False Localization Rates for Site Assignment of Post-Translational Modifications

Abstract

Accurate identification of post-translational modification (PTM) sites by tandem mass spectrometry (MS/MS) remains a major challenge in proteomics. While peptide identification has become largely automated, PTM site assignment often requires substantial manual curation and experimental follow up. Statistical validation of peptide identifications is commonly done using False Discovery Rates (FDR). The lack of an equivalent method for assessing False *Localization* Rates (FLR) has made it difficult to easily validate reported PTM localizations from mass spectrometry experiments. Current approaches localize PTMs using fixed score thresholds and do not enforce FLR on high-throughput searches. We propose one of the first generic approaches for calculating FLR using assignments of PTMs to incorrect residues as decoys while hits to valid residues as target. We also introduce a new scoring function which is able to explicitly model ambiguous or partially-localized PTMs and for co-elution of differently-modified variants of the same peptides with the same PTMs on different sites. We demonstrate our approach on MS/MS spectra on synthetic human phosphopeptides, Human lens, whole lysate with a diverse set of PTMs and high throughput *Saccharomyces cerevisiae* phosphopeptides where we are able to localize approximately 19% more MS3 and 33% MS2 spectra than Ascore.

Keywords: tandem mass spectrometry, post-translational modifications, site assignment, false discovery rate, quantification.

1 Introduction

Mass spectrometry is the most commonly used approach in high throughput proteomics and the dominant approach for large scale characterization of post translational modifications (PTMs) [1]. While large scale identification of modified peptides is common [2, 3, 4] and appropriately controlled by false discovery rates (FDR), the lack of an equivalent approach for PTM site assignments commonly results in the need for extensive manual validation. Probabilistic scoring methods have reduced the need for extensive manual validation, but still do not offer a global, data dependent measure of the quality of localization results. Similarly to when delta score thresholds were first introduced for database searches for peptide identification, to date the recommended fixed thresholds for site assignment are based upon empirical false localization rates estimated from small sets of synthetic peptides rather than re-estimated for each experiment. Moreover, since fixed thresholds may correspond to varying FLRs on different data, these also do not offer a way to compare competing fixed threshold scoring methods of the same datasets.

For peptide identification, false discovery rates (FDR) are widely used to allow estimation of the false identification rate for entire datasets from any experiment setup [5]. PTM localization does not currently have an equivalent data-dependent method [6]. A similar FDR approach is necessary for the experiment-wide significance assessment of PTM site assignments. While there have been attempts to approximate FLR using invalid modification amino acid sites for particular modifications as "decoy" results, such methods require that the "decoy" sites appear in similar frequency and in close proximity to the actual modification sites [7]. In datasets with a large number of possible modifications (such the human lens dataset used here), this is both impractical to automate and there may not be appropriate "decoy" candidates depending on the modification. Delta score approaches such as Mascot Delta Score [8] and SLIP scores [7] take the difference between the top two site assignments and rely on fixed thresholds based on synthetic datasets or the approach described above. Probabilistic score methods such as Ascore [9], PTM Score (from MaxQuant) [10], PhosphoRS [11], Phosphorylation Localization Score (PLScore) in Inspect [12], SloMo [13] and Phosphorinator [14] propose to estimate probabilistic scores proportional to the number of site determining peaks

between PTM sites. Since these use fixed thresholds which are independent from the dataset, they do not correct for multiple hypothesis testing and do not necessarily result in comparable false localization rates across datasets and experimental conditions.

Here we propose a new framework for estimation of false localization rates (FLRs) - data-dependent false discovery rates for PTM site assignments. First, in contrast to the current methods which assign delta scores between candidate PTM sites, our approach assigns scores to PTM *variants* of the same peptide where a variant is defined as a distinct combination of site assignments for the PTMs a particular peptide sequence. Second, we extend the concept of the Target/Decoy Approach [15] where invalid peptide sequences are used to estimate false identification rates against valid peptide sequences and use invalid (decoy) PTM site assignments, such as phosphorylation on Glycine, to estimate false localization rates on valid (target) sites. Third, we allow each MS/MS spectrum to be composed of more than one variant, thus allowing for identification of co-eluting variants and ambiguous and partially-ambiguous PTM site assignments.

To demonstrate our FLR framework, we further propose the first variant scoring scheme supporting ambiguity and co-elution and show how spectra of unmodified peptides can be used to assess the reliability of site assignments in spectra of the corresponding modified peptides. A variant scoring scheme should be able to distinguish between correct and incorrect PTM site assignments by addressing the challenges inherent to mass spectrometry of peptides: ambiguity due to incomplete MS/MS fragmentation and co-eluting peptides with the same modification on the same peptide sequence but with different modification sites, thus possibly resulting in multiple correct site assignments per spectrum (e.g., as in histone proteins [16]). Our proposed variant scoring scheme uses linear programming to decompose each MS/MS spectrum into all of its possible variants and assign a separate score to each distinct variant. Ambiguity is then addressed by grouping indistinguishable variants into *variant groups* and co-elution is automatically captured by co-occurrence of high-scoring distinct variants in the same MS/MS spectrum.

The proposed scoring scheme and FLR estimation method was validated by using a set of 180 synthetic phosphopeptides [8]. The R^2 value between the estimated and actual FLR, even for a relatively small dataset was between .633 and .809 indicating that our method of estimating FLR, especially in cases where there are larger numbers of results, will likely yield a reasonable estimation of FLR. The method also shows improvement over existing delta scoring methods in terms of the number of results localized at the same thresholds. In a high throughput dataset of *Saccharomyces cerevisiae* phosphopeptides we were able to demonstrate a 19% increase in localization over ascore in MS3 spectra and a 33% increase in localization in MS2 over Ascore at a 5% FLR threshold and an Ascore threshold of 13. In a human lens whole lysate sample with a diverse set of PTMs at an FLR of 5% we were able to unambiguously localize approximately 50% of non trivial results and partially localize 60% at an FLR of 5% even with 17 types of modifications allowed with up to two modifications per peptide.

2 Methods

After identifying spectra using an existing search tool, localization for spectra with PTMs consists of two steps. First, each modification variant (observed modifications and their position on a peptide) is given a score. Next, all modification variants in the entire dataset are sorted based on score and evaluated to estimate the false localization rate (FLR) at different score cutoffs. In the FLR step, all modifications which are assigned to an incorrect amino acid residue are considered invalid. The FLR approach relies on the fact that each possible modification site (even incorrect ones) are evaluated in isolation, so current delta score methods are unsuitable for this method. Our scoring method used for FLR generates scores for each possible modification variant based on how much the theoretical spectra for each variant contributes to the actual modified spectrum. This method of estimation works under the assumption that any particular expected theoretical peak in a particular variant will be scaled by its quantity in the modified spectrum. For example if a variant contributes 50% to the final spectrum we can expect that each peak intensity in the observed modified spectrum will be 50% of the predicted intensity for that ion. In this case we use the estimated quantity of each variant as the input score for FLR. Figure 1 provides an overview of our approach.

Scoring starts with a modified spectra and its associated peptide identification. All possible combinations of the observed modifications on the peptide and the amino acid sites are generated to create the possible

modification variants for the peptide. The unmodified version of the peptide identification is then used to find an unmodified spectrum of the unmodified peptide. A theoretical spectra for each variant is then generated from the spectrum of the unmodified peptide by taking the intensity of all identified ions (bs and ys for CID) and shifting the masses by the expected modification mass (note that this shift may be zero depending on the position of the modifications on the modification variant). Given the observed modified spectrum and the theoretical spectra for each variant, the relative abundances of variants are estimated using a linear programming (LP) formulation which estimates the quantity of each variant present in the final modified spectrum by minimizing the error between the observed modified spectrum and the expected intensities based on the quantities of each variant. Finally, we group the indistinguishable modification variants based on the modified spectrum. When there are no or very few peaks distinguishing two modification variants in the mixture spectrum, it is impossible for LP to differentiate between the two variants as well. In such cases, the individual variant abundances are not accurate but rather the summed abundance for the group. We output the relative abundances of variant groups instead of individual variant groups.

Following scoring, a target-decoy strategy is adopted to determine the global FLR for site assignments. The knowledge of site-specificity of the modification is introduced at this point. The variant groups which assign each PTM to at least one valid site are valid and are added to the target database. In cases with multiply modified peptides, all PTMs must have at least one valid site assignment in order to be considered valid. Similarly, variant groups which assign any PTM to invalid sites only are considered invalid and added to the decoy database. Each variant group is then scored by the total estimated abundance. FLR is then computed at each score threshold by estimating how often that threshold results in invalid/decoy site assignments. The grouping threshold can affect the ambiguity of results and the number of results output at different FLR cutoffs, so an alternate method to choosing a fixed cutoff is to choose successive grouping thresholds and calculating FLR at each threshold (See Section 2.5).

In the following subsections, we will discuss building the LP and grouping variants in more detail.

2.1 Enumeration of Modification Variants

Once a search has been performed and a particular spectra has been identified as containing PTMs, all modification variants must be generated to be scored by LP. Each variant of a peptide consists of the same sequence and modifications, the only difference is in the location of the modification types. At most one modification mass is allowed on each amino acid; multiply-modified amino acids are modeled using modifications of the resulting aggregate mass. For example, di-methylation is represented using nominal mass 28 Da instead of two methylations, each of mass 14 Da. Note that at this stage modification site specificity is ignored and all possible amino acid sites are considered candidates for modification. Given a peptide identification of length n and consisting of the multiset of modification masses $\mathcal{M} = m_1, \cdots, m_k$ with d duplicate masses, there are $\binom{n}{k} \cdot \frac{k!}{d!}$ possible modification variants.

The modification variants are exhaustively enumerated by generating all distinct permutations of length n consisting of $k+1$ elements (k modifications with a placeholder element for unmodified positions).

2.2 Inference of Fragment Ion Intensities

Consider a fragment ion of a peptide. We make the assumption that the intensity of an ion remains constant, even when the peptide is modified. This assumption is borne out by empirical observation of conservation of spectral shapes of a peptide sequence across different unmodified and modified tandem mass spectra. Previously, Sniatynski et al. [17] showed that delay-series correlation between the mass spectra of modified and unmodified peptides revealed significant spectral overlap at an offset indicative of the modification. Others, including us, make a similar observation [18].

We observe similar behavior in our test MS/MS datasets. We measure the similarity between two spectra using dot product (cosine score) which is a widely accepted measure for spectral similarity. [19, 20] For the purposes of this method, extracted ions are used so that like ions are compared with each other rather than comparing by mass windows. For example intensities of a modified b2 ion is compared to its unmodified b2 counterpart in a different spectrum. For simplicity, we define theoretical peptide masses $Masses(P)$ as the set of theoretical prefix and suffix masses from P based on its amino acid sequence. Since in this case,

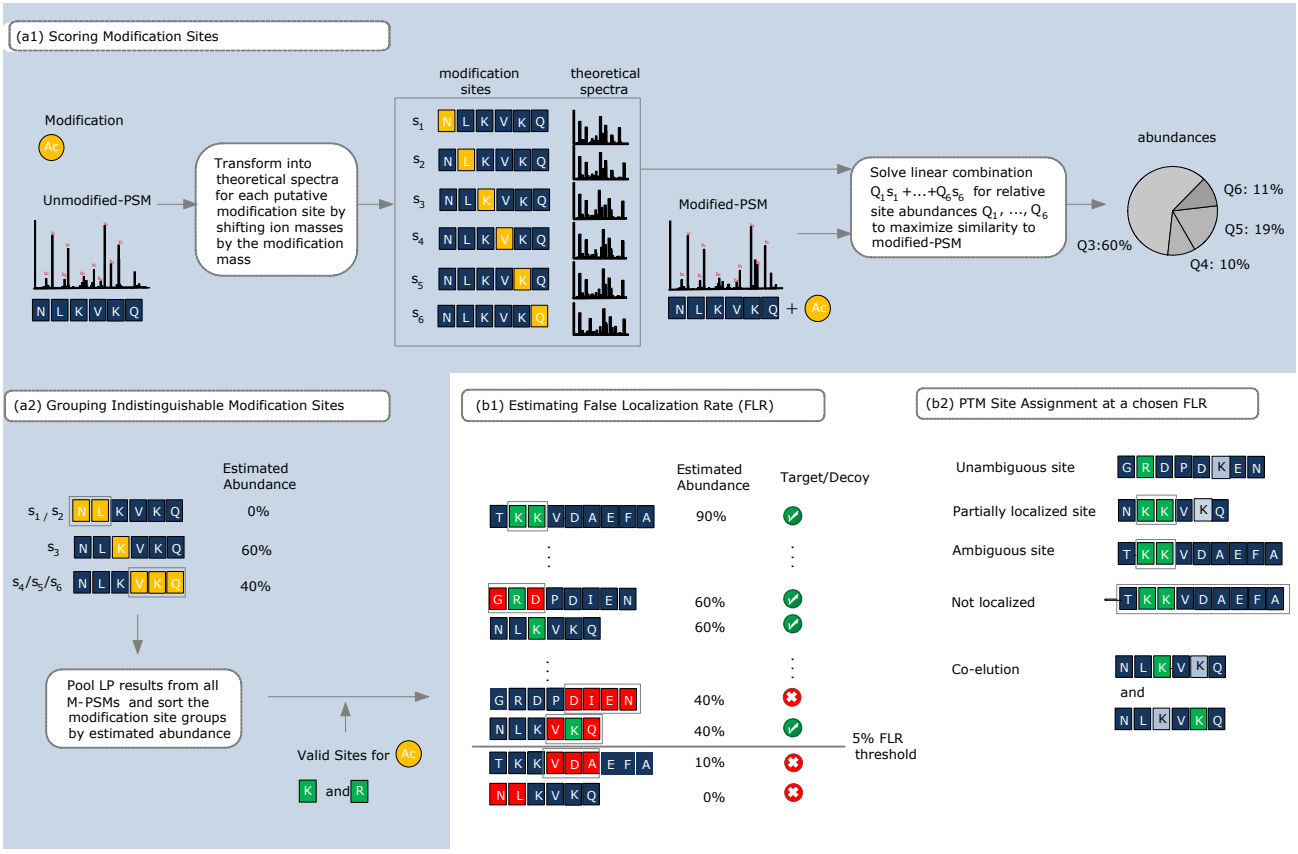


Figure 1: Overview of FLR approach: a) All modification sites are scored individually. b) Invalid site assignments are used as decoys for TDA-like calculation of FLR.
(a1) For each modified spectra, the first step in PTM localization is scoring the modification variants. The modified spectrum is modeled as a mixture of the theoretical spectra of the modification variants. By evaluating the contribution of each theoretical spectrum to the modified spectrum, the quantity of each variant can be estimated through linear programming (LP). (a2) Ambiguities in site assignment are handled by grouping indistinguishable sites into variant groups. Due to missing peaks in the modified peptide spectrum, some variants are indistinguishable in the LP. In such cases, we report the total abundance for the site groups instead of the abundance of the individual sites. (b1) A target-decoy strategy is adopted to determine the global FLR for site assignments. The knowledge of site-specificity of the modification is introduced at this point. The variant groups which assign each PTM to at least one valid site are valid and are added to the target database. Variant groups which assign at least one PTM to invalid sites only are invalid and added to the decoy database. Each variant group is then scored by the total estimated abundance. FLR is then computed at each score threshold by estimating how often that threshold results in invalid/decoy site assignments. (b2) PTM site assignments at a chosen FLR are reported as output. If the peptide has one reported group with one valid site only, then the modification is unambiguously assigned to that valid site. If the reported group has more than one possible valid site but does exclude sites, this is considered partially localized. If the reported group has multiple valid sites, then the site assignment is ambiguous. If the reported group contains all possible variants, we consider this as a non-localized result. Multiple groups passing the score threshold for the peptide are an indication of co-elution where the relative scores can be interpreted as the relative abundances of the co-eluting sites.

we only consider collision induced dissociation (CID) MS/MS, $Masses(P)$ consists of b and y single and doubly charged and b and y isomers with an extra dalton. However, depending on the dissociation method used, this can be easily configured to support other ion types. See Section Appendix A.1 for more details.

1. Given a spectrum S from a peptide P, we define $Intensities(S, P)$ as the intensities of the spectrum peaks in S at $Masses(P)$. Without loss of generality, the vector $Intensities(S, P)$ is always normalized to Euclidian norm 1.

2. Given a spectrum S from a peptide P and a spectrum S_v from a modified peptide variant P_v, we define the similarity between the extracted peak intensities as

$$\begin{aligned} Similarity(S, S_v) &= \cos(Intensities(S, P), Intensities(S_v, P_v)) \\ &= Intensities(S, P) \cdot Intensities(S_v, P_v) \end{aligned}$$

In Supplemental Figure A-1a, we show the distribution of the cosine score similarities of unmodified charge 2 human lens spectra pairs which are identified as the same peptide. 99.44% of the pairs have more than 0.4 cosine similarity confirming that the fragment ion intensities are conserved across unmodified spectra. The distribution of spectral similarities of modified-unmodified spectra pairs is shown in Figure A-1b. We see that even after modification, there is a significant spectral overlap between the spectra pairs of the same peptides: 96.2% of the pairs shows more than 0.4 cosine similarity. By contrast, as shown in Supplemental Figure A-1c, for unmodified spectra of the same length with parent masses within 20 Da of each other, only 65.8% of the spectra are above a cosine of 0.4.

Even for phosphorylation, which is known to strongly affect CID fragmentation [21, 22], this assumption seems to hold. For MS2 data in Supplemental Figure A-2b, 94.98% are above cosine .4. For unmodified vs modified MS3 data, shown in Supplemental Figure A-2c over 89.5% of spectra are above .4.

In cases where the unmodified spectrum is not available in the experiment itself, we can either use spectral libraries [23] or a prediction tool such as MassAnalyzer [24, 25, 26]. From these sets of unmodified spectra and identified spectra within the dataset used, the best candidate unmodified spectrum for a particular modified spectrum is chosen. For a modified spectrum S' annotated by P', we choose all spectra associated with the unmodified peptide P of the same sequence. If there is more than one candidate spectrum, we choose the unmodified spectrum with the highest cosine to the modified spectrum as our unmodified spectrum S. Note that there are other methods for choosing the best candidate, for example, giving preference to unmodified spectra from the same dataset as the modified spectrum. See Supplemental Materials Appendix A.1 for more details.

Following the selection of the best candidate spectrum, the theoretical spectrum for each variant is generated from the unmodified peak intensities.

1. Given a spectrum S from an unmodified peptide P, we want to predict a theoretical spectrum for P_v, a modified variant of P.

2. Extract peak intensities from S at $Masses(P)$ and use these to set the corresponding peak intensities in S_v at $Masses(P_v)$.

2.3 Quantitation of Modification Variants via Linear Programming

An experimental spectrum is modeled as a linear combination of the theoretical spectra for all possible variants of the same modified peptide sequence. By scaling peaks in each modification variant's theoretical spectrum based on the observed peaks in the modified spectrum, we can estimate the overall quantity of each modification variant using linear programming.

For a modified spectrum S' with a peptide identification P' with the unmodified peptide version of P, all variants are generated as described in Section 2.1. In order to build the LP, all possible theoretical peak masses for every variant need to be generated. The multiset of modifications $\mathcal{M}$ is extracted from P'. All possible combinations of $\mathcal{M}$ are generated and summed together to find all distinct mass shifts for

unmodified peaks in the spectrum. All peaks from $Masses(P)$ are then added to the theoretical ion vector T. All modification mass shifts are also added to each peak in $Masses(P)$ and added to T.

1. For each theoretical ion t_j in T, if a peak with intensity y from S' is within the expected peak tolerance, then observed peak intensity $O_j = y$ otherwise observed peak intensity $O_j = 0$.

2. For each theoretical ion t_j, if there is a variant v_i where $t_j \in Masses(v_i)$ in the expected peak tolerance, assume that Q_i is contributing to overall intensity O_j. We add all such variants to $\mathcal{V}_j \subset V$. The expected intensity of the peak theoretical peak t_j is determined as described in SectionAppendix A.1 and designated as d_j.

   For each theoretical ion, it is assumed that the observed peak is the sum of the contribution of all variants scaled by their quantity and expected intensity.

$$O_j \approx \sum_{i=1}^{|\mathcal{V}_j|} d_j \times Q_i$$

   From this, the error of each peak can be approximated as follows

$$\epsilon_j = O_j - \sum_{i=1}^{|\mathcal{V}_j|} d_j \times Q_i$$

An LP is then generated which minimizes the error of each peak.

Input	Output	Formulation
d_j for every ion j O_j for every ion j	Q_i for every variant v_i	$\min \quad \sum_{j=1}^{r} \lvert\varepsilon_j\rvert$ $\text{s.t.} \quad \varepsilon_j = O_j - \sum_{i=1}^{\lvert\mathcal{V}_j\rvert} d_j \times Q_i$ $Q_i \geq 0$ $Q_i s$ are normalized prior to output so that $\sum_i (Q_i) = 1$

Essentially, for each peak we seek to estimate how much each variant contributes to the overall abundance of that peak based on the expected intensity. We minimize error by comparing this predicted intensity to the modified spectrum. As a simple example, if we have a set of variants (P,16)EPT, P(E,16)PT, PE(P,16)T and PEP(T,16) we expect the $b3$ ion to have an intensity based on the intensity of the $b3$ ion from our unmodified spectrum. The $b3 + 16$ ion would be contributed to by the expected quantities of (P,16)EPT, P(E,16)PT, PE(P,16)T while the unmodified $b3$ ion would only be contributed to by PEP(T,16). If we do not see the unmodified version of the $b3$ ion and it is strong in the unmodified spectrum, it is likely that the PEP(T,16) variant is not present. We estimate the quantities of all three variants based on the intensities of the modified and unmodified versions of $b3$ and the other ions we observe and try to minimize the difference between our expected intensities and our actual $b3$ intensities. See Supplemental Materials Appendix A.2 for more details.

As we have linear constraints, and linear objectives, we can solve the problem efficiently using existing Linear programming techniques. [27] We are able to leverage existing LP solvers such as GLPK [1] and CPLEX [2] by using the generic CPLEX LP format to formulate our problem.

[1] http://www.gnu.org/software/glpk/

[2] www.ibm.com/software/integration/optimization/cplex-optimizer/

2.4 Grouping Modification Variants

The LP outputs an estimated abundance per modification variant. However, depending on the completeness of the fragmentation pattern, there might be no or very few *distinguishing peaks* between two modification variants. In that case, it is impossible for LP to distinguish between the variants in its abundance assignment. In the absence of distinguishing ions, some estimates for individual variant(s) will not be accurate, but for the *groups* of indistinguishable variants. Therefore, we group these variants that do not have enough/any distinguishing peaks in between, and report total estimated abundance per variant group instead of individual variants.

In order to minimize the effect that the order of grouping has on the groups which are output, a form of hierarchical clustering is used which recalculates the distance between groups every time pairs of groups are merged.

1. Form n clusters containing a single variant, $g_1 \cdots g_n$.
2. Find the distance between pairs of variant groups. To calculate the distinguishing intensity between g_i and g_j we consider all peaks identified by any variant in g_i and all peaks identified by any variant in g_j. If a peak in S' is identified by any variant in g_i and not by any variant in g_j we add that to the total distinguishing intensity. We then calculate the distance by dividing the sum of distinguishing intensity by the total identified intensity.
3. Find the two groups g_i and g_j with the lowest distinguishing intensity. If these groups have a distinguishing intensity above our grouping threshold, stop.
4. Merge the g_i and g_j and create new group g_k. Q_k is defined as the sum of Q_i and Q_j.
5. Recompute distances between g_k all other groups.

Grouping variants allows us to report ambiguous site identification of modifications when it is impossible to distinguish between the sites by the detected peaks in the spectrum. A higher grouping threshold will lead to more ambiguous results, while a lower grouping threshold will increase the granularity of the results, but will increase the number of false hits. As we can see in Figure 2, this threshold can have a very strong effect on the results which is highly dependent on the dataset. Successive thresholds can be used to avoid this problem as described in Section 2.5.

Note that this grouping method is only suitable for cases where there are a small number of modifications per peptide. As the number of modifications per peptide increases, there is a combinatorial explosion in the number of possible variants, much like the explosion in search space in peptide database search when more modifications are allowed per peptide. Since hierarchical clustering is essentially an n by n comparison of all the modification variants, this quickly becomes intractable. There are possible ways to address this problem in the future, including dynamically generating variants by iteratively grouping successive positions which do not have distinguishing intensity between them. Such methods will need to be investigated to address localization for highly modified peptides.

2.5 Evaluation of Dataset Results using FLR

In peptide identification, to evaluate the performance of search algorithms on a dataset, use of target and decoy databases is widely practiced. The target-decoy search strategy permits an impartial assessment of search results and by applying a score cutoff, the false discovery rate (FDR) can be controlled at a desired level. We adopt a similar target-decoy strategy to determine the global false-discovery rate (FDR) for modification variant identifications.

In order to calculate FLR, we take groups of variants and their associated quantities. During all previous scoring, amino acid site specificity has been ignored. At this stage we reintroduce the "known" amino acid specificity for the modification types. These variant groups construct our target and decoy database as we quantify variant groups instead of individual variants. A "valid variant" has all associated modifications associated with a valid amino acid residue. For example, $(P, 42)EP(T, 80)IDE$ is a valid variant if we

consider n-term acetylation and phosphorylation, $(P, 42)E(P, 80)TIDE$ would not be a valid variant. If a group has at least one valid variant, it is called *valid group* and it is added to the target. If a group does not have any valid variant, it is an *invalid group* and added to the decoy. Thus, an invalid group is considered a decoy identification while an valid group identified is considered a target identification.

The ratio of target to decoy will be influenced by which modifications are being considered and what their target amino acids are. In order to account for this, we adjust the incorrect hits by a scaling factor. We take the fraction of all valid groups, TP divided by fraction of invalid groups, FP from all of the returned variant groups, even those with a score of 0. This is our scaling factor ρ. FLR is determined by taking the number of incorrect hits I and correct hits T and calculating $\frac{I*\rho}{T}$.

FLR estimation by target-decoy strategy is coupled with our scoring scheme. Each instance in target and decoy is assigned a score. In our target-decoy approach for FDR, we use the estimated abundances assigned by LP to score each variant group along with the cosine of the theoretical spectra from our LP output to our original modified spectrum S'. This is a reasonable choice since we would have more confidence in the presence of a group if its estimated abundance is high. If the estimated abundance is low, it is more likely that the abundance is assigned due to contaminant peaks or noise. However, the granularity of scoring if only quantity is used is low. Most abundances are at 1 or 0 since the LP is able to unambiguously determine the site. This does not take into account the overall quality of the spectrum, such as a low number of peaks or a large number of low intensity peaks. In order to add an extra layer of granularity we sort by cosine between the estimated theoretical spectrum (based on the calculated quantity of the variant groups and the expected intensities of all theoretical peaks) and the modified spectrum.

Since variants which do not have enough distinguishing intensity are grouped together, there is another parameter that can affect the results: grouping threshold. If a high threshold is chosen, there will be fewer decoy results, but at the cost of higher ambiguity. (See Figure 2) In order to simplify the handling of grouping thresholds, we use a modified approach when calculating FLR. Instead of picking a static grouping threshold and calculating FLR at that single threshold, we use successive thresholds to maximize the granularity of results while still returning a high number of results.

First, FLR is calculated at the lowest grouping threshold. Then all results above our FLR cutoff are returned. Then for the next threshold, all spectra and their associated variants which were above are cutoff are removed and FLR is recalculated on the remaining set (including recalculating the scaling function). This continues until there are no more grouping thresholds to pass. As shown in Figure 2, this maximizes the number of results at every FLR threshold. Lower FLR thresholds will push towards more ambiguous results since higher grouping thresholds must be used to maintain that FLR, but this still yields much more robust results than simply picking a static threshold.

3 Results

3.1 Site Assignment Analysis on Synthetic Human Phosphoeptide Dataset

In this section, we validated our approach on an MS/MS dataset of peptides with known modification sites [8]. For our analysis we used the LTQ-Orbitrap XL CID data. Full scan MS spectra were acquired in the Orbitrap at a resolution of $60,000$ at m/z 400 after accumulating ions to a target value of $1 * 106$. The five most intense ions were selected for fragmentation by CID with a resolution of $.5$ Da for fragment tolerance . These $7,992$ CID spectra were then searched using $\mathrm{Inspect}$ [28] with 2 Da parent and $.5$ fragment tolerance against the 153 proteins from which the 180 peptides were synthesized plus 8 common contaminant proteins at a cutoff of 1% FDR. 966 charge 2 spectra and 281 charge 3 spectra were identified.

In order to ensure the presence of an unmodified version of each modified peptide, the NIST human ion trap spectral library was included. [23]. MassAnalyzer was also used to generate predictions for the unmodified versions of the modified peptides [24, 25, 26] at the appropriate charge states as well.

The identified spectra were then scored using the LP scoring method described above. Peptides which were originally identified by $\mathrm{Inspect}$ as having a modification on serine or threonine were assumed to have a modified fragment shift of -18 (β elimination of phosphate) [**?**]. Those which were identified as having a modification on tyrosine were assumed to have a modified fragment shift of $+80$ [**?**]. These results were then grouped as described in Section 2.4 using thresholds from $.015$ to $.45$ and steps of $.03$. FLR was then

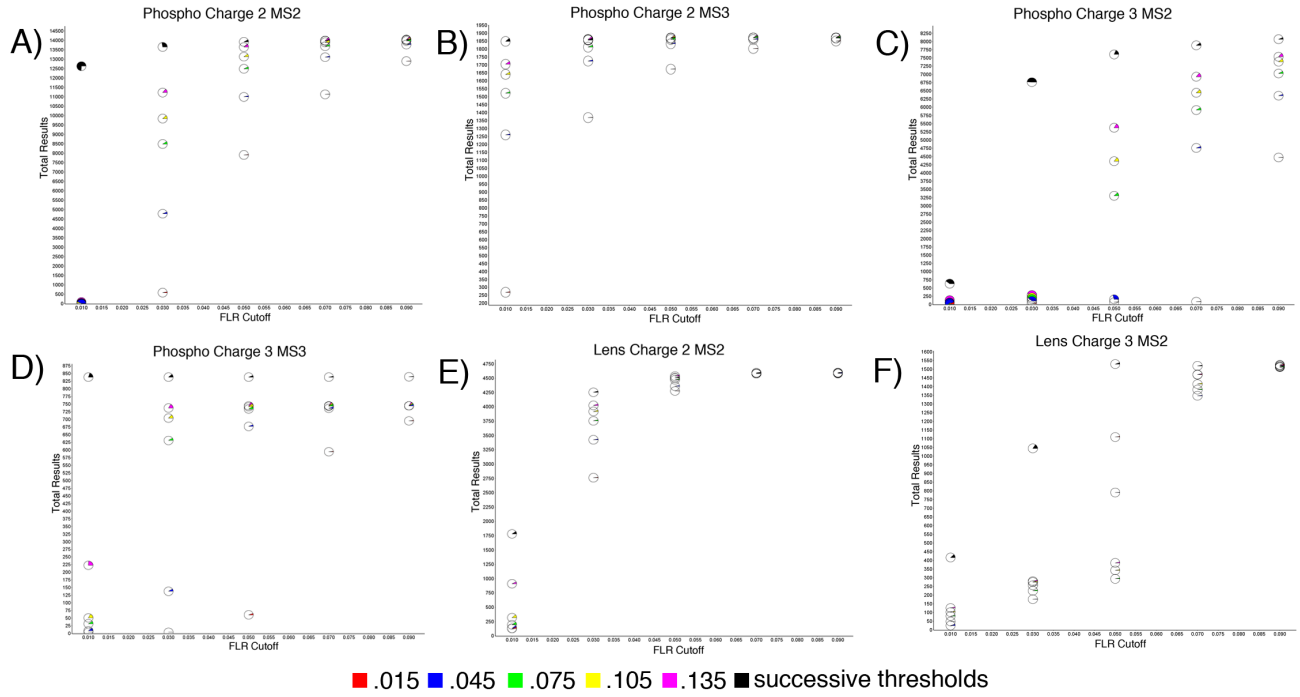


Figure 2: This figure illustrates the effect of adjusting grouping thresholds and our FLR on the proportion of total results and what percentage of those results are considered ambiguous. Decreasing the FLR threshold increases the number of ambiguous results as results with more variants grouped together are more likely to include a valid site. Increasing the grouping threshold increases the number of results at every threshold, but includes more ambiguous results. Using successive grouping thresholds means ambiguity is solely affected by FLR and the number of results at every threshold is maximized. Colored portions of each circle indicate the percentage of total results which are considered ambiguous by excluding any valid sites. a) Charge 2 MS2 phosphorylation data b) Charge 2 MS3 phosphorylation data c) Charge 3 MS2 phosphorylation data d) Charge 3 MS3 phosphorylation data e) Charge 2 MS3 lens data f) Charge 3 MS2 lens data

applied at successive thresholds as described in Section 2.5 assuminng that modifications of -18 on serine and threonine were valid and modifications of $+80$ on tyrosine were valid.

Based on the empirical FLR of this dataset we were able to compare our expected FLR values. All results at a particular threshold with at least one valid variant were considered. Any valid variant in a particular spectra which had the expected modification site as based on the original synthesis information was considered a target, an incorrect assignment was considered a decoy. As shown in Figure 3, for charge 2 data, the estimated FLR are close to the empirical FLR values. In charge 3, the estimated values do not match as well, most likely due to the fact that there were many many fewer results. In addition, there were a number of replicates of the same peptide which may have been a spurious misidentification. (See Appendix Appendix B.1). If we discard this specific case, the numbers for empirical and estimated FLR seem to be within a reasonable range of each other.

Even though this is a relatively small dataset, it is clear that FLR combined with our scoring system can provide a good approximation to the empirical FLR for phosphorylated data.

3.2 Site Assignment Analysis on Yeast Phospho-Peptides

Next, a discovery study was done on a larger dataset of phospho-peptides in order to to evaluate how well the LP and FLR combined approach approach performs for high throughput phosphorylation site assignment versus current methods such as Ascore [9].

The dataset consists of both neutral loss dependent MS3 spectra and MS/MS spectra of the de-phosphorylated peptides to increase the likelihood of having unmodified versions of modified peptides from the same dataset. The sample from S. Cerevisiae was first IMAC-enriched for phospho-peptides and neutral loss dependent

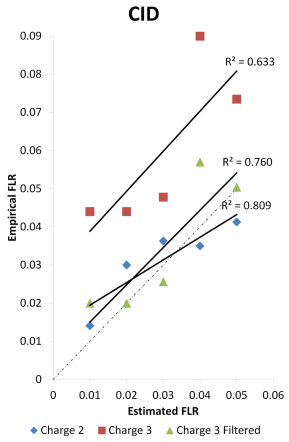


Figure 3: Empirical FLR based on expected residues vs. estimated FLR for the non decoy synthetic results. Only results which were either identical or superstrings of the expected peptide sequence from the original study were considered. For charge 2 nonunique spectra, the maximum number of results considered at .09 FLR were 956 spectra out of 966 modified spectra identified by Inspect. The charge 3 nonunique spectra had a maximum number of results of 252 out of 281 identified by Inspect. The filtered charge 3 nonunique spectra have Pep0001 A11 filtered. The maximum number of filtered results considered were 245.

MS3 spectra were acquired from the first half of the sample. The second half of the sample was CIP treated for removal of the phosphate group, HILIC fractionated into 11 fractions and MS/MS were acquired on the de-phosphorylated peptides. The CIP treated samples also were treated with SILAC K+8 and R+10 modifications. `where is the citation for this?`

The $1,669,605$ MS2 and $78,939$ MS3 spectra were identified using Inspect [28], against the Saccharomyces Genome Database from February 2011 [**?**] with parent mass tolerance 2 Da and fragment mass tolerance 0.5 at a 1% spectrum level FDR allowing for one phosphorylation or -18 per peptide. The data were searched twice, once with unmodified K and R and once with K+8 and R+10 fixed modifications. Since localization of the SILAC modifications do not matter, fixed mods were used to increase the total number of identifications without significantly increasing search time. In total, $139,628$ spectra were annotated. From these results, $30,292$ modified spectra (excluding those with SILAC mods) were found of which $19,481$ were charge 2 and $10,218$ of which were charge 3. From the MS2 spectra, $4,496$ phosphorylated peptides were identified (ignoring modification position), 927 of which also were identified in an MS3 spectrum.

Although the CIP treated sample contained many unmodified versions of the modified peptides, the NIST yeast ion trap spectral library was included in order to capture cases where the unmodified peptide was not identified [23]. Also, for the $4,496$ unique modified peptides, MassAnalyzer was used to generate predictions [24, 25, 26] at the appropriate charge states.

The identified spectra were then scored using the LP scoring method described above. Due to some unmodified results having SILAC modifications, altered values for K and R were used on some unmodified spectra. Phosphorylated peptides were assumed to have a modified fragment shift of -18 (β elimination of phosphate) [**?**].[1] These results were then grouped as described in Section 2.4 using thresholds from .015 to .45 and steps of .03. FLR was then applied at successive thresholds as described in Section 2.5 with the variant groups marked as valid if they assigned the modification to an S or T.

In order to compare with existing methods, Ascore [9] was run on the phosphorylation data as well. By taking results from Inspect and translating into pepxml, Ascore was run using the same input results as our FLR formula.[2] As shown in Figure 4 at equivalent FLR cutoffs (5% FLR and 13 for Ascore), the LP and FLR method represents a substantial improvement in the number of results. For MS2 data, this ranges from a 35% improvement in the number of results for charge 2 data to 41% improvement for charge 3 data.

[1] Due to the relatively small number of tyrosine phosphorylations (320 total), no special provisions were made to use +80 fragments for tyrosine phosphorylated peptides

[2] We were unable to perform a similar comparison with the lens dataset since Ascore was incorrectly allowing for modifications on serine and threonine for all modification types regardless of whether they were phosphorylations or not.

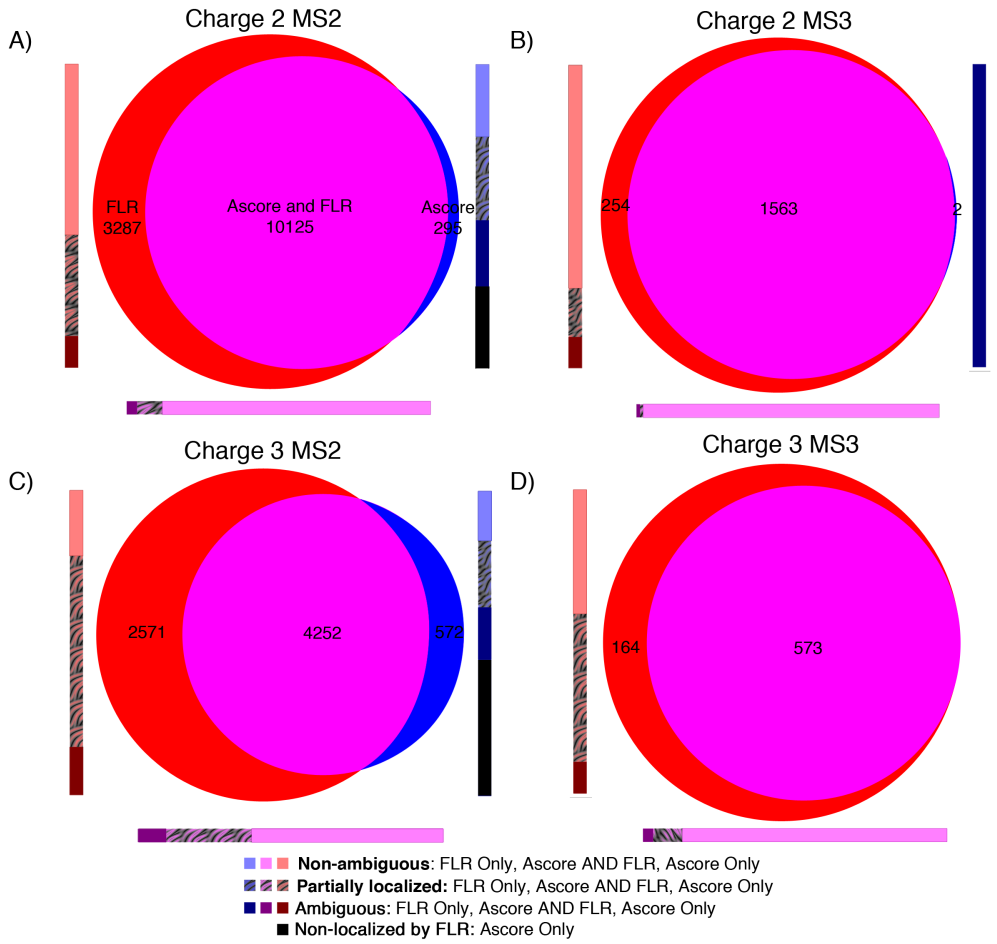


Figure 4: Ascore vs. FLR for non-trivial phosphorylation results. Non-ambiguous are FLR results which return only one valid variant. Partially localized results exclude possible valid sites, however there is more than one valid variant. Ambiguous does not exclude any valid site, however, it does exclude incorrect variant sites. Non-localized by FLR means that all possible variants are in one group, no valid or invalid sites are excluded. Non-localized results are included as Ascore only as we do not return any localization information. An FLR cutoff of .05 and an Ascore cutoff of 13 was used. a) Charge 2 MS2 results out of $13,881$ nontrivial results identified by Inspect. $13,412$ were below our FLR cutoff, $10,420$ were above the Ascore cutoff. b) Charge 2 MS3 results out of $1,830$ nontrivial results identified by Inspect. $1,817$ results were below FLR cutoff, $1,565$ results were above Ascore cutoff. c) Charge 3 MS2 results out of $7,797$ nontrivial Inspect results. $6,823$ were below FLR cutoff, $4,824$ were above Ascore cutoff. d) Charge 3 MS3 results out of 760 nontrivial Inspect results. 737 results were below FLR cutoff, 573 were above Ascore cutoff.

The improvement for MS3 data is less dramatic, although still striking from 16% for charge 2 and 28% for charge 3. The improvement is particularly noticable when the fragmentation is poor, as in charge 3 MS2, where the results often can be partially localized, but are not unambiguous.

3.3 Site Assignment Analysis on Lens Dataset

A high throughput human sample was evaluated in order to examine multiply modified peptides containing a larger set of modification types. The dataset consists of human lens proteins from multiple lens samples, from a variety of patients of different ages both those affected by cataracts and healthy. A major component of the lens proteome comprises of 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 [29]. Mass spectrometry data ($840,676$ spectra) from human lens proteins were acquired on a ESI ion trap mass spectrometer. This dataset was used by Wilmart et al. [30] and will be referred as 'lens dataset'.

The spectra were identified using Inspect [28], against a human lens protein database subset of the Human IPI database containing 57 proteins plus 8 common contaminant proteins with parent mass tolerance 2 Da and fragment mass tolerance 0.5 at a cutoff of 1% spectrum level FDR. Due to the large number of duplicate peptides, a peptide level FDR of 1% was also used. Two modification per peptide were allowed from the set of 17 most common modifications previously identified by Wilmarth et al [30] and Na et all [31] (See Supplemental Table 2). In total, 46,973 spectra were annotated and 11,456 were identified as modified peptides. Out of these results, only the 8,561 charge 2 and 2,139 charge 3 modified spectra were considered. From these there were 1,627 unique modified peptides and 592 unique peptides if modifications are ignored.

In order to ensure there was an unmodified version of each modified peptide, the NIST human ion trap spectral library was included. [23]. For the 592 unique peptides, MassAnalyzer was also used to generate predictions [24, 25, 26] at the appropriate charges state as well.

The identified spectra were then scored using the LP scoring method described above. These results were then grouped as described in Section 2.4 using thresholds from .015 to .45 and steps of .03. FLR was then applied at successive thresholds as described in Section 2.5 with the variant groups marked as valid or invalid according to their expected amino acid residues based on the original search parameters for Inspect. In some cases where two modifications are of similar mass (formylation and dimethylation, for example), all residues for both modification types are considered valid.

For results which passed the threshold, our localization method is able to unabiguously assign the site for almost all modification types, even those with a large number of possible modification sites such as methylation or formylation/dimehtylation. The only exception is with charge 3 phosphorylation, which is a pattern also seen in the Yeast dataset (see Section 3.2). Even in this case, the vast majority of sites exclude at least one possible site of phosphorylation. In total, out of 6,320 non-trivial spectra, at an FLR of 5% we are able to unambiguously localize 3,136 spectra (50%) and partially localize 3,136 spectra (59%).

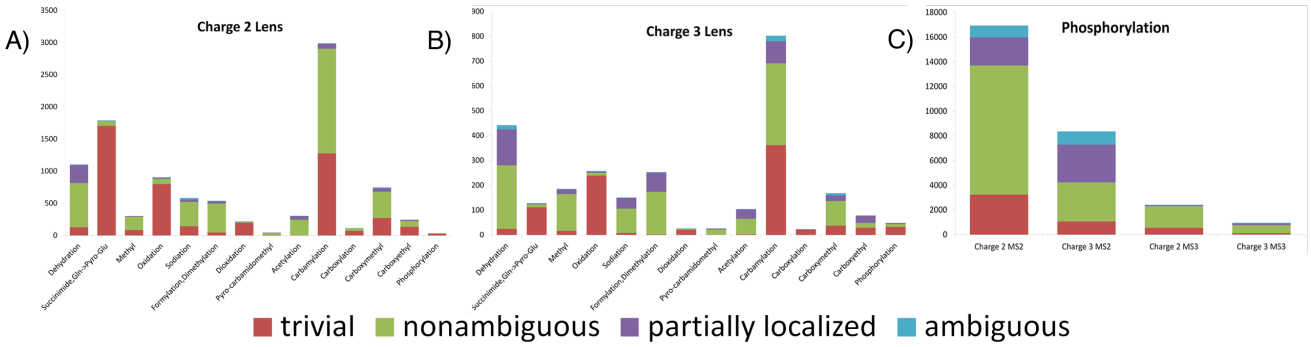


Figure 5: Lens and phospho modification breakdown at 5% FLR using successive grouping thresholds. Trivial results are results for which there is only one valid site. Non-ambiguous are FLR results which return only one valid variant. Partially localized results exclude possible valid sites, however there is more than one valid variant. Ambiguous does not exclude any valid site, however, it does exclude incorrect variant sites. a) Lens dataset at charge 2, 6,605 localized out of 8,561 modified spectra. b) Lens dataset at charge 3 1,927 out of 2,139 modified spectra c) Yeast dataset results.

4 Discussion

We have presented an accurate method for estimating false localization rates for post translational modifications in high throughput data. While our FLR method is shown in conjunction with our method for scoring modification sites, it would be possible to use a similar technique using a different scoring method as long as it is able to score each modification site independently. Although there have been previous attempts to estimate FLR [7], ours is the first approach which does not increase ambiguity in the initial search by allowing for "decoy" modification sites when the initial search is done and which also calculates FLR entirely independently of the initial search in a generic way. As the amount of mass spectrometry data containing post translational modifications increases, it is essential to have a generic method for assessing localization quality.

In addition to the clear advantages provided by being able to qualitatively assess the validity of site localizations, our scoring scheme provides advantages over current post translational scoring methods as well. It can capture information about site assignment even in cases where most other methods fail such as incomplete fragmentation or mixtures with the same modification on different sites. In complex samples this becomes especially clear. Being able to approximate localization even in cases where the answer is ambiguous can drastically increase the information that can be captured from the sample as we can see from the increase in the number of phosphorylation results for MS2 above Ascore (between 35 and 41%). While there are still improvements to be made, especially in handling grouping of variants when there are large numbers of modifications per peptide, these methods provide a step forward in assessing the quality of PTM assignments and streamlining downstream analysis of PTM expression.

References

[1] M. Mann and ON. Jensen. Proteomic analysis of post-translational modifications. *Nat Biotechnol*, 21:255–261, 2003.

[2] S A Beausoleil, M Jedrychowski, D Schwartz, J E Elias, J Villén, J Li, M A Cohn, L C Cantley, and S P Gygi. Large-scale characterization of hela cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A*, 101:12130–12135, 2004.

[3] N Dephoure, C Zhou, J Villén, S A Beausoleil, C E Bakalarski, S J Elledge, and S P Gygi. A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A*, 105(31):10762–10767, 2008.

[4] Jonathan C Trinidad, David T Barkan, Brittany F Gulledge, Agnes Thalhammer, Andrej Sali, Ralf Schoepfer, and Alma L Burlingame. Global identification and characterization of both o-glcnacylation and phosphorylation at the murine synapse. *Mol Cell Proteomics*, 11(8):215–229, Aug 2012.

[5] A I Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics*, 73:2092–2123, 2010.

[6] R J Chalkley and K R Clauser. Modification site localization scoring: Strategies and performance. *Mol Cell Proteomics*, page (in press), 2012.

[7] Peter R Baker, Jonathan C Trinidad, and Robert J Chalkley. Modification site localization scoring integrated into a search engine. *Mol Cell Proteomics*, 10(7):M111.008078, Jul 2011.

[8] M M Savitski, S Lemeer, M Boesche, M Lang, T Mathieson, M Bantscheff, and B Kuster. Confident phosphorylation site localization using the mascot delta score. *Mol Cell Proteomics*, 10:(in press), 2011.

[9] S A Beausoleil, J Villén, S A Gerber, J Rush, and S P Gygi. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol*, 24:1285–1292, 2006.

[10] Jesper V Olsen and Matthias Mann. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A*, 101(37):13417–13422, Sep 2004.

[11] T Taus, T Köcher, P Pichler, C Paschke, A Schmidt, C Henrich, and K Mechtler. Universal and confident phosphorylation site localization using phosphors. *J Proteome Res*, 10:5354–5362, 2011.

[12] Claudio P Albuquerque, Marcus B Smolka, Samuel H Payne, Vineet Bafna, Jimmy Eng, and Huilin Zhou. A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Mol Cell Proteomics*, 7(7):1389–1396, Jul 2008.

[13] Christopher M Bailey, Steve M M Sweet, Debbie L Cunningham, Martin Zeller, John K Heath, and Helen J Cooper. Slomo: automated site localization of modifications from etd/ecd mass spectra. *J Proteome Res*, 8(4):1965–1971, Apr 2009.

[14] Douglas H Phanstiel, Justin Brumbaugh, Craig D Wenger, Shulan Tian, Mitchell D Probasco, Derek J Bailey, Danielle L Swaney, Mark A Tervo, Jennifer M Bolin, Victor Ruotti, Ron Stewart, James A Thomson, and Joshua J Coon. Proteomic and phosphoproteomic comparison of human es and ips cells. *Nat Methods*, 8(10):821–827, 2011.

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

[16] Peter A DiMaggio, Nicolas L Young, Richard C Baliban, Benjamin A Garcia, and Christodoulos A Floudas. A mixed integer linear optimization framework for the identification and quantification of targeted post-translational modifications of highly modified proteins using multiplexed electron transfer dissociation tandem mass spectrometry. *Mol Cell Proteomics*, 8(11):2527–2543, Nov 2009.

[17] Matthew J Sniatynski, Jason C Rogalski, Michael D Hoffman, and Juergen Kast. Correlation and convolution analysis of peptide mass spectra. *Anal Chem*, 78(8):2600–2607, Apr 2006.

[18] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. Protein Identification via Spectral Networks Analysis. *Proc Natl Acad Sci U S A*, 104:6140–6145, 2007.

[19] 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*, 7:655–667, 2007.

[20] Stein and Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5:859–866, 1994.

[21] A. Tholey, J. Reed, and W. D. Lehmann. Electrospray tandem mass spectrometric studies of phosphopeptides and phosphopeptide analogues. *J Mass Spectrom*, 34(2):117–123, Feb 1999.

[22] Susanne C Moyer, Robert J Cotter, and Amina S Woods. Fragmentation of phosphopeptides by atmospheric pressure maldi and esi/ion trap mass spectrometry. *J Am Soc Mass Spectrom*, 13(3):274–283, Mar 2002.

[23] SE Stein and PA Rudnick, editors. *NIST Peptide Tandem Mass Spectral Libraries. Human Peptide Mass Spectral Reference Data, Yeast, ion trap, Official Build Date: Apr. 6, 2012, Human, ion trap, Official Build Date: May 30, 2012*. National Institute of Standards and Technology, 2012.

[24] Z. Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal Chem*, 76:3908–3922, 2004.

[25] Z Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal Chem*, 77:6364–6373, 2005.

[26] Z Zhang. Prediction of electron-transfer/capture dissociation spectra of peptides. *Anal Chem*, 82:1990–2005, 2010.

[27] George Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1998.

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

[29] D Tsur, S Tanner, E Zandi, V Bafna, and P A Pevzner. Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol*, 23:1562–1567, 2005.

[30] P A Wilmarth, S Tanner, S Dasari, S R Nagalla, M A Riviere, V Bafna, P A Pevzner, and L L David. Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility? *J Proteome Res*, 5:2554–2566, 2006.

[31] S. Na, N. Bandeira, and E. Paek. Fast multi-blind modification search through tandem mass spectrometry. *Mol Cell Proteomics*, 11:10.1074/mcp.M111.010199, 2011.

[32] N Bandeira. Protein identificatation by spectral network analysis. In C Chen and C Wu, editors, *Bioinformatics for Comparative Proteomics*, Methods in Molecular Biology. Springer Humana Press, USA, 2010.

Appendix A Supplemental methods

Appendix A.1 Inference of Fragment Ion Detectabilities

The idea behind our approach is that the observed intensity of a peak is directly related to the total abundance of the theoretical fragment ions that have the same corresponding m/z value, thus to the total abundance of the modification variants that contain those isobaric ions. Simply, if there is a peak in the spectrum which is associated with a single modification variant through a single ion, the peak intensity is a direct measurement for the abundance of the modification variant. However, the fragment ions have different intensities in the mass spectrometry depending on the physico-chemical properties of the fragment ions. For our purposes, only the relative contributions of the fragment ions to the peak intensities per unit abundance of its parent peptide are important.

For inference of expected ion intensities, we are motivated by conservation of spectral shapes of a peptide sequence across different unmodified and modified tandem mass spectra. Previously, Sniatynski et al [17]. showed that delay-series correlation between the mass spectra of modified and unmodified peptides revealed significant spectral overlap at an offset indicative of the modification. This observation has been confirmed in many other publications including ours [32]. We observe similar behavior in also our test MS/MS datasets.

We measure the similarity between two spectra using *cosine score* which is a widely accepted measure for spectral similarity [19]. Given a spectrum S from a peptide P, we define $Intensities(S, P)$ as the intensities of the spectrum peaks in S at $Masses(P)$. We can normalize the intensities in $Intensities(S, P)$ in two ways. First, we can calculate the Euclidian norm of the vector of intensities of S and then extract ions to generate $Intensities(S, P)$. Alternately, we can extract $Intensities(S, P)$ first and then normalize the vector. In cases where we wish to capture noise, we normalize before ion extraction. In cases where we do not care about background noise, we normalize after ion extraction.

For our purposes, since the neutral loss from the parent ion often dominates the intensity in phosphorylated CID spectra [21, 22], we normalize after extraction and ignore the parent ion m/z. Normalization is done simply by taking the euclidian norm of the vector. We consider only b and y ions (single and doubly charged and isotopic) for our fragment masses. For each theoretical mass we use the total summed intensity of peaks within our specified tolerance as the value for that ion type (note that this intensity can be zero if no peaks are found). For a pair of spectra S and S' with peptide annotations P and P' which are the same length, we calculate the cosine by computing the dot product of $Intensities(S, P)$ and $Intensities(S', P')$.

$$\begin{aligned} Similarity(S, S') &= \cos(Intensities(S, P), Intensities(S', P')) \\ &= Intensities(S, P) \cdot Intensities(S', P') \end{aligned}$$

In Figure A-1a, we show the distribution of the cosine score similarities of unmodified charge 2 human lens spectra pairs which are identified as the same peptide using Inspect. 99.44% of the pairs have more than 0.4 cosine similarity confirming that the fragment ion detectabilities are conserved across unmodified spectra. The distribution of spectral similarities of modified-unmodified spectra pairs is shown in Figure A-1b. We see that even after modification, there is a significant spectral overlap between the spectra pairs of the same peptides. 96.2% of the pairs shows more than 0.4 cosine similarity. By contrast, as shown in Figure A-1c, for unmodified spectra of the same length with parent masses within 20 Da of each other, only 65.8% of the spectra are above a cosine of 0.4.

Even for phosphorylation, which is known to affect CID fragmentation significantly [21, 22], this assumption seems to hold. For MS2 data in Figure A-2b, are above the cutoff. For unmodified vs modified MS3 data, shown in Figure A-2c over 89.5% of spectra are above the .4 cutoff. In cases where the unmodified spectrum of the peptide is unavailable from the dataset, there is no way to calculate detectibility. In these cases there are two approaches we can take. The unmodified spectra can be aquired from spectral libraries or theoretical predictions can be generated using an external tool. In our work, we use MassAnalyzer, which is a spectral prediction tool that uses a kinetic model to calculate expected ion fragmentation and intensity for various instrument types. [24, 25, 26].

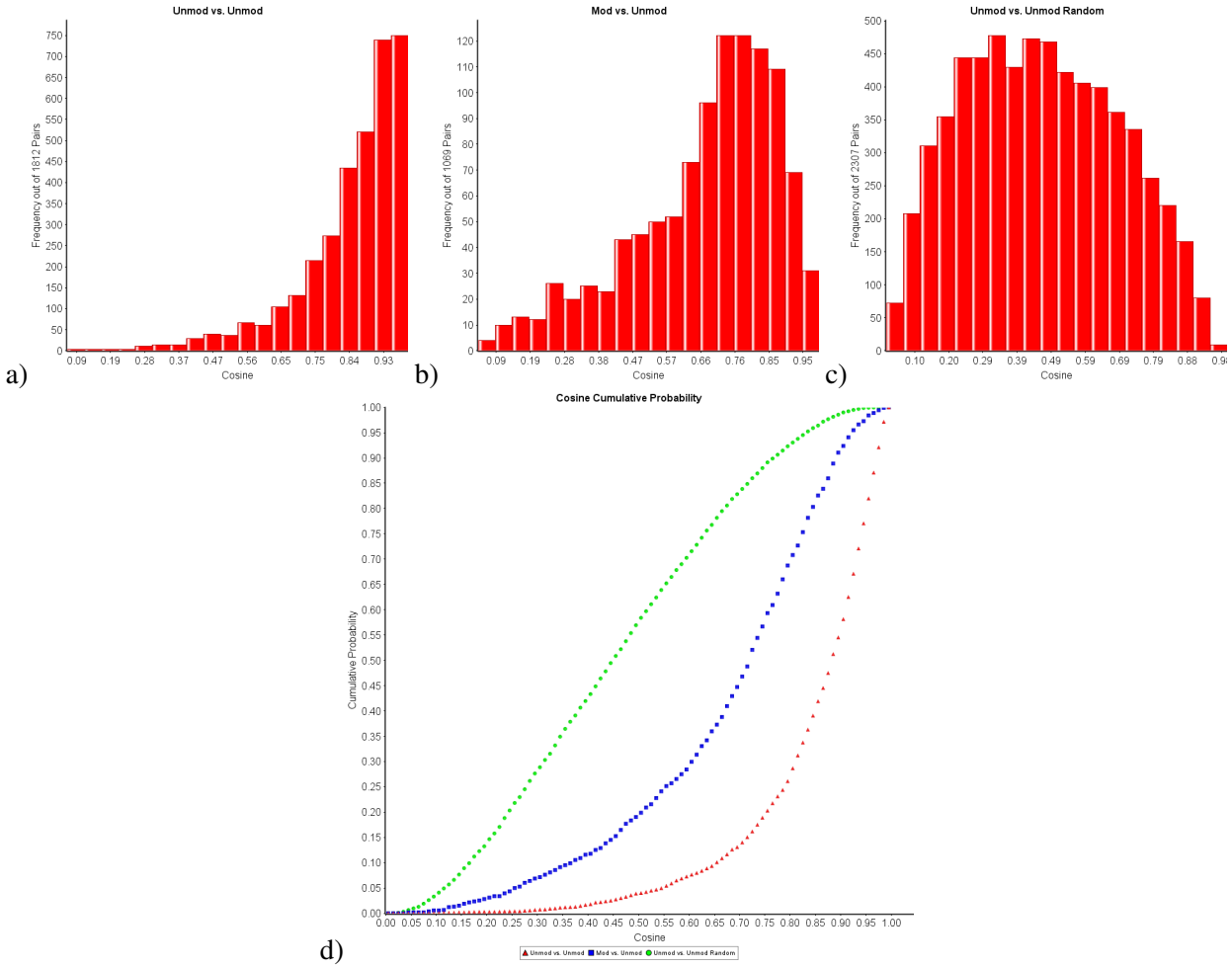


Figure A-1: Charge 2 Lens Dataset: Distribution of cosine score of (a) top scoring unmodified spectrum of a peptide vs. top five matching spectra with the same peptide annotation. (b) top scoring unmodified spectrum of a peptide vs. top five matching spectra with the same peptide sequence, but with modifications. (c) unmodified peptides of differing sequences of the same length and within 20 m/z (d) Cumulative probability distribution of cosine scores of charge 2 spectra of random peptides with similar mass and same length, the same unmodified peptide, unmodified and modified versions of the same peptide.

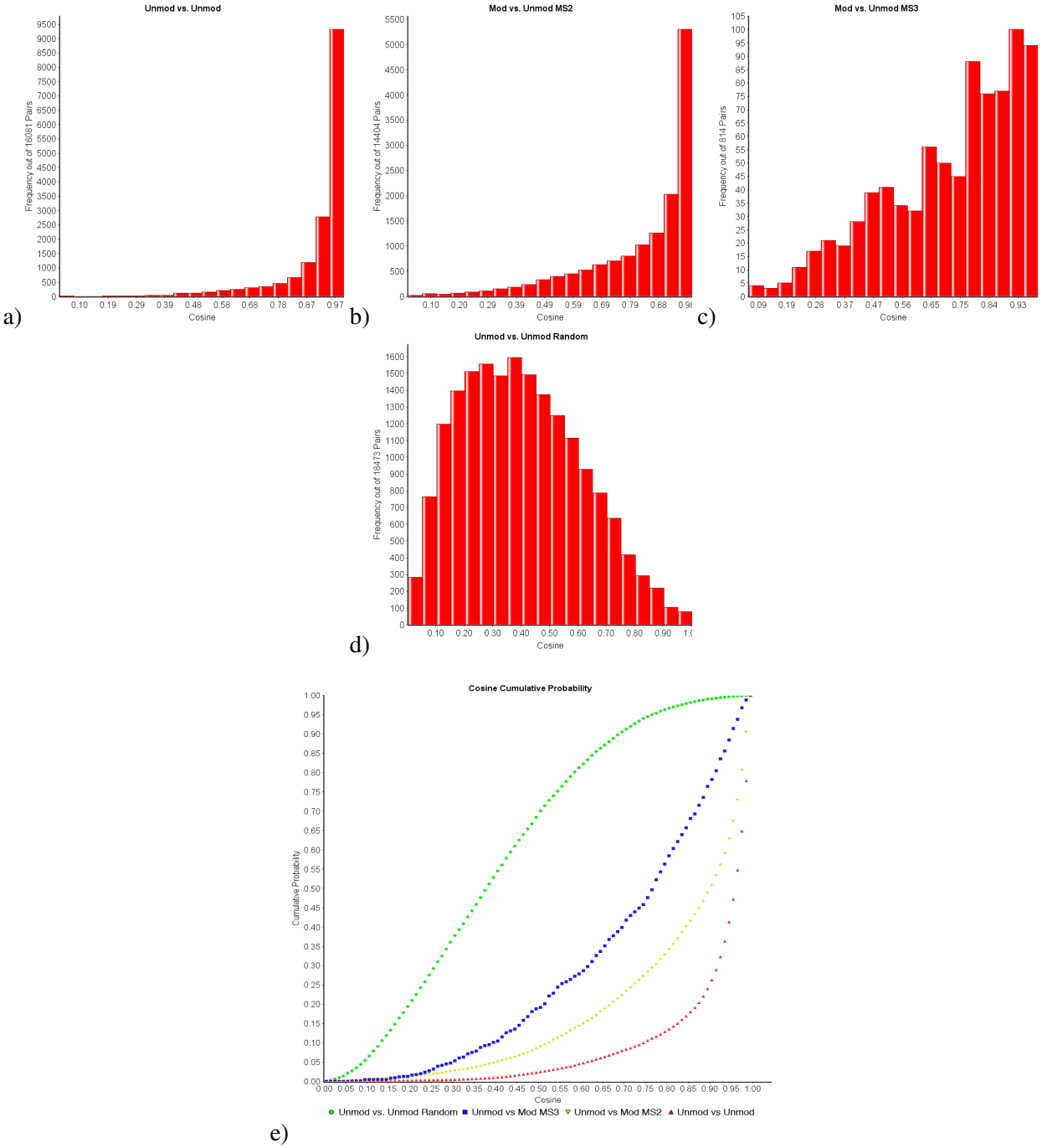
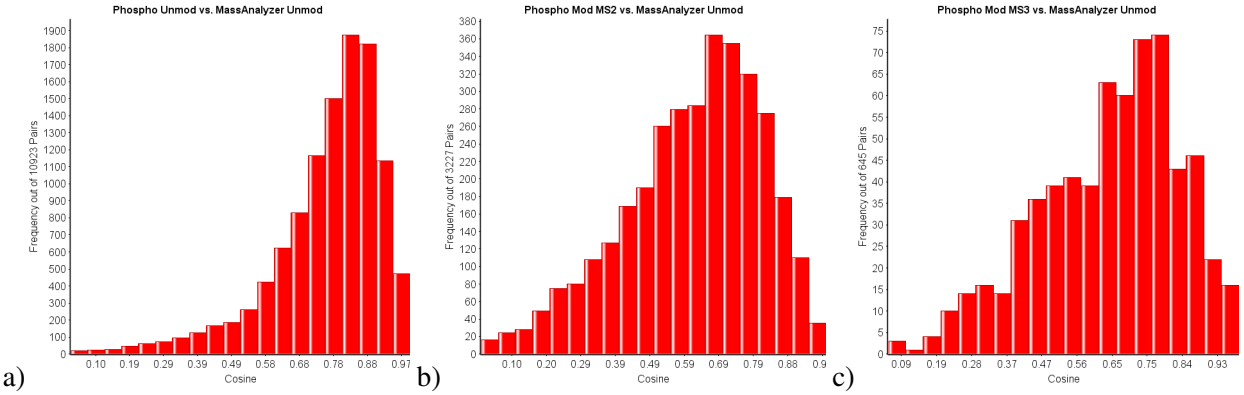


Figure A-2: Charge 2 Yeast dataset: Distribution of cosine score of (a)top scoring unmodified spectrum of a peptide vs. top five matching spectra with the same peptide annotation. (b) top scoring unmodified spectrum of a peptide vs. top five matching spectra with the same peptide sequence, but with modifications. (c)unmodified peptides of differing sequences of the same length and within 20 m/z (d) Cumulative probability distribution of cosine scores of spectra of random peptides with similar mass, the same unmodified peptide, unmodified and dehydrated versions of the same peptide for MS2 and MS3.

For our CID spectra for the yeast phospho datasets, the similarity of the predicted unmodified spectra to the real unmodified and modified spectra were very high. For MS2 modified spectra, 85.00% of the predicted MassAnalyzer unmodified spectra still had a cosine score of greater than $.4$. For MS3 modified spectra, 88.68% were above $.4$. While this isn't as high as the values for unmodified experimental spectra which were 94.98% for modified MS2 and 89.5% for modified MS3 respectively, in cases where the experimental unmodified spectra isn't available, we can see that MassAnalyzer still produces reasonable spectra. Figure A-3



a) b) c)

Figure A-3: Charge 2 Yeast dataset: Distribution of cosine score of (a) unmodified MS2 spectra vs. unmodified predicted MassAnalyzer MS2 spectra (b) modified MS2 spectra vs unmodified predicted MassAnalyzer MS2 spectra (c) modified MS3 spectra vs. unmodified predicted MassAnalyzer spectra.

For our synthetic human phosphorylated peptides, in Figure A-4, we can see that the similarity still quite high.

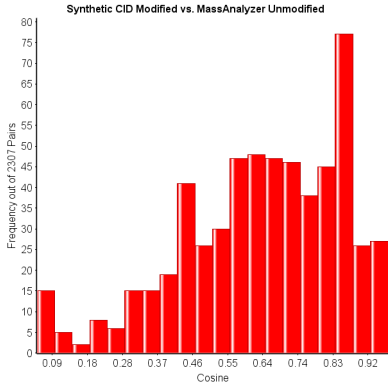


Figure A-4: Mascot synthetic dataset: Distribution of cosine score of MassAnalyzer CID unmodified spectral prediction vs. modified synthetic spectrum for CID spectra

Since it is possible to estimate peak intensities for modified spectra from unmodified spectra, it is possible to calculate theoretical peak intensities for modified variants of a peptide using the unmodified spectrum. Assuming we are given a modified spectrum S' from modified peptide P' whose unmodified version is P, we choose from a set of spectra $S_1 \cdots S_n$ all from peptide P. We then choose the spectrum S_i with the highest cosine similarity to S'. The S is normalized to a total intensity of 1000000. We then take $Intensities(S, P)$. If a peak from S is annotated with a single fragment ion, we assign the peak intensity as the expected intensity of that fragment ion. If a peak is annotated by multiple fragment ions, it is possible to choose from several strategies such as splitting the peak intensity among the ions according to their estimated ion probabilities or PRM scores, etc. Our results did not differ much with different strategies, so we adopted a simpler strategy. If a peak is annotated by multiple fragment ions, we assign the whole intensity to the ion with largest ion probability. For every other fragment ion not detected in the unmodified spectrum, we assign a detectability of $\epsilon > 0$. In our tests we used $\epsilon = 1$.

Appendix A.2 Quantification of Modification Variants via Linear Programming

The mixture spectrum of modification variants is the superposition of individual modified spectra from all of the modification variants. Therefore, a peak intensity does not necessarily correspond to a fragment ion from a single parent modification variant. Most often, multiple fragment ions from one or more variants have the same m/z value and contribute to the same peak. In the case of a singly modified peptide, for example, only y1 and b1 would be contributing to a single variant (the first and last modification positions). All other b and y ions would be shared between multiple variants.

For instance, in the simple example shown in Figure A-5, in the presence of all variants, the intensity of a b3 modified peak will be contributed to by three different variants. Note that since we have an equal mixture of all four variants, the intensity in the modified spectrum of b3 is $1 : 3$ between the unmodified and modified ion. It is crucial to see the mapping between the peaks and the theoretical fragment ions as well as the mapping between variants and the fragment ions. Each observed intensity value in the mixture spectrum at a theoretical m/z value gives information about the total abundance of the fragment ions from variants contributing to that peak.

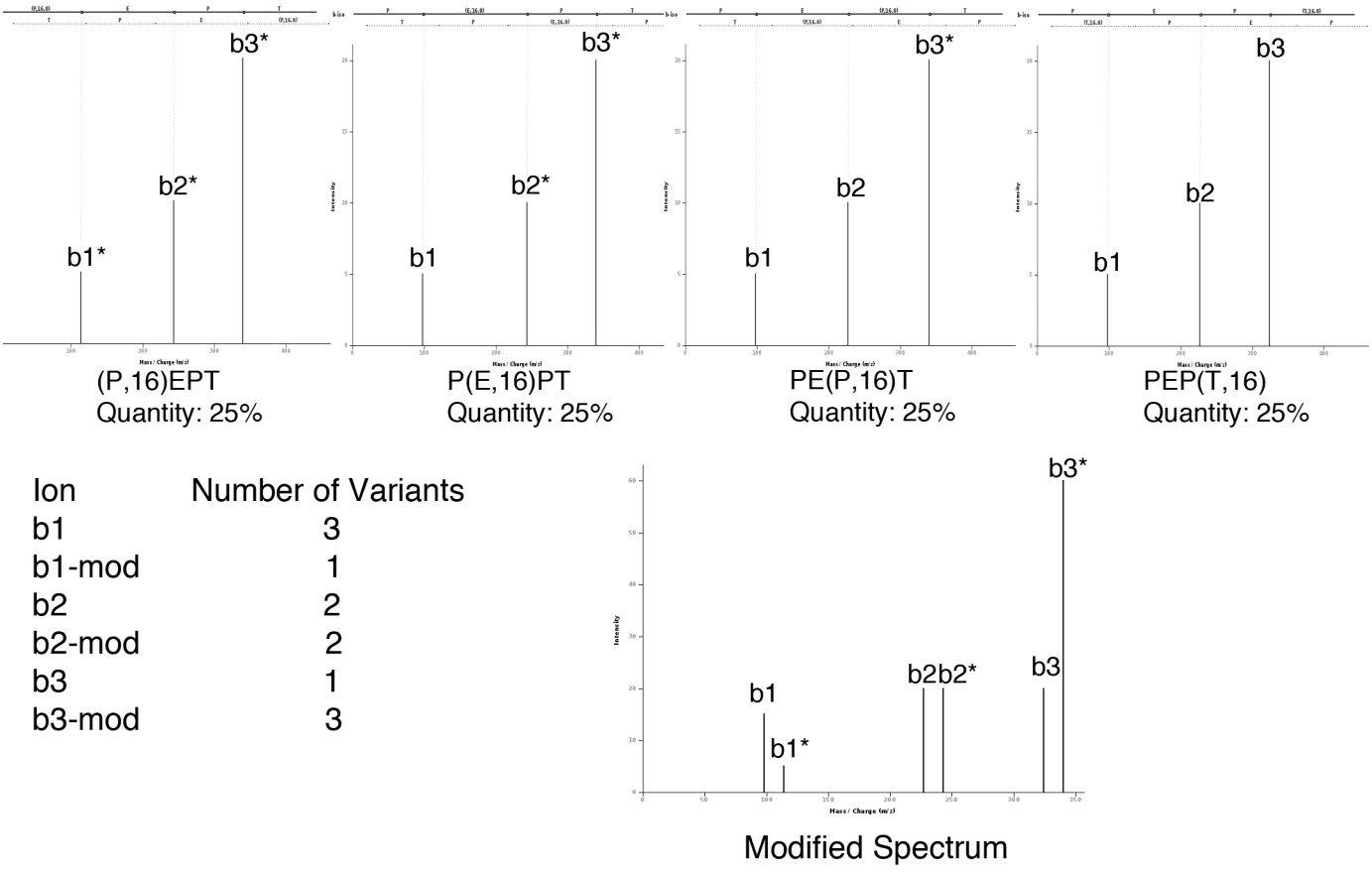


Figure A-5: Given a set of variants, (P,16)EPT, P(E,16)PT, PE(P,16)T and PEP(T,16), we assume that the spectra of all three modified peptides will be similar to the unmodified spectrum PEPT (not shown). If we have equal quantities of all possible variants, then the intensity of each peak can be contributed to by more than one variant. In this case, since b1 unmodified and b3 modified are only contributed to by one variant, the modified and unmodified peaks have a ratio of 3:1 and 1:3 respectively. The b2 peak has two variants contributing to both the unmodified and modified versions, so it has a ratio of 2:2.

Appendix B Synthetic Phosphorylation Results

Appendix B.1 High Scoring Decoy Results

In order to validate the FLR approach, we used spectra from synthetic peptides aquired on an LTQ-Orbitrap XL CID. [8] Full scan MS spectra were acquired in the Orbitrap at a resolution of $60,000$ at m/z 400 after accumulating ions to a target value of $1 * 106$. The five most intense ions were selected for fragmentation by CID with a resolution of $.5$ Da for fragment tolerance . These $7,992$ CID spectra were then searched using $\mathrm{Inspect}$ [28] with 2Da parent and $.5$ fragment tolerance against the 153 proteins from which the 180 peptides were synthesized plus 8 common contaminant proteins at a cutoff of 1% FDR. 966 charge 2 spectra and 281 charge 3 spectra were identified.

After performing our scoring For charge 3 data, as shown in Table 1 there were a number of repeated incorrect hits from the same plate and well number (plate 1, well 11). According to the synthesis information, this should be LQ(T,80)VHSIPLTINK, but we identify as LQTVH(S,80)IPLTINK. There are a few possibilities that may account for this. The most likely case is that our unmodified spectrum does not closely resemble the fragmentation present in the modified spectrum. In this particular case, while there is a NIST spectrum of the same peptide, the cosine was not very high, so the MassAnalyzer prediction was used. The other possibility is that there may have been well crossover. In plate 1 well B2, the modification peptide we identify is present. These two peptides are from different mixtures, however it is possible that since these are from the same synthesis plate, there was well crossover or another issue with contaminuation.

From manual inspection of the spectra both modification positions seem plausible (See Figure B-1), though the identification of doubly charged y9 and y10 modified fragments seems to push the identification in the direction of our LQTVH(S,80)IPLTINK assignment. Likewise, there is a possible phosphorylated b5 ion (which we do not use in our scoring scheme) that indicates that LQ(T,80)VHSIPLTINK might be present.

Table 1: Incorrect assignments at 1% FLR

File	Index	LP quantity	LP valid variants	Synthesis sequence	mixture	Synthesis plate	Synthesis Plate Position
ppeptidemix1_CID_Orbi.mgf	981	LQTVH(S,-18.0106)IPLTINK	LQ(T,-18.0106)VHSIPLTINK	1	Pep0001	A11	
ppeptidemix1_CID_Orbi.mgf	979	LQTVH(S,-18.0106)IPLTINK	LQ(T,-18.0106)VHSIPLTINK	1	Pep0001	A11	
ppeptidemix1_CID_Orbi.mgf	996	LQTVH(S,-18.0106)IPLTINK	LQ(T,-18.0106)VHSIPLTINK	1	Pep0001	A11	
ppeptidemix1_CID_Orbi.mgf	991	LQTVH(S,-18.0106)IPLTINK	LQ(T,-18.0106)VHSIPLTINK	1	Pep0001	A11	
ppeptidemix1_CID_Orbi.mgf	975	LQTVH(S,-18.0106)IPLTINK	LQ(T,-18.0106)VHSIPLTINK	1	Pep0001	A11	
ppeptidemix1_CID_Orbi.mgf	1005	LQTVH(S,-18.0106)IPLTINK	LQ(T,-18.0106)VHSIPLTINK	1	Pep0001	A11	
ppeptidemix3_CID_Orbi.mgf	352	AGIH(T,-18.0106)SGSLSSR	AGIHT(S,-18.0106)GSLSSR	3	Pep0002	A11	
ppeptidemix5_CID_Orbi.mgf	337	ETTT(S,-18.0106)PKKYYLAEK	ET(T,-18.0106)TSPKKYYLAEK	5	Pep0008	A9	
ppeptidemix4_CID_Orbi.mgf	530	ETTT(S,-18.0106)PKKYYLAEK	E(T,-18.0106)TTSPKKYYLAEK	4	Pep0008	A8	
ppeptidemix1_CID_Orbi.mgf	441	ETTT(S,-18.0106)PKKYYLAEK	ETT(T,-18.0106)SPKKYYLAEK	1	Pep0008	A10	
ppeptidemix5_CID_Orbi.mgf	354	ETTT(S,-18.0106)PKKYYLAEK	ET(T,-18.0106)TSPKKYYLAEK	5	Pep0008	A9	

As shown in Figure B-1, this is primarily due to the fact that the doubly charged y6 fragment is not present in the unmodified spectra and therefore is labeled as a b3 ion.

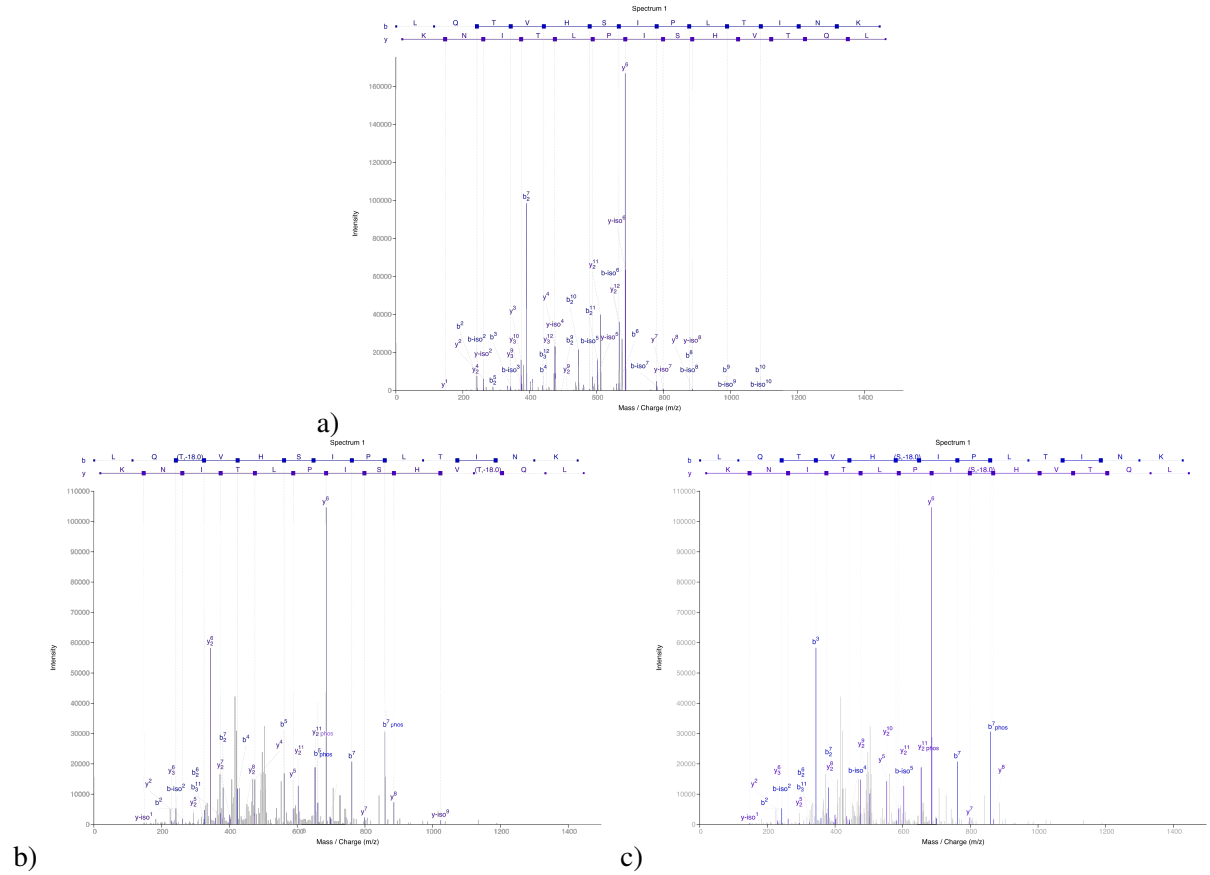


Figure B-1: Comparison of unmodified and modified spectra with two possible modification assignments. Note that while phosphorylated fragments are labeled, they are not included in the LP score. a) Unmodified charge 3 spectrum used for LP b) Scan 975 from ppeptidemix1_CID_Orbi.mgf with expected synthetic modification site c) Scan 975 from ppeptidemix1_CID_Orbi.mgf with site assigned by LP

Appendix C Supplemental results for the Lens dataset

Table 2: Lens Modifications

δ Mass	Residues	Putative Modification
-18.010565	D,E,S,T	Dehydration
-17.026549	N, N-terminal Q	Succinimide, Gln $\rightarrow$ Pyro-glu
14.01565	C,H,K	Methyl
15.994915	H,M,W	Oxidataion
21.981943	D,E	Sodiated
27.994915	H,S,T	Formylation
28.031300	K	Dimethylation
31.989829	W	Dioxidation
39.994915	N-terminal C	Pyro-carbamidomethyl
42.010565	N-terminal, K	Acetylation
43.005814	N-terminal, K	Carbamylation
43.989829	W	Carboxylation
58.0361	K	Carboxymethyl
72.021129	K	Carboxyethyl
79.966331	S,T,Y	Phosphorylation