Published in final edited form as:
J Proteome Res. 2008 October ; 7(10): 4422–4434. doi:10.1021/pr800400q.

Improved Sequence Tag Generation Method for Peptide Identification in Tandem Mass Spectrometry

Xia Cao[1] and **Alexey I. Nesvizhskii**[1,2]

[1]Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA

[2]Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

The sequence tag-based peptide identification methods are a promising alternative to the traditional database search approach. However, a more comprehensive analysis, optimization, and comparison with established methods are necessary before these methods can gain widespread use in the proteomics community. Using the InsPecT open source code base (Tanner et al., Anal Chem. 2005, 77:4626–39), we present an improved sequence tag generation method that directly incorporates multi-charged fragment ion peaks present in many tandem mass spectra of higher charge states. We also investigate the performance of sequence tagging under different settings using control datasets generated on five different types of mass spectrometers, as well as using a complex phosphopeptide-enriched sample. We also demonstrate that additional modeling of InsPecT search scores using a semi-parametric approach incorporating the accuracy of the precursor ion mass measurement provides additional improvement in the ability to discriminate between correct and incorrect peptide identifications. The overall superior performance of the sequence tag-based peptide identification method is demonstrated by comparison with a commonly used SEQUEST/PeptideProphet approach.

Keywords

Proteomics; Tandem Mass Spectrometry; Peptide Identification; Database Searching; De Novo Sequencing; Algorithms; Statistical Analysis

Introduction

Tandem mass spectrometry (MS/MS) has become the method of choice for identifying peptides and proteins from complex biological samples.[1–3] Application of mass spectrometry (MS) in a high throughput setting brings significant computational data analysis challenges.[4] In particular, peptide identification from MS/MS spectra can be time consuming and error-prone, especially when applied for the identification of post-translational modifications (PTMs) such as phosphorylation [5–13].

Computational methods for peptide identification from MS/MS spectra can be roughly divided into two categories: database searching and *de novo* sequencing. Database search-based methods take an experimental MS/MS spectrum as input and compare it against theoretical fragmentation patterns constructed for peptides from the searched database to find a match.[2] Representative computational tools that automate this process include

Corresponding author: Alexey I. Nesvizhskii, Department of Pathology, University of Michigan, 4237 Medical Science I, Ann Arbor, MI, 48109, nesvi@med.umich.edu, Tel: +1 734 764 3516.

SEQUEST,[8] Mascot,[7] X! Tandem,[6] OMSSA,[10] for a review see.[14] The limitations of this method include restricted nature of the search, i.e., it can find the exact peptide sequences from the specified protein sequence database only. The computational time also becomes an issue when the set of candidate peptides is large, as in the case of phosphopeptide analysis or genomic database searches. *De novo* peptide sequencing method, exemplified by computational tools such as Lutefisk[15, 16], Sherenga[17], PEAKS[11] and PepNovo,[9] reconstructs the peptide sequences directly from the mass spectra without referring to a sequence database for help[2]. This method allows identification of peptides that are not present in the searched sequence database. However, it is also computationally intensive and requires high quality MS/MS spectra, which makes it unpractical for large scale analysis.

To address these limitations, hybrid computational strategies utilizing the idea of “sequence tags” have been proposed. A tag is a short amino acid sequence with a prefix mass and a suffix mass value which designate its position in the peptide. The database search time is reduced significantly by only searching those candidate peptides that contain the tags extracted from the MS/MS spectrum. Sequence tagging was first introduced by Mann and Wilm,[18] and further developed in recent years.[12, 13, 19–23] InsPecT[13] is an example of a freely available open-source peptide identification tool that use tags as a filter to conduct the peptide identification.

While the sequence tag-based method is clearly a promising alternative, a more comprehensive analysis and comparison with established methods is necessary before these methods can gain widespread use in the proteomics community. For example, previous tests were largely carried out using older control dataset such as the original ISB 18 protein mix, which may no longer be representative of data generated using the current generation of instruments. The comparisons need to be done not directly with SEQUEST or Mascot, but with the results of those tools after additional validation by PeptideProphet[4] or similar statistical methods. Furthermore, most previous studies focused on the analysis of doubly charged MS/MS spectra only [12, 13, 24], whereas new instruments such as LTQ-FT, and new fragmentation mechanisms such as electron-transfer dissociation (ETD), acquire a significant proportion of MS/MS spectra on peptide ions of charge state 3+ or higher.

In this work, we first present a method, based on the tag generation algorithm of InsPecT, to generate an improved set of tags for spectra of charge 3+ or higher using multi-charged fragment ion peaks present in those spectra. Although suggested previously[25, 26], to the best of our knowledge, this is the first work that discusses in detail the utilization of the multi-charged peaks in the tag construction process, and investigates their effect on peptide identification. We also investigate the performance of the sequence tag-based method using control datasets generated on different types of mass spectrometers and using a complex phosphopeptide-enriched sample. Furthermore, we demonstrate that additional modeling of InsPecT search scores using semi-parametric mixture modeling approach incorporating the mass accuracy of the precursor peptide m/z measurement[27] provides additional improvement in the ability to discriminate between correct and incorrect peptide assignments.

Experimental Procedures

Experimental Data

Protein mix data—The first group of data sets used in this study was obtained from a mixture of purified trypsin digested proteins[28]. About 200 fmol of the standard mixture (Mix 3) was analyzed following separation by high pressure liquid chromatography (HPLC) followed by electrospray ionization (ESI) tandem mass spectrometry using the following MS instruments: Thermo Electron LTQ-FT, Thermo Electron LTQ, Agilent XCT Ultra, Thermo Electron LCQ Deca, and Waters/Micromass QTOF Ultima. For LTQ-FT, in addition to using the whole data set (40376 MS/MS spectra), a more detailed analysis was performed using a subset of spectra (20377 spectra in total from 5 spectral files: B06-11071, B06-11072, B06-11073, B06-11074, and B06-11075). For the other four mass spectrometer types, the following subsets of the full data set were used: LTQ (13432 MS/MS spectra, 3 spectral files LT20060324_S_60min18mix_03, LT20060324_S_60min18mix_04, LT20060324_S_60min18mix_05), Agilent (46391 MS/MS spectra, 3 spectral files: C066-000005, C066-000007, C066-000009), LCQ (11784MS/MSspectra, 4spectralfilesLQ20060324_s_60min18mix_03, LQ20060324_s_60min18mix_04, LQ20060324_s_60min18mix_05, LQ20060324_s_60min18mix_06), and QTOF (4017 MS/MS spectra, 3 spectral files QT20060328_Den18mix_01, QT20060328_Den18mix_02, QT20060328_Den18mix_03).

Phosphopeptide-enriched sample—The data set of *Drosophila melanogaster* trypsin digested proteins enriched for phosphopeptides using immobilized metal affinity chromatography (IMAC) was described in[29]. The spectra were acquired on a Thermo Electron LTQ instrument in MS-MS^2-only mode (no MS^3 spectra). This data set contained 7228 MS/MS spectra from 2 spectral files (A07_5206_c and A07_5208_c). Due to high efficiency of phosphopeptide-enrichment (89% of the identified peptides were phosphorylated), this data set was used to evaluate the performance of the method on modified peptides.

InsPecT settings

Default parameter settings in InsPecT input file were used except when noted. The tag length was set to 3 and the number of top scoring tags was 100. Trypsin was specified as the enzyme used to digest proteins. With the protein mix data, no modifications were specified except a fixed modification of 57.0 Da (iodacetamide alkylation) for cysteine. The searched sequence database was constructed from the sequences of proteins known to be present in the mixture, common contaminants, and a large number of decoy entries derived by reversing the sequences from the human IPI database. In the case of phosphopeptide-enriched data set, a maximum of three modifications were allowed which could either be optional phosphorylation modification (+80.0 on S, T, and Y) or a fixed modification of 57 for cysteine. The searched database for the phosphopeptide-enriched sample consisted of all *Drosophila melanogaster* sequences exported from the UniProt database, 26311 entries total, to which the reversed set of sequences were appended.

Charge state assignment

Since the charge state of a peptide ion which produced MS/MS spectrum cannot always be accurately determined, it creates ambiguities in the analysis of the data.[17, 30, 31] InsPecT uses a support vector machine (SVM) model to determine the charge state (1+, 2+ or 3+) of a spectrum before the tag generation is conducted. This charge determination function is effective in that it helps to reduce the database search running time since most spectra are searched only once using the charge state determined by the SVM model. However, it can also lead to incorrect assignments of the charge state resulting in incorrect peptide identification. An alternative approach is to search multiply charged spectra allowing all possible charge states (referred to below as “forced charge” option). In the case of high mass resolution instruments such as LTQ-FT and QTOF, the charge state was taken directly from the mzXML file, i.e. as determined by the instrument software.

SEQUEST analysis

MS/MS spectra were extracted from mzXML files in *.dta format using mzXML2Other tool *. Resulting *.dta files were searched with SEQUEST using the following parameter. Protein mix data: peptide mass tolerance of 3.0 Da; b- and y-ion series; partial trypsin digestion, allowing for one missed cleavage site; a fixed modification of 57.0 was specified for cysteine. Phosphopeptide-enriched sample: peptide tolerance of 3.0 Da; partial trypsin digestion, one possible missed cleavage; fixed modification of 57.0 for cysteine; variable modifications of 80.0 Da (phosphorylation) were specified for S, T, and Y; a maximum 4 PTMs per peptide. The same sequence databases were used as described above.

First, SEQUEST results were used to derive a benchmark data set of spectra with known peptide assignments to evaluate the accuracy of tags derived using different settings. SEQUEST assignments were analyzed using PeptideProphet[4] and ProteinProphet[32]. The list of proteins was first filtered using ProteinProphet probability of 0.9, and then all MS/MS spectra assigned a peptide from one of the high confidence proteins and with PeptideProphet probability above 0.1 were extracted. A sequence tag generated from a spectrum in the benchmark data set was considered correct if the tag segment was in the SEQUEST assigned peptide sequence and its prefix and suffix masses were both within a 3 Da mass tolerance from the actual prefix and suffix masses. For the comparison between SEQUEST/PeptideProphet and sequence tagging, SEQUEST search results were analyzed using PeptideProphet and filtered using various probability thresholds. PeptideProphet was run under the most common settings, with a high mass accuracy binning option (‘-A’ option) specified for QTOF and LTQ-FT data[33]. In the case of phosphopeptide-enriched sample, PeptideProphet was run with ‘-l’ option which is beneficial in the case of phosphorylated peptides.[29]

Results and Discussion

The sequence-tag based peptide identification strategy involves the following steps:[13] (1) spectrum filtering to remove noise; (2) generation of tags from the filtered spectrum using partial *de novo* sequencing; (3) database filtering to limit the set of candidate peptides to those containing one of the extracted tag; (4) tag extension in which the tag is extended in the sequences to complete the peptide selection process; (5) computation of a score for each candidate peptide and selection of the top scoring match. The subsequent sections of the manuscript investigate in detail the sequence tag generation part of the algorithm (illustrated in Figure 1), and the effect of various sequence tagging options on the entire peptide identification process. It should also be emphasized that one of the main advantages of sequence tag-based peptide identification methods is the computational speed of the analysis. The improved computational efficiency of sequence-tag based peptide methods compared to straightforward database searching using SEQUEST or similar tools has been extensively discussed in the literature[13]. Thus, the discussion below will only consider how the time of the analysis is affected by various sequence tagging options and parameters investigated in this work, with regular InsPecT’s performance under default settings taken as a reference point.

Exploratory analysis: spectrum filtering and fragment ion statistics

The key step in the process is generation of a small set of sequence tags for each spectrum that satisfy the following property: at least one tag in the set of generated tags is present in the sequence of the peptide that generated the spectrum, so that the peptide will not be filtered out.[24] The performance of the tag generation algorithm depends in part on the details

*http://tools.proteomecenter.org/software.php of spectrum processing and the types of fragment ions considered. By default, InsPecT considers only 6 most intense peaks in the spectrum within a 50 Da window (the number of peaks per interval considered will be referred to as peak depth below). Another assumption during tag generation in InsPecT, as well as in most other programs, is that observed peaks represent singly charged fragment ions. However, spectra of charge 3+ or higher are expected to contain a significant number of multi-charged fragments. Some doubly charged fragment ions are also present in 2+ spectra, e.g. peaks corresponding to a neutral loss of or one or more residues at the peptide N-terminus.[34]

To investigate the fragment ion statistics for spectra of different charge state, an exploratory analysis was performed using 2+, 3+, and 4+ charged spectra from the protein mix LTQ-FT dataset (the 5 LC-MS/MS run subset, see Experimental Data). The spectra were filtered using the default InsPecT peak depth of 6 (window size 50 Da), as well as using peak depth 3 and 12. Table 1 shows statistics for the most common ion types assigned to all correctly identified peptides (based on SEQUEST analysis) in this data set. Fragment ions were assigned to peaks in a ranked order based on their overall likelihood, i.e., a peak that can be explained by more than one fragment ion would be assigned to a more common ion type. A mass tolerance of 0.8 Da was used when assigning fragment ions to peaks. Multiple peaks within a 0.8 Da of a predicted fragment ion, if present in the filtered spectrum, were counted as one. For each peak depth, the Table shows the average number of peaks assigned an ion of a given type in a spectrum of a particular charge state. Also shown are the percentage of total assigned peaks labeled as given type, and the ion type propensity. Propensity is defined here as the percentage of all possible ions of a given type that were observed in a spectrum on average.

Table 1 shows several interesting trends. In the case of doubly charged spectra, the singly charge ions b^+ and y^+ dominate (labeled as b and y in Table 1). The observed distributions of ion types and propensities are largely in agreement with previous observations[17, 35]. At peak depth 6, a spectrum of charge state 2+ contains on average 7.1 (propensity 0.64) and 8.7 (propensity 0.79) peaks corresponding to b^+ and y^+ fragment ions, respectively. Combined, b^+ and y^+ ions account for about 40% of all assigned peaks. Note that this percentage may be slightly underestimated due to assignment of peaks to other fragment ion types by random change. The frequency of random assignment can be estimated to be between 0.5 and 1 peak per ion type per spectrum (propensity ~ 0.05 – 0.1) based on the observed counts for ions not expected to be present in doubly charged MS/MS spectra, such as b^{+++} and y^{+++} (b3 and y3) ions. Doubly charged fragment ions are far less abundant, 1.1 (propensity 0.1) and 2.7 (propensity 0.24) b^{++} and y^{++} ions per spectrum, respectively. Since the propensity y^{++} is just above the noise, and at the noise level for b^{++}, their contribution to tag generation is not expected to be significant.

The situation is different for spectra of higher charge state. In the case of triply charged MS/MS spectra, doubly charged fragment ions are almost as frequent as singly charged ions. For example, the propensity of y^{++} ions in the case of 3+ spectra is 0.39, close to the propensity of 0.5 observed for y^+ ions. In the case of spectra of charge state 4+, b^{++} and y^{++} ions become the most frequently observed ions, with an additional substantial contribution from triply charged fragment ions. This suggests that in the case of high charge MS/MS spectra multi-charged fragment ions should be considered by the tag generation algorithms.

In considering the importance of multi-charged fragment ions, it is also necessary to consider the effect of spectrum filtering, and in particular the peak depth parameter (Table 1). As peak depth increases, less intense peaks start contributing to the ion statistics. For example, at peak depth 3, in the case of 2+ spectra there are on average 5.6 and 8.1 peaks per spectrum labeled as b^+ and y^+, respectively. Combined, they explain close to half of all assigned peaks. At peak depth 12, the average number of peaks assigned these two ion types increases to 7.8 and 9, respectively. However, inclusion of less intense peaks has a more significant impact on other ion types, such as loss of water or ammonia ions and multi-charged ions. As a result, at peak depth 12 only approximately 30% of all assigned ions are b^+ and y^+ ions. Similar trends were observed for spectra of 3+ and 4+ charge states: at peak depth 3, just a few ion types dominate, but at higher peak depth less common ion types start catching up. This suggests that the comparative analysis of various sequence tagging methods, and in particular the effect of multi-charged ions on tag generation, should consider the peak depth parameter.

Spectrum graph generation and edge constraints

To incorporate multi-charged fragment ions, a new tag generation algorithm was implemented using the InsPecT source code. The outline of the method is shown in Figure 1 using an example of a spectrum acquired on a triply charged peptide ion. As the first step, the experimental spectrum is processed using InsPecT filtering function. The default peak depth of 6 peaks per 50 Da window is used, except when noted. The analysis starts with a standard assumption of singly charged fragment ions. Given a peak with mass to charge ratio M in a spectrum acquired on a peptide ion with a singly protonated mass PM, a node of mass M - H^+ (b^+ node) and its complementary node of mass PM - M (y^+ node) are created, where H^+ is the proton mass. These two nodes are inserted in the spectrum graph and labeled as type “0” and “1”, respectively. To include multi-charged fragment ions, the procedure is extended by assuming that each peak in the spectrum may instead represent a doubly charged fragment ion. Thus, two additional nodes of mass 2(M-H^+) and PM-2(M-H^+)-H^+ are created and labeled as “2” and “3” (b^{++} and y^{++} nodes), respectively. The triply charged fragment ions can be accommodated in the same way (nodes of type “4” and “5”).

After all the nodes are inserted in the spectrum graph, edges are created between any two nodes that have the mass difference between them corresponding to the mass of an amino acid (or amino acid with a modification) within a certain mass tolerance (edge mass tolerance constraint). The spectrum graph is then searched and each sub-path of a fixed length (length 3 used in this work) is extracted as a tag. Each tag is represented as a triplet <prefix mass, tag segment, suffix mass>. Connecting all node types, denoted as [0123] in the 3+ spectrum example used in Figure 1, leads to a large number of edges, making the spectrum graph very complex. This increases the number of sub-paths in the spectrum graph, which in turn results in a loss of sensitivity (see below).

To reduce the spectrum graph complexity, additional constraints on edge creation are introduced. The most stringent constraint would be to allow creation of sequence tags from nodes of the same type only, denoted as [0][1][2][3] in Figure 1, e.g. 0-0-0-0 tag (all b^+ ions) or 1-1-1-1 tag (all y^+ ions). Another option is to allow a mix of different node types, but limit the kind of connections allowed. For example, it has been observed that peptide fragments are more likely to appear together in a spectrum as a pair of complementary peaks, e.g. b^+ ion peak and its complement y^{++} ion peak, or b^{++} and its complement y^+ in the case of 3+ spectra.[31, 36] Thus, a reasonable assumption is to allow nodes of not more than two different types within the same tag.

Furthermore, as an additional constraint, the number of node type mismatches can be limited, e.g. allowing tags with 1 mismatch (e.g. 0-3-0-0 tag) but not 2 mismatches (as in 0-3-3-0 tag). Figure 2 shows the results of the analysis of the number mismatches in the case of the phosphopeptide-enriched data set. For each MS/MS spectrum of charge 3+ that was identified with SEQUEST (see Experimental Methods), 100 randomly selected incorrect tags and the two highest scoring correct tags were extracted. The tags then were divided into categories based on the number of mismatches. It is evident that a large portion of correct tags have 0 mismatches (i.e., tags derived from fragment ions of the same type, e.g. all y^+ ions). Allowing not more than 1 mismatch covers 99% of all correct tags. In contrast, incorrect tags are distributed across all three categories, with a substantial number (30%) of tags containing 2 mismatches.

Finally, when connecting nodes of type 2 or higher, it is reasonable to allow a higher mass tolerance than what is used for singly charged fragment ions (0.5 Da). The same fragment ion m/z mass measurement error effectively results in a larger error in the determination of the corresponding peptide fragment mass due to multiplication of the m/z value by the charge value.

The nomenclature for describing various constraints is illustrated in Figure 1. For example, (03)(12) denotes a sequence tag generation method that allows connection between nodes of type 0 and 3, or type 1 and 2 only, edge mass tolerance of 0.5 Da, and not more than 1 mismatch. The increase in the edge mass tolerance to 1.0 Da from 0.5 Da, when applied, is explicitly indicated, e.g., (03)(12)1.0. For reference, the tagging method of InsPecT considers nodes of type 0 and 1 only, no restriction on the number of mismatches, and uses 0.5 Da mass tolerance, and thus can be denoted as [01].

Tag scoring

The tag scoring function implemented in InsPecT uses a Bayesian network approach that effectively captures the relationship between the peak intensity and the fragment ion type. The tag score is a combination of the so-called node cut score, flank score, and edge mass error calculated from the tag.[37] The retraining of the Bayesian scoring function given the modified tag generation process was not attempted in this work. Instead, several alternative scoring functions were tested. An SVM model and a logistic regression model were trained using multiple spectral features computed for each sequence tag. These features included the summed fragment ion intensities, sum of the intensity ranks, sum of squared mass error for all edges in the tag, and other features. However, when implemented within the InsPecT source code, the results were not consistently better (data not shown). Thus, all the subsequent analysis was performed using the existing InsPecT scoring function. While performing this analysis, it was observed that sequence tags exhibited a heterogeneous distribution of tag scores depending on the types of nodes used to create the tag. This was partially addressed in this work by selecting 50 high scoring tags separately for each group, e.g. in the case of (03)(12) setting, 50 (03) tags and 50 (12) tags, for a total of 100.

Evaluation of tag generation methods

Sensitivity analysis—An ideal tagging method should generate the tags fast and with high sensitivity. The sensitivity (referred to as 'tag accuracy' in [24]) is defined here as the number of spectra in the SEQUEST-derived benchmark data set for which at least one correct tag was in the list of N highest scoring tags (N=100 here). The sensitivity of a tagging method depends on both the coverage of the peptide sequence by generated correct tags and also the on discriminating power of the tag scoring function. Tag coverage is defined here as the number of amino acids on the peptide covered by all the extracted correct sequence tags (regardless of the tag score) to the length of the peptide. For example, if for a peptide "WQDESDDEEGDQK" the tag extraction method generates five correct tags: DES, ESD, SDD, DEE and EEG, covering 8 of the 13 residues, the tag coverage is 8/13=61.5%. A good tag scoring function would assign higher score for correct tags and lower score for incorrect tags, ensuring that at least one of the highest scoring tags is correct.

From the practical point of view, the sensitivity is the most important parameter. However, the total number of tags that can be extracted from the spectrum is informative as well since it affects the computational time of the tag generation algorithm. The tag coverage parameter is less important in the context of this work since it does not matter what fraction of the peptide sequence can be reconstructed by assembling extracted short sequence tags, as long as at least one of the tags is correct (which is measured by the sensitivity parameter). However, the tag coverage analysis is important for understanding the relationship between various tagging options, the complexity of the spectrum graph, the sensitivity, as presented below. Furthermore, it may be of interest in related applications, such as *de novo* peptide sequencing.

To investigate the performance of various constraints applied during the tag generation process, different tagging methods were compared in terms of the tag sequence coverage, the number of all possible sequence tags that can be extracted from the spectrum graphs, and the sensitivity. The analysis was first conducted using 2+ and 3+ spectra from the phosphopeptide-enriched data set. Figure 3 shows the total number of sequence tags, tag coverage, and the sensitivity for several different tagging methods and using charge state determination function. The first two points in the plot represent methods that do not use multi-charged peaks, with the second point, [01], representing default InsPecT settings. In terms of the spectrum graph complexity, 2+ and 3+ spectra demonstrate a similar behavior. As expected, creating and connecting all the four types of nodes makes the spectrum graph more complex, leading to a much larger number of possible sequence tags compared to using singly charged peaks only (Fig. 3a). The number of tags also increases with larger edge mass tolerance. The sequence coverage trends, however, are very different. For 3+ spectra, the sequence coverage (Fig. 3b) increases when multi-charged fragment peaks are included. In contrast, in the case of 2+ spectra the sequence coverage does not benefit from inclusion of multi-charged peaks, reflecting the presence of only a few doubly charged fragment ions in those spectra (Table 1).

The sensitivity measures the ability to obtain a sufficiently high score for at least one of the correct tags given a much larger background of incorrect (random) tags that can be extracted from the spectrum graph. Thus, the goal here is to define the set of conditions representing the optimal tradeoff between the increase in the sequence coverage (positive effect on sensitivity) and the total number of sequence tags (negative effect). Figure 3c shows that for 3+ spectra allowing multi-charged fragment ions increases the sensitivity, with (01)(23)1.0 being the optimal point in this data set. Connecting nodes of all types without any restriction further increases the sequence coverage, however, the sensitivity drops due to a more substantial increase in the number of sub-paths in the spectrum graph. In the case of 2+ spectra, the optimal performance is observed under [01] setting, again confirming that in those spectra doubly charged fragment ion peaks are relatively rare.

The results of the sensitivity analysis for 3+ spectra in the phosphopeptide data set are shown in more detail in Table 2. Inclusion of multi-charged peaks increases the sensitivity from 77.6% or 76.6% with [01] method to 82.7% or 82.5% for tagging method (01)(23)1.0 using either charge determination or charge forced option, respectively. It should be noted that the charge state determination function of InsPecT performed incorrect assessment on approximately 5% of 3+ spectra. When searched with SEQUEST, 457 spectra were determined (via assignment of a high probability peptide) to be of 3+ charge state, whereas only 433 of them were called 3+ by the charge determination function of InsPecT.

As mentioned above, computational efficiency is an important advantage of sequence-tag based methods. Thus, it is important that the algorithmic improvements, such as inclusion of multi-charged fragment ions, do not lead to a substantial increase in the computational time. The computational time of the sequence-tag extraction part of the method depends in part on the complexity of the spectrum graph. Compared to [01] method, (03)(12) or (01)(23) methods are not significantly slower, with the increase in time not exceeding 5%. Allowing connection among nodes of any type, such as [0123], is an order of magnitude slower, as expected from Fig. 3a.

Effect of mass spectrometer type—The sensitivity of various tagging methods was further investigated using 3+ spectra generated on different MS platforms (Table 3). As with the phosphopeptide data set (Table 2), the charge state determination function of InsPecT resulted in misclassification for a substantial number of spectra. Thus, Table 3 shows the results obtained using the forced charge option only (the charge state is taken from the mzXML files in the case of LTQ-FT data). The results obtained using the charge state determination function of InsPecT are shown in the Supplementary Table 1. It should be noted that using the forced charge option instead of the charge determination function of InsPecT increases the overall time of the analysis by close to 50% due to the need to process the same spectrum multiple times. Still, this increase is acceptable given a substantial loss in the number of correct identifications observed when using the charge state determination function.

Regardless of the MS platform, inclusion of doubly charged fragments resulted in higher sensitivity. Overall, the best performance was observed with LTQ-FT data set (96.9% sensitivity), followed by QTOF and LTQ (95.7%), and then LCQ and Agilent ion trap, 83.4% and 81.4%, respectively. For QTOF data set the best performance was achieved by the tagging method with the edge mass tolerance of 0.5 Da, whereas it was 1.0 Da (for connecting node types 2 and 3) for the other four instrument types. This reflects the superior fragment ion mass accuracy of QTOF instruments.

It should also be noted that the analysis of QTOF data was likely performed in a suboptimal way. MS/MS spectra generated using a QTOF instrument are of a higher mass accuracy and resolution than those generated using ion trap instruments. Thus, these spectra can be deconvoluted and de-isotoped, allowing reliable identification of the mass and charge state of the fragment ion.[38] Neither the commonly used conversion program ReAdW used in this work to create mzXML files from raw instrument files, nor InsPecT program itself perform these functions. Thus, additional data processing of QTOF data, e.g. using Mascot Distiller[†] or a similar tool, should improve the performance of the sequence tag extraction algorithm, and the entire peptide identification process, by reducing the number of nodes in the graph given the knowledge of the fragment ion charge state, and by eliminating spurious nodes arising from isotope-resolved (first, second isotope) peaks. Advanced spectrum processing methods may improve the analysis of data generated on other instrument types as well.[39]

Effect of spectrum quality—The MS/MS spectrum quality is one of the most important factors determining the likelihood of successful peptide identification. The Agilent data set was used here to explore the correlation between the spectrum quality and the ability to extract at least one correct sequence tag. In constructing the benchmark dataset (see Experimental Methods), those MS/MS spectra were selected for which SEQUEST assigned a peptide from a protein identified with high probability as determined by subsequent ProteinProphet analysis. Among those spectra, some were identified with low PeptideProphet probability and are likely to be of lower quality. Table 4 compares the performance of different tag generation methods on spectra with PeptideProphet probability between 0.1 and 0.9 (lower quality spectra) and those with probability equal or greater than 0.9 (higher quality). For all settings, the sensitivity is significantly higher for higher quality spectra than for lower quality spectra. Within each spectrum quality group, the tagging

[†] www.matrixscience.com

methods using multi-charged peaks outperform the ones using singly charged peaks only. However, the improvement is far more significant for spectra of lower quality.

Analysis on spectra of charge 4+—The analysis was extended to spectra of charge 4+ present in the protein mix LTQ-FT data set, in which case triply charged fragment ions were also considered in tag generation (node types 4 and 5). Table 5 lists the sensitivity for different settings. To evaluate the variability across technical replicates of the same sample, the results are shown separately for each LC-MS/MS run (mzXML file), with the combined data set statistics shown at the bottom of the Table. The tagging method using only singly charged peaks, [01], performed the worst compared to other tagging methods. However, the use of triply charged fragment ions did not lead to further improvement compared to the settings found optimal for 3+ spectra. The tagging method (03)(12),1.0 achieves the highest sensitivity of 85.1%, a significant improvement compared to only 66.4% with the [01] method. The results for the individual runs are highly consistent, indicating that the increase in the sensitivity is due to algorithmic improvements and not due to run to run variability of the mass spectrometer.

Effect of the number of top scoring tags considered—The 4+ LTQ-FT data set was also used to investigate the sensitivity as a function of the number of top scoring tags generated for each spectrum (Figure 4). The results are shown for three tagging methods, [01], (01)(23)1.0, and (03)(12)1.0. In the case of (01)(23)1.0 setting, the algorithm first selects 50 highest scoring (01) tags, followed by selection of 50 highest scoring (23) tags (tags 51–100). As a result, [01] and (01)(23)1.0 curves are largely identical for the first 50 tags. The plot shows that the gain in sensitivity by generating more tags of the same kind starts diminishing after the selection of the first 25 highest scoring tags. In fact, extracting 100 tags instead of 50 under [01] setting improved the sensitivity by less than 10%. On the other hand, the addition of (23) tags allows a much higher increase in the sensitivity by bringing in a new set of tags involving y^{++} and b^{++} fragment ions. The (03)(12)1.0 method, also shown in Fig. 4 starts slower due to suboptimal performance of (03) tags (b^{+} and y^{++} fragment ions) compared to (01) tags, but after addition of (12) tags (y^{+} and b^{++} ions) gives the best performance overall. Similar results were observed for 3+ spectra (data not shown).

The number of top scoring tags extracted from a spectrum in this work was fixed at 100, which is a default parameter in InsPecT. Figure 4 indicates that the number of tags can potentially be reduced to 50 without a substantial loss in the sensitivity. A range of 50–100 tags was also identified as an optimal range in the previous studies[13]. The optimal number of tags represents a trade-off between the sensitivity and the computational time, and thus is data set dependent. When the time of the analysis becomes the main consideration, e.g. in the case of very large datasets or phosphopeptide analysis, as low as 25 tags may be used. However, regardless of the total number of tags extracted from the spectrum, (01)(23)1.0 or (03)(12)1.0 methods provide a substantial improvement over [01] method in the case of MS/MS spectra of charge state 3+ and higher.

Effect of spectrum filtering—The results presented above were obtained using the default spectral filtering setting of InsPecT (peak depth 6). However, the effects of spectral filtering and inclusion of multi-charged fragment ions are interrelated, as discussed above. Inclusion of multi-charged ions is beneficial only if (1) their propensity is well above the noise level and (2) the propensity of singly charged ions is well below 1, since otherwise multi-charged ions would only add redundant information. Since increasing the peak depth has an effect of increased propensity of b^{+} and y^{+} ions (Table 1), the influence of multi-charged ions is expected to diminish with increasing peak depth. Table 6 shows the sensitivity of the [01] and (03)(12)1.0 tag generation methods for spectra of charge 3+ in the protein mix LTQ-FT data set (5 run subset also used in Table 3) as a function of peak depth:

retaining 3, 6, or 12 most intense peaks per 50 Da window. Indeed, increasing the peak depth results in increased sensitivity in the case of [01] tagging option. At peak depth 12, [01] option gives a sensitivity of 93.4%, compared with 91.1% when using the default peak depth of 6, and 84.5% with the peak depth reduced to 3. Thus, the sensitivity of the [01] method gets closer to, although does not reach, the 96.9% sensitivity achievable with the use of multi-charged ions.

Furthermore, using [01] tagging option and increasing the peak depth from 6 to 12 leads to more than a two-fold increase in the computational time due to doubling of the number of nodes in the spectrum graph. In contrast, and as mentioned above, the computational time of methods such (03)(12)1.0 remain similar to that of [01] method (at the same peak depth) since similar doubling in the number of nodes due to consideration of multi-charged ions is compensated by additional constraints on edge creation. Using [0][1] option with peak depth 12 (also shown in Table 6) gives computational times comparable to [01] and (03)(12)1.0 with default peak depth 6, but with a lower sensitivity of 92.9%.

Peptide Identification

Generation of sequence tags from MS/MS spectra, the main focus of this work, represents only the first step in the peptide identification process. The more relevant metrics of the performance are the number of identified peptides at a certain fixed false discovery rate (FDR)[14] and the overall computational time. These performance characteristics will be discussed in the remainder of the manuscript using the protein mix LTQ-FT and the phosphopeptide data sets.

After top scoring tags are generated for a set of spectra, they are organized in a trie structure,[40] and the protein database is scanned to get matches for each tag. If there is a match, the tag is extended from the matching region in both directions to attempt matching the tag’s prefix mass and suffix mass (allowing for user specified modifications). For all candidate peptides selected this way, the composite match quality score (MQScore) is computed using a combination of several features, including the number of b and y ion matched, the summed intensity of b and y peaks, and the number of tryptic termini (NTT). To distinguish between correct and incorrect peptide assignments, InsPecT calculates a probability score (referred to as p-value) using a parametric mixture modeling approach similar to that of PeptideProphet[4]. The p-value is computed as the complement (1 – probability) of the posterior probability that the assignment is correct determined from the distributions of the FScore among correct and incorrect identification. The FScore is calculated from MQScore and Delta MQScore in a way resembling the SEQUEST discriminate function[4].

Making a direct comparison between InsPecT and SEQUEST/PeptideProphet may not be in favor of InsPecT, especially in the case of LTQ-FT data, since PeptideProphet utilizes the mass accuracy of the precursor ion measurement dM, i.e. the difference between the measured and calculated precursor ion mass. This information is known to significantly improve the power to discriminate between correct and incorrect identifications. Thus, the InsPecT FScore and the mass accuracy dM distributions were also modeled using the recently described semi-parametric mixture modeling approach[27]. This model uses decoy peptides for the estimation of the negative mixture component, which removes parametric assumptions making it applicable to any distribution of scores without the need to worry about the shapes of the underlying distributions when fitting the data. The results of filtering the data using probabilities computed by the model were compared with that using InsPecT p-values (which do not take dM into account), as well as with SEQUEST/PeptideProphet.

Figure 5 shows the receiver operating characteristic (ROC) plots for the protein mix LTQ-FT data set. In the case of 4+ and 3+ spectra (Fig. 5a, b), four methods were compared: 1) SEQUEST/PeptideProphet (labeled 'SEQUEST+PP'); 2) Sequence tag based identification using singly charge (SC) fragment ions only, [01] method, followed by p-value filtering ('Sequence tagging (SC)') 3) Sequence tagging using multi-charged (MC) fragment ions, (03)(12)1.0 method, followed by p-value filtering ('Sequence tagging (MC)'); 4) Sequence tagging, (03)(12)1.0 method, followed by semi-parametric modeling with inclusion of the mass accuracy parameter dM ('Sequence tagging (MC) + Mass Acc'). Note that 'Sequence tagging (SC)' essentially represents the unmodified InsPecT program, except that the peptide ion charge state was taken from the mzXML file instead of using the charge determination function. In the case of 2+ spectra (Fig. 5c), the tags were generated using [01] method only.

Since the size of the decoy subset of the searched database was very large compared to the number of target sequences (proteins present in the protein mix sample), all identifications of non-decoy peptides were taken as correct. Peptide identifications were filtered using varying p-value or probability cut-offs, and the number of correct identifications and the false discovery rate (fraction of peptide assignments passing the threshold that are incorrect) were plotted for each cut-off. Only the most relevant region of FDR, i.e. less than 5%, is shown. Several trends are apparent. First, for 3+ and 4+ spectra, incorporation of multi-charged fragment ions in the sequence tagging algorithm significantly increases the number of identified peptides, especially in the case of 4+ spectra. Second, inclusion of the mass accuracy parameter provides additional improvement, helping the sequence tag-based method outperform SEQUEST/PeptideProphet.

The increase in the sensitivity due to inclusion of the mass accuracy parameter observed here is another example of how auxiliary information, i.e. variables other than the search scores, can significantly improve discrimination when incorporated in the statistical model[14, 33]. This is further illustrated in Figure 6, which plots the distributions of FScore (Fig. 6a) and dM (Fig. 6b) among correct and incorrect identifications for 3+ spectra. The correct identifications are clustered in a narrow range of dM values close to 0, whereas incorrect identifications are distributed across the entire 5 Da interval (the mass tolerance used by InsPecT tag extension algorithms in the process of selecting candidate peptides is 2.5 Da). As a result, the mass accuracy dM provides a boost in discriminating power in addition to the main FScore itself.

Figure 7 shows the results of a similar analysis performed using 2+ and 3+ spectra from the phosphopeptide-enriched data set. Since the searched protein database for this data set included an equal number of target and decoy sequences, the number of correct peptide identifications cannot be determined directly. Instead, for each probability score cut-off it was estimated as the number of matches to target sequences minus the number of matches to decoys. Similarly, FDR was estimated as the ratio of the number of decoy matches to the number of matches to target sequences. Although the LTQ mass accuracy is lower than that for LTQ-FT, the mass accuracy parameter dM still improved discrimination. Interestingly, in this data set the sequence tag-based method outperforms SEQUEST/PeptideProphet to a greater degree. One possible explanation for this is that allowing phosphorylation as a variable modification significantly increases the number of candidate peptides that have to be scored by SEQUEST, which in turn increases the likelihood that the correct peptide is masked (not a top scoring match) by one of the incorrect peptides. Restricting the set of candidate peptides to those sequences that contain one of the generated sequence tags brings its size back to manageable levels. Second, the InsPecT peptide scoring function was specifically optimized for phosphopeptide analysis[41], which is another advantage compared to the cross-correlation score of SEQUEST.

Conclusion

We presented an improved sequence tag generation method that directly incorporates multi-charged fragment ions often observed in MS/MS spectra of high charged state. Based on a comprehensive analysis, the method was optimized to improve the sensitivity without a substantial increase in the computational time. The improved performance was demonstrated with spectra collected using a protein mix sample on five different mass spectrometers (Thermo LTQ-FT, LTQ and LCQ, Waters/Micromass QTOF, and Agilent XCT), as well as using a complex phosphopeptide-enriched sample analyzed using an LTQ instrument. We also demonstrated that additional modeling of InsPecT search scores and inclusion of auxiliary information (e.g. mass accuracy) provides additional noticeable improvement in the sensitivity of peptide identification. In summary, this work confirms the potential of the sequence tagging as sensitive and computationally efficient peptide identification strategy, especially in the case of phosphopeptide analysis. The new tagging method was implemented as an extension of the InsPecT open-source program.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by NIH/NCI Grant CA-126239 and NIH/NCRR - National Resource for Proteomics and Pathways Grant #P41-18627. We thank Jignesh Patel, Pavel Pevzner, Kang Ning, and Samuel Payne for helpful discussions, Huyngwon Choi for help with semi-parametric modeling, and Peter Ulintz for providing scripts for generating ion statistics.

References

1. Hernandez P, Müller M, Appel RD. Automated protein identification by tandem mass spectrometry: Issues and strategies. Mass Spectrometry Reviews. 2006; 25(2):235–254. [PubMed: 16284939]
2. Nesvizhskii AI. Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol Biol. 2007; 367:87–119. [PubMed: 17185772]
3. Xu C, Ma B. Software for computational peptide identification from MS-MS data. Drug Discovery Today. 2006; 11(13–14):595–600. [PubMed: 16793527]
4. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. Anal Chem. 2002; 74(20):5383–5392. [PubMed: 12403597]
5. Bern M, Cai Y, Goldberg D. Lookup Peaks: A Hybrid of de Novo Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry. Anal Chem. 2007; 79(4):1393–1400. [PubMed: 17243770]
6. Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Communications in Mass Spectrometry. 2003; 17(20):2310–2316. [PubMed: 14558131]
7. Perkins, David N.; Pappin, DJC.; Creasy, DM.; Cottrell, JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20(18): 3551–3567. [PubMed: 10612281]
8. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry. 1994; 5(11):976–989.
9. Frank A, Pevzner P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. Anal Chem. 2005; 77(4):964–973. [PubMed: 15858974]
10. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open Mass Spectrometry Search Algorithm. J Proteome Res. 2004; 3(5):958–964. [PubMed: 15473683]

11. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Communications in Mass Spectrometry. 2003; 17(20):2337–2342. [PubMed: 14558135]

12. Tabb DL, Saraf A, Yates JR. GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model. Anal Chem. 2003; 75(23):6415–6421. [PubMed: 14640709]

13. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra. Anal Chem. 2005; 77(14):4626–4639. [PubMed: 16013882]

14. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nature Methods. 2007; 4(10):787–797. [PubMed: 17901868]

15. Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Communications in Mass Spectrometry. 1997; 11(9):1067–1075. [PubMed: 9204580]

16. Taylor JA, Johnson RS. Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry. Anal Chem. 2001; 73(11):2594–2604. [PubMed: 11403305]

17. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De Novo Peptide Sequencing via Tandem Mass Spectrometry. Journal of Computational Biology. 1999; 6(3–4):327–342. [PubMed: 10582570]

18. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem. 1994; 66(24):4390–4399. [PubMed: 7847635]

19. Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. J Bioinform Comput Biol. 2005; 3(3):697–716. [PubMed: 16108090]

20. Searle BC, Dasari S, Wilmarth PA, Turner M, Reddy AP, David LL, Nagalla SR. Identification of Protein Modifications Using MS/MS de Novo Sequencing and the OpenSea Alignment Algorithm. J Proteome Res. 2005; 4(2):546–554. [PubMed: 15822933]

21. Sunyaev S, Liska AJ, Golod A, Shevchenko A, Shevchenko A. MultiTag: Multiple Error-Tolerant Sequence Tag Search for the Sequence-Similarity Identification of Proteins by Mass Spectrometry. Anal Chem. 2003; 75(6):1307–1315. [PubMed: 12659190]

22. Savitski MM, Nielsen ML, Zubarev RA. New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. Molecular & Cellular Proteomics. 2005; 4(8):1180–1188. [PubMed: 15911534]

23. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA. The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Molecular & Cellular Proteomics. 2007; 6(9):1638–1655. [PubMed: 17533153]

24. Frank A, Tanner S, Bafna V, Pevzner P. Peptide Sequence Tags for Fast Database Search in Mass-Spectrometry. J Proteome Res. 2005; 4(4):1287–1295. [PubMed: 16083278]

25. Chong KF, Ning K, Leong HW, Pevzner P. Modeling and characterization of multi-charge mass spectra for Peptide sequencing. J Bioinform Comput Biol. 2006; 4(6):1329–52. [PubMed: 17245817]

26. Ning, K.; Chong, KF.; Leong, HW. A Database Search Algorithm for Identification of Peptides with Multiple Charges Using Tandem Mass Spectrometry. In: Li, J.; Yang, Q.; Tan, A-H., editors. BioDM Singapore 2006. Springer-Verlag Berlin Heidelberg; Singapore: 2006. p. 2-13.

27. Choi H, Ghosh D, Nesvizhskii AI. Statistical Validation of Peptide Identifications in Large-scale Proteomics using Target-Decoy Database Search Strategy and Flexible Mixture Modeling. J Proteome Res. 2008; 7:286–292. [PubMed: 18078310]

28. Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, Letarte S, Gafken PR, Katz JE, Mallick P, Lee H, Schmidt A, Ossola R, Eng JK, Aebersold R, Martin DB. The Standard Protein Mix Database: A Diverse Data Set To Assist in the Production of Improved Peptide and Protein Identification Software Tools. J Proteome Res. 2008; 7:96–103. [PubMed: 17711323]

29. Ulintz PJ, Bodenmiller B, Andrews PC, Aebersold R, Nesvizhskii AI. Investigating MS2-MS3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence. Mol Cell Proteomics. 2008; 7:71–87. [PubMed: 17872894]

30. Colinge J, Magnin J, Dessingy T, Giron M, Masselot A. Improved peptide charge state assignment. PROTEOMICS. 2003; 3(8):1434–1440. [PubMed: 12923768]

31. Sadygov RG, Eng J, Durr E, Saraf A, McDonald H, MacCoss MJ, Yates JR. Code Developments to Improve the Efficiency of Automated MS/MS Spectra Interpretation. J Proteome Res. 2002; 1(3):211–215. [PubMed: 12645897]

32. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. Anal Chem. 2003; 75(17):4646–4658. [PubMed: 14632076]

33. Choi H, Nesvizhskii AI. Semi-supervised Model-based Validation of Peptide Identifications in Mass Spectrometry-based Proteomics. J Proteome Res. 2008; 7:254–265. [PubMed: 18159924]

34. Martin DB, Eng JK, Nesvizhskii AI, Gemmill A, Aebersold R. Investigation of neutral loss during collision-induced dissociation of peptide ions. Analytical Chemistry. 2005; 77(15):4870–4882. [PubMed: 16053300]

35. Huang YY, Triscari JM, Tseng GC, Pasa-Tolic L, Lipton MS, Smith RD, Wysocki VH. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. Analytical Chemistry. 2005; 77(18):5800–5813. [PubMed: 16159109]

36. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R. Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data: Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides. Mol Cell Proteomics. 2006; 5(4): 652–670. [PubMed: 16352522]

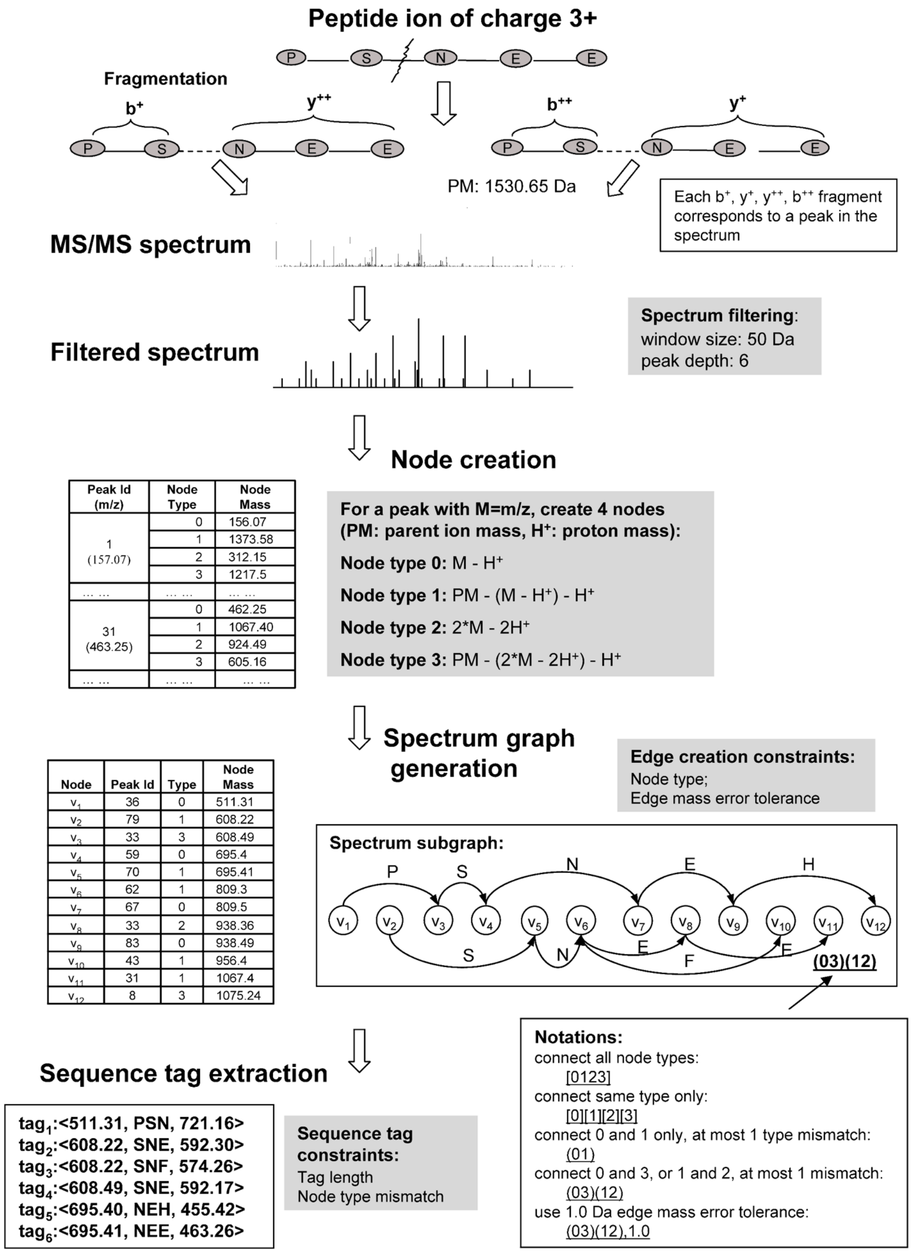
37. Tanner, SW. PhD Dissertation. University of California; San Diego: 2007. Efficient and Accurate Bioinformatics Algorithms for Peptide Mass Spectrometry.

38. Gentzel M, Kocher T, Ponnusamy S, Wilm M. Preprocessing of tandem mass spectrometric data to support automatic protein identification. Proteomics. 2003; 3(8):1597–1610. [PubMed: 12923784]

39. Mujezinovic N, Raidl G, Hutchins JRA, Peters JM, Mechtler K, Eisenhaber F. Cleaning of raw peptide MS/MS spectra: Improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. Proteomics. 2006; 6(19):5117–5131. [PubMed: 16955515]

40. Aho A, Corasick M. Efficient string matching: an aid to bibiographic search. Communications of the ACM. 1975; 18:333–340.

41. Payne SH, Yau M, Smolka MB, Tanner S, Zhou H, Bafna V. Phosphorylation-Specific MS/MS Scoring for Rapid and Accurate Phosphoproteome Analysis. Journal of Proteome Research. 2008



Each b^+, y^+, y^{++}, b^{++} fragment corresponds to a peak in the spectrum

Spectrum filtering:
window size: 50 Da
peak depth: 6

Peak Id (m/z)	Node Type	Node Mass
1 (157.07)	0	156.07
	1	1373.58
	2	312.15
	3	1217.5
… …	… …	… …
31 (463.25)	0	462.25
	1	1067.40
	2	924.49
	3	605.16
… …	… …	… …

For a peak with M=m/z, create 4 nodes (PM: parent ion mass, H^+: proton mass):

Node type 0: $M - H^+$

Node type 1: $PM - (M - H^+) - H^+$

Node type 2: $2*M - 2H^+$

Node type 3: $PM - (2*M - 2H^+) - H^+$

Edge creation constraints:
Node type;
Edge mass error tolerance

Node	Peak Id	Type	Node Mass
v_1	36	0	511.31
v_2	79	1	608.22
v_3	33	3	608.49
v_4	59	0	695.4
v_5	70	1	695.41
v_6	62	1	809.3
v_7	67	0	809.5
v_8	33	2	938.36
v_9	83	0	938.49
v_{10}	43	1	956.4
v_{11}	31	1	1067.4
v_{12}	8	3	1075.24

Notations:
connect all node types:
[0123]
connect same type only:
[0][1][2][3]
connect 0 and 1 only, at most 1 type mismatch:
(01)
connect 0 and 3, or 1 and 2, at most 1 mismatch:
(03)(12)
use 1.0 Da edge mass error tolerance:
(03)(12),1.0

tag_1:<511.31, PSN, 721.16>
tag_2:<608.22, SNE, 592.30>
tag_3:<608.22, SNF, 574.26>
tag_4:<608.49, SNE, 592.17>
tag_5:<695.40, NEH, 455.42>
tag_6:<695.41, NEE, 463.26>

Sequence tag constraints:
Tag length
Node type mismatch

Figure 1.

Overview of the method and different tag generations settings.

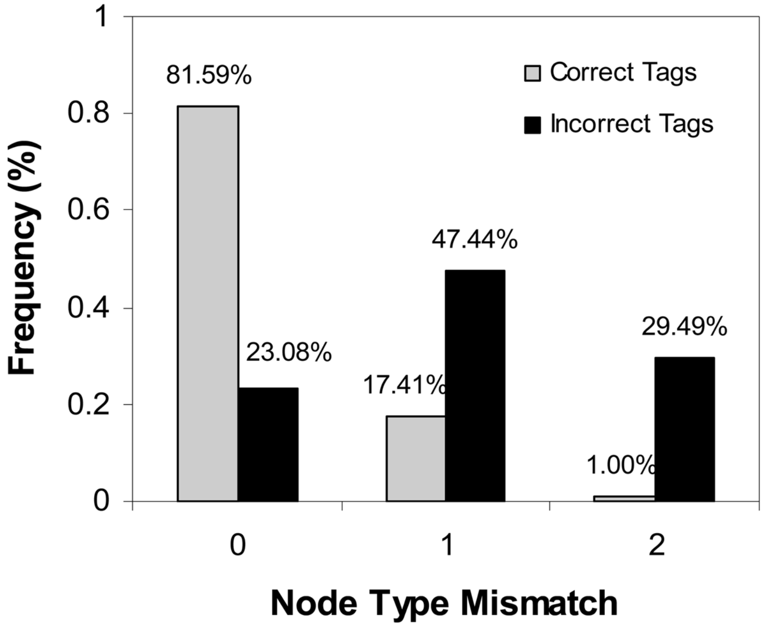


Figure 2.

The frequency of observing one the two highest scoring correct tags or one of the 100 randomly selected incorrect tags extracted from 3+ spectra in phosphopeptide-enriched data set having 0, 1, or 2 mismatches of node type.

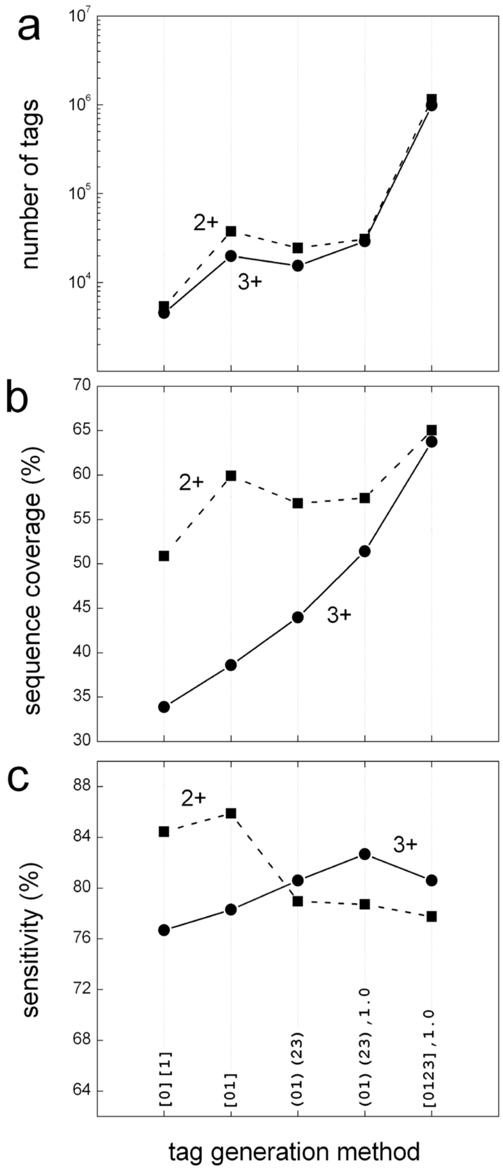


Figure 3.

Optimal sensitivity as a trade-off between the total number of sequence tags that can be extracted from the spectrum graph and the coverage of the peptide sequence by generated correct tags. **(a)** Average number of sequence tags (correct or incorrect) that can be extracted from 2+ (dashes) or 3+ (solid curve) spectra in the phosphopeptide data set using different tag generation options. **(b)** Coverage of the peptide sequence by generated correct tags. (c) Sensitivity, i.e. the percentage of spectra for which one of the top 100 scoring tags is correct

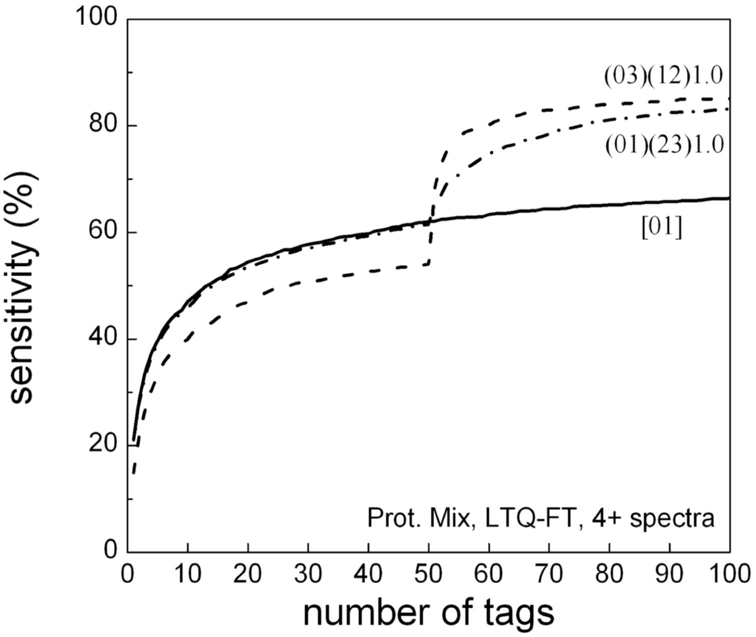


Figure 4.

Sensitivity of different tag extraction methods as a function of the number of top scoring tags considered for each spectrum (up to 100, the default value) in the protein mix LTQ-FT 4+ charge state spectra data set.

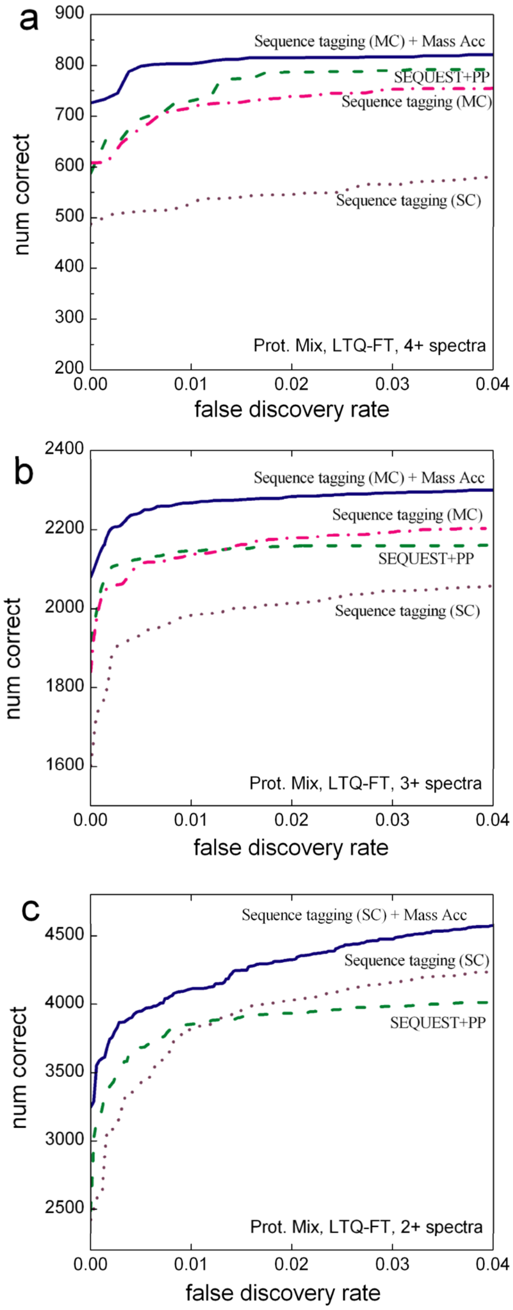


Figure 5.

The number of correct identifications as a function of false discovery rate for spectra from the protein mix LTQ-FT data set. **(a)** 4+ spectra. Shown are the results of SEQUEST/PeptideProphet analysis (dashed green curve); sequence tagging using singly charge (SC) fragment ions only, [01] method, followed by p-value filtering (dotted purple curve); sequence tagging using multi-charged (MC) fragment ions, (03)(12)1.0 method, followed by p-value filtering (dash dot magenta curve); sequence tagging, (03)(12)1.0 method, followed by semi-parametric modeling with inclusion of the mass accuracy parameter dM (solid blue curve). **(b)** 3+ spectra, same as above **(c)** 2+ spectra. The tags were generated using [01] method only.

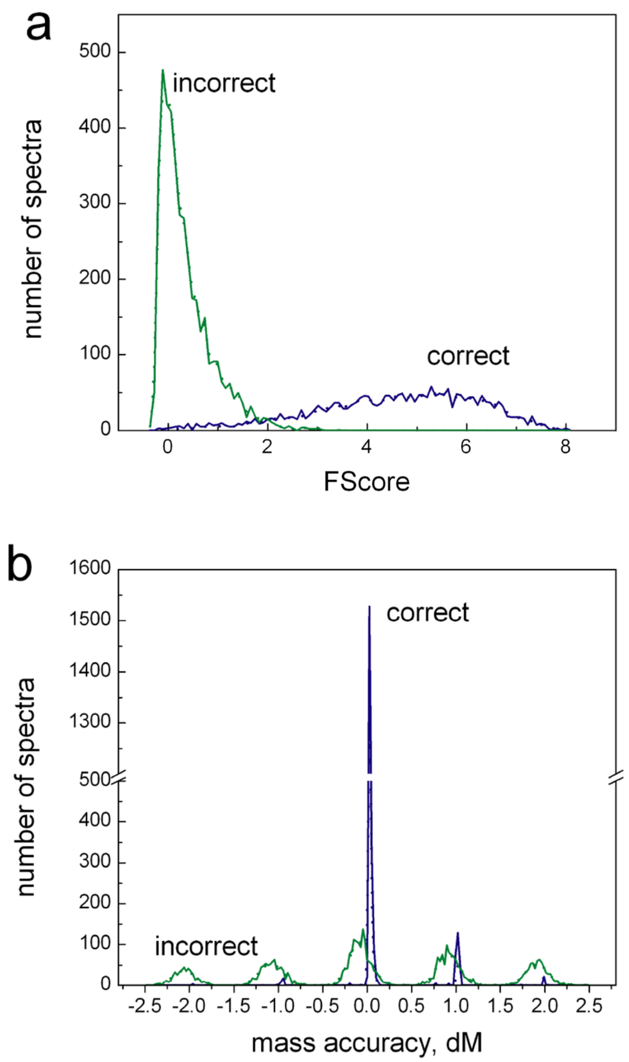


Figure 6.

Histogram of InsPecT **(a)** FScore and **(b)** mass accuracy score dM plotted separately for correct (solid blue) and incorrect (solid green) peptide identifications for 3+ spectra in the protein mix LTQ-FT data set. Also shown are distributions fitted by the semi-parametric model (dashed curves).

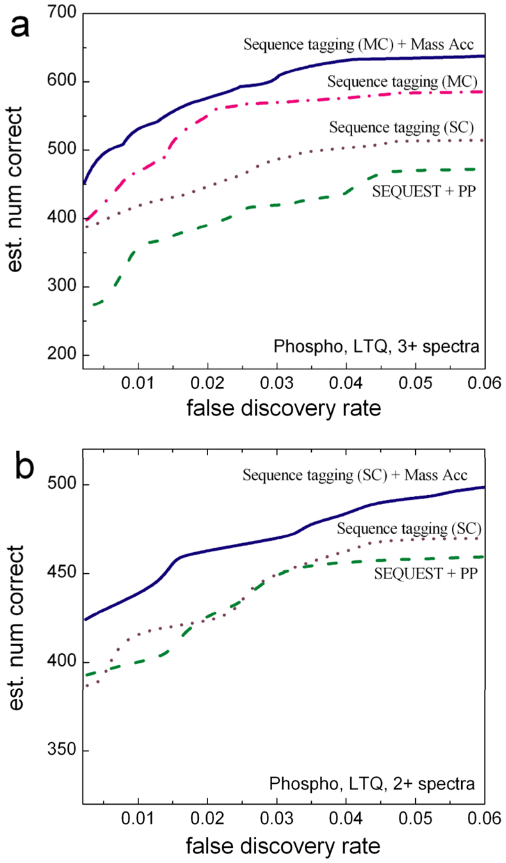


Figure 7.

The estimated number of correct identifications as a function of false discovery rate in the phosphopeptide data set **(a)** 3+ spectra, SEQUEST/PeptideProphet analysis (green dashed curve), sequence tagging using singly charge (SC) fragment ions only, [01] method, followed by p-value filtering (dotted purple curve); sequence tagging using multi-charged (MC) fragment ions, (03)(12)1.0 method, followed by p-value filtering (dash dot magenta curve); sequence tagging, (03)(12)1.0 method, followed by semi-parametric modeling with inclusion of the mass accuracy parameter dM (solid blue curve). **(b)** 2+ spectra. The tags were generated using [01] method only.

Table 1

Fragment ion statistics for protein mix LTQ-FT dataset. Statistics computed separately for peptides identified from MS/MS spectra of different charge state and at different peak depth: using 3, 6, or 12 most intense peaks per 50 Da window. For each ion type shown are the average number of ions observed per spectrum, the percentage of total assigned peaks labeled as given type, and the propensity (percentage of all possible ions of a given type that were observed in the spectrum on average).

Peak depth	**3**			**6**			**12**		
Ion type	**# ions**	**%**	**propensity**	**# ions**	**%**	**propensity**	**# ions**	**%**	**propensity**
				charge 2+					
a	0.84	3.06	0.085	1.52	3.42	0.15	2.56	4.14	0.2425
a-h2o	0.62	2.27	0.061	1.27	2.86	0.12	2.37	3.83	0.222
a-nh3	0.91	3.29	0.087	1.59	3.58	0.15	2.39	3.86	0.2217
b	**5.65**	**20.5**	**0.512**	**7.14**	**16.1**	**0.64**	**7.81**	**12.62**	**0.703**
b-h2o	2.07	7.52	0.185	3.79	8.53	0.34	5.13	8.28	0.4514
b-h2o-h2o	0.67	2.42	0.06	1.56	3.5	0.14	2.69	4.35	0.2343
b-h2o-nh3	0.94	3.42	0.083	2.22	4.99	0.19	3.64	5.89	0.3193
b-nh3	1.55	5.62	0.138	3.45	7.77	0.3	4.83	7.81	0.4253
b2	**0.58**	**2.12**	**0.053**	**1.1**	**2.48**	**0.1**	**1.95**	**3.15**	**0.1731**
b2-h2o	0.02	0.07	0.002	0.07	0.16	0.01	0.19	0.3	0.0168
b2-nh3	0.08	0.28	0.007	0.21	0.47	0.02	0.55	0.89	0.0473
b2-nh3-h2o	0.08	0.28	0.007	0.22	0.5	0.02	0.57	0.92	0.0497
b3	0.31	1.13	0.029	0.65	1.47	0.06	1.16	1.87	0.1053
y	**8.05**	**29.3**	**0.727**	**8.7**	**19.6**	**0.79**	**8.98**	**14.51**	**0.8123**
y-h2o	1.33	4.85	0.127	3.57	8.04	0.33	5.53	8.94	0.4998
y-nh3	1.17	4.25	0.11	2.41	5.42	0.22	3.29	5.31	0.2953
y2	**1.65**	**5.99**	**0.152**	**2.69**	**6.06**	**0.24**	**3.88**	**6.26**	**0.3488**
y2-h2o	0.02	0.09	0.002	0.09	0.19	0.01	0.23	0.37	0.0198
y2-nh3	0.46	1.67	0.04	0.99	2.24	0.09	1.65	2.66	0.1452
y2-nh3-h2o	0.16	0.6	0.015	0.51	1.15	0.05	1.28	2.07	0.1135
y3	0.35	1.28	0.033	0.68	1.53	0.06	1.21	1.96	0.1111
				charge 3+					

Peak depth	**3**			**6**			**12**		
Ion type	**# ions**	**%**	**propensity**	**# ions**	**%**	**propensity**	**# ions**	**%**	**propensity**
a	0.71	2.2	0.042	1.22	2.31	0.07	2.17	2.7	0.1233
a-h2o	0.54	1.67	0.031	1.13	2.13	0.06	2.16	2.68	0.1208
a-nh3	0.64	2	0.036	1.18	2.22	0.07	2.01	2.49	0.1122
b	**5.04**	**15.7**	**0.285**	**6.75**	**12.7**	**0.38**	**8.09**	**10.02**	**0.4505**
b-h2o	1.01	3.13	0.056	2.21	4.16	0.12	3.68	4.57	0.207
b-h2o-h2o	0.46	1.42	0.025	0.96	1.81	0.05	1.95	2.42	0.1072
b-h2o-nh3	0.73	2.28	0.041	1.48	2.78	0.08	2.76	3.42	0.1541
b-nh3	0.88	2.74	0.05	2.11	3.97	0.12	3.67	4.55	0.206
b2	**3.39**	**10.5**	**0.182**	**5.34**	**10.1**	**0.29**	**7.51**	**9.31**	**0.4115**
b2-h2o	0.07	0.2	0.004	0.27	0.5	0.01	0.65	0.8	0.0338
b2-nh3	1.2	3.72	0.065	2.43	4.57	0.13	4.02	4.98	0.2201
b2-nh3-h2o	0.26	0.8	0.015	0.91	1.72	0.05	2.54	3.15	0.1397
b3	0.45	1.4	0.024	0.98	1.85	0.05	2.01	2.49	0.1101
y	**7.59**	**23.6**	**0.431**	**8.8**	**16.6**	**0.5**	**9.43**	**11.69**	**0.5308**
y-h2o	1	3.1	0.058	2.45	4.61	0.14	4.35	5.39	0.2495
y-nh3	0.64	1.98	0.038	1.55	2.92	0.09	2.61	3.24	0.1484
y2	**4.95**	**15.4**	**0.279**	**7.02**	**13.2**	**0.39**	**8.99**	**11.15**	**0.498**
y2-h2o	0.08	0.24	0.004	0.35	0.66	0.02	0.8	0.99	0.041
y2-nh3	1.31	4.07	0.077	3.15	5.92	0.18	5.16	6.4	0.2869
y2-nh3-h2o	0.29	0.89	0.016	1.1	2.07	0.06	3.06	3.79	0.1734
y3	0.96	2.98	0.054	1.72	3.25	0.1	3.04	3.77	0.1696
				charge 4+					
a	0.67	2.18	0.029	1.09	2.07	0.05	2	2.43	0.0842
a-h2o	0.44	1.42	0.018	1.05	2	0.04	2.09	2.53	0.0852
a-nh3	0.58	1.88	0.025	1.06	2.02	0.05	1.77	2.15	0.0743
b	**3.67**	**11.9**	**0.151**	**5.21**	**9.9**	**0.21**	**6.74**	**8.18**	**0.2763**
b-h2o	0.43	1.4	0.018	1.2	2.28	0.05	2.37	2.87	0.0968
b-h2o-h2o	0.33	1.08	0.013	0.72	1.37	0.03	1.55	1.88	0.0634
b-h2o-nh3	0.65	2.1	0.029	1.25	2.37	0.05	2.33	2.82	0.098

Peak depth	**3**			**6**			**12**		
Ion type	**# ions**	**%**	**propensity**	**# ions**	**%**	**propensity**	**# ions**	**%**	**propensity**
b-nh3	0.64	2.07	0.026	1.43	2.72	0.06	2.72	3.3	0.1107
b2	**4.09**	**13.2**	**0.164**	**6.5**	**12.4**	**0.26**	**9.31**	**11.28**	**0.3771**
b2-h2o	0.08	0.25	0.003	0.24	0.46	0.01	0.65	0.79	0.0263
b2-nh3	0.86	2.79	0.035	2.16	4.1	0.09	4.06	4.93	0.1648
b2-nh3-h2o	0.28	0.91	0.012	1.03	1.96	0.04	2.64	3.2	0.111
b3	**1.68**	**5.44**	**0.065**	**3**	**5.7**	**0.12**	**5.01**	**6.08**	**0.2016**
y	**5.38**	**17.4**	**0.227**	**7.01**	**13.3**	**0.29**	**8.18**	**9.92**	**0.3406**
y-h2o	0.71	2.29	0.031	1.73	3.29	0.07	3.18	3.86	0.1355
y-nh3	0.43	1.38	0.018	1.02	1.93	0.04	1.89	2.29	0.079
y2	**5.89**	**19.1**	**0.245**	**8.46**	**16.1**	**0.35**	**10.7**	**12.98**	**0.4402**
y2-h2o	0.1	0.32	0.004	0.38	0.73	0.02	0.92	1.11	0.0366
y2-nh3	1.09	3.52	0.047	2.84	5.39	0.12	5.02	6.09	0.2096
y2-nh3-h2o	0.4	1.29	0.017	1.12	2.12	0.05	2.88	3.49	0.1209
y3	**2.49**	**8.05**	**0.104**	**4.14**	**7.87**	**0.17**	**6.45**	**7.83**	**0.2645**

Table 2

Sensitivity analysis of different tag generation methods for spectra of charge 3+ from phosphopeptide-enriched data set.

	Sensitivity (%)							
Charge state option	**# of Spectra**	**[01]**	**(03)(12)**	**(01)(23)**	**(03)(12), 1.0**	**(01)(23), 1.0**	**[0][1] [2][3]**	**(0123)**
Forced	457	76.6	81.4	80.3	81.6	82.5	77.0	80.3
Determined	433	77.6	82.5	80.6	82.7	82.7	78.3	81.1

Table 3

Sensitivity analysis of different tag generations methods for spectra of charge 3+ acquired using different mass spectrometry instruments.

	Sensitivity (%)					
Instrument	**# of Spectra**	**[01]**	**(03)(12)**	**(01)(23)**	**(03)(12), 1.0**	**(01)(23), 1.0**
LTQ-FT	4206	91.1	96.5	96.0	96.9	96.5
QTOF	325	89.2	92.6	95.7	92.6	93.5
LTQ	798	92.9	94.6	95.6	94.7	95.7
LCQ	507	76.5	82.6	79.9	83.4	82.6
Agilent	1940	76.1	78.9	79.7	79.8	81.4

Table 4

Sensitivity analysis of different tag generation methods for spectra of charge 3+ in protein mix Agilent data set. Spectra were divided into two groups based on the identification confidence score: PeptideProphet probability between 0.1 and 0.9 (lower quality spectra) and equal or greater than 0.9 (higher quality).

		Sensitivity (%)	
Quality	**# of Spectra**	**[01]**	**(01)(23), 1.0**
lower	409	63.3	73.8
higher	1176	83.3	87.1

Table 5

Sensitivity analysis of different tag generation methods for spectra of charge 4+ in protein mix LTQ-FT data set.

	Sensitivity (%)									
File	**# of Spectra**	**[01]**	**[0][1] [2][3]**	**(03)(12)**	**(01)(23)**	**(03)(12), 1.0**	**(01)(23), 1.0**	**(0123)**	**[0][1][2] [3][4][5]**	**(012345)**
B06_11071	152	65.8	80.9	87.5	82.2	87.5	83.6	84.8	80.9	82.9
B06_11072	172	68.0	79.7	85.5	84.9	87.2	86.1	81.4	79.1	80.8
B06_11073	153	70.6	83.0	87.6	85.0	88.9	86.3	86.9	83.0	86.9
B06_11074	156	63.5	77.6	83.3	84.6	85.3	85.3	82.1	77.6	81.4
B06_11075	155	68.4	78.1	82.6	80.7	85.8	83.2	82.6	78.1	81.9
B06_11076	144	63.2	78.5	84.7	82.0	84.7	81.9	78.5	79.2	80.6
B06_11077	149	69.1	77.9	84.6	81.2	85.2	83.2	81.9	77.9	83.9
B06_11078	138	66.7	73.2	78.3	79.7	79.7	80.4	76.8	73.9	77.5
B06_11079	139	64.8	81.4	89.2	88.4	88.4	87.6	86.1	81.4	87.6
B06_11080	126	63.5	73.0	83.3	80.2	83.3	79.4	79.4	73.0	81.0
Combined	1484	66.4	77.9	84.1	82.4	85.1	83.2	81.5	78.0	81.9

Table 6

Effect of spectrum filtering on sequence tag generation. Sensitivity analysis of three different tag generation methods for spectra of charge 3+ in protein mix LTQ-FT data set (4206 spectra in total). Spectra were filtered using different peak depth: retaining 3, 6, or 12 most intense peaks per 50 Da window.

	Sensitivity (%)		
Peak depth	**[01]**	**[0][1]**	**(03)(12), 1.0**
3	84.5	79.2	93.5
6	91.1	89.8	96.9
12	93.4	92.9	96.9