Modification Site Localization Scoring: Strategies and Performance

Robert J. Chalkley†¶ and Karl R. Clauser§

Using enrichment strategies many research groups are routinely producing large data sets of post-translationally modified peptides for proteomic analysis using tandem mass spectrometry. Although search engines are relatively effective at identifying these peptides with a defined measure of reliability, their localization of site/s of modification is often arbitrary and unreliable. The field continues to be in need of a widely accepted metric for false localization rate that accurately describes the certainty of site localization in published data sets and allows for consistent measurement of differences in performance of emerging scoring algorithms. In this article are discussed the main strategies currently used by software for modification site localization and ways of assessing the performance of these different tools. Methods for representing ambiguity are reviewed and a discussion of how the approaches transfer to different data types and modifications is presented. Molecular & Cellular Proteomics 11: 10.1074/mcp.R111.015305, 3–14, 2012.

Cells respond to elements in their extracellular milieu via a variety of signaling mechanisms, which may be generated in response to environmental cues, electrical stimuli, or chemical messengers. Mostly the signals are propagated by alteration of pre-existing proteins through the addition of post-translational modifications (PTMs)1 to key residues to change their function. These signals are then relayed and amplified until they ultimately manifest in changes in protein expression. It is increasingly clear that mis-regulation of PTMs is a major basis of disease, whether it causes aberrant signaling in cancer, or cytotoxic protein aggregation in neurodegenerative diseases. For this reason, the study of protein post-translational modifications is vital to understanding biological regulation (1).

Mass spectrometry-based proteomics has revolutionized the characterization of protein PTMs, in that it has created the first unbiased strategies to identify which proteins are being modified, with what types of modifications, and on which specific residue/s. Of the many types of PTMs used, protein phosphorylation has justifiably received the greatest attention (2, 3). However, there are many other regulatory modifications such as serine and threonine O-GlcNAcylation (4), lysine or arginine methylation (5), acetylation of lysine side-chains (6), or lysine ubiquitination (7) that are all transient and important to study and understand. There are many other PTMs used by the cell, both transient and stable, that affect protein activity. Indeed, different PTMs do not operate independently of each other, so by studying only a single modification type it is impossible to deconvolute their contributions to signaling mechanisms and cellular responses (8, 9).

The large-scale analyses of all of these modifications follow a similar strategy, in which an enrichment step (antibody affinity, metal affinity, lectin affinity) is followed by tandem mass spectrometric analysis of the resulting mixture (10). Thousands of modified peptides may be reliably identified in these studies, but the challenge of determining modification site localizations among these peptides is less well addressed.

An MS/MS spectrum enables modification site localization by presence in the spectrum of one or more ions of unambiguously assignable ion type that are derived from fragmentation between two amino acids in the peptide that can bear the modification. When such an ion is not present, or the ion cannot be readily distinguished from noise, the modification cannot be confidently localized to either site. When only one amino acid in the peptide can bear the modification there can be no ambiguity as to the localization. Fig. 1 illustrates two spectra of phosphopeptides; one in which a site assignment is unambiguous and one in which there is no evidence to distinguish between two potential sites.

The above rules make assessment of site localization reliability seem straightforward, and indeed it is for some spectra, but it can be complicated by deciding what is a real peak and what is noise; whether peaks that are suggestive of a modification site (e.g. a peak that could correspond to either a phosphate loss from a modified fragment or water loss from an unmodified: the former is more commonly seen) should be
given any weighting; and deciding which residues in a peptide should be considered as potentially bearing a given modification; e.g. for a methylation whether one also considers potential amino acid substitutions that would lead to the same mass change.

For peptide identification two statistical measures of reliability are routinely determined for results. The first of these is a measure of reliability for individual peptide identifications and is usually a measure of how likely a given quality of match will have been achieved by chance; i.e. the smaller this number (probability or expectation value) the more reliable the identification. The second reliability measure, a false discovery rate (FDR), is at the data set level, and estimates the total number of incorrect results being reported. The FDR is normally calculated by a target-decoy database searching strategy in which an assumption is made that the frequency of matching to the normal database and decoy database will be the same (11). Unfortunately, no equivalent reliability measure; i.e. a false localization rate (FLR), can be easily calculated for modification site localization results, although approaches to estimate this have been employed and are discussed in a later section. The only situation in which a true FLR can be measured is when the correct modification site localizations are known. The most obvious situation in which this is the case is through the use of synthetic peptides (12), but another strategy that has been employed was to assess decoy residue site localizations in a phosphopeptide subset of data in which there was only one serine, threonine or tyrosine present per peptide (13).

Although all search engines have a score that measures the certainty of peptide identification, only a few currently have integrated an additional score to measure the reliability of modification site localization (13–16). In the meantime, a variety of different post-search engine tools have emerged to address modification site localization (12, 17–23). Prior to the availability of site localization scoring programs the only option for assessing the site localization reported was “manual verification”; i.e. the researcher looking at each spectrum in turn and making a judgment on whether they deemed the site localization reliable. This is obviously a subjective process, heavily reliant on the expertise of the researcher (and their patience to thoroughly assess as many as a thousand spec-
tra). Indeed, prior to journal publication guidelines forcing
researchers to assess site localization reliability (24) there
were many PTM studies published in which this question was
not even addressed; i.e. they reported results as returned by
the relevant search engine with no further analysis. The effect
of this was the listing of some site localizations that upon
closer inspection of the data should not have been reported.
This would be bad enough in itself, but PTM databases have
extracted these results to populate their resources and it is
not immediately apparent from these databases which site
localizations were determined based on these generally less
stringent standards. This is discussed in more detail in a later
section of this manuscript.

When the proteome informatics research group (IPRG) of
the Association of Biomolecular Resource Facilities (ABRF)
conducted a study in 2010 on identifying phosphopeptides
and localizing phosphorylation sites, the 22 participants who
attempted to assess modification site localization reported
the use of nine named pieces of software and a further six
custom or in-house tools. (25) Despite the range of tools
deployed, most implement one of two basic strategies. These
two main strategies for scoring site localization reliability ei-
ther try to assess the chance of a given peak that allows site
determination to have been matched at random, or calculate
a search engine score difference between peptide identifica-
tions with different site localizations. Tools employing the
former strategy are A-Score (17), PTM Score (MaxQuant/
Andromeda) (18), the Phosphorylation Localization Score
(PLS) in Inspect (15), SLoMo (20), Phosphinator (26), Phos-
phoRS (22), whereas examples of the latter strategy include
Mascot Delta Score (12), the SLIP score in Protein Prospector
(13), and the variable modification localization (VML) score in
Spectrum Mill[Agilent]. PepArML (27) provides experimental
site localization scores in its results using an approach related
to the latter strategy, in which site localization scores are
calculated for each post-translational modification observed
on a peptide by summing together the confidence scores of
its peptide identifications and normalizing by the total Pep-
ArML confidence associated with the peptide (analogous to
how PTM Score converts their probability estimates into site
scores) (N. Edwards, personal communication). Table I sum-
marizes features of some of these tools, and the remaining
sections of this article will contrast these different site local-
ization strategies, compare their performance and discuss
their applicability for data types other than phosphorylation
data acquired in low resolution ion traps, which have been the
main data type assessed so far using these tools.

**Peak Picking**—An often under-appreciated crucial step in
both peptide identification and modification site localization is
deciding which masses in a peak list file to use for analysis.
Within the peak list there will be a mixture of masses that are
fragment ions from the peptide of interest, and other masses
that are produced by chemical or electrical noise. As the noise
peaks are generally of lower intensity, the decision of which

| Software | Peak Picking | Scoring: Peak Probability (PP) or Difference Score (DS) | Score Has Same Meaning for High Mass Accuracy Data? | Score Uses High Mass Accuracy? | Represents Ambiguity? | Reports best scoring site; does not indicate next best alternative site? | Lists all potential sites within score threshold? |
|----------|--------------|--------------------------------------------------------|--------------------------------------------------|--------------------------------|---------------------|---------------------------------------------------------------|-----------------------------------------------|
| A-Score  | n peaks per 100 Th | PP | N | Y | A given score becomes more reliable for higher mass accuracy data | A given score becomes more reliable for higher mass accuracy data |
| PTM Score | n peaks per 100 Th | PP | N | Y | A given score becomes more reliable for higher mass accuracy data | A given score becomes more reliable for higher mass accuracy data |
| Mascot Delta Score | n peaks per 100 Th | PP | N | Y | A given score becomes more reliable for higher mass accuracy data | A given score becomes more reliable for higher mass accuracy data |
| PTM Score | Andromeda | n peaks per 110 Th | DS | N | A given score becomes more reliable for higher mass accuracy data | A given score becomes more reliable for higher mass accuracy data |
| Mascot Delta Score | Andromeda | n peaks per 110 Th | DS | N | A given score becomes more reliable for higher mass accuracy data | A given score becomes more reliable for higher mass accuracy data |
| SLIP Score | Protein Prospector | 20 most intense peaks in each half of observed m/z range | DS | N | A given score becomes more reliable for higher mass accuracy data | A given score becomes more reliable for higher mass accuracy data |
| VML Score | Spectrum Mill | 25 peaks with the highest signal/noise following removal of isotopes, precursor, and noise | DS | N | A given score becomes more reliable for higher mass accuracy data | A given score becomes more reliable for higher mass accuracy data |
peaks to use is normally on the basis of intensity. Most tools employ an intensity threshold, in which only peaks above this are considered, although an intensity-based cross-correlation approach similar to that employed by the search engine Sequest has also been employed (19). Different methods are employed by tools for thresholding. The simplest approach is to employ a universal intensity threshold across the whole spectrum. However, not all fragment ion peaks encode the same amount of information. Sequence ions (for CID, b and y ions) are more informative than internal ions or immonium ions, and ions that are formed by fragmentation nearer the middle of the peptide are more information-rich, in that they are defining the mass of a longer stretch of amino acids. For these reasons, a peak thresholding strategy that ensures peak representation over a wide m/z range can lead to better sensitivity in peptide identification and site localization. Batch-Tag in Protein Prospector splits the observed m/z range in half to create a peaklist for searching with the same number of peaks listed in each half of the m/z range (typically the 20 most intense peaks in each half to create a 40 m/z peak list). A third strategy is to split the spectrum into m/z bins, then use the “n” most intense peaks within each bin for searching. This is the strategy employed by A-Score (100 Th bins), PTM Score (100 Th bins), and Mascot (110 Th bins). This strategy guarantees peak representation throughout the m/z range, but will lead to low intensity peaks being used in “quiet” parts of the spectrum. This is exemplified in Fig. 2, which plots peak lists generated using the 20 + 20 and 4 per 100 Th strategies for a spectrum of the phosphopeptide RGTPhosphoVEGSVQEVQEEK. In each strategy there are a similar number of peaks in the list (40 peaks versus 42 peaks, respectively). However, there are several peaks unique to each approach, including some that correspond to b and y ions of the correct peptide identification. In this example each peak list contains one peak that can be used to localize the phosphorylation to the threonine residue, although it is a different peak in each example; b4 ion in the 20 + 20 peak list and y10 in the 4 per 100 peak list.

The Protein Prospector peak list creation strategy has been compared with the 100 Th binning strategy on the same data set (13). The differences were modest in comparison to a 4 peaks per 100 Th strategy, but the 20 + 20 peak list led to slightly more correct and slightly fewer incorrect site localizations, thereby returning more correct results at a lower FLR. Considering more peaks (5 peaks per 100 Th) produced a measurably higher FLR, mainly because of reporting incorrect site localizations for spectra in which using the other peak list generation strategies the site localizations were deemed ambiguous.

Site Localization Scoring: Probabilities of Random Peak Matching—The first two significant site localization scoring tools, A-Score and PTM Score, use similar approaches to score assignments. In both cases the tools were developed for assessing phosphorylation site identifications in low mass-accuracy ion trap CID data. Both tools treat observed peaks as integer masses, which as this type of data is typically measured with an m/z accuracy of roughly ± 0.5 Th is not an unreasonable step to take. They then make the assumption that if every peak mass is equally likely to be observed at random, then if e.g. four peaks are considered per 100 Th, the chances of randomly matching one of these peaks is 4 in 100. After calculating their probabilities, they both then convert the values into $-10\log_{10}(p)$ scores.

The differences between A-Score and PTM Score are that PTM Score uses this approach to calculate probability scores.
for the peptide identification as a whole with each possible site localization, then converts these scores into probabilities for each potential site localization by making the sum of all of the scores equal a probability of 100% (as the peptide is definitely modified somewhere) and allocating probabilities to each site in the peptide on this normalized probability scale (18).

In contrast, A-Score calculates their probability score only based on matching potential "site-determining" b and y ions; i.e. peaks that would be the same for all site localizations are ignored. Instead of reporting site localization scores for all sites, it only reports a score for the best site localization, which is a difference score between this site localization and the next best possibility (17). In the original publication of A-Score, the authors proposed using a threshold score of 19, which would mathematically correspond to a probability P of ~0.01 or a site being localized with 99% certainty (17). Later phosphoproteomic papers from the authors’ laboratory often use a more liberal threshold (p < 0.05, A-Score > 13) (28). If one tests a range of values into the A-Score probability equation, it is clear that achieving a score >13 requires the best localization to have at least two more site determining ions matched in the MS/MS spectrum regardless of the number of amino acids separating two candidate localization sites.

The accuracy of the probabilities calculated by these methods is dependent on the assumptions used to fit MS/MS data into the binomial probability model. One of the unrealistic assumptions is that all masses are equally likely to be observed at random. Amino acids use a limited range of elements and several only differ by a methyl group from each other, so some masses are in practice much more likely to be observed than others. For example, a peak at mass 201 could be a b2 ion formed by combining the amino acids EA, LS, IS or TV. On the other hand a peak at m/z 206 cannot be formed by any amino acid combination. However, as in each case they are comparing or normalizing to other site localization scores this may mitigate this issue. Nevertheless, instead of treating these probability-based scores as accurate measures of site-localization certainty, they should be considered more simply as just scores subject to a threshold determined by the user to suit individual certainty objectives.

In contrast, the scoring in PhosphoRS attempts to overcome the issue of peak distribution across the mass range of the spectrum by replacing the core probability calculation of N peaks per 100 Th with \((N \times d) / w\), where N is the total number of extracted peaks, d is the specified fragment ion mass tolerance, and w is the full mass range of the MS/MS spectrum (22). Not only does this probability adjustment allow for different regions of an MS/MS spectrum to contain vastly different numbers of peaks with different optimal peak depths for distinct m/z windows, but also directly allows for the use of narrow mass tolerances appropriate for data generated on high resolution instruments.

**Site Localization Scoring: Search Engine Difference Scoring**—A database search engine is going to consider all of the potential site localizations in a peptide when performing peptide identification. Hence, there is information in the search results that can provide a measure of site localization reliability simply by extracting the difference in score/expectation value for peptide identifications with different site localizations. This is the principal that is employed by Mascot Delta Score (12), SLIP scoring in Protein Prospector (13), and variable modification localization scoring in Spectrum Mill (16). The first two both report scores that are the difference in either probability or expectation value scores (the difference will be identical for both measures), reported on a log_{10} scale; i.e. a difference in expectation value of one order of magnitude would score 10; two orders of magnitude would score 20. In Spectrum Mill the identification score is not probabilistic, but instead based on the number of matching peaks, their ion type assignment, and the relative height of unmatched peaks. The score threshold for confident site localization (difference in score between the top two localizations >1.1) corresponds to at least 1 b or y ion located between two candidate sites that has a peak height >10% of the tallest fragment ion (neutral losses of phosphate from the precursor and related ions are excluded from the relative height calculation). There are two significant differences between these tools. Batch-Tag (SLIP scoring) and Spectrum Mill mostly use the same number of peaks per spectrum (typically 40 peaks in Batch-Tag, and typically in Spectrum Mill the 25 peaks with the highest signal/noise following removal of isotopes and noise), whereas the number of peaks used by Mascot (Mascot Delta Score) changes from spectrum to spectrum (and is typically lower than the number used by Batch-Tag), as the search engine varies this value for optimal confidence in peptide identification. Second, SLIP scoring is integrated directly into the Batch-Tag searching as is variable modification localization scoring in Spectrum Mill, whereas the Mascot Delta score is calculated by separate software using the Mascot.dat results file as input. In addition to the convenience of having the scores automatically reported in Protein Prospector and Spectrum Mill, it means that site localization is assessed for all results from the search engine, whereas for other search results software the additional effort involved in transferring results to a second program for extra analysis means that site localization reliability will not be routinely assessed. Most search engines would need to be adapted slightly to be able to calculate site localization difference scores for all results. Mascot stores the top ten search results for each spectrum. However, it is sometimes the case that the same peptide with a different site localization does not score in the top ten results; e.g. if a long string of y ions are observed that are only matched with one site localization. In this situation, no site localization score can be reported. Hence, Batch-Tag and Spectrum Mill were altered so that the scores for the next best site localization are always stored to allow scores for all site localizations to be reported.
Performance of Current Localization Scoring Tools—In 2010 the ABRF-iPRG conducted a collaborative LC-MS/MS data analysis study focused on the evaluation of proteomics laboratories in identifying phosphopeptides and localizing phosphorylation sites (25). The 35 participants were provided with the same phosphoproteomic LC-MS/MS data set generated from the tryptic digest of a lysate of human K562 cells following strong cation exchange (SCX) and immobilized metal affinity chromatography (IMAC) phosphopeptide enrichment. Of 12 fractions collected across the SCX gradient, the data set for the study constituted only fractions 3, 4, and 12. Participants were asked to identify the phosphopeptides present in the sample with ≤ 1% false discovery rate (FDR), and assess the certainty or ambiguity in localizing the site(s) of phosphorylation to particular amino acid residues. Because of the conventional target-decoy measure of FDR (11) the identification portion of the study could be kept on a level playing field. However, because of the absence of a standard metric for false localization rate (FLR) participants could not be given specific guidance on a threshold measure of localization certainty. They were instead requested to make yes/no localization decisions for each peptide spectrum match and provide a description of their scoring mechanisms and thresholds. The results of the 22 participants who assessed site localization were then evaluated based on consensus agreement on localization. Figs. 3A, 3B shows that 79% of the time the participants unanimously agreed on the site of localization (2487 of 3136 PSMs). This reflects only cases in which ambiguity was possible (i.e. number of STY residues in peptide); 119 spectra had unanimous agreement on localization being ambiguous; and with 220 spectra only 1 participant designated the site localization as certain. B, For the 3136 spectra with >2/22 participants willing to designate localization as certain they unanimously agree on localization of phosphosites for 79% of spectra. C, In the remaining 21% of cases (649 spectra), the disagreement was not found to be limited to a single or select few participants. D, The participants who were most likely to make localization decisions (rather than declare ambiguity) were more likely to disagree with the consensus view. The x axis of (C) and (D) is sorted in descending order of # localized/# identified.

**Fig. 3.** Phosphosite localization agreement for SCX IMAC fraction 4 in the 2010 ABRF-iPRG study on identifying phosphopeptides and localizing phosphorylation sites. A, For 3932 spectra the identification was designated as certain by ≥ 2/22 participants. Of these, 3136 spectra had possible ambiguity in site localization that was designated as certain; 457 spectra had no possibility of site localization ambiguity (#phosphosites equal to #STY residues in peptide); 119 spectra had unanimous agreement on localization being ambiguous; and with 220 spectra only 1 participant designated the site localization as certain. B, For the 3136 spectra with >2/22 participants willing to designate localization as certain they unanimously agree on localization of phosphosites for 79% of spectra. C, In the remaining 21% of cases (649 spectra), the disagreement was not found to be limited to a single or select few participants. D, The participants who were most likely to make localization decisions (rather than declare ambiguity) were more likely to disagree with the consensus view. The x axis of (C) and (D) is sorted in descending order of # localized/# identified.
disagree with the consensus view (leftmost participants).

Consequently, much of the divergence from consensus could be attributed to inordinately liberal thresholds, with decisions made based on marginal localization evidence present in the MS/MS spectra. An example can be seen in Fig. 4. Although 79% unanimous agreement is very promising, it also highlights the need for a more general reporting standard.

**False Localization Rate (FLR)**—Comparison of peptide identification results from different software tools is now most widely done by subjecting each tool to the universal metric of false discovery rate (FDR), calculated using a target-decoy database searching strategy. With properly constructed decoys this puts all tools on a level playing field because the frequency of random matching to target peptides and decoy peptides will be the same (11). More detail on measuring the reliability of peptide and protein identification can be found in a recent review by Nesvizhskii (29). Unfortunately, for modification site localization no equivalent reliability measure; i.e. a FLR, can currently be readily calculated. Part of the reason is that peptide identifications with incorrect site localizations are not random matches; they are very similar to the correct answers, so decoy sequences do not provide an error estimate. The current leading candidate FLR approaches all involve computationally allowing modification of amino acid residues that biologically cannot bear the modification (13). The decoy site localization step should ideally be performed independent of the decoy peptide identification, as the inclusion of extra modifiable residues will affect the peptide identification step, producing more false positives or false negatives (depending on where the acceptance threshold is drawn) (13). For a decoy localization based FLR to be accurate, the frequency of the decoy residues would need to be the same as the targets, and the proximity of the decoy residues to the correct site localization would also have to follow a similar pattern to target residues: the closer a residue is to the correct site of modification, the more likely it is to be incorrectly interpreted as being the modification location.

Although the combined codon frequency of the biologically phosphorylatable residues S, T, and Y is ~20%, their frequency in the UniProt human sequence database when requiring tryptic peptides of length 8–40 residues with no missed cleavages allowed is 17.2%. Similar residue frequency in the UniProt human database can be achieved by allowing decoy residue combinations of E, V, and N (17.1%) or D, E, and I (16.5%). However, for individual tryptic peptides the relative frequencies would rarely be matched using such a
constant selection of decoy residues. AA frequencies can be calculated for several databases at http://proteomics.
broadinstitute.org/millhtml/faindexframe.htm by selecting the Calculate statistics utility.

In addition to having a similar compared frequency of 14.1% in the UniProt human database compared with serine and threonine (14.5%), proline (P) and glutamic acid (E) are interesting candidates for decoy amino acid localization scoring because they are each found in the consensus motifs for many kinases (as are other serines and threonines). Hence, they can be expected to provide optimal proximity to correct phosphorylation sites. Both residues produced similar FLR results when used as decoys in one study (13). If both target and decoy residues are tested together (i.e., any S, T, Y, P, E residues allowed to bear phosphorylation in a single peptide spectrum match), then this must lead to more total ambiguous localization decisions in a data set, as the number of modifiable residues is being increased by roughly two-thirds (from 3 to 5 residue types). Considering only one decoy residue type at a time, but performing multiple searches, will reduce this issue. Nevertheless, these types of FLR measurements are likely to overestimate the actual FLR. An alternative would be to separately allow site localization among target residues and decoy residues for a particular peptide spectrum match, and then compare the top two scores.

Although frequency and proximity issues may prevent accurate assessment of the false localization status of an individual peptide spectrum match, counting the number of decoy residue localizations across an entire data set may yield an approximate global FLR, that even if not accurate, may be adequate to serve as a universal measure of differences in algorithm performance and allow classifying the individual score thresholds used with a particular algorithm as more liberal or conservative than others. The absence of an accepted FLR metric is likely to lead to continued confusing and spurious claims to superiority by each localization algorithm developed and ongoing uncertainty as to the reliability of localizations alleged in high-throughput LC-MS/MS PTM reports.

A more sophisticated approach to trying to assign a weighting to the importance of observing a particular site-determining ion has been proposed by Nuno Bandeira and coworkers (personal communication; manuscript in preparation). By determining the intensity of the equivalent backbone cleavage ions in the unmodified spectrum of the same peptide, one can be savorier about whether a given site-determining ion should be expected to be observed. A potential limitation of this approach is the need to observe the unmodified peptide, which is not likely to be the case if a modification enrichment step was employed prior to sample analysis. However, with the rapid population of public spectral libraries (30), if equivalent fragmentation data is deposited in these resources, they could be used to extract peak intensity information. Combining this with decoy residue searching may provide a more accurate FLR measure.

More research in the area of FLR metric calculation is critical to the field.

_Harvesting MS/MS based Modification Localizations from the Literature_—In recent years a few knowledgebase web sites have emerged that harvest the thousands of modification site localizations listed in the supplementary tables of peer-reviewed journal articles on phosphoproteomic studies driven by LC-MS/MS. The most well-developed of these web sites include PhosphoSitePlus (www.phosphosite.org) (31) and Phospho.ELM (phospho.elm.eu.org) (32). These web sites currently rely upon the harvested papers’ authors to have gotten the localization correct. Users of the websites then have limited ability to filter the results by relying on corroborating site observations in multiple papers or excluding high-throughput data altogether. PHOSIDA (www.phosida.
.com) disseminates modification sites identified and localized in publications emerging from research in the laboratory of Matthias Mann (33). As a result, all MS/MS spectra contributing to the identified sites have been analyzed through a common software platform and subject to consistent scoring thresholds. Because published modification localizations generated by LC-MS/MS are harvested by knowledgebases and potential errors hinder downstream use of the information by other researchers, the community would be better off if the researchers who produce and analyze LC-MS/MS data were to err on the conservative side when setting identification/localization thresholds. Consequently, an ambiguous modification localization decision for a particular peptide spectrum match is far preferable to getting it wrong. As more raw LC-MS/MS data from PTM studies is deposited in the public domain, either by journal requirements or voluntary action of authors, it becomes increasingly possible for knowledgebases to undertake efforts to reprocess the data with the most recent algorithms and scoring metrics and enforce uniform quality standards on the information that they disseminate.

_Benchmarking and Comparison of Site Scoring Tools_—As previously discussed, the only reliable way to benchmark the performance of site localization software is through the analysis of a data set in which the answers are known. Such a data set was created by the analysis of 180 synthetic phosphopeptides, where many MS/MS spectra were acquired of each peptide to create a large spectral data set (12). In this study they compared Mascot delta score to A-score for analysis of data acquired on a Q-TOF Micro mass spectrometer that produced low mass accuracy (±0.4 Da) quadrupole-type CID data. In a subsequent study SLIP scoring of Protein Prospector was also evaluated on this same data set (13). The comparisons showed that A-Score was the most conservative in terms of assigning a score to a site localization (it reported a site localization score for 73% of peptide IDs, compared with 85% with Mascot delta scores and 88% with SLIP scores). However, of the sites reported, SLIP scoring had the lowest FLR (6.3%, compared with 8.7% using A-score and 10.9% using Mascot delta score).
It should be noted that these values represent results with any score, whereas in practice researchers will normally employ a score threshold for each of these tools, so the reliability of results will be higher than these values (and the percentage of peptides with site localizations reported will be lower). However, these numbers probably represent a trend in the relative performance of these tools. A-score (and PTM-score) only consider b and y ions, assuming that the modification will be present on fragments. SLIP scoring and Mascot Delta scoring may also consider other ion types (depending on the instrument type specified); e.g. SLIP scoring and Mascot Delta scoring will give some weight to water or phosphate loss peaks. This may be the explanation for why Mascot Delta score and SLIP scoring report site localizations for a higher percentage of phosphopeptide spectra than A-score in the comparison. Reassuringly, at least in the case of SLIP scoring, the use of these extra peaks did not seem to adversely affect the overall reliability of the site localizations.

In a recent paper describing the development of PhosphoRS and benchmarking it against other phosphosite localization programs the authors attempted to put the various programs on the same playing field in examining data generated by various dissociation methods including CID, HCD, and ETD (22). Data sets generated from 179 synthetic phosphopeptides were used to establish the score thresholds necessary to achieve specific FLR values. Those score thresholds were then applied to substantially larger data sets (thousands of peptides) derived from phosphopeptide enrichment of a tryptic digestion of a HeLa cell lysate to show that PhosphoRS outperformed the other programs by localizing as much as –7% more phosphosites. Although one might criticize this comparison by suggesting that the synthetic peptide data set may be too small and may not adequately represent the diversity of sequences in the biological model system, in the absence of a widely accepted metric for calculating an FLR on each experimental data set these authors have clearly attempted to make a fair comparison in a reasonable manner. Furthermore, when reading a paper describing a new scoring algorithm, it would be difficult to evaluate in the absence of some sort of comparison to pre-existing algorithms.

Performance for High Mass Accuracy Data—Higher quality spectra are required to identify modification sites than are necessary for peptide identification. One way to improve the quality of the data is to measure the fragment masses with higher mass accuracy. The higher mass accuracy allows better differentiation of real peaks from noise peaks; i.e. it is less likely that a noise peak will be assigned as a fragment ion. It also allows charge-state determination, which further reduces false positive peak matches.

Clearly, the unit mass resolution model for A-score and PTM-score is less appropriate for this data type. One could increase the number of mass bins within the 100 Th bins and adjust the probability accordingly. This is essentially what Phosphinator does (26). However, the assumption that all mass bins can be observed becomes increasingly flawed as you shrink the size of mass bins. If the granularity of the bins is increased then the significance of a given score will also change dramatically for high resolution data (higher scores will be produced for the same number of peaks identified). If the model of assuming unit mass resolution is not altered then a score may have the same meaning, but you would be throwing away all of the benefits of the mass accuracy for site localization.

The effect on search engine difference scoring is more complicated to predict as search engines may implicitly use mass accuracy in their scoring. If they made perfect use of mass accuracy in their scoring then an e-value difference should have the same measure of reliability for all data types. This topic was partly assessed for the Mascot Delta score by a comparison of searching HCD with ±0.5 Da to ±0.02Da (12). It was found that a lower Mascot Delta score threshold could be employed for the higher mass accuracy search (and that more sites could be reliably identified). In our own hands we have found that SLIP scoring thresholds are a lot more similar for higher and lower mass accuracy data than the Mascot delta score results; the higher mass accuracy data typically leads to larger SLIP scores, as matching a given peak has more statistical significance at higher mass accuracy, but the reliability of a given score appears to be relatively similar.

One thing to note is that with higher mass accuracy data it may be safer to use a lower threshold/larger peak list for searching, as noise peaks are less likely to be an issue in false positive peak matching.

Performance for Different PTMs—The majority of the results that site localization software have been employed upon are phosphopeptide data sets. Is there likely to be a difference in performance of these software when other PTMs are analyzed? As the scores are based on the number of peaks observed that distinguish between sites, then the answer depends on whether the modification creates additional ion types. If a scoring system gives weight to e.g. phosphate loss peaks, then scores could potentially have a slightly different meaning when used for a modification that does not produce loss ions. Unfortunately, it is very difficult to test this out, as there are currently no large data sets of synthetic peptides with PTMs other than phosphorylation.

As previously mentioned, another strategy to create a data set of known modification sites is to filter a PTM data set to peptides that only contain one modifiable residue, then perform searches allowing for modification of decoy residues. This was attempted in-house on an O-GlcNAc modification data set where over six thousand modified spectra were identified (manuscript describing this data set is under review). However, when extracting out only those spectra that contain a single modifiable residue (serine or threonine) there were only 170 spectra, of which six reported decoy residue matches with a SLIP score, which is clearly not enough to derive any meaningful statistics from. A significant factor to
this problem is the sequence preference around sites of O-GlcNAcylation. Although the O-GlcNAc transferase does not have a clearly defined motif, the most common amino acids found on either side of the modification site are other serines and threonines. Only 3.1% of GlcNAc modified peptides in this data set contained a single modifiable residue, whereas when this strategy was employed for a phosphopeptide data set 7.5% of phosphopeptide sequences contained a single serine, threonine or tyrosine. This highlights that the sequence motif of a modification can make a significant difference to the ability to pinpoint modification sites.

So what types of modifications create the greatest challenges? The modifications when there is generally the least information for site localization are: (1) modifications that are very labile in the mass spectrometer, commonly O-linked modifications such as glycosylation and sulfation (although these are less of an issue in ECD and ETD data); (2) modifications that can occur on many common amino acids; e.g. phosphorylation; (3) modifications that can occur on both peptide termini and amino acid side-chains; e.g. acetylation can occur on protein and peptide N termini as well as lysine side-chains: these are particularly problematic in ion trap CID fragmentation spectra because of the missing low mass region. An important point to make about these modification localization tools is that the user indicates which amino acid types can bear a particular modification, so if the modification is on a residue that was not indicated as a possibility; e.g. if a user was looking for lysine methylation, but did not allow for arginine, aspartate, or glutamate methylation, then the results may be unreliable.

**Representing Ambiguity**—For some spectra it is not possible to reliably pinpoint the site of modification and different strategies are employed to report this type of result. In the case of PTM Score and PhosphoRS a confidence measure is reported for each potential site and it is left up to the user to determine what reliability threshold they wish to accept. In the case of SLIP scoring all potential site localizations that are within the score threshold set as acceptable are listed, without indicating which one of these is more likely than other/s. In the case of A-score or Mascot Delta score the most likely localization is labeled with a low score, without indicating what the next best site localization would be.

A unique feature of the Protein Prospector output is that for spectra with ambiguous site localizations if the user clicks on the peptide sequence it will automatically plot the different site localizations onto the same spectrum, allowing visual comparison of the results and highlighting any discriminatory peaks (see Fig. 5). This allows easy assessment of ambiguous spectra, although if user intervention is allowed as part of the acceptance criteria it will introduce inconsistent reliability measures for the reported results.

**CONCLUSIONS**

The availability of modification site scoring software has dramatically improved the consistency and reliability of sites being published. Of the two most common types of strategies
employed for calculating these measures, the search engine-approaches should be better able to make use of mass accuracy and the different fragmentation behavior of different modifications, so should be more sensitive. However, the meaning of a given score using these strategies has more potential to differ from data set to data set. In all cases, as there is currently no easy way to measure a FLR for the results, one has to make the assumption that scores have a consistent meaning when transferred from a standard data set to experimental data.

Hopefully in the near future more standard PTM data sets will become available that will allow for more rigorous benchmarking of site localization tools for other types of data and PTMs. Integration of these tools into the database search engine results is a significant advance, as most of the other tools that have been written to perform this task are using a strategy similar to one described in this article, so software re-implementing a strategy has often been developed for the sake of dealing with different file formats from alternative search engines or fragmentation types. If the proposed standard file format for search engine results (35) becomes more widely used, this will also allow consolidation of tools and will make it easier for all researchers to report reliability measures for their modification results.

Acknowledgments—We would like to thank Paul Rudnick, Katalin Medzihradzsky, and Peter Baker for useful discussions.

* This work was funded by the NIH National Center for Research Resources P41RR001614, the National Institute of General Medical Sciences P41M103481, the National Cancer Institute (U24CA126476 to Steven A. Carr), the National Heart Lung and Blood Institute (HHSN268201000033C to Steven A. Carr), and the Vincent Coates Foundation.

[1] To whom correspondence should be addressed: University of California San Francisco, 600 16th Street, Genentech Hall N-474A, San Francisco, California 94158. Tel.: 415–476-5189; Fax: 415–502-6000; E-mail: chalkley@cgl.ucsf.edu.

REFERENCES

1. Pawson, T., and Scott, J. D. (2005) Protein phosphorylation in signaling-50 years and counting. Trends Biochem. Sci. 30, 286–290
2. Boersma, P. J., Mohammed, S., and Heck, A. J. (2009) Phosphopeptide fragmentation and analysis by mass spectrometry. J. Mass Spectrom. 44, 861–878
3. Grimsrud, P. A., Swaney, D. L., Wenger, C. D., Beaucene, N. A., and Coon, J. J. (2010) Phosphoproteomics for the masses. ACS Chem. Biol. 5, 105–119
4. Hart, G. W., Housley, M. P., and Slawson, C. (2007) Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. Nature 446, 1017–1022
5. Ratnath, P., Dhayalan, A., Ma, H., and Jeltsch, A. (2008) Specificity of protein lysine methyltransferases and methods for detection of lysine methylation of non-histone proteins. Mol. Biol. Synd. 4, 1186–1190
6. Zhao, S., Xu, W., Jiang, W., Wu, Y., Lin, Y., Zhang, T., Yao, J., Zhou, L., Zeng, Y., Li, H., Li, Y., Shi, J., An, W., Hancock, S. M., He, F., Qin, L., Chen, J., Yang, P., Chen, X., Lei, Q., Xiong, Y., and Guan, K. L. (2010) Regulation of cellular metabolism by protein lysine acetylation. Science 327, 1000–1004
7. Kirkpatrick, D. S., Denison, C., and Gygi, S. P. (2006) Weighing in on ubiquitin: the expanding role of mass-spectrometry-based proteomics. Nat. Cell Biol. 7, 750–757
8. Strahl, B. D., and Allis, C. D. (2000) The language of covalent histone modifications. Nature 403, 41–45
9. Zeidan, Q., and Hart, G. W. (2010) The intersections between O-GlcNAcylation and phosphorylation: implications for multiple signaling pathways. J. Cell Sci. 123, 13–22
10. Zhao, Y., and Jensen, O. N. (2009) Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. Proteomics 9, 4632–4641
11. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207–214
12. Savitski, M. M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M., and Kuster, B. (2011) Confident phosphorylation site localization using the Mascot Delta Score. Mol. Cell Proteomics
13. Becker, H., Trinidad, J. C., and Chalkley, R. J. (2011) Modification site localization scoring integrated into a search engine. Mol. Cell Proteomics M111.008078
14. Cox, J., Neuhauser, N., Michiganlsa, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. 10, 1794–1805
15. Albuquerque, C. P., Smolka, M. B., Payne, S. H., Bafna, V., Eng, J., and Zhang, H. (2008) A multidimensional chromatography technology for in-depth phosphoproteome analysis. Mol. Cell Proteomics 7, 1389–1396
16. Spectrum Mill - Agilent Technologies Inc.; Available from: http://www.chem.agilent.com/en-US/Products/software/chromatography/ ms/spectrummillformasshunterworkstation/pages/default.aspx
17. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Biotechnol. 24, 1285–1292
18. Olsen, J. V., Blagoev, B., Gnud, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127, 635–648
19. Ruttenberg, B. E., PLisa, T., Nepper, M. A., and Hoffert, J. D. (2008) PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. J. Proteome Res. 7, 3054–3059
20. Bailey, C. M., Sweet, S. M., Cunningham, D. L., Zeller, M., Heath, J. K., and Cooper, H. J. (2009) SLOMo: automated site localization of modifications from ETD/ECD mass spectra. J. Proteome Res. 8, 1965–1971
21. Swaney, D. L., Wenger, C. D., Thompon, J. A., and Coon, J. J. (2009) Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. Proc. Natl. Acad. Sci. U.S.A. 106, 995–1000
22. Taus, T., Köcher, T., Pashiere, P., Paschoke, C., Schmidt, A., Henrich, C., and Mechler, K. (2011) Universal and confident phosphorylation site localization using phosphoRS. J. Proteome Res. 10, 5354–5362
23. Edwards, N., Wu, X., and Tseng, C. W. (2009) An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra. Clin. Proteomics 5, 23–36
24. Braddishaw, R. A., Burlingame, A. L., Carr, S., and Aebersold, R. (2006) Reporting protein identification data: the next generation of guidelines. Mol. Cell Proteomics 5, 787–798
25. Rudnick, P. A., Askenazi, M., Clausen, K. R., Lane, W. S., Martens, L., McDonald, W. H., Mertins, P., Meyer-Arendt, K., Searle, B. C., and Kowalak, J. A. Proteome Informatics Research Group 2010 Study. Available from: http://www.abrf.org/index.cfm/group/show/ProteomeInformaticsResearchGroup.53.htm
26. Phanstiel, D. H., Brumbaugh, J., Wenger, C. D., Tian, S., Probasco, M. D., McDonald, W. H., Mertins, P., Meyer-Arendt, K., Searle, B. C., and Kowalak, J. A. Proteome Informatics Research Group 2010 Study. Available from: http://www.abrf.org/index.cfm/group/show/ProteomeInformaticsResearchGroup.53.htm
27. PapArlsH. Available from: https://edwardslab.bmbc.georgetown.edu/pymsis/
28. Huttlin, E. L., Jedyrchoski, M. P., Elias, J. E., Goswami, T., Rad, R., Beausoleil, S. A., Villén, J., Haas, W., Sowa, M. E., and Gygi, S. P. (2010) A tissue-specific atlas of mouse protein phosphorylation and expression. Cell 143, 1174–1189
29. Nemighzski, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteome Res. 73, 2092–2123
30. Deutch, E. W. (2011) Tandem mass spectrometry spectral libraries and library searching. Methods Mol. Biol. 696, 225–232
31. Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* 40(Database issue), D261–70

32. Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., and Diella, F. (2011) Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.* 39(Database issue), D261–7

33. Gnad, F., Gunawardena, J., and Mann, M. (2011) PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res.* 39(Database issue), D253–60

34. Chalkley, R. J., Thalhammer, A., Schoepfer, R., and Burlingame, A. L. (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8894–8899

35. Eisenacher, M. (2011) mzIdentML: an open community-built standard format for the results of proteomics spectrum identification algorithms. *Methods Mol Biol.* 696, 161–177