Anal. Chem. **2005,** *77,* 4626−4639

InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra

Stephen Tanner,*,† Hongjun Shu,‡ Ari Frank,§ Ling-Chi Wang,|| Ebrahim Zandi,|| Marc Mumby,‡ Pavel A. Pevzner,§ and Vineet Bafna§

Department of Bioengineering and Computer Science Department, APM 3832, University of California—San Diego, 9500 Gilman Drive, La Jolla, California 92093-0114, Protein Chemistry Laboratory, Alliance for Cellular Signaling, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9196, and Keck School of Medicine, University of Southern California, 1441 Eastlake Avenue, Los Angeles, California 90033

Reliable identification of posttranslational modifications is key to understanding various cellular regulatory processes. We describe a tool, *InsPecT*, to identify posttranslational modifications using tandem mass spectrometry data. InsPecT constructs *database filters* that proved to be very successful in genomics searches. Given an MS/MS spectrum *S* and a database *D*, a database filter selects a small fraction of database *D* that is guaranteed (with high probability) to contain a peptide that produced *S*. InsPecT uses peptide sequence tags as efficient filters that reduce the size of the database by a few orders of magnitude while retaining the correct peptide with very high probability. In addition to filtering, InsPecT also uses novel algorithms for scoring and validating in the presence of modifications, without explicit enumeration of all variants. InsPecT identifies modified peptides with better or equivalent accuracy than other database search tools while being 2 orders of magnitude faster than SEQUEST, and substantially faster than X!TANDEM on complex mixtures. The tool was used to identify a number of novel modifications in different data sets, including many phosphopetides in data provided by Alliance for Cellular Signaling that were missed by other tools.

Fueled by recent improvements in instrumentation as well as software for interpreting MS data, tandem mass spectrometry has become the tool of choice for protein identification. Database search, typified by tools such as Sequest[11] and Mascot,[28] is a popular approach to peptide identification. While the underlying algorithms are effective and used extensively for identification, many spectra remain unidentified by these searches. This can be attributed to several factors including poor quality of fragmentation/ionization and the presence of modifications and mutations that are not explicitly represented in the database.

Another approach to identification is the de novo sequencing approach, in which the peptide is reconstructed solely from the mass spectrum, without the use of a peptide database.[1] The de novo algorithms include Lutefisk[37,38] (a publicly available tool), SHERENGA[8] (part of the Spectrum Mill by Agilent), and Peaks[21] (Bioinformatics Solutions, Inc.), PepNovo[12] (publicly available tool), and approaches by Chen et al.[5] and Bafna and Edwards[4] (see ref 20 for a recent review). These algorithms model the problem by generating *prefix residue mass* (PRM) spectra (scored versions of the MS/MS spectra) and constructing a spectrum graph. Until recently, the application of these algorithms was limited to high-quality spectra or spectra of peptides that were not already in a database. Recent algorithmic improvements, as well as improvement in instrumentation, have led to a resurgence of interest in these tools. Improvements in de novo sequencing notwithstanding, database search algorithms remain the workhorse of peptide identification. The database search approach considers theoretical fragmentation spectra from database peptides and finds one that best matches the input spectra. In a sense, this approach is not that different from de novo sequencing, since the latter can be viewed as a search in the (virtual) database of all peptides. The key in both approaches is a score function that ranks candidate peptides according to their likelihood of generating the spectrum. In the database search scenario, the set of candidates is limited. With incomplete fragmentation and low signal-to-noise ratio, some spectra may not provide enough information to differentiate (de novo) between two or more candidate peptides. In such cases, the database search approach simply selects the one that is present in the database and thereby obtains the correct identification.

Posttranslational modifications level the playing field for the two approaches. Even a single change in the peptide sequence due to posttranslational modifications and mutations shift spectral

* To whom correspondence should be addressed. E-mail:stanner@ucsd.edu.
† Department of Bioengineering,, APM 3832, University of California.
‡ University of Texas Southwestern Medical Center.
§ Computer Science Department, University of California—San Diego.
|| University of Southern California.

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422* (6928), 198−207.
(2) Aho, A. V.; Corasick, M. J. *Commun. ACM* **1975**, *18*, 333−340.
(3) Bafna, V.; Edwards, N. *Bioinformatics* **2001**, *17 (*Suppl. 1), 13-21.
(4) Bafna, V.; Edwards, N. *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology*; 2003; pp 9−18.
(5) Chen, T.; Kao, M. Y.; Tepel, M.; Rush, J.; Church, G. M. *J. Comput. Biol.* **2001**, *8* (3), 325−337.
(6) Craig, R.; Beavis, R. C. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2310−2316.
(7) Creasy, D. M.; Cottrell, J. S. *Proteomics* **2002**, *2* (10), 1426−1434.
(8) Dancík, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6* (3−4), 327−342.
(9) Day, R. M.; Borziak, A.; Gorin, A. In *Proceedings of 2004 IEEE Computational Systems in Bioinformatics (CSB 2004);* 2004; pp 505−508.
(10) DeGnore, J. P.; Qin, J. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 1175−1188.

10.1021/ac050102d CCC: $30.25
Published on Web 06/09/2005

peaks significantly. Enumerating all possible mutations and modifications in the database makes the database prohibitively large but does not seriously affect de novo approaches. Consequently, the current approach to identifying modifications is based upon a virtual "on-the-fly" enumeration of modifications of all candidate peptides. This approach is still computationally expensive, so that a search for modifications remains limited to smaller databases or spectra of particular interest.

The problem is similar to one faced by the genomics community in their search for sequence similarities. To find distant homologues , the alignment scoring had to be more sophisticated (and computationally expensive), thus making the database search too expensive to be a routine tool in the laboratory. The problem was solved using *database filters* that quickly eliminated much of the database, while retaining all true hits (see Figure 1). We argue that filtration is key to identification of modified peptides using database search. At first glance, this is counterintuitive since there is no apparent connection between reducing the number of candidates and identifying modified peptides. Note, however, that aggressive (but accurate) filtration allows us to apply more sophisticated and computationally intensive scoring to the few remaining candidates. Indeed, if the database is reduced to a few peptides, one can afford to consider a rich set of posttranslational modifications (PTMs) for every peptide in the filtered database. Also, the decreased number of candidates reduces the possibility of a high score being achieved by chance.

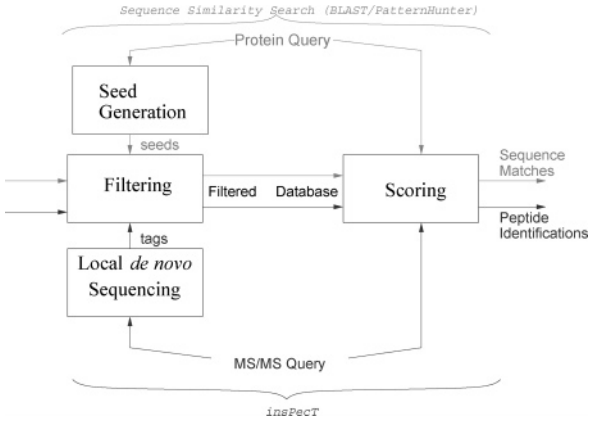


Figure 1. Filtering paradigm for protein identification using tandem mass spectrometry, compared with sequence similarity search. Our new protein identification tool InsPecT uses de novo sequencing to produce text-based filters similar to seeds in BLAST/PatternHunter. Carefully chosen filters dramatically speed up MS/MS database searches and make rigorous PTM searches feasible.

The notion of filtering has been embraced by the mass spectrometry community. Mann and Wilm[23] were early proponents of using sequence tags in a database search, and this idea has been greatly extended by Tabb et al.,[35] Searle et al.[31] Sunyaev et al.,[34] Day et al.,[9] and others. The main idea is that while correlating masses is expensive, sequence-based searches are efficient. This leads to a new paradigm for tandem MS interpretation, as shown in Figure 1. The tandem mass spectrum is interpreted de novo to generate partial sequence information. The database is scanned using this information, and candidate peptides that match sequence tags are then correlated with the spectra. Tabb et al.[35] did not focus extensively on identifying modifications, but pointed out that the tags can be used to find possible candidates with modifications. Sunyaev et al.[34] followed a different path, with an aggressive de novo approach to identify much longer tags, followed by a BLAST-like search to identify peptides related (but not identical) to the peptide that generated the query. Searle et al.[31] performed a breadth-first search to match the candidate peptide to the query spectrum allowing for unanticipated modifications. Their search resulted in candidate peptides that must then be validated. A different approach to filtering has been proposed by Craig and Beavis.[6] They focused on the notion of filtering proteins rather than peptides. In a fast first pass, they scanned the database for tryptic unmodified peptides that match the spectra. The second pass is restricted to only those proteins that matched at least one peptide in the first pass. This approach is based on the assumption that each protein in the sample contains at least one tryptic peptide with at most minor modification. In practice, this leads to a very efficient implementation with only a slight loss in sensitivity.

Indeed, while conceptually simple, the filter-based paradigm has a number of technical difficulties that must be overcome before searching for modifications becomes routine. Much work has been done recently on improved de novo sequencing and tagging algorithms.[13] However, these algorithms have been tested mainly on high-quality spectra of unmodified peptides, and the

(11) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.
(12) Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77*, 964−973.
(13) Frank, A.; Tanner, T.; Pevzner, P. Peptide sequence tags for fast database search in mass-spectrometry. To appear in proceedings of Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB).
(14) Halmer, L.; Gruss, C. *Nucleic Acids Res.* **1996**, *24* (8), 1420−1427.
(15) Hohmann, P. *Mol. Cell. Biochem.* **1983**, *57* (1), 81−92.
(16) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74* (20), 5383−5392.
(17) Keller, A.; Purvine, S.; Nesvizhskii, A.; Stolyar, S.; Goodlett, D. R.; Kolker, E. *OMICS* **2002**, *6* (2), 207−212.
(18) Lagarias, J. C.; Reeds J. A., Wright M. H., Wright, P. E. *SIAM J. Optimization* **1998**, *9*, 112−147.
(19) Lu, B.; Chen, T. *Bioinformatics* **2003**, *19* (Suppl. 2), 113−113.
(20) Lu, B.; Chen, T. *BIOSILICO.* In press.
(21) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337−2342.
(22) MacCoss, M. J.; Wu, C. C.; Yates, J. R. *Anal. Chem.* **2002**, *74* (21), 5593−5599.
(23) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.
(24) Mascot (ms/ms ion search. http://www.matrixscience.com/.
(25) Miller, B. S.; Zandi, E. *J. Biol. Chem.* **2001**, *276* (39), 36320−36326.
(26) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75* (17), 4646−4658.
(27) Nowak, Scott J.; Corces, Victor G. *Trends Genet.* **2004**, *20* (4), 214−220.
(28) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20* (18), 3551−3567.
(29) Sadygov, R. G.; Yates, J. R. *Anal. Chem.* **2003**, *75* (15), 3792−3798.
(30) Schroeder, M. J.; Shabanowitz, J.; Schwartz, J. C.; Hunt, D. F.; Coon, J. J. *Anal. Chem.* **2004**, *76* (13), 3590−3598.
(31) Searle, B. C.; Dasari, S.; Turner, M.; Reddy, A. P.; Choi, D.; Wilmarth, P. A.; McCormack, A. L.; David, L. L.; Nagalla, S. R. *Anal. Chem.* **2004**, *76* (8), 2220−2230.
(32) Shu, H.; Chen, S.; Bi, Q.; Mumby, M.; Brekken, D. L. *Mol. Cell. Proteomics* **2004**, *3*, 279−286.
(33) Shu, H.; Mazouni, F.; Lin, K.; Lyons, K.; Sethuraman, D.; Mumby, M.; Brekken, D. L. In preparation.
(34) Sunyaev, S.; Liska, A. J.; Golod, A.; Shevchenko, A.; Shevchenko, A. *Anal. Chem.* **2003**, *75* (6), 1307−1315.
(35) Tabb, D. L.; Saraf, A.; Yates, J. R. *Anal. Chem.* **2003**, *75* (23), 6415−6421.
(36) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. *Anal. Chem.* **2003**, *75* (5), 1155−1163.
(37) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11* (9), 1067−1075.
(38) Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73* (11), 2594−2604.

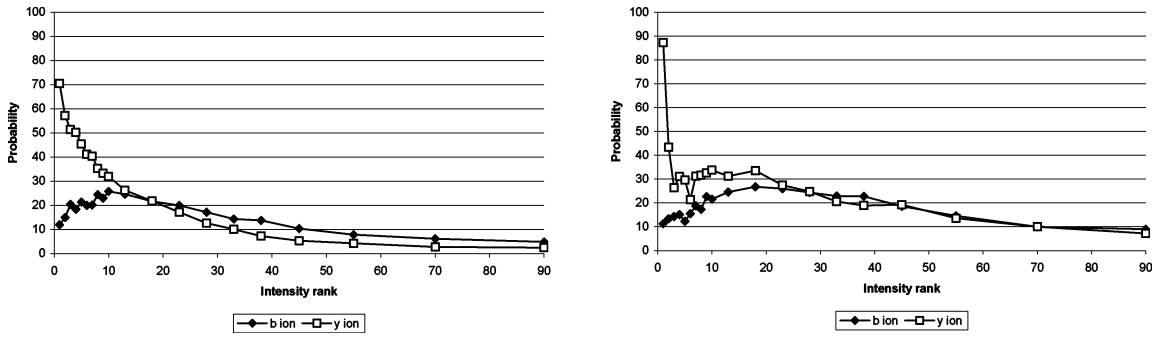


Figure 2. Odds that a peak is a b or y fragment, based on its intensity rank, for peaks in the center sector (left plot) and right sector (right plot) of charge-2 spectra. The most intense sectral peaks are very likely to represent y ions, and these odds drop rapidly as the rank falls. Low-intensity peaks in the center sector are unlikely to be b or y ions, as illustrated by the more rapid tailing off of the curves on the left.

possibility of filtering out a true modified peptide remains real. Next, while sequence tags can be used to efficiently filter a sequence database, in practice, most implementations are slower than regular database searching. Also, after matching the tags, identification of candidate peptides is not trivial. Tabb et al.[35] relaxed the restriction on the flanking masses to include modifications, whereas Searle et al.[31] employed a computationally expensive breadth-first search algorithm to identify candidate peptides. Efficient generation of (modified) peptide candidates remains a challenging part of filter design. Finally, the candidate peptides must be scored against the query spectrum. Much attention has been devoted to scoring[3,7,19,28,29,36,40,41] and on reliability of the peptide assignments.[16,22,26] However, many scoring schemes make the implicit assumption that peptides are of similar lengths. This is reasonable for unmodified peptides as the parent mass restricts the range of peptides that can be scored. However, the presence and absence of multiple modifications can change the parent mass of the candidate peptide, making it necessary to normalize for different sequence lengths. Additionally, the presence and absence of modifications can change fragmentation patterns, and the scoring function must be tailored to accommodate such changes. For all these reasons, reliable and automated identification of modified peptides remains a challenging problem.

In this paper, we describe an automated tool, *InsPecT* (Interpretation of Spectra with PT modifications) that searches large databases for possibly modified peptides by improved algorithms for each of the modules in Figure 1. It has the following features: (a) *tag-based filters* using a novel de novo interpretation algorithm that works in the presence of modifications and poor spectral quality; (b) a fast *trie-based search* for scanning the database with sequence tags; (c) a dynamic programming technique to identify candidate peptides *with modifications* without explicit enumeration of peptides; (d) a scoring algorithm that reflects peptide fragmentation patterns, and a novel quality score based on several complementary features.

Our tool can search for modified and unmodified peptides significantly faster than other database search tools. We identified a large number of modifications in a search of several large data sets. Additionally, our approach is modular, along the lines of Figure 1. As each of these areas is in active development, we anticipate replacing some of the modules with superior algorithms.

METHODS

Our approach to peptide identification follows the scheme in Figure 1. Local de novo sequencing is used to generate sequence tags. These tags are used as PTM-aware text-based filters that reduce a large database to a few peptide candidates. We then carry out probabilistic scoring of these candidates, extending previous approaches[3,8] to both rank the peptide candidates and assess their quality.

Preprocessing. Low-intensity peaks are filtered from spectra in order to remove noise and speed later computations. Each peak is compared with peaks in the neighborhood, centered around itself, with radius 25 Da. Any peak that is not one of the top six peaks in its neighborhood is dropped. Parent mass correction is performed, correcting the parent mass by up to 2 Da to maximize b/y ion pairing. On a typical set of spectra (the SimMod_0 data set, described in Data), this parent mass correction decreased the mean parent mass error from 0.5 to 0.2 Da.

Tag Generation. Tag generation is a modified (but not simpler) version of de novo spectrum interpretation. We construct a directed acyclic graph where each node corresponds to a PRM, and each edge corresponds to a (possibly modified) amino acid. A path in the graph is a possible tag.

Given a peak of mass M for a spectrum of parent mass P, we produce a node at mass $M - |H|$ (for b fragments) and at $P - M$ (for y fragments). Two additional "goalpost" nodes are placed at zero mass and at the parent residue mass. Scoring parameters are based upon odds empirically derived from an annotated set of spectra. The PRM range of a spectrum is divided into three equal sectors, and tagging parameters are derived for each sector and parent ion charge individually. A PRM node's score, $S(N)$, is derived from the following three parameters, scored as log-odds ratios.

1. Intensity Rank. Peaks are ranked from the most intense to the least. Nodes receive a score based on the odds that a peak of a given rank is a b or y ion, shown in Figure 2. These odds differ significantly by sector and charge. For instance, the odds that a rank-3 peak from the first or third sector of a charge-2 spectrum is a b ion are 69 and 14%, respectively. We take the log ratio of these odds with the odds that a randomly chosen peak matches a b or y ion, and obtain score $S_1(N)$.

(39) Tsur, D.; Tanner, S.; Zandi, E.; Bafna, V.; Pevzner, P. A. *Identification of Posttranslational Modifications via Blind Search of Mass-Spectra*. Submitted.
(40) Yates, J. R.; Eng, J. K.; McCormack, A. L. *Anal. Chem.* **1995**, *67* (18), 3202−3210.
(41) Yates, J. R.; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67* (8), 1426−1436.

```
Extend_tag_hit(S,n)
    (* Find a peptide sequence that matches suffix mass S, starting at position n and allowing feasible modifications *)
    p = n
    d = MaxMod (*maximum index of DM*)
    R = 0
    repeat
        while (R > S − DM[d] + ε)
            d = d-1;
            if (d < 0) EXIT;
        if (|S − R − DM[d]| < ε)
            MATCH
        p = p + 1
        R = R + M[p]
    until (d < 0)
```

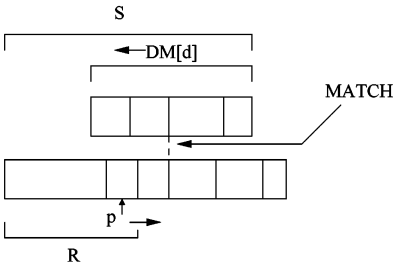


Figure 3. Algorithm to check if the suffix of a tag-hit is a candidate, allowing for modifications.

2. PRM Support. For each PRM, we enumerate the ion types that support the node by checking for peaks at expected masses. The set of ion types identified is a witness set. Each node receives a score based upon empirical probability that its witness set represents a true peak. The optimal (and largest) witness set includes b, y, b − H_2O, y − H_2O, b − NH_3, y − NH_3, a, a − H_2O, a − NH_3, b2 (doubly charged b), and y2 (doubly charged y). Witness set odds are affected by charge and sector. For instance, b2 and y2 ions are more abundant in charge-3 spectra (21% odds of appearing, versus only 8% in charge-2 spectra). We compare witness set odds with the odds that a randomly chosen peak matches a b or y ion; the log-odds ratio is $S_2(N)$.

3. Isotope Pattern. For each atomic mass up to 1750, we compute the expected relative intensity of the first heavy isotopic peak based upon isotope frequencies of atoms C, H, O, N, and S. Each spectral peak is classified as a primary isotopic peak (with "child" peak at +1 Da), secondary isotopic peak (with "parent" peak at −1 Da), or lone peak. Using secondary peaks in a path is undesirable, particularly since there are pairs of amino acids whose masses differ by roughly 1 Da. For example, a peak classified as a primary isotopic peak is over 5 times more likely to be a y ion than is a peak classified as a secondary isotopic peak. The log-odds ratio for this feature gives us score $S_3(N)$.

We connect two nodes of mass m_1 and m_2 with an edge if the distance ($m_2 - m_1$) is within ϵ (by default, $\epsilon = 0.5$) of a legal jump. Legal jumps include all amino acid masses and (if PTMs are allowed) all singly modified amino acids. We assign a score, $S(E)$, to each such edge based upon its skew (the difference between the edge length and the expected mass). In addition, edges containing PTMs receive a penalty derived from the expected frequency of that PTM.

Tags are scored additively. Given a path connecting nodes $N_0 \ldots N_n$ via edges $E_0 \ldots E_n - 1$, with node scores $S(N_k)$ and edge scores $S(E_k)$, assign the tag a score of $\sum_{k=1}^{n} (S_1(N_k) + S_2(N_k) + S_3(N_k)) + \sum_{k=1}^{n-1} S(E_k)$. The three score factors S_1, S_2, and S_3 are listed in order of significance; they have Fisher criterion scores of 1.63, 1.03, and 0.37, respectively.

Currently, we generate tags of length 3 and retain up to 50−100 top-scoring tags for subsequent database searching. While this is effective, we have found that the top 10−25 tags already show high sensitivity (see Results). We plan to reduce the final number of tags using additional features and increase the length of the tags in future work.[13]

Database Search and Candidate Generation. We search the database for sequence tags using a trie-based data structure.[2] After a rapid construction of the trie automaton, all matches to any tag in the trie can be computed in a single scan of the database, *independent* of the number of tags in the automaton. We note that this is one of many ways to search for tags. For instance, the database could be indexed by precomputing the locations of each tag. In practice, however, this approach has problems. The index must be recomputed each time the database or parent mass tolerance is changed, and the index size grows explosively with the number of PTMs considered. With the automaton, we must scan the entire database at least once. However, the cost is easily amortized by combining tags from multiple spectra into a single automaton. Each node of our trie contains zero or more tags; leaf nodes all contain at least one tag. We preprocess the database to facilitate fast scanning and are working to distribute processing to further speed the search. InsPecT scans the database with the trie, looking for any peptide that matches a tag.

When such a peptide is found (an initial match, or tag hit), we attempt to extend it to a full match by finding flanking sequences which match the tag's prefix and suffix masses, allowing for potential modifications. Different combinations of allowed modifications generate many different molecular masses. To search efficiently, we use the following approach: Define a *decoration* to be the mass of a possible set of posttranslational modifications. For instance, if our search allows up to two phosphorylations and two methylations, we generate a total of 9 decorations (including the "empty decoration"), [0, 14 (1 methylation, 0 phosphorylation), 28, ...]. We order the decorations according to increasing mass and generate an array *DM* to store the mass values. This is done once, at the beginning of the search. Consider a tag with suffix value *S*. The tag matches a peptide in the database with decoration *d* if the peptide contains the tag, and has a suffix of mass *R*, such that $|S - (R + DM[d])| < \epsilon$. Figure 3 describes an efficent algorithm to search for candidate peptides that takes time linear in the size of the array *DM*.

Prefix extension is performed similarly. Successful extensions are filtered to remove unfeasible decorations (for instance, oxidation requires the presence of a methionine). Next, given a candidate peptide, the optimal attachment positions of PTMs on the flanking regions are determined by dynamic programming. This technique allows us to determine the optimal attachment

positions without explicit enumeration and scoring of all possibilities. We consider a set of attachment sites to be optimal if the resulting set of PRM nodes has the highest possible score, as assigned during tag generation.

We select attachment sites by computing $S[d, j]$, which is the optimal score attainable by attaching decoration d to the first j residues $r_1, ..., r_j$ of the candidate peptide. For an amino acid r, let $D(r)$ represent the set of all possible decorations that can be applied to that residue. Let $M[d, j]$ be the PRM created by attaching decoration d to the first j residues. Then, $S[d, j]$ is computed by the following recurrence

$$S[d, j] = \text{NodeScore}(M[d, j]) + \max_{d' \le d, d' \in (r_j)} \{S[d - d', j - 1]\} \quad (1)$$

The corresponding dynamic program is used to compute the optimal attachment points. NodeScore($M[d, j]$) equals the score $S(N_k)$ of the PRM node at that mass or a penalty if no PRM node was generated nearby. Multiple PTMs at one amino acid—such as double oxidation of a methionine residue—are permitted where chemically reasonable. Note that $D(r)$ remains a small set. The peptide, with appropriately attached modifications, forms a candidate peptide that must now be scored.

Candidate Peptide Scoring. The candidate peptides must be scored and ranked according to their relative likelihoods of generating the mass spectrum. Clearly, scoring is the mainstay of all database search algorithms and the focus of intensive research.[3,7,19,28,29,36,40,41] When posttranslational modifications are considered, the score function must be modified. A key issue here is normalization for peptides of different lengths. In the absence of modifications, all candidate peptides have identical mass and, therefore, a small variation in peptide length. When putative modifications are considered, the variation in length can be considerable, particularly if multiple modifications must be considered. As longer peptides have more putative fragments that can match, as well as more fragments that might be missing, certain score functions (such as the total ion current explained by the peptide) have a peptide length bias. Here, we describe a score function that extends earlier approaches[3,8] and also normalizes for peptide length. Dynamic programming is used to compute the score.

We assume a probabilistic model in which a peptide generates fragments that show up as spectral peaks. Peaks can possess skew in m/z due to instrument error and have an intensity proportional to the likelihood of the fragment ion being seen. We also consider a null model, in which the same number of spectral peaks are generated at random m/z values. Consider a spectrum $\mathcal{S}' = S_1, S_2, ..., S_k$ with the k spectral peaks sorted by mass/charge ratio. For a candidate peptide P, let $\mathcal{T}(P) = f_1, f_2, ..., f_n$ denote the candidate fragments (theoretical peaks) sorted by mass/charge ratio. We will define a probability $Pr(\mathcal{S}|\mathcal{T}(P))$, that the spectrum S is generated by the peptide P. We compare this to $Pr(\mathcal{S}|\ n)$, the probability that the spectrum was generated from the null model. The score for the candidate peptide is the log-odds function

$$\log\left(\frac{Pr(\mathcal{S}|\mathcal{T}(P))}{Pr(\mathcal{S}|\ n)}\right) \quad (2)$$

To compute $Pr(\mathcal{S}|\mathcal{T}(P))$, we use a mapping between theoretical ions and spectral peaks. The (hardware-dependent) probability $Pr(f_j|\ P)$ that a theoretical ion f_j was generated by a peptide P is computed based on empirical observations of peptide composition. For example, the probability of seeing a neutral H_2O loss is higher in the presence of acidic residues. The probability of not seeing a fragment ion f_j, $Pr(\bar{f}_j|\ P)$, is simply $1 - Pr(f_j|\ P)$. Ion probabilities are computed separately for each sector, reflecting the richer sets of peaks in the center of typical spectra. The probability of matching S_i with f_j also depends on the difference in m/z values (skew), and is modeled by $Pr(S_i \simeq f_j)$. In the null model, the probability $Pr(f_j|\ n)$ that a randomly generated peak matches fragment f_j is simply a function of the number of peaks in its sector and the mass tolerance allowed. Similarly, $Pr(\bar{f}_j|\ n)$ equals $1 - Pr(f_j|\ n)$.

An unassigned spectral peak S_i is assumed to be noise and is generated with a probability $Pr(S_i \simeq \phi)$, which depends on the intensity of the peak. In the null model, the probability $Pr(N|\ n)$ that a peak is noise (unaccounted for by the n "true" nodes) depends somewhat on peak count, but is generally near 1.

Since many assignments of peaks to ions are possible, we choose one that maximizes the log-odds score. This can be done using a dynamic programming computation under the reasonable assumption that assigned peaks do not *cross*. In other words, if peaks S_{i1}, S_{i2}, with $S_{i1} < S_{i2}$ are assigned to f_{j1}, f_{j2}, then $f_{j1} \le f_{j2}$. Let $\mathcal{S}_i$ represent the first i spectral peaks $S_1, ..., S_i$, and $\mathcal{T}_j(P)$ denote the first j theoretical fragments. Denote

$$\psi(i, j) = \log\left(\frac{Pr(\mathcal{S}_i|\mathcal{T}_j(P))}{Pr(\mathcal{S}_i|\ n}\right)$$

Clearly, it is sufficient to compute $\psi(i, j)$ for all i, j. The computation is given by the recurrence

$$\psi(i, j) = \max \left\{ \begin{array}{ll} \psi(i-1, j) + \log\left(\frac{\Pr(S_i \simeq \sigma)}{\Pr(N|n)}\right) & \text{peak } i \text{ is noise} \\ \psi(i-1, j-1) + \log\left(\frac{\Pr(f_j|P)}{\Pr(f_j|n)}\right) + \log(\Pr(S_i \simeq f_j)) & \text{peak } i \text{ is matched with fragment } j \\ \psi(i, j-1) + \log\left(\frac{\Pr(\bar{f}_j|P)}{\Pr(\bar{f}_j|n)}\right) & f_j \text{ is not generated} \end{array} \right\}$$

This algorithm returns the score of the optimal assignment, as well as the mapping from theoretical fragments to spectral peaks. The main advantage of this model is that the fragmentation probabilities can be recomputed for different instrumentation types, and the effects of PTMs on fragmentation (as in phosphopeptides) can be explicitly modeled. The main disadvantage is that we do not explicitly handle dependencies between different fragments. Therefore, we refine the initial interpretation, primarily by removing unacceptable ion interpretations, where secondary ions are present but a primary ion is not. For instance, no theoretical peak of ion type b $- H_2O$ can be assigned to a spectral peak unless the corresponding b peak has been assigned. Any such spectral peaks are reinterpreted as noise. A second refinement step rescues peaks that are interpretable as isotopes of other (non-noise) peaks. Finally, depending on search options, a small bonus is added for matching protease digestion rules (no missed cleavage, and the ends match the protease digestion specificity).

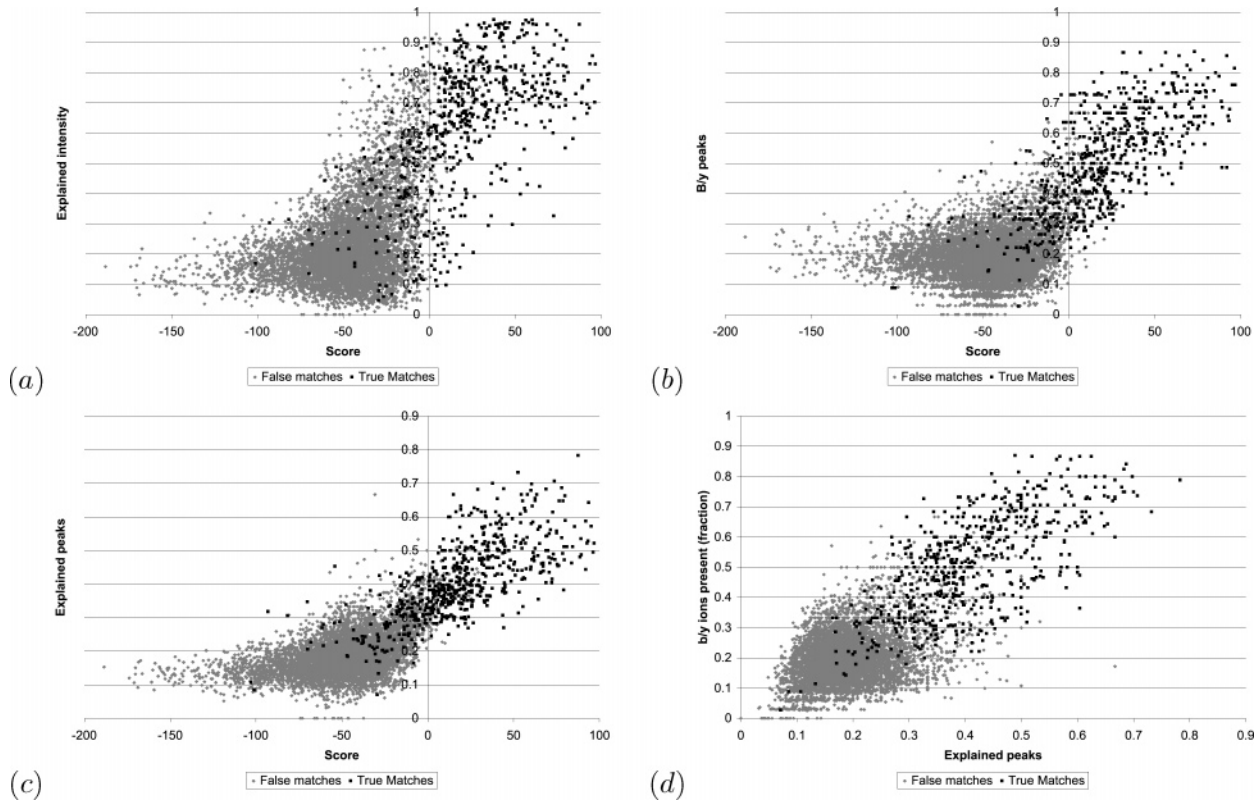


Figure 4. Other score features that complement the candidate score in distinguishing a correct peptide from a false one. (a) Explained intensity versus score. (b) b/y score versus score. (c) Explained peaks versus score (d) b/y score versus explained peaks. The measures all help distinguish correct top matches from incorrect top matches and complement each other.

In ongoing work, we plan to replace these refinements with more sophisticated scoring algorithms that include fragment dependencies.[12]

***P*-Value Computation.** The scoring procedure outlined above ranks the candidate peptides. This is not sufficient as the top-scoring peptide might not be the correct one even if the rank order of peptides is correct, if the true peptide is not in the database. This is a common issue with database search algorithms. Different criteria, both automated (such as the Sequest Xcorr score) and manual, can be applied to separate true positives from high-scoring spurious matches. Based on optimization of several features, we devised a match quality score to estimate the probability that the top match is correct, and not just the best of various poor alternatives. Besides the *candidate score*, we considered several features: *explained intensity* (*I*), *explained peaks* (*P*), *δ-score*, and *b/y ion score* (*B*). The explained intensity of a candidate is the fraction of total ion current belonging to annotated spectral peaks. Similarly, the explained peak score is the fraction of the (filtered) peaks that are annotated. The *δ*-score is the difference in score between a candidate peptide and the next-best candidate. The b/y ion score is the fraction of b and y ions found in the spectrum. As shown in Figure 4, these measures help distinguish true peptides and complement the score computation. The *δ*-score is affected by database size (a small database will give artificially large *δ*-score values to poor matches). Therefore, we do not consider it in computing *p*-value but report it separately. In the future, we plan to incorporate it into a more sophisticated confidence calculation calibrated to an MS run.[26]

We compute an optimal linear combination of the remaining four factors, to

$$Q = w_1 S + w_2 I + w_3 P + w_4 B \qquad (3)$$

obtain a quality score of a match as The scores *S*, *I*, *P*, and *B* are normalized by subtracting the mean and dividing by standard deviation. The weights w_i have been chosen to optimally discriminate correct matches from incorrect matches. Let *T* be the set of true matches whose score *Q* is below 1, and let *F* be the set of false matches whose score *Q* is above −1. (For a perfect scoring scheme, *T* and *F* would both be empty sets.) We used an objective function, which penalizes these poorly resolved cases:

$$\sum_{m \in T} B(1 - Q_m)^2 + \sum_{m \in F} (1 + Q_m)^2$$

The constant *B* is the ratio of total incorrect to correct matches, used so that false negatives and false positives will carry the same weight. Assignment of weights was carried out using the Nelder–Mead simplex method.[18] Weights were tuned on a training set consisting of 612 correct and 6 471 incorrect (but top-scoring) matches on the ISB data set and tested on a separate data set containing 590 correct and 6 373 incorrect spectra. Repeated runs from different starting conditions produced the same optimized weights, indicating that the optimum is global. The optimized weights indicate that success in finding b and y ion series, and annotating many of the peaks of the spectrum, is more indicative

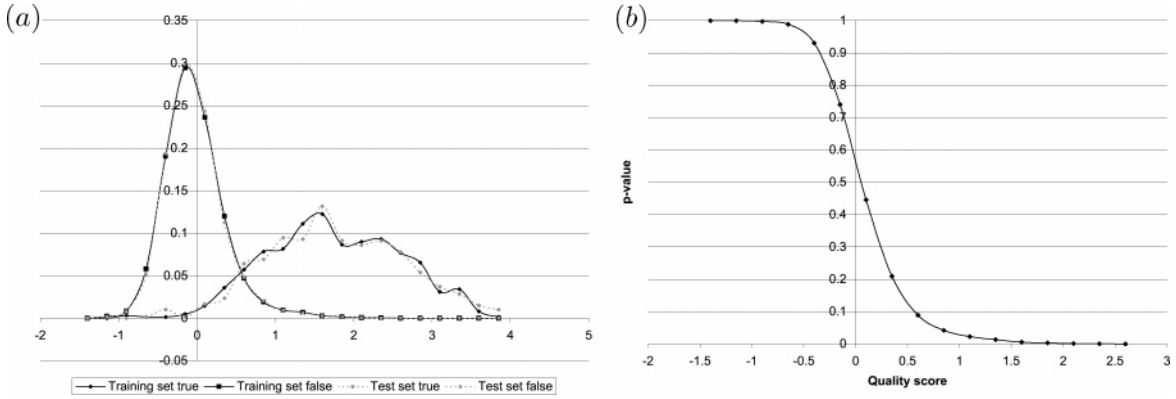


Figure 5. (a) Histogram of match quality scores, separated into correct and incorrect matches, showing the performance of quality scores in separating true matches from false matches. (b) Empirically generated *p*-values. False matches rarely attain quality scores, and so higher quality scores give low *p*-values.

of a true match than explaining a large fraction of the total ion current. The *p*-value for a candidate peptide is computed by comparing the match quality score to the distribution of quality scores for incorrect matches. Figure 5a shows the performance of the optimal quality score in separating between the two sets for both training and test data. We use the distribution of quality scores on the incorrect set as a *p*-value (Figure 5b). This *p*-value was tested on the second set of spectra, and sensitivity and specificity were as good. In the future, we plan to reconsider the *p*-value computations when searching (as is usual) a full run of spectra as a batch.

RESULTS

Data. We used four different data sets. The first two, ISB and SimMod, are control data sets on a limited number of proteins and were used to validate InsPecT's performance and compare it to other tools. The Mus-IMAC data set represents a complex mixture enriched for phosphopeptides, but also other modifications. Finally, the IKKb data set has a smaller set of proteins but contains peptides digested with multiple enzymes and has a richer set of modifications.

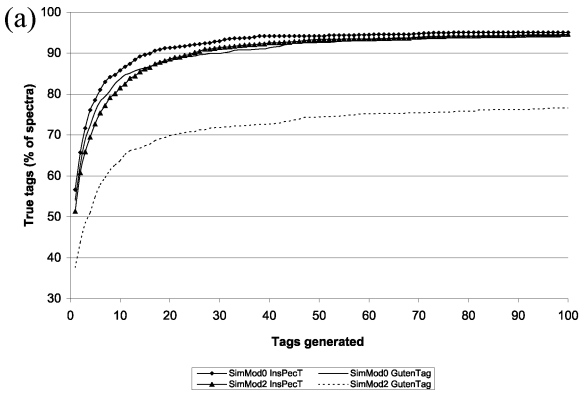
ISB: The ISB data set[17] is a well-known collection of MS/MS spectra from 22 separate LC−MS runs on a ThermoFinnigan ESI-ITMS. Two mixtures were prepared by combining a set of purified proteins. This list of pure proteins, together with known contaminants such as keratin, constitutes a *valid* set. We created an *ISB-search* database containing 93 000 human proteins from the nr database, together with the valid proteins, for a total size of 31 Mb. As a first approximation, an identification is considered correct if the top match in the search database is a valid protein, and incorrect otherwise. Also, we consider a peptide match to be correct if it substitutes *I* for *L* or *Q* for *K* (or vice versa).

SimMod: We selected the 1000 annotated, doubly charged spectra from the ISB data set with greatest total intensity. We constructed a data set of modified spectra by adding feasible modifications to each peptide and shifting the spectral peaks appropriately. Given a theoretical peak at mass m_1 that shifts in mass by δ, we shift all peaks in the neighborhood of m_1 by δ. We also swap peaks in a neighborhood of $m_1 + \delta$ left by $-\delta$; carrying out swaps rather than shifts avoids introducing gaps in the spectrum. Swaps of major ion types (b and y) take precedence over swaps of secondary ion types (neutral losses). SimMod_i refers to the data set with i modifications randomly selected from the set of feasible modifications. Thus SimMod_0 is the original data set, while SimMod_1 and SimMod_2 represent spectra with one and two modifications. In the absence of curated data sets with real modifications, the SimMod data sets are useful in testing methods, allowing the measurement of the false negative rate. The set of allowable modifications was hydroxylation of proline or lysine, sulfation of tyrosine, and oxidation of methionine. The SimMod data set was searched against the same *ISB-search* database used for the standard ISB data set. (We note that the performance on simulated data sets has its caveats and is not a generally accepted aproach for MS data. However, by simulating modifications, we can test for *false negatives*, which is otherwise a very difficult task. Such simulated data sets are often used in Bioinformatics, and they measure performance on specific facets of data, thereby enabling improved tool development.)

Mus-IMAC Data Set: The Mus-IMAC data set represents partially annotated spectra of peptides from the Protein Chemistry Laboratory of the Alliance for Cellular Signaling. Murine RAW 264.7 cells were treated with the serine/threonine phosphatase inhibitor calyculin-A and mass spectra of enriched phosphopeptides obtained as described previously.[32] Proteins were extracted using the Tripure reagent (Roche Applied Science), digested with trypsin, and subjected to immobilized metal affinity chromatography (IMAC) to enrich for phosphopeptides. The enriched samples were analyzed by LC−MS/MS using a ThermoFinnigan LCQ Deca mass spectrometer. The data set consists of three mass spectrometry runs (14 061 spectra).[33] An *nr-mus* database was created by extracting proteins that matched the keywords "mouse", "mus musculus", or "m. musculus" from the current NCBI nonredundant database. The nr-mus database contained 78 500 proteins (30 Mb).

IKKb: GST-tagged human IKKb was expressed in yeast and in *Escherichia coli* as described previously[25] and purified on glutathione Sepharose (E. Zandi and T. Higashimoto, unpublished data). To produce overlapping peptides, SDS−PAGE-purified IKKb was digested with different combinations of trypsin, elastase, or Glu-C in the presence or absence of 10−15% 2-propanol. The digested peptides were analyzed by LC−MS/MS using a Thermo Finigan LTQ mass spectrometer. A total of 45 500 spectra were acquired from this peptide mixture at USC Medical School.

(a)



(b)

Tags	Tagging Accuracy (%)				
	1	10	20	50	100
SimMod0 Inspect	56.6	85.8	91.3	94.2	95.1
SimMod0 GutenTag	54.0	83.7	88.1	92.6	94.3
SimMod1 Inspect	52.0	83.5	89.1	93.1	94.8
SimMod1 GutenTag	48.3	76.6	81.9	87.4	89.8
SimMod2 Inspect	51.4	81.5	88.5	93.4	94.6
SimMod2 GutenTag	37.6	63.7	69.9	74.4	76.6

Figure 6. Tagging performance of InsPecT. The tagging accuracy is defined as the percentage of spectra with at least one correct tag in the top *n* predictions. (a) Comparison with GutenTag. (b) Performance across the SimMod data set. Adding modifications has a minor effect on accuracy.

Table 1. (a) InsPecT Search Speed in the Presence of Posttranslational Modifications and (b) InsPecT Peptide Identification Accuracy in the Presence of Posttranslational Modifications

(a) InsPecT Search Speed[a]

	speed (s)			
database size (MB)	tagging only	0.66	5.7	5.4
SimMod0 GutenTag	0.85	1.54	3.00	n/a
SimMod0 Inspect	0.01	0.02	0.10	0.84
SimMod1 Inspect	0.01	0.03	0.15	1.41
SimMod2 Inspect	0.01	0.04	0.21	2.05

(b) InsPecT Peptide Identification Accuracy[b]

	accuracy	
	top candidate	top 10
SimMod0 GutenTag	90.0	93.8
SimMod0 Inspect	94.2	94.6
SimMod1 Inspect	92.2	93.3
SimMod2 Inspect	88.3	88.9

[a] Trie-based searching allows the identification of many spectra in one pass through the database. [b] The accuracy results are reported for the 5.7-Mb database.

Spectra from this set were searched against a database containing 16 valid proteins (IKKb, GST, several keratins, and several proteases) and 2000 invalid proteins (randomly selected from Swiss-Prot).

Hardware: Timing statistics were gathered on a desktop PC with 2.8-GHz Intel processor with 1GB of RAM. The operating system was RedHat Linux, kernel version 2.6.5.

Performance on Control Data Sets: Tagging, Speed, and Accuracy. The tagging accuracy of InsPecT was tested on the SimMod data sets and is shown in Figure 6. All results are for tags of length 3, along with flanking masses that are correct to a prespecified tolerance. Figure 6a plots the fraction of spectra with a correct tag against the number of tags generated. The data are tabulated in Figure 6b. As the number of tags predicted goes from 1 to 100, the fraction of spectra tagged correctly rises to ~95%. A manual examination reveals that many of the remaining 5% of spectra for which tags are missed are of low quality. Note that these tests are on spectra in which the parent mass can be off by up to 2 Da, and fragment ions may be off by 0.5Da. With improvements in instrumentation w.r.t fragmentation, and mass accuracy, we should be able to search with longer and fewer tags, greatly improving the efficiency.

The ability of InsPecT to incorporate modifications into a tripeptide tag is important for high accuracy. GutenTag does not generate such tags. When modifications are restricted to the N- and C-termini of a peptide, so that more unmodified valid tags are attainable, GutenTag's performance is significantly better. For peptides with one PTM on the N- or C-terminal residue, GutenTag's accuracy for 100 tags is 90% (as compared to 82% when PTMs are not restricted to the ends).

Next, we tested the speed and accuracy of InsPecT (combined tagging, filtering, and scoring) in identifying peptides in the SimMod data sets. The results are also compared against GutenTag (run in database scan mode). See Table 1. The increased running time is primarily due to the many candidate peptides found in the virtual database. The time for tag generation is independent of the size of database and the number of allowed modifications. Thus, it accounts for 16% of the total running time when searching with 0 modifications but only 4% of the total running time for two modifications. The increase in running time is primarily due to the larger number of candidate peptides.

Through extensive empirical tests, we find that the number of candidate peptides (and therefore, the running time) increases as a linear function of each of the following: number of tags allowed, number of possible modifications, and size of the database. As the number of tags increases from 10 to 100, the search time (with 0 modifications) on a 54-Mb database increases from 0.1 to 0.87 s/spectrum. Correspondingly, the time to search (25 tags, 54-Mb database) for a list of 0, 5, and 10 possible modifications is 0.38, 2.65, and 6.3 s/spectrum, respectively. We note that these times are independent of enzyme specificity rules, and we do not restrict peptides to be tryptic. Even with a modest list of two modifications, allowing for nontryptic peptides, InsPecT is ~2 orders of magnitude faster than Sequest (see Mus-IMAC results). We also compared its speed against X!Tandem. Recall that X!Tandem uses a clever two pass filter for proteins. In the

Table 2. Most Prevalent Posttranslational Modifications over the ISB Data set[a]

protein	peptide	modified	unmodified
PPB_ECOLI	A.RTPEM+16PVLENR.A	23	25
CAH2_BOVIN	K.DFPIAN+1GER.Q	21	30
TRFE_BOVIN	R.AAANFFSASCVPC-2ADQSSFPK.L	21	0
CAH2_BOVIN	K.YGDFGTAAQQPD-18GLAVVGVFLK.V	20	70
CASB_BOVIN	A.RELEELN+1VPGEIVESLSSSEESITR.I	14	15
TRFE_BOVIN	K.SVTDCTSNFC-2LFQSNSK.D	14	0
LACB_BOVIN	K.VLVLD+22TDYKK.Y	14	41
CAH2_BOVIN	K.YAAELHLVHWN+1TK.Y	14	9
CAH2_BOVIN	K.EPISVSSQQM+16LK.F	13	35
LACB_BOVIN	K.VAGTWYSLAMAASD+22ISLLDAQSAPLR.V	13	23

[a] The number of spectra with the modification (modified) and without (unmodified) is indicated.

first pass, a subset of proteins is selected with at least one tryptic, unmodified peptide to the spectra. The search for modifications and semitryptic peptides is limited to these protein sequences in the second pass. The ISB data set, with a small mixture of true proteins, largely unmodified, is an ideal data set. Correspondingly, X!Tandem is fast, requiring ~0.12 s/spectrum, comparable to an InsPecT search with only 10 tags. The running time of X!Tandem increases with the complexity of the mixture, as the number of proteins in the second pass increases. On the Mus-IMAC data set (where InsPecT identified 280 proteins), the running time for X!Tandem is ~4.4 s/spectrum against a 30-Mb database. In contrast, InsPecT (with 25 tags) takes ~0.8 s/spectrum and identifies phosphopeptides that are missed by X!Tandem (see Mus-IMAC results). For low-complexity mixtures, the idea for a two-pass protein filtration is certainly appealing, and future versions of InsPecT will include this as an option.

Accuracy was tested by searching the spectra against a large database. A search is considered correct if the top hit is a valid protein, and incorrect otherwise. As expected, the accuracy falls upon increasing allowed modifications, mainly because of a failure to generate correct tags. However, as Table 1 shows, InsPecT shows good speed and accuracy with increasing numbers of modifications and increasing database size. GutenTag was not designed to produce candidate peptides with PTMs, so we report its database results only for SimMod_0. Note that the increase in sensitivity is very modest in going from 25 to 100 tags. Nevertheless, for maximum sensitivity, we keep the default size at 100 tags in subsequent experiments, as the overall running time is not too large.

Sensitivity on Controlled Data Sets (ISB). We searched the set of ISB spectra against the database using InsPecT. The data set has been interrogated with a number of tools and can be used to measure InsPecT's sensitivity. The posttranslational modifications permitted were disulfide linkage ($C - 2$), methionine oxidation ($M + 16$), dehydration ($D - 18$, $E - 18$, $T - 18$), deamidation of asparagine ($N + 1$), DOPA ($Y + 16$), thioproline conversion of N-terminal cysteines ($C + 12$), and incorporation of sodium ($D + 22$, $E + 22$).

The top match was valid for a total of 3215 spectra, as compared with 2756 originally found by SEQUEST. A total of 947 new annotations were gained, and 488 annotations were lost. The average search time was 0.9 s/spectrum. Many of the differences from SEQUEST results are explained by the greater emphasis our scoring scheme places on peak intensity. Tag-based filters are robust against low-intensity noise but have difficulty with low-quality spectra where the b and y peaks have low intensity relative to noise. The average explained intensity for the new annotations is 55%, while for the lost SEQUEST annotations, it is only 32%. Those spectra where InsPecT and SEQUEST were both able to annotate are particularly high quality, with average explained intensity of 64%. Of the spectra correctly identified by SEQUEST, the majority were missed due to errors in tagging that filtered out the correct peptide. Refinements to our tagging algorithm may decrease the number of such cases by improving filter sensitivity. We have found that performing a (more sensitive) second search pass against the proteins identified in the first pass can also boost the number of spectra annotated (data not shown).

A large number of spectra (530) contained a posttranslational modification. Table 2 shows the most commonly modified residues in this sample. Not surprisingly, there are a large number of oxidized methionines, but other modifications are also prevalent. Annotation of spectra containing cysteine is of interest on this data set, since almost none of the validated Sequest annotations contain cysteine residues. In contrast, a total of 80 spectra containing cysteine were annotated by InsPecT. We also found 14 peptides containing a −2 modification on cysteine, with 0 unmodified cases. Without carboxamidomethyl protection, cysteines from the sample can form disulfide bonds. Those peptides that contain multiple cysteines preferentially form intramolecular disulfide bonds (albeit not necessarily the disulfide linkages found in the native protein), which show up as a −2 modification. See Figure 7 for an example. In addition, N-terminal cysteines may be converted to thioproline, protecting them from promiscuous disulfide bond formation.

The average number of initial tag hits per spectrum was 263 000. The number of candidate peptides per spectrum (produced by successful tag extension) was 700. Nontryptic peptides were not excluded from scoring (and indeed, many nontryptic matches were found). Given our initial database size of 31 Mb, and the fact that peptides of various lengths are considered, the filtration efficiency due to tags is roughly 2×10^{-6}.

Most PTMs in this data set likely represent chemical damage to the sample processing rather than regulation in vivo. However, many peptide identifications are not possible without considering PTMs. Additionally, detection of chemical damage is important in MS-based quantification, where the relative intensity of peptides

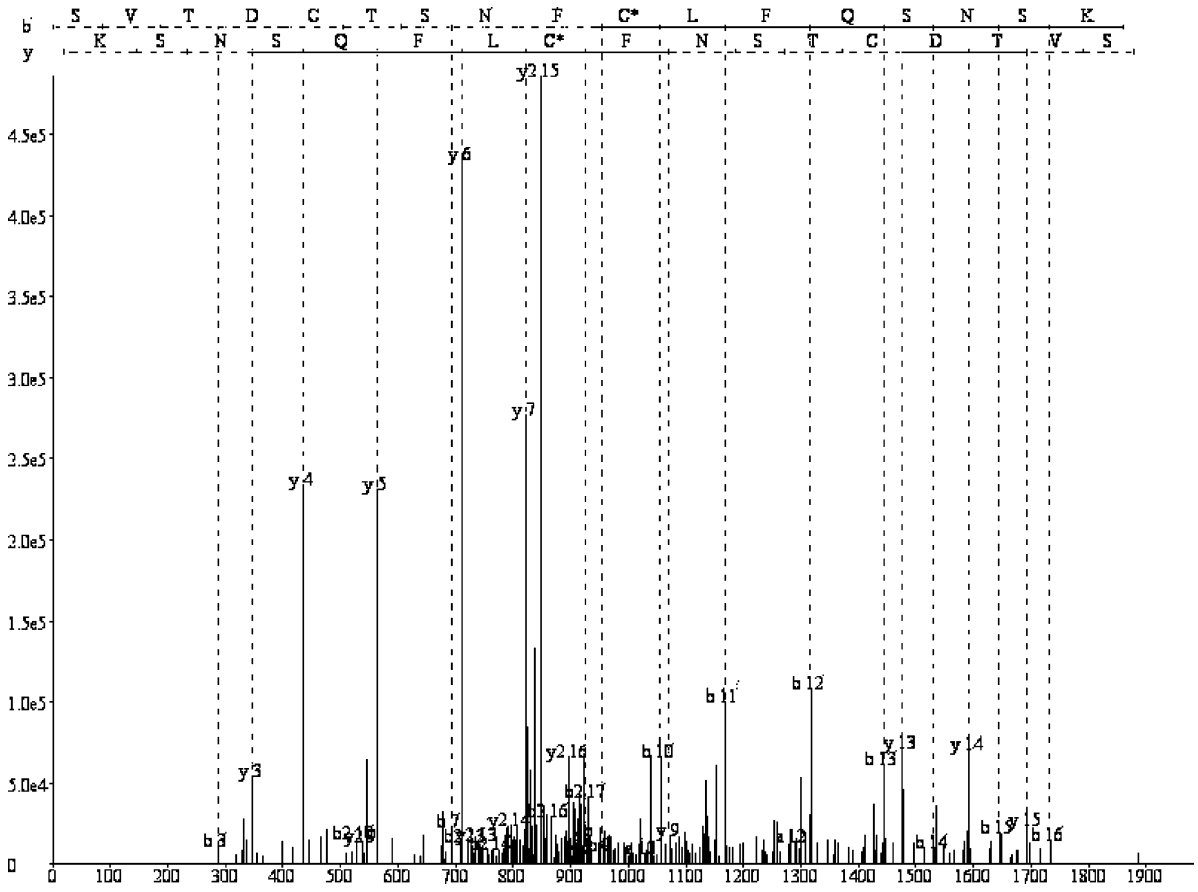


Figure 7. Annotated spectrum from the ISB data set for a peptide containing an intramolecular disulfide bond. Note that breaking peptide bonds between the two cysteines does not disconnect the molecule, so b and y peaks in this range are suppressed.

in different samples (ex: diseased versus normal) is a proxy for relative expression levels of the proteins. Unknown differences in oxidation levels in the two samples might lead to spurious conclusions on relative protein expression.

We searched this data set for other putative modifications using InsPecT as well as another search tool that searches in a blind mode.[39] In both searches, we found very few additional modifications. A detailed analysis of these modifications is omitted from this paper.

Performance on Complex Mixtures: Mus-IMAC. The Mus-IMAC spectra were searched against the mus-nr database (78 500 proteins, 30 Mb) using InsPecT. The search revealed a large number of modified and unmodified hits with reliable (p-value <0.01) annotations on 1335 spectra. These high-confidence annotations include a total of 91 phosphorylated and 310 non-phosphorylated peptides, from 277 proteins in all. The searches were conducted allowing a maximum of four modifications (≤ 2 oxidations, ≤ 2 phosphorylations), a parent mass tolerance of 2.5 Da, nontryptic peptides, and missed cleavages. Even with these permissive settings, the search took an average of 0.4 s/spectrum using 25 tags. Using these same settings with Sequest led to a search time of 99 s/spectrum.

For comparison, these spectra were also searched using Sequest and Mascot, as run off the web server.[24] We note that the comparisons of different search tools is difficult for complex data sets where the true answer is not known. The answers of different tools depend on the cutoffs, which are not directly comparable across different algorithms. Manual validation is possible but is subjective. Therefore, we chose to validate the top hits using match quality critera that can readily be compared between tools and are generally accepted. We considered the top phosphopeptide annotations (100 or 500) generated by each tool, as ordered by that tool's scoring scheme. Each peptide is annotated depending upon the number of b/y peaks present, the percentage of total intensity explained, and the percentage of top 25 peaks explained. The performance of various tools is summarized in Figure 8a.

By these criteria, InsPecT performs very well. In over 95% of the cases, the top 100 phosphopeptides explain at least 30% of the total intensity, 30% of all b/y ions, which provides confidence that the answer is correct. The corresponding numbers are much lower for Mascot and Sequest. The numbers drop for all tools when we look at the top 500 phosphopeptide predictions. This is not surprising, as InsPecT only reported 91 phosphopeptides as having a significant score. We further subjected these identifications to a very stringent manual validation. Figure 8b describes the results of that validation. While reiterating that manual validation is inherently subjective, we assert that InsPecT finds a number of novel phosphorylation sites missed by the other tools. At the same time, the numbers shown here are a conservative estimate, and it is likely that many other InsPecT annotations are indeed correct. Another interesting aspect of Figure 8b is that InsPecT finds everything that is found by both Sequest and Mascot while the other tools often miss sites that are found by the other two. Several sites obtained by InsPecT are validated by a large number of spectra for the same, and overlapping peptides.

(a)

Filter	InsPecT	Mascot	Sequest
30% of b/y peaks present	100 (89)	57 (36)	46 (38)
50% of b/y peaks present	75 (42)	23 (7)	6 (3)
30% of intensity explained	95 (76)	60 (45)	26 (20)
30% of peaks explained	100 (89)	57 (36)	46 (38)

(b)

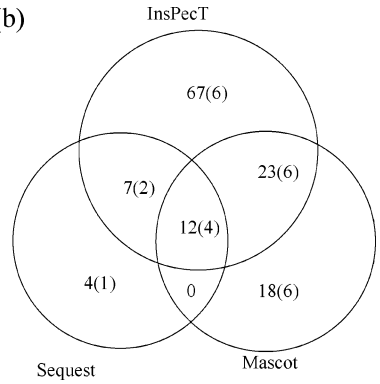


Figure 8. Comparison of phosphopeptide annotations from three database search tools. (a) The percentage of the top 100 (500) phosphopeptide annotations meeting each filter is shown. A b or y fragment containing a phosphorylation site is considered to be present if a peak is seen at either the correct theoretical mass M or $M - 98$. (b) Manually validated phosphopeptides in the Mis-IMAC data set. A Venn diagram compares the number of spectra and (in parentheses) phosphorylation sites identified by Mascot, InsPecT, and Sequest.

Table 3. Phosphopeptides Identified on the Mus-IMAC Data Set

peptide	gi number	protein	peptides	spectra	*p*-value
APTSTD*p* TPIYSQVAPR	2584837	nonreceptor protein tyrosine phosphatase	1	3	0.0001
ASGQAFELIL*p* SPR	9789995	stathmin 1	1	5	0.003
DLLHP*p* SPEEEK	26024336	ribosomal protein S27	1	1	0.001
DTGK*p* TPVEPEVAIHR	51830766	similar to 40S ribosomal protein S20	1	43	0.0001
GSGIFDES*p* TPVQTR	39573709	HN1-like protein	1	4	0.0001
IVN*p* SLSLLPK	13561075	nuclear export factor-b	1	6	0.005
LHKVI*p* YITQQ	51768221	similar to hypothetical protein FLJ32685	1	36	0.001
LPNIQAVLLPKK*p* TESHHK	817939	histone H2A	3	4	0.002
MASNIFG*p* TPEENPPSWAK	6680237	hematological and neurological expressed sequence 1	1	5	0.0001
PAAPAAPAPAEK*p* TPVKK	356168	histone H1b	1	2	0.001
PAAPAAPAPVEK*p* TPVK	34328365	histone 1, H1d	3	10	0.0001
RADLNQGIGEPQ*p* SPSR	31981086	EF hand domain containing 2	1	2	0.003
RP*p* SVYLPTR	38089310	RIKEN cDNA 1500034J01	1	2	0.0002
SETAPAETAAPAPVEK*p* SPAK	21426893	histone 1, H1b	1	2	0.0006
SKESVPDFPL*p* SPPK	9789995	stathmin 1	2	10	0.0001
SQE*p* TYETLKHEKPPQ	119869	FceRI gamma	1	6	0.003
V*p* YIYIYIL *p* Y	18043693	C630007L23Rik protein	1	1	0.005
YFD*p* SGDYNMAK	9624979	endosulfine alpha	2	15	0.0001

Several sites are witnessed by a single high-quality spectrum for the protein; such sites must be found in a single pass through the database rather than in a two-pass refinement approach.

Among the novel phosphopeptides, we find phosphothreonine and phosphoserine sites in histones, including PAAPAAPAPVEKpTPVKK and PNIQAVLLPKKTEpSHHK. (See Table 3, and Figure 9.) Phosphorylation has been described as a key mechanism for regulation of histones H1 and H3.[14,15,27] A large number of unmodified histone peptides were matched as well (6 of the top 10 peptides come from histone proteins), indicating that histones are plentiful in the cell extract. This gives us further confidence in the identification of the histone phosphorylations.

Many of the novel phosphopeptides include missed cleavage sites or nontryptic end points, indicating that efficiency considerations may have forced the other tools to overlook them in a more restrictive search. Many phosphorylation sites are identified in multiple peptides, which differ by a missed cleavage or methionine oxidation. These multiple witnesses provide strong evidence that the phosphorylation site is correct. We note that several phosphopeptides come from proteins with no other spectral matches; such phosphorylation sites pose a particular challenge for two-pass searching, since they must be found in the first search pass or not at all.

Phosphorylations are among the most widespread and biologically significant posttranslational modifications. In addition to shifting peak masses, PTMs (and phosphorylations in particular) affect ion fragmentation patterns. As the data set of phosphorylated spectra grows, we can mine it to obtain empirical estimates of fragmentation probabilities, which in turn will improve the score function, leading to more identifications. Our search identified a total of 157 spectra matching 18 manually verified phorylation sites (43 phosphoserine spectra, 77 phosphothreonine, and 37 phosphotyrosine). Even with this limited data set, we can start computing fragmentation probabilities. Table 4 summarizes our findings. For example, it has been suggested[10] that the loss of neutral ion H_3PO_4 is very common for phosphoserine. For multiply charged peptides, this prouces a dominant $M - H_3PO_4$ ion in the spectrum, as well as neutral losses on fragment ions. The effect is much less pronounced for phosphothreonine and phosphotyrosine.

Additionally, a b prefix ion, formed by a break immediately N-terminal of the phosphorylation site, has an 71% chance of being

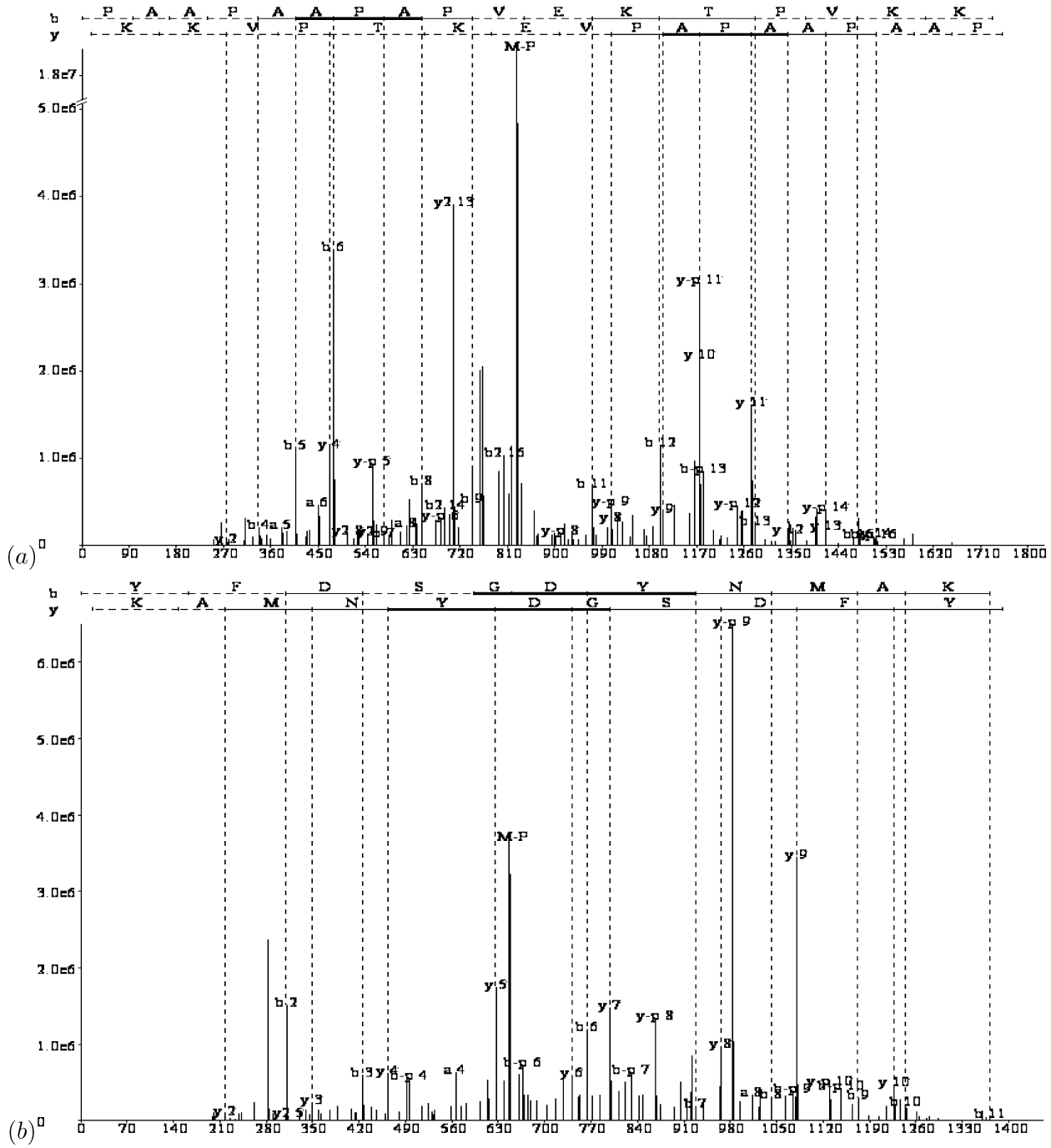


Figure 9. Phosphorylated peptides from the Mus-IMAC data set. (a) is novel, and (b) is a known peptide. Bold segments of the b and y ion ladders indicate the tripeptide tag used in to locate the candidate.

observed (as compared with 46% for b ions overall). A y suffix ion, formed by a break immediately C-terminal of the phosphorylation site, has an 81% chance of being observed (as compared with 53% for y ions overall). This reflects the fact that breakages are particularly likely to leave positive charge on the side without the (negatively charged) phosphate group. This is similar to the favored breakage on the N-terminal (rather than C-terminal) side of proline residues. However, the counterpart ions (b suffix, and y suffix) are not significantly suppressed in phosphopeptides. As we train our model of fragmentation preferences, our ability to identify PTMs will improve.

Test for Rich Modifications: IKKb. The IKKb data set (45 500 spectra) was searched against the database of 16 proteins and 2000 random proteins (total size 0.65 Mb) using a rich set of 10 permissible modification masses. Up to two instances of each PTM were permitted per peptide, and up to two total PTMs were allowed. This produces a list of 61 possible decorations in all. The list of PTMs included three fictitious modifications to measure false positive rates. The average search time with this rich set of PTMs was 0.36 s/spectrum.

A total of 13 041 spectra received valid protein annotations, excluding all annotations with a spurious PTM. Of these spectra, 8579 (66%) were unmodified, 3124 (24%) carried one PTM, and 1338 (10%) carried two PTMs. As the odds of matching a valid protein by chance are low (less than 1%), the estimated accuracy of these annotations is ~97%. Table 5 summarizes the PTM

Table 4. Percentage Frequency of b and y Ions, as Well as Phosphate Losses, in Several Environments[a]

phosphorylation	b	b − 98	b − 80	y	y − 98	y − 80
containing pS	46	51	10	53	54	11
containing pT/pY	36	32	8	31	22	6
pre-pS	71	n/a	n/a	53	75	12
post-pS	42	51	6	81	n/a	n/a
pre-pTY	50	n/a	n/a	27	28	11
post-pTY	33	43	10	77	n/a	n/a

[a] pre-pS refers to ions formed by breaking the N-terminal peptide bond of a phosphorylated residue, post-pS to the C-terminal bond. Phosphate loss peaks are particularly prevalent in the presence of phosphoserine, even more so than standard b and y peaks. Phosphate losses are particularly prevalent for breaks adjacent to the phosphorylation site.

Table 5. Posttranslational Modifications Found in the IKKb Data Set[a]

residue	mass Δ	putative modification	spectra	rate	mean score
M	16	oxidation	2110	65.53	4.00
C	14	CAM vs PAM	440	59.10	3.87
N	1	deamidation	1274	12.36	3.56
W	16	oxidation	133	6.66	3.20
M	32	double oxidation	212	6.58	5.72
K	28	dimethylation	465	5.15	2.64
K	14	methylation	268	2.97	2.41
W	32	double oxidation	59	2.96	1.8
T	−18	water loss	234	2.72	1.84
D	−18	water loss	146	1.49	1.31
E	−18	water loss	177	1.48	0.83
E	22	sodium	173	1.45	2.08
D	22	sodium	136	1.39	0.91
P	40	spurious	98	1.05	−0.27
A	10	spurious	80	0.96	−1.00
I	25	spurious	7	0.70	−0.90
L	25	spurious	101	0.48	−0.97

[a] The rate reported is the percentage of amino acid residues from valid proteins which carry the PTM. The mean match quality score shows a clear separation between valid and spurious PTMs.

findings. For each posttranslationally modified peptide identified in the data set, the unmodified equivalent is matched as well. This suggests that these PTMs are not constituitive modifications. Because this data set incorporates many scans from a highly purified protein sample, relatively rare chemical events (such as water losses from acidic residues before fragmentation) are detectable.

The large number of spectra allows us to estimate the "rate" at which PTMs are incorporated, by considering the percentage of amino acids from valid proteins that carry the particular modification. Note that spurious PTMs have a low rate, as one would expect, implying a low rate of false PTM identifications. Only 2% of all modified spectra are decorated with a spurious modification. In addition to being rare, the spurious PTMs are more likely to occur in annotations with a low match quality score. The quality score is −0.7 (p-value 0.07) on average, as compared to 3.6 (p-value 0.0002) for annotations incorporating only valid PTMs). For example, while +22 (sodium) is a rare modification on E and D, comparable to the rate at which a spurious modification might occur, the average quality scores are 2.08 and 0.83, distinctly higher than those for spurious modifications. In addition, if we require that a modified peptide be confirmed by multiple spectra, only 4 of 427 such peptides include a spurious PTM.

One could question the value of the posttranslational modifications identified here, given that a human protein is expressed in yeast. The expression in yeast cells allows us to generate the large amounts of protein needed to comprehensively interrogate for modifications. Indeed, many PTMs are conserved between human and mouse. Another unexpected finding is the absence of phosphopeptides, even though IKKb is known to be phosphorylated. We found that the regions known to be phosphorylated are not represented in the MS coverage of the protein (data not shown). This points to possible difficulties in spectral acquisition from intact phosphopeptides (see also ref 30). Overall, our results demonstrate that InsPecT is a viable tool for rapidly identifying a large number of putative modifications with low error.

DISCUSSION

We demonstrate that it is possible to identify modifications in peptides from complex mixtures using desktop computer resources. We focus extensively on search efficiency using sequence tag filters. Efficient searching is often not at the forefront of proteomic research, as one can compensate by "adding more computers". In practice, however, efficiency considerations often dictate which peptides get identified. As PT modifications lead to a combinatorial explosion, they are often searched for under very restrictive conditions. For example, one might limit the search to proteins from which unmodified peptides have been found or forbid multiple modifications or missed trypsin cleavages. We demonstrate that it is feasible to search efficiently without such restrictions.

Our approach is modular and will improve with improvements in the modules for tagging, filtering, and scoring. With improvements in instrument mass accuracy, and de novo sequencing algorithms, sequence tags should become the dominant filter, identifying all but a few of the true peptides. As other results, and our own results on the ISB data set show, the tag-based approaches do very well for high-intensity spectra, where the signal peak intensities dominate. We also exploit the full power of sequence-based search using tags by combining the tags from multiple spectra in a single scan. With a linear time preprocessing of the tags to construct a trie-automaton,[2] a single pass through the database handles many tags and spectra. Correspondingly, our search is faster than other database search algorithms, even after allowing for modifications. With improvements in tag generation, we can use longer (and fewer) tags, leading to further speed improvements. Finally, a dynamic programming technique allows us to generate modified peptide candidates without explicit enumeration of all modifications. In future work, we intend to extend our algorithm to handle "blind" modification search (for modifications of unspecified size), as well as mutation-tolerant search. Our tool is available by contacting the authors.

A web interface to the InsPecT tool is provided at http://peptide.ucsd.edu/. In addition, the tool is available from the authors by request.

ACKNOWLEDGMENT

We thank Lien Chung for running GutenTag benchmarks. We benefitted greatly from discussions with Helge Weissig and Nuno Bandeira. We are grateful to Alexei Nesvizhskii for a careful reading of this paper and for many useful discussions. This project was supported by NIH grant NIGMS 1-R01-RR16522. Production of the IKKb dataset was supported by NIH grant R01GM65325 and by the PEW Scholars Program.

Received for review January 18, 2005. Accepted May 11, 2005.

AC050102D