in the IKKb dataset, retaining matches with p-value below 0.05. MS-Alignment identified 8,641 unmodified peptides, 3,571 spectra with a single PTM and 334 spectra with two PTMs. Table 2 shows the PTM frequency matrix while Table 3 presents the ranked list of PTMs. Most large entries in PTM frequency matrix (highlighted in gray) do indeed correspond to known modifications, validating our approach.

The activation of the inhibitor kappaB kinase (IKK) complex and its relationships to insulin resistance was subject to intensive studies recently. However, the only known PTMs of IKK are phosphorylations, which does not explain mechanisms of signaling and activation/inactivation of

[1] We are unable to compare MS-Alignment against other algorithms, as none of the other approaches quantify their accuracy and sensitivity.

Offset	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	(N)	(C)	Total
-18		1	19*	10	1*		5	3	1	1		3	4	4	9	9	56	1			44	20	127
1	12		24	24*	17	40	17	34*	4*	77*	2	191*	43	29*	3	19	30*	13*	11	23	4	2	613
14	6			3				4*	93		40		1			7*		12*			90	49	166
16	1*		12*	26	1*	11	2	39*		24*	613*	4*	1*	9		1*		1*	84	1	62	81	830
17			1*	1		3		3*		8*	71*	10*		8		1		1	17	1	1	3	125
22	7		11*	18	8	24	5	16		9	1	11	1	18		16	17	18		8	1	1	188
28	1*	2*						9	154*	21	1					30*		42*			157	33	260
32		2						1			224*	1	1*		5				20		186	1	254
53			36	41		4		17	1	12	1	8	14	3		1	2	5		2			147
71		297								1	1		1								14	1	300
156	2							12		77			6			3	1				85		101

Table 2: PTM frequency matrix for IKKb dataset (4,239 PTM annotations total). The first column indicates the mass shift of a modification. The entry for modification Δ, and amino acid a refers to the number of times a appeared with modification of mass Δ in a top-scoring spectral interpretation. Columns (N) and (C) give the total number of times the modification occurred on the N-terminal and C-terminal amino acids. Only rows with a total of 100 or more are shown. Entries of 50 or more are shown in grey, and entries that can be explained by mutations are starred.

IKK by over 200 different stimuli, including cytokines, chemicals, ionizing and UV radiation, oxidative stress, etc. Moreover, the mechanisms of stimulations by these different stimulants are not known yet. It is likely that different stimuli use different mechanisms of signaling involving yet unknown PTM sites. Table 3 illustrates that even a single protein (IKKb) from the IKK complex has an unusually rich set of PTMs. Further investigation of these PTMs may lead to the discovery of novel signaling mechanisms, for example by oxidative stress.

Table 4 lists some of the modification sites with strong support from multiple overlapping annotations. Because this sample was digested with multiple proteases (including trypsin, proteinase K, and elastase), overlapping peptides are available for most positions. The IKKb spectra suggest the presence of mutations, as well as previously unreported PTM types. Most of these modifications are missed by traditional database search.

The above results were obtained by running MS-Alignment with up to 2 modifications. Peptides with more than 2 modifications were never reported in the past since such peptides are difficult to find and validate. MS-Alignment with up to 3 modifications found two peptides with strong support for 3 modifications: GM[16]QNLAPND[22]LPLLAME[16] and K[28]IITHPNFN[1]GNTLDNDIM[16]LI. With the exception of D[22], each of these modification sites has been confirmed by multiple singly- or doubly-modified spectra.

3.4 Results on Lens

Our blind search of Lens dataset identified 5,616 unmodified spectra, 3,027 spectra with a single modification, and 244 spectra with two modifications. Table 5 summarizes the PTMs obtained through iterative selection. As with IKKb, the majority of annotations use known chemical modifications while most of the mis-annotations are modifications of mass +1. While most modifications in Table 5 are known, we also find some novel ones, including putative carboxylation of tryptophan.

After the blind search discovers the set of PTMs present in a data-set, it is not unreasonable to perform a restricted search pass with a list of identified PTMs using InsPecT [31]. This second

AA	Mass	Spectra	Sites	Putative Annotation
M,W	16	803	50	oxidation
C	71	332	9	PAM-cys
M,W	32	248	15	double oxidation
N	1	225	36	deamidation
K	28	184	3	dimethylation
non-specific	22	176	64	sodium
K,M	14	154	4	methylation
E,D,P	53	130	28	Fe(III)
T,E,D	-18	117	16	dehydration
V	28	56	2	dimethylation
K	-57	46	1	mutation to alanine
S	28	30	1	mutation to aspartate
M	38	23	4	potassium
C	76	22	3	beta-mercaptoethanol
M	-2	21	3	mutation to glutamate
non-specific	1	355	133	isotopic peaks
L	156	92	1	truncated K+28
I	16	49	3	misplaced oxidation
L	17	27	6	misplaced oxidation
non-specific	2	22	17	isotopic peaks
I	44	20	2	misplaced K+28 and M+16
L	54	19	7	shadow of Fe(III)
L	34	17	3*	mutation to phenylalanine
V	15	18	2*	mutation to asparagine
M	-32	15	5*	mutation to valine

Table 3: Ranked list of modifications for the IKKb data-set (according to the number of spectra with modification (Δ, a)). Modifications below the horizontal line were judged to be spurious. Modifications were selected iteratively, halting when the next PTM would annotate fewer than 15 new spectra. Each modification in the table was supported by multiple peptides, except for the starred entries. Each modification other than S+28 was supported by the presence of the same peptide lacking the modification. Modifications on different amino acids with the same shift Δ are combined in a single line. Modification with offset 53 is not listed in ABRF database but is described in Hunyadi-Gulyas and Medzihradszky, 2004 [15] as Fe(III) adduct.

pass can use knowledge of fragmentation effects of identified PTMs, such as phosphorylation and oxidized methionine. This second pass also has the benefit of generating few δ-correct annotations, due to the much smaller search space. We re-searched the Lens data-set using the PTMs from Table 5. We obtained a total of 11,278 spectral annotations with p-value < 0.05, resulting in a 40% increase in the number of spectral annotations of modified peptides. The results include 6,672 unmodified spectra, 3,526 spectra with one modification, and 1,080 spectra with two modifications.

Peptide	count	
28 on K (dimethylation or mutation to R)		
K*LSSPATL	17	0
K*LSSPATLN	6	0
K*LSSPATLNS	55	0
K*LSSPATLNSR	19	0
K*LSSPATLN+1SR	5	0
IK*LSSPATLNSR	1	0
IMLIK*LSSPATLN	1	0
IMLIK*LSSPATLNSR	17	0
IMLIK*LSSPATLN+1SR	1	0
DNDIMLIK*LSSPATLNSR	1	0
TLDNDIMLIK*LSSPATLNSR	1	0
IITHPNFNGNTLDNDIMLIK*LSSPATLNSR	3	0
K*IITHPNFNGN	21	0
K*IITHPNFN+1GN	1	0
K*IITHPNFNGNTL	6	0
K*IITHPNFN+1GNTL	2	0
GNEQFINAAK*IITHPN	1	0
LGEHNIDVLEGNEQFINAAK*	4	20
-57 on K (unknown)		
WHNQETGEQIAIK*	25	26
WHNQETGEQIAIK*Q	22	0
14 on K (methylation)		
K*LSSPATL	9	0
K*LSSPATLN	1	0
K*LSSPATLNS	36	0
K*LSSPATLNSR	8	0
IMLIK*LSSPATLNSR	1	0
TLDNDIM+16LIK*	4	11
IITHPNFNGNTLDNDIMLIK*	4	6
IITHPNFN+1GNTLDNDIMLIK*	2	2
IITHPNFNGNTLDNDIM+16LIK*	4	24
22 on G (sodium)		
NYPNG*GFTAE	2	0
AFPSAINQDNYPNG*GFTAE	4	3
AFPSAINQDNYPN+1G*GFTAE	1	21
76 on C (beta-mercaptoethanol)		
SC*ILQEPK	13	37
PQPESVSC*ILQEPK	3	4
32 on M (double oxidation)		
THPNFNGNTLDNDIM*LI	3	3
IITHPNFNGNTLDNDIM*LI	2	0
IITHPNFNGNTLDNDIM*LIK	1	6
M*MALQTDIVDLQ	10	0
M*MALQTDIVDLQR	160	3
M*MALQTD+53IVDLQ	3	0
M*MALQTDIVD+22LQ	3	0

Peptide	count	
16 on M (oxidation)		
EM*EQAVELCGR	6	5
WREM*EQAVEL	3	0
AWREM*EQAVEL	10	0
LPEM*LK	3	0
DFLSKLPEM*	9	0
DFLSKLPEM*L	38	35
71 on C (PAM-cys)		
C*ISDGKLNEGH	9	1
PDKPATQC*ISDGK	1	0
IPDKPATQC*ISDGK	22	0
LIPDKPATQC*ISDGK	2	0
AGLALIPDKPATQC*ISDGK	104	0
AGLALIPDKPATQC*ISDGKLNE	12	0
AGLALIPDKPATQC*IS+1DGKLNE	1	0
1 on N (deamidation)		
TYLN*GDHVTHPDFM	2	0
TYLN*GDHVTHPDFM+16	2	0
TYLN*GDHVTHPDFM+16L	8	0
WHN*QETGEQI	1	39
WHN*QETGEQIA	5	55
WHN*QETGEQIAIK	1	26
-18 on T (dehydration)		
T*GGFGNVIR	50	0
LGT*GGFGNVIR	5	184
LGT*GGFGN+1VIR	1	4
53 on D,E (unknown)		
LGEHNID*VLE	1	119
LGEHNID*VLEGNEQ	2	35
LGEHNID*VLEGNEQFINAAK	2	20
NIDVLE*GNEQ	7	5
NIDVLE*GNEQFI	1	14
NIDVLE*GNEQFINAA	2	15
LGEHNIDVLE*GNEQ	1	35
LGEHNIDVLE*GNEQFINAAK	1	20
IQQDTGIPE*EDQE	2	0
IQQDTGIPE*EDQELL	6	15
IQQDTGIPE*EDQELLQ	1	2
IQQDTGIPEE*DQELL	7	15
28 on S (mutation to D)		
GPGTS*ILSTWIGGSTR	3	0
FGPGTS*ILSTWIGGSTR	1	0
DIFGPGTS*ILSTWIGGSTR	21	0
DIFGPGTS*ILSTWIGGSTRSISGT	2	0
DIFGPGTS*ILSTWIGGSTRSISGTSMATPHVAGLA	3	0

Table 4: Validation of PTMs described by multiple occurrences of identical or overlapping peptides (IKKb data-set). The second column is the number of occurrences of the spectra of modified peptide, while the third column is the count of the spectra of unmodified peptides.

A total of 519 modification sites were confirmed by two or more spectra. We further filtered this list to modifications which had the modification confirmed by successive "rungs" in the b and/or y ion ladder. In addition, we rejected any modification of +1 not confirmed by at least one QTOF spectrum. This heavily filtered list contains 378 modification sites supported by 4,442 spectra.

As we are using the same data-set as Searle et al., 2005 [26], we can compare some of the findings. They report a total of 80 modified sites (44 previously known, 36 novel). On the same data-set, our algorithm found 57 out of the previously reported 80 sites and discovered 322 new PTM sites. In addition, we were able to identify a wider range of modifications as shown in Table 6. The list of all modified sites can be found in the supplemental tables. Of the 23 sites annotated by OpenSea, but not by us, 12 were identified but did not pass our strict validation test (p-value $< .05$). We therefore argue that while MS-Alignment is more conservative in validating the found PTMs, it still finds many more PTMs than OpenSea. Carboxylation (see 2) is an added modification believed to be relevant to disease progression.

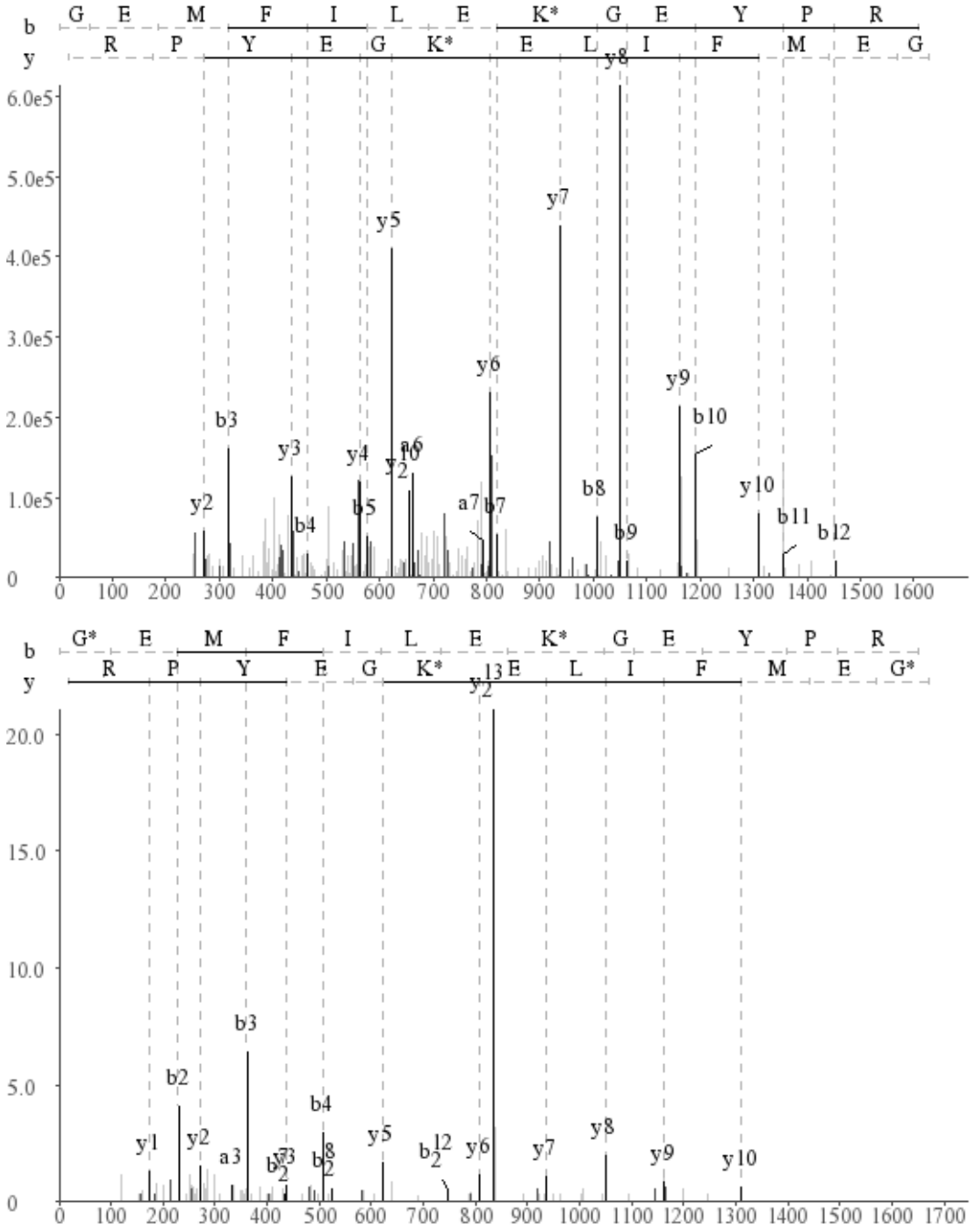

Figure 2: Two spectra from the lens data-set supporting placement of carboxylation at lysine K117 of cysstallin beta 1. Support from two peptide species (one with a carbamylation on the N-terminal glycine, one without) provides additional support.

AA	Mass	Spectra	Sites	Putative annotation
Q,N	1	499	100	deamidation
N-terminus	42	446	12	acetylation
N-terminus	43	308	98	carbamylation
non-specific	38	229	102	potassium
C	14	160	14	methylation
M,W	16	143	37	oxidation
Q	-17	103	14	deamidation
S,E,T	-18	94	27	dehydration
M	58	53	2	oxidation+acetylation
K	43	47	17	carbamylation
W	44	23	6	carboxylation
N	-17	22	5	deamidation
S	28	21	14	formylation
S	80	20	5	phosphorylation
K	58	17	6	carboxylation
K	72	14	6	unknown
W	-2	10	3	cross-linking
R	55	9	5	unknown

Table 5: Ranked list of modifications for the Lens data-set. Iterative selection was stopped when no additional PTMs could explain more than eight additional spectra. Each PTM was confirmed by two or more overlapping peptides. Each PTM, with the exceptions of +42 (constitutive acetylation), was confirmed by the presence of unmodified annotations of the same peptide. Carbamylation (+43) can be seen on the N-terminus of any peptide fragment, while acetylation (+42) is specific to the N-terminus of the entire protein. The modification with shift 58 on K is not listed in the ABRF database, it was discovered in [2] and recently reported in lens proteins [8].

AA	Mass	OpenSea	InsPecT
Q,N	1	34	94
M,W	16	25	37
S,H	28	5	11
C,H	14	8	10
S,T	80	2	6
N-terminus	42	6	5
N-terminus	43	a	75
non-specific	38	a	71
S,T	-18	a	29
Q	-17	a	16
K (non-terminal)	43		10
K	58		9
W	44		7
R	55		6
W	-2		1
K	72		1

Table 6: Summary of validated modification sites over Lens proteins, compared with results reported in [26]. Modification types marked a were previously reported on these spectra, but the number of confirmed sites was not reported.

3.5 Results on ISB

Despite the fact that ISB dataset is arguably the most studied collection of MS/MS spectra to date, very few modifications were reported in this dataset. We found many modifications in the ISB dataset (Table 7) including a surprisingly large number of spectra with $PTM(-2, C)$. All the peptides with modification with shift -2 on Cysteine contain two Cysteine residues, providing strong evidence that these modifications correspond to cross-link formation. Moreover, very few of the original annotations from the ISB dataset (14 out of over 2500) contain cysteines, which is unexpected. We propose the following explanation of the above phenomenon: disulfide linkages in the sample were re-established and most Cysteine containing peptides ended up as part of some cross-linked peptide pair. This non-specific bond formation produced a variety of different molecules, each with tiny concentration and unusual fragmentation properties. However, the few

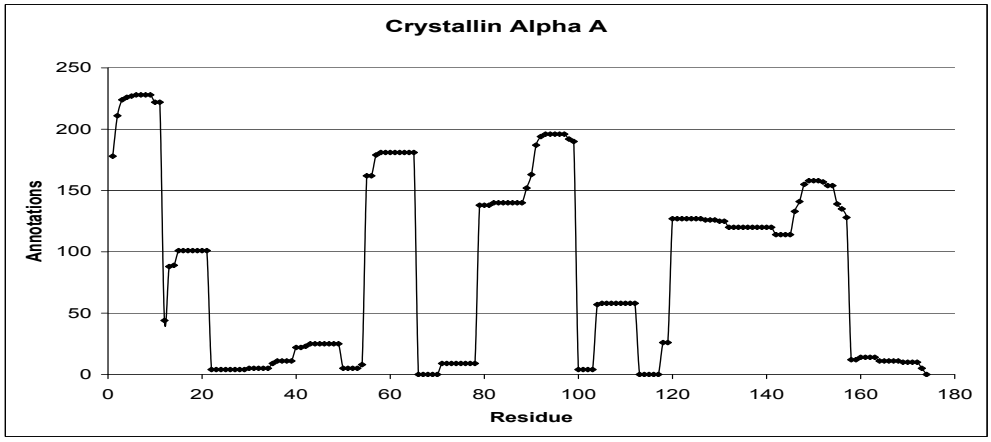

Figure 3: Number of annotations (c_s) over residues in crystallin α A.

AA	Mass	Spectra	Sites	Putative Annotation
non-specific	22	74	24	sodium
M	16	43	3	oxidation
C	-2	39	7	cross-linking
R	-28	39	1	mutation to lysine
Y	16	37	10	DOPA
N	1	35	5	deamidation
D	-18	17	5	dehydration
C	12	7	2	thioproline

Table 7: Ranked list of modifications for the ISB data-set. Each was supported by overlapping peptides. Iterative selection was stopped when no additional PTMs could explain more than fifteen additional spectra. In addition, the modification C+12 was manually verified and added to the list.

peptides which contained two Cysteines preferentially formed intra-molecular disulfide bridges. Such peptides are present in reasonable concentrations and undergo mostly-normal fragmentation, and so were discovered in blind search. Indeed, we found that cysteine-containing peptides had a strong b-ladder with a gap corresponding to the interval between two cysteines and a strong y-ladder with the same gap, an overwhelming evidence in favor of our explanation. A few additional annotations were obtained with modification mass +12 on an N-terminal cysteine, corresponding to a thioproline conversion which would also protect the residue from promiscuous cross-linking.

We found modifications of -28 on Arginine. This modification always occurs on the consecutive pair of Arginine residues in CAH2_BOVIN. These residues were never seen without the modification, suggesting this is a mutation of Arginine to Lysine rather than chemical damage.

We took the ranked list of identified PTM types and re-searched the ISB spectra using InsPecT against a large database containing the correct proteins as well as all human proteins from the nr database. This search was able to annotate 16% more spectra and 20% more peptides than currently known ISB annotations. We therefore argue that a combination of MS-Alignment and InsPecT significantly increases the number of annotated spectra in typical MS/MS samples.

(a)

Protein	Residue	Δ	c_s	μ_s
IKKb	K44	-57	46	54
IKKb	C370	14	152	97
IKKb	C412	14	121	67
IKKb	C716	14	29	91
GST	M1	16	21	100
GST	M69	16	15	100
GST	M129	16	50	59
Ig Gamma 1	M106	16	16	94
IKKb	M83	16	58	88
IKKb	M317	16	71	81
IKKb	M384	16	49	98
IKKb	M517	16	19	58
IKKb	M636	16	75	79
IKKb	M676	16	42	63
IKKb	M688	16	48	60
Proteinase K	S312	28	30	100
Trypsin	K97	28	132	58
Trypsin	K133	28	25	96

(b)

Protein	Residue	Δ	c_s	μ_s
BFSP2_HUMAN	Q299	-17	10	91
CRBB1_HUMAN Beta 1	Q235	-17	225	55
CRGC_HUMAN Gamma C	Q142	-17	20	59
CRYAA_HUMAN Alpha A	Q104	-17	52	77
CRYAB_HUMAN Alpha B	Q108	-17	37	97
CRBA1_HUMAN Beta A3	N120	1	24	73
CRBS_HUMAN Beta S	Q120	1	158	60
CRGC_HUMAN Gamma C	N24	1	14	50
CRGD_HUMAN Gamma D	N24	1	16	53
CRBA1_HUMAN Beta A3	C117	14	21	55
CRBA1_HUMAN Beta A3	C82	14	37	53
CRYAA_HUMAN Alpha A	M1	16	21	100
CRYAB_HUMAN Alpha B	M1	16	42	100
CRBB2_HUMAN Beta 2	A1	42	72	100
CRYAA_HUMAN Alpha A	M1	42	219	100
CRYAB_HUMAN Alpha B	M1	42	283	100
CRBB1_HUMAN Beta 1	K186	58	12	100

Table 8: Heavily-modified sites ($c_s \geq 10$ and $\mu_s \geq 50\%$) from (a) the IKKb data-set, and (b) the Lens data-set.

4 Modification Rates

MS-Alignment gives us a rare opportunity to quantify the *modification-rate* at each site. We define coverage c_s of a site s as the total number of modified or unmodified peptide spectra that encompass the site. Figure 3 plots c_s over crystallin α A protein. The graph approximates a step function with its discontinuities at tryptic cleavage sites. The *modification-rate*, μ_s at site s is given by

$$\mu_s = \frac{\text{\#spectra with modification on } s}{c_s}$$

The vast majority of modification events are rare ($\mu_s < 10\%$).

Table 8 lists sites over the IKKb data-set with $c_s \geq 10$ modified with rate $\mu_s > 50\%$. Most entries are sites susceptible to chemical damage. However, the high rate of modification of residue K44 in IKKb suggests the presence of a mutation to alanine, since post-translational modification would require the chemically infeasible truncation of the Lysine side chain to a methyl group. The high rates of modification of residues K97 and K133 of trypsin are likely due to chemical dimethylation performed by the supplier to prevent autodigestion.

We also examined sites with high modification rate in the Lens dataset. We find constitutive acetylation at the N-terminus of most crystallins. We also find that (-17, Q) has high μ_s for glutamine residues that are N-terminal in a tryptic fragment.

5 Conclusions

Recently, a number of methods have been described for identifying PTMs using tandem MS. With few exceptions, these methods generate putative candidates, which must then be manually validated. By incorporating a robust scoring function, and automated validation, we take the first step towards fully automated discovery of PTMs (and the extent and frequency of such modifications) from large data-sets of spectra. We expect that even larger datasets will provide further validation of our PTM frequency matrix approach and will lead to discoveries of new types of PTMs. Also,

as the tools mature, and more modified peptides are identified, we can begin to differentiate between different types of modifications by mining these data-sets. As an example, Tanner et al. [31] successfully promote the use of phosphate-loss ions as signatures of phosphorylation. Other modifications undoubtedly will lead to differences in fragment ion propensities which can be used to train PTM-specific score functions.

6 Acknowledgments

This project was supported by NIH grant NIGMS 1-R01-RR16522. We are grateful to Brian Searle and Larry David for making their Lens dataset available and to Larry David, Katalin Medzihradszky, and Philip Wilmarth for many useful discussions. Production of the lens data-set was supported by National Eye Institute grant EY007755. This research was supported in part by the UCSD FWGrid Project, NSF Research Infrastructure Grant Number EIA-0303622. Production of the IKKb data-set was supported by NIH grant R01GM65325 and by the Pew Scholars Program.

References

[1] Abrf delta mass reference, http://www.abrf.org.

[2] M. U. Ahmed, S. R. Thorpe, and J. W. Baynes. Identification of n epsilon-carboxymethyllysine as a degradation product of fructoselysine in glycated protein. *J. Biol. Chem.*, 261:4889–4894, 1986.

[3] D. C. Anderson, W. Li, D.G. Payan, and W.S. Noble. A New Algorithm for the Evaluation of Shotgun Peptide Sequencing in Proteomics: Support Vector Machine Classification of Peptide MS/MS Spectra and SEQUEST Scores. *Journal of Proteome Research*, 2(2):137–46, 2003.

[4] V. Bafna and N. Edwards. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17 Suppl 1:13–21, 2001.

[5] B.A. Ballif, J. Villen, S.A. Beausoleil, D. Schwartz, and Gygi S.P. Phosphoproteomic analysis of the developing mouse brain. *Mol Cell Proteomics*, 3(11):1093–101, 2004.

[6] N. Bandeira, H. Tang, V. Bafna, and P.A. Pevzner. Shotgun protein sequencing by tandem mass spectra assembly. *Analytical Chemistry*, (in press), 2004.

[7] G.T. Cantin and J.R. Yates. Strategies for shotgun identification of post-translational modifications by mass spectrometry. *Journal of Chromatography A*, 1053:7–14, 2004.

[8] J.W. Crabb, M. Miyagi, X. Gu, K. Shadrach, K.A. West, H. Sakaguchi, M. Kamei, A. Hasan, L. Yan, M.E. Rayborn, R.G. Salomon, and J.G. Hollyfield. Drusen proteome analysis: an approach to the etiology of age-related macular degeneration. *Proc Natl Acad Sci U S A.*, 99(23):14682–7, 2002.

[9] R. Craig and R.C. Beavis. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom.*, 17(20):2310–6, 2003.

[10] D.M. Creasy and J.S. Cottrell. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics*, 2(10):1426–1434, Oct 2002.

[11] J.K. Eng, A.L. McCormack, and J.R. Yates. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal Of The American Society For Mass Spectrometry*, 5(11):976–989, Nov 1994.

[12] A. Frank, S.W. Tanner, V. Bafna, and P.A. Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *J. of Proteome Research*, (in press), 2005.

[13] Y. Han, B. Ma, and K. Zhang. SPIDER: Software for Protein identification from Sequence Tags with *De Novo* Sequencing Error. In *IEEE Computational Systems Bioinformatics Conference (CSB)*, pages 206–215, 2004.

[14] B.T. Hansen, S.W. Davey, A.J. Ham, and Liebler. D.C. P-mod: an algorithm and software to map modifications to peptide sequences using tandem ms data. *J Proteome Res.*, 4(2):358–68, 2005.

[15] E. Hunyadi-Gulyas and K. Medzihradszky. Factors that contribute to the complexity of protein digests. *DDT: targets - mass spectrometry in proteomics supplement*, 3(2), 2004.

[16] A. Keller, A.I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*, 74(20):5383–5392, Oct 2002.

[17] A. Keller, S. Purvine, A. Nesvizhskii, S. Stolyar, D. R. Goodlett, and E. Kolker. Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis. *OMICS*, 6(2):207–212, 2002.

[18] B. Lu and T. Chen. A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics*, 19 Suppl 2:113–113, Oct 2003.

[19] M.J. MacCoss, C.C. Wu, and J.R. Yates. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal Chem*, 74(21):5593–5599, Nov 2002.

[20] A.I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75(17):4646–4658, Sep 2003.

[21] D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, Dec 1999.

[22] P.A. Pevzner, V. Dancík, and C.L. Tang. Mutation-tolerant protein identification by mass spectrometry. *J Comput Biol*, 7(6):777–787, 2000.

[23] P.A. Pevzner, Z. Mulyukov, V. Dancik, and C.L. Tang. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res*, 11(2):290–299, Feb 2001.

[24] J. Razumovskaya, V. Olman, D. Xu, E. Uberbacher, N.C. VerBerkmoes, R.L. Hettich, and Y. Xu. A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with sequest. *Proteomics*, 4:961–969, 2004.

[25] R.G. Sadygov and J.R. Yates. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem*, 75(15):3792–3798, Aug 2003.

[26] B. S. Searle, S. Dasari, P.A. Wilmarth, M. Turner, A. P. Reddy, L.L. David, and S.R. Nagalla. Identification of protein modifications using ms/ms de novo sequencing and the opensea alignment algorithm. *Journal of Proteome Research*, 4:546–554, 2005.

[27] B.C. Searle, S. Dasari, M. Turner, A.P. Reddy, D. Choi, P.A. Wilmarth, A.L. McCormack, L.L. David, and S.R. Nagalla. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal Chem*, 76(8):2220–2230, Apr 2004.

[28] H. Shu, S. Chen, Q. Bi, M. Mumby, and D.L. Brekken. Identification of phosphoproteins and their phosphorylation sites in the wehi-231 b lymphoma cell line. *Molecular and Cellular Proteomics*, 3:279–286, 2004.

[29] D.L. Tabb, L.L. Smith, L.A. Breci, V.H. Wysocki, D. Lin, and J.R. Yates. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem*, 75(5):1155–1163, Mar 2003.

[30] W.H. Tang, B.R. Halpern, I.V. Shilov, S.L. Seymour, S.P. Keating, A. Loboda, A.A. Patel, D.A. Schaeffer, and L.M. Nuwaysir. Discovering known and unanticipated protein modifications using ms/ms database searching. *Anal Chem*, 77(13):3931–46, 2005.

[31] S. Tanner, H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P.A. Pevzner, and V. Bafna. Inspect: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77(14):4626–4639, 2005.

[32] J.R. Yates, J.K. Eng, and A.L. McCormack. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*, 67(18):3202–3210, Sep 1995.

[33] J.R. Yates, J.K. Eng, A.L. McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, 67(8):1426–1436, Apr 1995.

A The MS-Alignment algorithm

In this section, we formulate the *Modified Peptide Identification Problem* and describe an algorithm that solves it.

A.1 Preliminaries

We begin with some definitions. Let $A = \{a_1, \ldots, a_{20}\}$ be the set of amino acids, each with molecular mass $\mathrm{mass}(a_i)$. A *peptide* $P = p_1 \cdots p_n$ is a sequence of amino acids, with mass $\mathrm{mass}(P) = \sum_{i=1}^{n} \mathrm{mass}(p_i)$. For an experimental spectrum S, $PM(S)$ is the mass of the spectrum, which is equal to the mass of the peptide that generated the spectrum.

Peptide fragmentation in a tandem mass spectrometer can be characterized by a set of numbers $\{s_1, \ldots, s_k\}$ representing the different types of ions that correspond to the removal of a certain chemical group from a peptide fragment (for example, $s = 17$ corresponds to loss of water). For tandem mass spectrometry, the *theoretical* spectrum $T(P)$ of a peptide P can be calculated by subtracting all possible ion types $s_1, \ldots, s_k$ from the masses of all prefixes and suffixes of P.

The Shared Peak Count between an experimentally measured spectrum S and a peptide P is the number of masses in S that are equal to masses in $T(P)$. In reality, peptide sequencing algorithms use more sophisticated scoring functions than a simple shared peaks count, incorporating different weighting functions for the matching masses and taking into account intensities of peaks. Let $\mathrm{Match}(S, P)$ be a function that scores the likelihood that a spectrum S is produced by a peptide P.

Computationally, a *modification* Δ of the peptide $P = p_1 \cdots p_n$ at position i results in a modified peptide $\hat{P}$ with the mass of residue p_i increased by Δ. We emphasize that this operation defines a theoretical spectrum of any modified peptide $\hat{P}$ and allows one to compute $\mathrm{Match}(S, \hat{P})$. We study the following problem:

Modified Protein Identification Problem
Input: A database of proteins, an experimental spectrum S, and a parameter k capping the number of modifications.
Output: A modified peptide $\hat{P}$ with the best match $\mathrm{Match}(S, \hat{P})$ to the spectrum S that is at most k modifications away from a peptide P that appears in the database (namely, P is a substring of some protein in the database).

Any modified peptide that is at most k modifications away from a peptide in the database is called a *candidate*. As described in Section 2, our approach consists of several steps. One of these steps is candidates genetation: We find the top scoring candidates according to some scoring function Match_1. These candidates are then rescored using a more sophisticated scoring function as described in Section 2.

In the next section, we describe a general algorithm for generating candidates. Then, in Sections A.3 and A.4 we show how to solve the problem more efficiently when there are at most 2 modifications.

To simplify the presentation, we make the following assumptions on the input. We assume that if two modifications appear on two consecutive amino acids of the peptide, then either the b-ion or the y-ion that corresponds to the cleavage site between these two amino acids appears in the spectrum S. Moreover, we assume that there are no two consecutive cleavage sites whose b and y

ions are missing from S. We also assume that the masses of all amino acids and all modifications are integers, and that there are no measurement errors in the spectrum S. We note that our actual algorithm does not need these assumptions.

A.2 Algorithm for arbitrary number of modifications

To start, low-intensity peaks are filtered from spectra as described in Bandeira et al. [6]. The score $\mathrm{Match}_1(S, \hat{P})$ of a candidate $\hat{P}$ consists of two parts: scores for the masses of the prefixes of $\hat{P}$, and scores for the modifications in $\hat{P}$. More precisely, let $\mathrm{MassScore}(v)$ be a scoring function for every mass v. Let $\mathrm{PTMScore}(\Delta, a) \leq 0$ be a penalty for having a modification Δ on the amino acid a. For every candidate $\hat{P}$, the score of $\hat{P}$ is the sum of $\mathrm{MassScore}(v)$ for every mass v of a prefix of $\hat{P}$ (including the entire peptide $\hat{P}$ and the empty prefix), plus the sum of scores of the modifications of $\hat{P}$ according to $\mathrm{PTMScore}$.

We now show how to compute the score of the highest scoring candidate. ¿From S we build a prefix residue mass (PRM) spectrum S', namely, for every mass $v \in S$, we add to S' the masses $v - 1$ and $PM(S) - (v - 19)$, Furthermore, we add to S' the masses 0 and $PM(S)$.

Denote the protein database q as a single sequence $p_1 \cdots p_n$, and let m be the size of the set S'. For every $j \leq n$ and $v \in S'$, let $D_k(j, v)$ be the maximum score of a peptide $\hat{P}$ with exactly k modifications whose unmodified peptide P is a substring of q that ends at p_j, and whose mass is v. Note that the size of the table D_k is $n \times m$.

The table D_0 can be easily computed in $O(mn)$ time. To compute a value $D_k(j, v)$ for $k \geq 1$ we need to consider five cases: (1) The optimal peptide $\hat{P}$ for $D_k(j, v)$ does not have a modification on its last amino acid, and the mass v' of the prefix of $\hat{P}$ of length $|\hat{P}| - 1$ is in S' (2) $\hat{P}$ does not have a modification on its last two amino acids and $v' \notin S'$, (3) $\hat{P}$ has a modification on its last amino acid and $v' \in S'$, (4) $\hat{P}$ has a modification on its last amino acid and $v' \notin S'$, and (5) $\hat{P}$ has a modification on its penultimate amino acid but not on its last, and $v' \notin S'$. Formally, the recurrence formula for $D_k(j, v)$ is given by the following lemma.

Lemma A.1. $D_k(j, v) = \max\{d_{j,v,k,1}, d_{j,v,k,2}, d_{j,v,k,3}, d_{j,k,v,4}, d_{j,k,v,5}\} + \mathrm{MassScore}(v)$, *where* $d_{j,v,k,1}$, $d_{j,v,k,2}$, $d_{j,v,k,3}$, $d_{j,k,v,4}$, *and* $d_{j,k,v,5}$ *are defined in Figure 4.*

After computing $D_k(j, v)$ for all j and v, we can find the score of the highest scoring candidate with at most k modifications by computing $\max_{k' \leq k, j} D_{k'}(j, PM(S))$. Each value $D_{k'}(j, PM(S))$ is the maximum score of a candidate with k' modifications that ends at p_j. This candidate can be found by traversing the dynamic programming tables starting at $D_{k'}(j, PM(S))$. By performing this process on the T k', j pairs with highest $D_{k'}(j, PM(S))$ values (for some parameter T), we obtain a set of candidates, which is passed to the second stage.

The time complexity of the algorithm above is $O(km^2n)$. This is expensive for typical values of the parameters, and can be improved for two special cases of practical importance. The first case is when $\mathrm{PTMScore}(\Delta, a)$ is a constant C, namely, it does not dependent on the modification or the residue a. In that case, we can compute cases 3–5 of the algorithm in constant time by maintaining additional information. Define $M_k(j, v) = \max_{w<v} D_k(j, w)$. It is easy to see that in case 3, $d_{j,v,k,3} = M_{k-1}(j - 1, w)$. Cases 4 and 5 can be modified in a similar fashion. M_k can be computed in constant time per entry, leading to an $O(kmn)$ time algorithm.

$$d_{j,v,k,1} = \begin{cases} D_k(j-1, v-\text{mass}(p_j)) & \text{if } v-\text{mass}(p_j) \in S' \\ -\infty & \text{otherwise} \end{cases}$$

$$d_{j,v,k,2} = \begin{cases} D_k(j-2, v-\text{mass}(p_{j-1}p_j)) + \text{MassScore}(v-\text{mass}(p_j)) & \text{if } v-\text{mass}(p_{j-1}p_j) \in S' \\ -\infty & \text{otherwise} \end{cases}$$

$$d_{j,v,k,3} = \max\left(\{D_{k-1}(j-1, w) + \text{PTMScore}(v-(w+\text{mass}(p_j)), p_j) \middle| \forall w \in S', w < v\} \cup \{-\infty\}\right)$$

$$d_{j,v,k,4} = \max\left(\left\{ \begin{array}{l} D_{k-1}(j-2, w) + \text{PTMScore}(v-(w+\text{mass}(p_{j-1}p_j)), p_j) \\ \quad + \text{MassScore}(w+\text{mass}(p_{j-1})) \end{array} \middle| \forall w \in S', w < v \right\} \cup \{-\infty\}\right)$$

$$d_{j,v,k,5} = \max\left(\left\{ \begin{array}{l} D_{k-1}(j-2, w) + \text{PTMScore}(v-(w+\text{mass}(p_{j-1}p_j)), p_{j-1}) \\ \quad + \text{MassScore}(v-\text{mass}(p_j)) \end{array} \middle| \forall w \in S', w < v \right\} \cup \{-\infty\}\right)$$

Figure 4: Definitions of $d_{j,v,k,1}$, $d_{j,v,k,2}$, $d_{j,v,k,3}$, $d_{j,k,v,4}$, and $d_{j,k,v,5}$.

In the next section, we describe an efficient algorithm for an arbitrary function PTMScore, but we limit the number of modifications to 2. As the results with 3 or more modifications are not very reliable, this restriction is reasonable. The handling of 1 modifications is simpler, and is described in Section A.4.

A.3 Algorithm for two modifications

We denote by M_1 (resp., M_2) the minimum (resp., maximum) mass offset of one modification. Instead of using the tables D_0, D_1, and D_2, our algorithm will use D_1 and the following two tables: For every $i \leq j$ such that $\text{mass}(p_i \cdots p_j) \leq PM(S)+M_2$, $\text{PrefixScore}(i,j)$ is the score of $p_i \cdots p_j$, namely $\text{PrefixScore}(i,j) = \text{MassScore}(0) + \sum_{k=i}^{j} \text{MassScore}(\text{mass}(p_i \cdots p_k))$. Similarly, for every $j \leq i$ such that $\text{mass}(p_j \cdots p_i) \leq PM(S) - M_1$, $\text{SuffixScore}(j,i) = \text{MassScore}(PM(S)) + \sum_{k=j}^{i} \text{MassScore}(PM(S) - \text{mass}(p_k \cdots p_i))$. Computing the table PrefixScore is done by going over all i, and accumulating the scores $\text{MassScore}(\text{mass}(p_i \cdots p_j))$ for all j. Computing SuffixScore is done similarly.

Define $\Delta_{i,j,v} = v - \text{mass}(p_i \cdots p_j)$ for every $i \leq j \leq n$ and $v \in S'$. To compute the table D_1, we use the following lemma (note that there are only four cases in the lemma, while Lemma A.1 has five cases).

Lemma A.2. $D_1(j,v) = \max\{d_{j,v,1,1}, d_{j,v,1,2}, d_{j,v,3}, d_{j,v,4}\} + \text{MassScore}(v)$, *where* $d_{j,v,1,1}$ *and* $d_{j,v,1,2}$ *are defined in Figure 4, and* $d_{j,v,3}$ *and* $d_{j,v,4}$ *are defined in Figure 5.*

After computing $D_1(j,v)$ for all j and v, we can find the maximum score of a candidate as follows: Define $\hat{\Delta}_{j,i,v} = PM(S) - \text{mass}(p_{j+1} \cdots p_i) - v$ for all j, i, and v. The maximum score of a candidate is $\max\left(\cup_{j,v}\{b_{j,v,1}, b_{j,v,2}\}\right)$, where $b_{j,v,1}$ and $b_{j,v,2}$ are defined in Figure 5. Like in the previous algorithm, we generate the highest scoring candidates by traversing the table D_1.

Time complexity Denote $M = \max\{|M_1|, |M_2|\}$. Computing the the tables PrefixScore and SuffixScore takes $O(m'n)$ time, where $m' = \lfloor (PM(S)+M)/57 \rfloor + 1$ is an upper bound on the number of cells $\text{PrefixScore}(i,j)$ (or $\text{SuffixScore}(j,i)$) that we need to compute for a fixed i. This upper bound follows from the fact that the minimum mass of an amino acid is 57 Da. The upper

$$d_{j,v,3} = \max\left(\{\text{PrefixScore}(i, j-1) + \text{PTMScore}(\Delta_{i,j,v}, p_j) | \forall i \text{ s.t. } M_1 \le \Delta_{i,j,v} \le M_2\} \cup \{-\infty\}\right)$$

$$d_{j,v,4} = \max\left(\left\{ \begin{array}{l} \text{PrefixScore}(i, j-2) + \text{PTMScore}(\Delta_{i,j,v}, p_{j-1}) \\ + \text{MassScore}(v - \text{mass}(p_j)) \end{array} \middle| \forall i \text{ s.t. } M_1 \le \Delta_{i,j,v} \le M_2 \right\} \cup \{-\infty\}\right)$$

$$b_{j,v,1} = \max\left(\left\{ \begin{array}{l} D_1(j, v) + \text{PTMScore}(\hat{\Delta}_{j,i,v}, p_{j+1}) \\ + \text{SuffixScore}(j+2, i) \end{array} \middle| \forall i \text{ s.t. } M_1 \le \hat{\Delta}_{j,i,v} \le M_2 \right\} \cup \{-\infty\}\right)$$

$$b_{j,v,2} = \max\left(\left\{ \begin{array}{l} D_1(j, v) + \text{PTMScore}(\hat{\Delta}_{j,i,v}, p_{j+2}) \\ + \text{SuffixScore}(j+3, i)\} + \text{MassScore}(v + \text{mass}(p_{j+1})) \end{array} \middle| \forall i \text{ s.t. } M_1 \le \hat{\Delta}_{j,i,v} \le M_2 \right\} \cup \{-\infty\}\right)$$

Figure 5: Definitions of $d_{j,v,3}$, $d_{j,v,4}$, $b_{j,v,1}$, and $b_{j,v,2}$.

bound is pessimistic since only one amino acid has a mass of 57 Da. In practice, the average number of cells that we need to compute for each i is $\bar{m} \simeq (PM(S) + M)/100$. We note that $\bar{m}$ is usually much smaller than m.

The time complexity of the algorithm is $O(dmn + m'n)$ where $d = \lfloor (M_2 - M_1)/57 \rfloor + 1$ is an upper bound on the number of values of i we need to consider in order to compute a single value of $d_{j,v,3}$, $d_{j,v,4}$, $b_{j,v,1}$, or $b_{j,v,2}$ (this follows from the fact that for all i, $\Delta_{i+1,j,v} - \Delta_{i,j,v}$ is the mass of some amino acid). The average time complexity of the algorithm is $O(\bar{d}mn + \bar{m}n)$, where $\bar{d} \simeq (M_2 - M_1)/100$. Typical values are $M_1 = -100$ and $M_2 = 160$, implying $\bar{d} \simeq 3$.

A.4 Algorithm for one modification

The maximum score of a candidate is $\max_{i,j} b_{i,j}$, where

$$b_{i,j} = \max\left(\left\{ \begin{array}{l} \text{PrefixScore}(i, j) + \text{PTMScore}(\hat{\Delta}_{i,i',0}, p_{j+1}) \\ + \text{SuffixScore}(j+2, i') \end{array} \middle| \forall i' \text{ s.t. } M_1 \le \hat{\Delta}_{i,i',0} \le M_2 \right\} \cup \{-\infty\}\right).$$

Generating the corresponding candidate is trivial. The average time complexity of the algorithm is $O(\bar{d}\bar{m}n)$. We note that the algorithms presented here are more efficient than the straightforward exhaustive search algorithm. The time compelxity of the exhaustive search algorithm is roughly $O(\bar{d}\bar{m}^2 n)$ for one modification, and $O(\bar{d}(M_2 - M_1)\bar{m}^3 n)$ for two modifications.

A.5 Implementation Details

In the implementation of the algorithms given in previous sections, we use the following scoring functions: For the mass scoring function MassScore, we use the scoring function from Tanner et al. [31]. The function PTMScore is defined as follows:

$$\text{PTMScore}(\Delta, a) = \begin{cases} C & \text{if } M_1 \le \Delta \le M_2 \\ & \text{and } \text{mass}(a) + \Delta \ge 50\ , \\ -\infty & \text{otherwise} \end{cases}$$

where $C < 0$ is some constant. This function forbids implausible interpretations, and gives better results than a constant penalty function.

B P-value computation

Computation of p-values is performed using a support vector machine (SVM) trained over correct and incorrect matches from the ISB data-set. The SVM uses a radial basis function (RBF) kernel, and a set of six features:

- Raw match score, computed by alignment of theoretical and spectral peaks as described in the Inspect paper (Tanner 2005).
- Percentage of intensity explained.
- Percentage of top 50 peaks explained.
- Percentage of theoretical b/y peaks (within dynamic range) found in the spectrum
- Difference in match score (δ-cn) between this match and the runner-up. If multiple matches to the same locus were found, compute the difference in match score to the first runner-up from a distinct locus. (For example: The match EAM#MAPK and EAMM#APK likely have similar scores, but their δ-cn can still be large)
- Number of candidates scored. (This feature compensates for the fact that δ-cn is affected by database size)

All features are calculated after performing window-based filtering on the spectral peaks. The output of the SVM is a real number, where values above zero are likely 'correct' and values below zero likely 'incorrect'. We constructed the histogram of SVM outputs over our entire training set, and used this to construct a lookup table for deriving p-values.

C Supplementary Tables

Annotated spectra for post-translational modifications, as well as raw search results, are available from http://bioinfo2.ucsd.edu/.

Offset	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Total
-18		1	3*	3	1*		2	2	1	1		3	1	2	2	6	6	1			35
1	8		12	11*	11	16	9	15*	4*	30*	2	33*	16	20*	3	15	13*	8*	4	10	240
14	2			3				3*	2		2		1			1*		1*			15
16	1*		1*	6	1*	3	2	4*		4*	39*	3*	1*	4		1*		1*	9	1	81
17			1*	1		2		1*		3*	12*	1*		4		1		1	4	1	32
22	3		8*	10	6	11	2	5		4	1	6	1	6		9	5	4		2	83
28	1*	2*						3	3*	2	1					1*		2*			15
32		1						1			5*	1	1*		1				6		16
53			7	10		3		6	1	6	1	1	5	2		1	2	3		1	49
71		9								1	1		1								12
156	1							2		1			1			1	1				7

Table 9: PTM site count matrix for IKKb dataset (1,072 sites total). The first column indicates the mass shift of a modification. The entry for modification Δ, and amino acid a refers to the number of distinct residues of type a for which a top-scoring spectral annotation assigned modification mass Δ at that residue. Only rows from the frequency matrix are repeated. Entries that can be explained by mutations are starred.

Search	Total of grid entries	Entropy	Entropy / ln(n)
WrongDB	320 (2312)	4.98 (6.98)	0.86 (0.90)
MutDB	1485 (2110)	4.44 (5.12)	0.61 (0.67)
IKKB	4248 (15410)	4.66 (7.11)	0.56 (0.74)
Lens	3986 (20097)	5.07 (7.51)	0.61 (0.76)

Table 10: Entropy of the PTM frequency matrix for several searches. Results are given for hits with p-value < 0.05, and (in parentheses) for all top hits. Searches "MutDB" and "WrongDB" used the same collection of previously annotated spectra. The WrongDB search was run against the Lens database, generating no valid annotations. Incorrect annotations produce random noise in the PTM frequency matrix

DeltaMass	Protein	Residue	Spectra	Peptide species	Rate
14	CRBA1_HUMAN Beta A3	C82	37	10	52.86
14	CRBA1_HUMAN Beta A3	C117	21	6	55.26
14	CRBA1_HUMAN Beta A3	C185	41	9	43.62
14	CRBS_HUMAN Beta S	C24	79	5	33.76
14	CRBS_HUMAN Beta S	C26	117	7	49.16
14	CRGD_HUMAN Gamma D	C110	5	2	55.56
-18	CRYAA_HUMAN Alpha A	S42	2	2	4.65
-18	CRYAA_HUMAN Alpha A	T55	2	2	1.12
-18	CRYAA_HUMAN Alpha A	S59	4	1	1.9
-18	CRYAA_HUMAN Alpha A	S81	2	1	1.18
-18	CRYAA_HUMAN Alpha A	D84	3	2	1.55
-18	CRYAA_HUMAN Alpha A	D91	4	2	1.89
-18	CRYAA_HUMAN Alpha A	D92	2	2	0.91
-18	CRYAA_HUMAN Alpha A	D105	7	3	8.64
-18	CRYAA_HUMAN Alpha A	D106	11	4	13.58
-18	CRYAA_HUMAN Alpha A	T148	11	5	5.98
-18	CRYAA_HUMAN Alpha A	D151	3	1	1.55
-18	CRYAB_HUMAN Alpha B	S59	3	3	1.84
-18	CRYAB_HUMAN Alpha B	D109	4	1	9.3
-18	CRBB1_HUMAN Beta 1	S80	3	2	3.57
-18	CRBB1_HUMAN Beta 1	T125	8	2	7.84
-18	CRBB1_HUMAN Beta 1	S137	2	2	2
-18	CRBB1_HUMAN Beta 1	S151	7	2	5.26
-18	CRBB2_HUMAN Beta 2	S174	2	2	1.98
-18	CRBA1_HUMAN Beta A3	D37	2	1	1.13
-18	CRBA1_HUMAN Beta A3	D97	2	1	0.52
-18	CRBA1_HUMAN Beta A3	S100	5	2	1.21
-18	CRBA1_HUMAN Beta A3	S102	4	2	0.95
-18	CRBA1_HUMAN Beta A3	S200	35	9	25.93
-18	CRBA1_HUMAN Beta A3	T204	6	4	4.29
-18	CRBA1_HUMAN Beta A3	S205	5	3	3.62
-18	CRBS_HUMAN Beta S	S166	13	1	3.86
-18	CRBA4_HUMAN Beta A4	S180	16	2	21.33
-18	CRBA4_HUMAN Beta A4	T184	2	1	2.63
-18	BFSP2_HUMAN Beaded filament structural protein 2	S91	2	1	14.29
-17	CRYAA_HUMAN Alpha A	Q50	4	1	36.36
-17	CRYAA_HUMAN Alpha A	Q90	6	2	3.53
-17	CRYAA_HUMAN Alpha A	Q104	52	4	76.47
-17	CRYAA_HUMAN Alpha A	Q147	11	4	7.05
-17	CRYAB_HUMAN Alpha B	Q108	37	2	97.37
-17	CRYAB_HUMAN Alpha B	R149	20	3	8.44
-17	CRBB1_HUMAN Beta 1	Q69	7	1	1.46
-17	CRBB1_HUMAN Beta 1	Q235	225	17	54.88
-17	CRBB2_HUMAN Beta 2	Q12	4	2	4.65
-17	CRBA1_HUMAN Beta A3	Q203	4	3	3.13
-17	CRBS_HUMAN Beta S	Q106	10	3	4.74
-17	CRBS_HUMAN Beta S	R157	4	2	50
-17	BFSP2_HUMAN Beaded filament structural protein 2	Q299	10	3	90.91
-17	CRGD_HUMAN Gamma D	Q142	16	2	48.48
-17	CRGC_HUMAN Gamma C	Q142	20	3	58.82
-17	CRGC_HUMAN Gamma C	R146	5	2	11.9
-2	CRBB1_HUMAN Beta 1	W236	8	2	1.85
1	CRYAA_HUMAN Alpha A	Q6	41	8	14.8
1	CRYAA_HUMAN Alpha A	Q50	3	3	30
1	CRYAA_HUMAN Alpha A	Q90	19	5	10.38
1	CRYAA_HUMAN Alpha A	N123	9	7	4.21
1	CRYAA_HUMAN Alpha A	Q126	27	16	12.68
1	CRYAA_HUMAN Alpha A	Q147	14	7	8.81
1	CRYAB_HUMAN Alpha B	Q26	6	4	9.84
1	CRYAB_HUMAN Alpha B	N78	3	2	7.69
1	CRYAB_HUMAN Alpha B	N146	59	12	24.89
1	CRBB1_HUMAN Beta 1	N57	8	4	13.79
1	CRBB1_HUMAN Beta 1	N67	49	4	9.8
1	CRBB1_HUMAN Beta 1	Q69	20	4	4.05
1	CRBB1_HUMAN Beta 1	N81	8	5	9.41
1	CRBB1_HUMAN Beta 1	Q105	15	3	5.47
1	CRBB1_HUMAN Beta 1	N107	9	2	3.28
1	CRBB1_HUMAN Beta 1	N124	16	2	15.69
1	CRBB1_HUMAN Beta 1	N157	13	6	9.63
1	CRBB1_HUMAN Beta 1	N161	17	6	11.49
1	CRBB1_HUMAN Beta 1	Q166	19	5	12.75
1	CRBB1_HUMAN Beta 1	Q204	11	2	7.75
1	CRBB1_HUMAN Beta 1	N216	25	9	20.83
1	CRBB1_HUMAN Beta 1	Q222	12	6	10.17
1	CRBB1_HUMAN Beta 1	Q224	21	5	17.5
1	CRBB1_HUMAN Beta 1	Q226	5	3	4.2
1	CRBB1_HUMAN Beta 1	Q235	30	9	13.95
1	CRBB2_HUMAN Beta 2	Q7	10	2	11.36
1	CRBB2_HUMAN Beta 2	Q12	5	3	5.75
1	CRBB2_HUMAN Beta 2	Q54	5	4	6.49
1	CRBB2_HUMAN Beta 2	Q63	13	6	15.85
1	CRBB2_HUMAN Beta 2	N65	9	4	10.98
1	CRBB2_HUMAN Beta 2	Q70	3	2	6.67
1	CRBB2_HUMAN Beta 2	N113	13	8	7.6
1	CRBB2_HUMAN Beta 2	N115	7	4	4.19
1	CRBB2_HUMAN Beta 2	Q137	12	2	7.14
1	CRBB2_HUMAN Beta 2	Q162	7	2	15.56
1	CRBB2_HUMAN Beta 2	Q182	7	4	4.9
1	CRBA1_HUMAN Beta A3	Q38	18	5	10.17
1	CRBA1_HUMAN Beta A3	N40	22	5	12.43
1	CRBA1_HUMAN Beta A3	Q42	22	8	12.5
1	CRBA1_HUMAN Beta A3	N54	11	6	20
1	CRBA1_HUMAN Beta A3	N103	101	12	23.93
1	CRBA1_HUMAN Beta A3	N120	24	7	72.73
1	CRBA1_HUMAN Beta A3	K131	16	4	5.48
1	CRBA1_HUMAN Beta A3	N133	50	8	16.56
1	CRBA1_HUMAN Beta A3	Q149	11	6	19.64
1	CRBA1_HUMAN Beta A3	N155	5	3	7.46
1	CRBA1_HUMAN Beta A3	N156	10	5	14.93
1	CRBA1_HUMAN Beta A3	Q172	22	2	10.63
1	CRBA1_HUMAN Beta A3	Q180	4	4	4.65
1	CRBA1_HUMAN Beta A3	Q203	12	6	8.82

AA	Mass	Spectra	Putative annotation
K	16	135	hydroxylation
P	16	207	hydroxylation
Y	80	124	sulfation
K	32	50	double hydroxylation (one misplaced)
Y	96	35	sulfation + (misplaced) hydroxylation
P	32	24	double hydroxylation (one misplaced)
M	16	19	oxidation
P	17	7	hydroxylation (incorrect mass)
F	80	6	misplaced sulfation
Q	16	6	misplaced hydroxylation
A	103	6	LA+103M... instead of SLAM+16...
F	-88	5	YGVSGGGF-88SSASNR versus Y+80Y+80GYTGAFR
P	147	5	...P+147IP instead of ...P+16IPN
Y	160	5	double sulfation (one misplaced)

Table 12: Full list of modifications chosen over the ModSpec data-set. Iterative selection identifies the valid modifications (hydroxylation on P or K, oxixdation on M, sulfation on Y).

Mutant residue	Putative reversion	Residue	Mutation mass δ	Species	Spectra	Actual mutation(s)
I	V	140	14	3	83	
E	S	49	42	8	61	
F	M	151	16	2	55	D152E(+14)
K	S	51	41	8	54	
M	L	205	18	3	45	
P	Q	156	-31	1	33	K156P (-31)
F	L	419	34	2	30	I419F (+34)
F	M	64	16	1	27	N60Q(+14)
N	T	479	13	1	22	
S	A	625	16	2	21	V623I(+14)
M	V	58	32	3	21	
Y	F	18	16	1	20	
H	L	306	24	2	20	
S	A	173	16	5	20	
A	S	351	-16	1	19	
I	V	403	14	1	18	
M	L	220	18	1	15	L221M(18)
M	Q	728	3	2	15	
T	A	53	30	1	15	
S	V	58	-12	1	15	

Table 13: The PTM selection procedure was run on the MutDB search results, and putative revertant point mutations were assigned to modification sites. Results are shown for sites with good coverage (15 or more spectra) Of 20 mutations, 14 are recovered exactly, 2 recovered modulo I/L or K/Q substitution, and 4 are assigned δ-correct annotations (mass off by up to two daltons, position off by up to four residues).