Published in final edited form as:
Proteomics. 2010 October ; 10(19): 3515–3524. doi:10.1002/pmic.200900727.

Integrated data management and validation platform for phosphorylated tandem mass spectrometry data

Anna-Maria Lahesmaa-Korpinen[1], **Scott M. Carlson**[2], **Forest M. White**[2,3], and **Sampsa Hautaniemi**[1,*]

[1] Genome-Scale Biology Program, Institute of Biomedicine, University of Helsinki, Haartmaninkatu 8, Helsinki 00014, Finland [2] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [3] Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract

Tandem mass spectrometry (MS/MS) is a widely used method for proteome-wide analysis of protein expression and post-translational modifications (PTMs). The thousands of MS/MS spectra produced from a single experiment pose a major challenge for downstream analysis. Standard programs, such as Mascot, provide peptide assignments for many of the spectra, including identification of PTM sites, but these results are plagued by false positive identifications. In phosphoproteomics experiments only a single peptide assignment is typically available to support identification of each phosphorylation site, so minimizing false positives is critical. Thus, tedious manual validation is often required to increase confidence in the spectral assignments.

We have developed phoMSVal, an open-source platform for managing MS/MS data and automatically validating identified phosphopeptides. We tested five classification algorithms with 17 extracted features to separate correct peptide assignments from incorrect ones using over 3000 manually curated spectra. The naive Bayes algorithm was among the best classifiers with an area under the ROC curve value of 97% and positive predictive value of 97% for phosphotyrosine data. This classifier required only three features to achieve a 76% decrease in false positives as compared to Mascot while retaining 97% of true positives. This algorithm was able to classify an independent phosphoserine/threonine dataset with area under ROC curve value of 93% and positive predictive value of 91%, demonstrating the applicability of this method for all types of phospho-MS/MS data. PhoMSVal is available at http://csbi.ltdk.helsinki.fi/phomsval

Keywords

bioinformatics; data management; feature selection; machine learning; phosphoproteomics

1 Introduction

Liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) enables the quantitative analysis of protein expression and the site-specific analysis of protein post-translational modification (PTM) on a proteome-wide scale. For example, mass

Correspondence: Sampsa Hautaniemi, DTech, Biomedicum Helsinki, Rm. B524b, PO Box 63, 00014 University of Helsinki, Finland, sampsa.hautaniemi@helsinki.fi, Fax: +358 9 191 25610.

The authors have declared no conflict of interest.

spectrometry based phosphoproteomics has been implemented in many labs, and has been applied to many biological systems, including yeast, mammalian cell culture and mammalian tissues [1,2,3]. In a typical experiment, proteins from biological sample(s) are digested to peptides and fractionated or enriched for particular PTMs, including phosphorylation or acetylation. Enrichment methods for phosphorylation include immobilized metal affinity chromatography (IMAC) [1], titanium dioxide (TiO_2) [4], or phosphospecific immunoprecipitation [5] prior to LC-MS/MS analysis. Often this analysis is combined with an isotopic labeling step, e.g. table isotope labeling with amino acids in cell culture (SILAC) [6], or isobaric tags for relative and absolute quantitation (iTRAQ) [7] to enable relative quantification of phosphorylation or protein expression across multiple biological samples.

For these experiments, a single LC-MS/MS analysis will typically yield thousands of spectra that are used to identify the associated peptides and any sites of PTMs. Given the large number of MS/MS spectra, programs such as Mascot [8], SEQUEST [9], and ProteinPilot [10] are typically used to compare the spectra to databases of peptides from the appropriate species, thereby automating the process of peptide identification and PTM site assignment.

Although the scoring system and database comparison method employed by each identification algorithm are different, each program generates a list of peptide assignments with corresponding scores. It is typically left to the user to decide the proper cutoff to minimize false positive and false negative identifications. A common strategy for choosing the threshold is to search spectra against reverse- or randomized sequence databases [11,12,13]. It has been shown, however, that there are several types of false positive spectra that cannot be identified using this method due to similarity of false positives and their matched actual spectra [14]. This result indicates that even after using a peptide identification program, the resulting MS/MS dataset consists of significant fraction of false positives.

Since the ultimate goal of most phosphoproteomic experiments is to identify phosphorylation sites which regulate biological processes, the fact that some percentage of the peptide assignments are incorrect can be very costly, especially since it may take months or years of experimentation to determine the biological role of any individual phosphorylation site. In order to improve on the accuracy of peptide and PTM assignments it is possible to manually validate MS/MS spectra [14,15], but this process is tedious and potentially error-prone (depending to a great extent on the experience of the person performing the validation). In order to improve the selection of accurate phospho-MS/MS assignments, and thereby decrease false positive identifications, we have developed an approach that automatically validates the peptide identification.

A number of quality prediction methods for mass spectrometry have been introduced. These can be divided roughly into two categories [16]: *a priori* approaches analyze spectrum quality before applying peptide identification software, thereby eliminating poor quality spectra prior to database searching; while *a posteriori* approaches make the quality assessment after peptide identification, and can therefore evaluate the quality of the spectrum in the context of a given peptide assignment. InsPecT is an example of *a priori* approach and it combines local *de novo* sequencing and filtering with sequence tags to reduce the size of the searched database, resulting in faster and more accurate peptide identifications [17]. Since *a priori* methods use only features directly extracted from the spectra [16,18], features that depend on the peptide assignment cannot be used. One such feature, introduced here, is the number of peaks that are not assigned to an expected fragment ion. Our results show that this is a key feature for assessing phospho-MS/MS spectrum assignments.

Algorithms that identify the positions of phosphorylation sites within a peptide after peptide identification generally function by assigning scores for each possible arrangement of phosphorylation sites [19,20]. For instance, the Ascore algorithm for phosphorylation site assignment quantifies the probability of the correct phosphorylation site based on the presence of site-determining ions in the spectrum [19]. Another tool for phosphorylation site assignment, PhosphoScore, uses a tree algorithm to create all possible phosphorylated versions of a peptide and then matches the experimental spectrum to these theoretical peptide sequences to find the most likely phosphorylation sites [20]. In addition, machine learning methods that use the peptide sequence to calculate features such as similarity to known sequences, predicted protein structure and sequence conservation have been developed recently [21,22]. Lu and colleagues developed a support vector machine (SVM) based method, DeBunker, with features extracted from the spectral data and peak identification information to reduce the false positive rate of phosphorylation site identification to 2% from approximately 5% with decoy database searching [23]. These methods, however, depend on having a correct initial peptide assignment, and do not directly address the question of separating correct from incorrect assignments.

In order to facilitate preprocessing and downstream analysis of phosphorylated LC-MS/MS data, we have implemented phoMSVal for management and automated validation of data from tandem mass spectrometry experiments. PhoMSVal imports data into a MySQL relational database, extracts features for classification, and assigns a classification label to each spectrum designating whether the given peptide assignment is likely to be correct. As success of a prediction algorithm depends on the features used, we characterized the impact of 17 quality features in discriminating correct assignments. Further, we used five different classification algorithms for all combinations of features. Our results demonstrate the optimal combination of features required for evaluating assignments and show that correct and incorrect assignments can be discriminated with excellent specificity and sensitivity, thus reducing the need for manual validation of spectra.

2 Methods

A single MS/MS experiment can easily produce thousands of spectra. In order to facilitate management of these data we have implemented a library of Python scripts, phoMSVal, for systematic storing and retrieval of spectra, automatic feature extraction and evaluation of phosphopeptide assignments, resulting in automation of manual validation. Included is a graphical user interface, where the user can select the classifier, select the dataset to classify, import new data and get results of the classification. The overall schematic of our approach is illustrated in Figure 1.

Briefly, data for MS/MS spectra consisting of peak lists (*m/z* and intensity values), peptide assignment, and assignment score are stored to a MySQL database along with metadata describing the experiment. Peaks within each spectrum are matched to expected fragments from the assigned peptide, and quality features for each spectrum are extracted to the database for input to a classification algorithm. PhoMSVal is open-source and can be downloaded with user manual at http://csbi.ltdk.helsinki.fi/phomsval.

We have three major objectives: (i) implement a relational database for phospho-MS/MS data storing, retrieval and analysis, (ii) analyze the relevance of 17 features in classification, and (iii) construct a classifier using the proper features that can reliably separate incorrect assignments from correct assignments, thus speeding up the manual validation of targets.

2.1 Data handling and database

Due to the importance and complexity of mass spectrometry data, a need for systematic data management platforms has been recognized [24,25]. However, no single system has gained widespread usage due to their complexity [26]. Here, we use an easy-to-use and flexible MySQL relational database with a graphical user interface for handling LC-MS/MS data (Supplementary Material figures S1 and S2). The database contains five tables responsible for storing data about the spectra and a sixth table that allows annotating the peptides to proteins. Peptides may map to different protein isoforms, which may result in incorrect annotations [27]. As a default, the annotations in PhoMSVal are taken directly from the search engine results. However, if more informative peptide annotations are known, such as peptide classifications described in [28], these can be uploaded to the database and used to annotate the peptides.

2.2 Description of datasets

Eleven phosphotyrosine peptide datasets were collected from cell culture experiments using a range of cell types and stimulations. Each experiment included four cell types and/or stimulation conditions labeled with isobaric mass tags for relative and absolute quantitation (iTRAQ, Applied Bioystems). Three datasets consisted of lysates of lung cancer cell lines (H529, H2073, H2122 and Calu-6) (ATCC), four datasets consisted of MCF7 breast cancer cells over-expressing HER2 and/or with tamoxifen resistance induced by long-term low-dose exposure and four datasets consisted of breast cancer lines T47D, A549 and Met2a (with or without c-Met over-expression).

As an independent validation data set, we used a phosphoserine/threonine dataset. This dataset was collected from rat liver tissue as described in [3]. All datasets are available at the phoMSVal website and ProteomeCommons (https://proteomecommons.org/dataset.jsp?i=74545).

2.2.1 Sample processing—For each of the cell lines, the cells were grown to 80% confluence in 15-cm cell culture plates. The lung cancer cell lines (H529, H2073, H2122 and Calu-6) (ATCC) grown to 80% confluence and serum-starved overnight prior to lysis. Each MCF7 variant was serum-starved overnight and then stimulated with heregulin for 0, 5, 15 and 30 minutes prior to lysis and labeling with iTRAQ. The breast cancer cell lines (T47D, A549 and Met2a) were serum-starved overnight and then stimulated for 0, 5, 15 or 30 minutes with hepatocyte growth factor. Following stimulation they were lysed in 3mL 8M urea + 1mM sodium orthovanadate. Protein content was determined using the bicinchoninic acid protein assay (Pierce). Lysates were diluted treated with 10mM dithiothreitol for 1 hour at 56°C, then cysteines were blocked with 55mM iodoacetamide for 45 minutes at room temperature on a rotor in foil. Lysates were diluted to 2M urea in 100mM ammonium acetate pH 8.9 and digested overnight with sequencing-grade trypsin (Promega) 1:100 w/w overnight at room temperature with rotation. Digested lysates were acidified with 1mL glacial acetic acid and desalted on a C18 cartridge (Waters) and divided into aliquots of 400ug peptide. Peptides solutions were reduced to approximately 1mL in a vacuum centrifuge and then lyophilized to dryness. Peptides were labeled with iTRAQ reagents according to the manufacturer instructions, then combined and reduced to dryness in a vacuum centrifuge. Labeled samples were resuspended in 100mM Tris 0.03% NP40 pH 7.4 and immunoprecipitated overnight with 4G10 anti-phosphotyrosine antibody pre-coupled to 20uL Protein A-agarose. Phosphopeptides were enriched from the immunoprecipitation by immobilized metal affinity chromatography (IMAC) and analyzed by HPLC-MS/MS as previously described [29].

The sample processing for the phosphoserine/threonine dataset was done as described in [3]. Briefly, the sections collected from the rat tissue were homogenized (PowerGen 700, Fisher Scientific) and protein was extracted according to the manufacturer's directions and digested with trypsin (Promega). To enrich for phosphopeptides, the sample was loaded on an Fe3+-charged IMAC column. The enriched peptides were analyzed by HPLC-MS/MS as described in [3].

2.2.2 Data processing—Raw data was converted to Mascot files (.mgf) using the Mascot module of Analyst 2.0 (Applied Biosystems). Mascot files were searched against appropriate species databases using Mascot 2.1 (Matrix Science) with the following parameters: tryptic digest, 1 missed cleavage, fixed modification carbamidomethyl, and D/E/N-term methylation for the rat liver data, optional modifications phosphoserine/phosphothreonine, phosphotyrosine, oxidation of methionine. Peptide tolerance was 1 Da and MS/MS tolerance was 0.15 Da. Every peptide hit with rank 1 was analyzed by manual inspection.

The spectra and corresponding peptide identifications were input into a database using phoMSVal. Manual curation was done by experienced LC-MS/MS users and then independently inspected by a single user to ensure that uniform criteria had been applied. Peptide assignments from Mascot were classified as "correct" or "incorrect", and phosphorylation site positions were verified and manually corrected if necessary. A false positive, or "incorrect" assignment, was given if either the peptide sequence was incorrect or the PTM position was incorrect.

Spectra were excluded from the analysis if the peptide was not phosphorylated. This left a total of 2662 spectra with manually validated phosphopeptide assignments for use in classification. During manual curation of these spectra, it was found that 271 were incorrectly assigned by Mascot, thus the Mascot assignments contained almost 13% false positive identifications (334 out of 2662 spectra). The 11 phosphotyrosine datasets were used for training and feature selection (2309 spectra), while the spectra from an additional phosphoserine/threonine dataset (353 spectra) were used for validation. The data were mostly singly phosphorylated, though about 12% (279 spectra) were doubly or triply phosphorylated. Of these, about 35% of the Mascot assignments were found to be false positives in the manual validation.

2.3 Feature extraction

For all stored spectra, 17 quality-features were extracted and used to evaluate peptide assignments. Here we have selected 16 quality features that have been previously described and propose a novel feature describing the percent of unidentified peaks in the spectrum. An example spectrum with illustration of features extracted is shown in Figure 2. The 16 previously described features are based on standard spectrum statistics (peak mean intensity, I_{avg}, standard deviation, σ, total intensity, I_{tot}, number of peaks, N, number of very low peaks, n, intensity value of most intense peak, I_{max}, m/z ratio of most intense peak, mz_{Imax}, maximum m/z value, mz_{max}). Intensity balance (I_{bal}) is a feature introduced by Bern *et al.* [30] to describe how data are distributed across a spectrum. Mascot calculates a score for all assigned peptides and we use this score (S) as a feature to measure how well a spectrum matches its peptide assignment [8]. It is worth noting that Mascot scores for correct peptide assignments may not always exceed the thresholds used by Mascot, and that scores for incorrect peptide assignments may exceed these same thresholds. This is why the Mascot score alone may not be enough to result in trustworthy validation of a spectrum. It is also important to note that scores from other database searching algorithms can be used in lieu of the Mascot score, although re-training of the classifier will be required.

Features based on peak identifications have been used previously with an SVM method [23]. The peak assignments given by Mascot are typically incomplete. This is because Mascot searches a potentially enormous library of possible proteolytic fragments (based on the protease used for digestion) and PTMs. Therefore, the peptide fragmentation scheme used by Mascot is, in general, kept fairly simple to decrease the risk of misassigning low probability fragment ions to incorrect peaks in the spectrum. To provide more complete fragment assignments we developed an additional tool to be used following the initial peptide assignment, which can consider a broader fragmentation scheme, including neutral losses and internal fragmentation. Trying to exhaustively explain each fragment ion mimics the process of manual validation, and, if done well, can therefore improve the accuracy of peptide assignments. Implementation of this tool does not artificially inflate scores or increase the number of false positives during the initial Mascot search because each spectrum has already been assigned to a particular peptide. The peak assignments generated by this tool are stored into the database during uploading of data.

The more complete peak identifications are used to calculate seven features, six of which were used together previously (averages of intensities of b-ions, I_{bIons}, y-ions, I_{yIons}, and unidentified peaks, I_{noID}, the number of fragment ion neutral losses, N_{NL}, average intensity of fragment ion neutral losses, I_{NL}, the percent of unidentified peak intensities explained by neutral losses, $noID_{NL}$) [23]. Our novel feature, the percent of unidentified high intensity peaks (N_{noID}), is based on the fact that correctly assigned spectra should have most or all high intensity peaks matched to expected fragmentation events. If there are several unassigned high intensity peaks then is it likely that the spectrum may have been incorrectly assigned. This novel feature significantly enhances the ability to distinguish between correct and incorrect peptide assignments.

2.4 Classification methods

In order to test whether the set of 17 quality features provides enough information to separate correct peptide assignments from incorrect ones, we used five classifiers implemented in the Weka machine learning workbench [31]. The classifiers used here were logistic regression, decision tree, random forest, artificial neural network (ANN) and naïve Bayes classifier. More details on the methods and their parameters are provided in the Supplementary Material.

3 Results

Although several features for MS/MS quality assessment have been previously published, their importance and synergism are still poorly understood. Furthermore, strongly correlating variables may bias the prediction method performance. In order to identify a set of features to use for classifications, we first analyzed all 17 features for their correlations and distributions. The correlations of the features are shown in Figure 3. As expected, several of the features were highly correlated, such as number of peaks (N) with number of low peaks (n) and maximum intensity (I_{max}) with mean intensity (I_{avg}). Also the I_{bal} feature has strong negative correlation with 10 features, which is due to the fact that when the overall intensity of the spectrum increases, the intensity balance becomes lower.

The distributions of the features were also examined and 15 out of 17 features do not obey normal distribution, with the exceptions of N_{noID} and I_{bal} (data not shown). Thus, the Fisher criterion score that has been used to identify the most important features [32] is not valid here as it strongly depends on the normality assumption.

In order to comprehensively characterize the impact of spectral features to classification, we reduced computational complexity by retaining only one feature of the feature-pairs having

>95% correlation. This resulted in 13 features for further analysis. All correlation values are provided in Supplementary Material (Table S1).

To find the best classification algorithm we tested five different classifiers with all 13 features. Performance of the classification algorithms was first measured with cross-validation (see Supplementary Material for details), followed by an analysis of the independent validation dataset of phosphoserine and -threonine data. The algorithms were compared using the receiver operating characteristic (ROC). In an ROC plot, sensitivity is plotted as a function of (1 - specificity), which corresponds to the fraction of true positives vs. the fraction of false positives [33]. In our case study, the area under the ROC curve (AUC) was used as the primary metric to quantify quality of each classifier.

The positive predictive value (PPV) was used as the secondary metric to quantify quality of each classifier. PPV is defined as the fraction of true positive assignments (“correct” that are classified as “correct”) over all positive assignments (all “correct” classifications) or *true positives/(true positives + false positives)*. The PPV value states how well a classifier is able to minimize the number of false positives while maximizing the number of true positives. This minimization is important, because in automated validation it is important to be able to automatically accept all instances classified as ”correct” with as few false positives as possible.

Table 1 contains AUC and PPV values for each algorithm for cross-validation and the independent validation set, a large dataset with phosphoserine and -threonine proteins. The ROC curves for cross-validation can be seen in Figure 4. Four of the five algorithms show good classification accuracies. For reference, the ROC curve for the Mascot identifications is also plotted. Notably, all classifiers except decision tree outperform the Mascot identification alone. The high PPV and AUC values in the independent validation show that the features tested are relevant for all kinds of phosphorylation types (phosphotyrosine, phosphoserine and phosphothreonine).

The random forest algorithm was used to calculate the importance of each feature as described in Section 2.3.3. This analysis determined the Mascot score as the most important feature, followed by percent of unidentified peaks as shown in Table 2. The random forest variable importance calculation is inherently univariate, *i.e.*, the decrease of classification accuracy is calculated using one quality feature at a time. This may result in spurious results if subsets of the features are strongly correlated, as they are here (Figure 3).

To overcome this univariate approach and identify the optimal set of features that gives the highest AUC and PPV values in the training and validation datasets, we trained each of the five classifiers with all the 8191 (2^{13}-1) feature combinations. Feature combinations that resulted in the highest AUC values are listed in Table 3 along with the best AUC and PPV values. Searching through all feature combinations improved the AUC for all algorithms, with the decision tree classifier having the most dramatic effect. All optimal quality feature sets include the Mascot score (S) and the percentage of unassigned peaks (N_{noID}). Four of the five optimal sets also contain the maximum observed m/z value (mz_{max}).

The results in Table 3 demonstrate that ANN, logistic regression, random forest and naïve Bayes classifiers performed remarkably well. In the cross-validation analysis the random forest classifier achieved the best overall performance with the cross-validation AUC value of 97.8% and the PPV value of 96.5%, though the results with logistic regression, ANN and naïve Bayes are practically equally good. The validation with an independent pS/pT dataset shows that naïve Bayes achieved the best result in terms of AUC (92.8%) and PPV (91.3%) followed by ANN, logistic regression and random forest. These results demonstrate that the machine learning approach is able to validate both pY and pS/pT data.

In the cross-validation, the naïve Bayes classifier reduced the number of false positives compared to original Mascot identifications by 76% (from 271 to 57) while retaining 97% of true positives (1968 out of 2038). Similar results were obtained when using the independent validation set (51% reduction of false positives (from 63 to 26); and 94% retention of true positives (273 out of 290)). The differences between the prediction methods are not significant, which indicates that the features used are informative for classifying the MS/MS spectra.

The classifiers are able to process multiply phosphorylated spectra. In the independent validation dataset, there were 52 multiply phosphorylated spectra, of which Mascot incorrectly assigned 21 and correctly 31 spectra (40% false positives). PhoMSVal with the naïve Bayes method reduced the number of false positives to 4 (18% false positives) while retaining 18 (58%) of the true positives.

The majority of the currently available peptide identification methods, such as Ascore, PhosphoScore and DeBunker are designed for SEQUEST data and cannot be directly used with Mascot data. However, to mimic DeBunker we used the SVM classifier in Weka with eight features and parameters delineated in [23]. The classification accuracy for the cross-validation was 86.1% (AUC 69.4% and PPV 92.9%) and for the independent dataset 79.6% (AUC 58.4% and PPV 84.9%).

4 Discussion

High-throughput and quantitative phospho-MS/MS data are increasingly important for systems level modeling of signaling and metabolomic pathways. MS/MS data are often analyzed with data-driven methods such as clustering and regression methods [34,35] or integrated into a mathematical model to describe signaling pathway kinetics [36]. Accordingly, tools to manage and validate a large number of MS/MS spectra are crucial to gain trustworthy results.

One of the most frequently used algorithms for automatic peptide identification is Mascot, most likely due to its reported high specificity for peptides and the best identification for a specified false positive rate [37]. However, it has been shown that Mascot identifications contain a significant number of false positives [14]. The false positives are not restricted to Mascot, but are present in all reverse- or random database searching algorithms, as these cannot identify all false positives due to their similarity to correct assignments [14]. Simply using a score threshold to differentiate true positive assignments from false positives is not optimal since it comes at the cost of losing correct assignments. We tested this strategy by classifying the data using a decision tree classifier and only the Mascot score as a feature, resulting in a tree with a split using a cutoff of the Mascot score at 17 for correct identifications. The classification resulted in an AUC of 90% and PPV value of 95%, indicating 5% false positives. False positives can confound downstream analyses and computational modeling efforts; and would seriously hinder follow-up experiments focusing on individual proteins.

We have addressed the challenges of MS/MS data management and validation of phosphopeptide peak assignments by developing phoMSVal, an open-source platform enabling storage, query and automatic validation of phospho-MS/MS data. The performance of phoMSVal was demonstrated with more than 2,600 manually curated phospho-MS/MS peptide identifications assigned by Mascot, including 13% (334) assignments that were considered to be incorrect assignments based on manual validation. To automate the determination of correct and incorrect peptide identifications, we first chose 16 features that are shown to be informative to the phosphopeptide spectra quality assessment and introduced a novel feature. These 17 features were studied for their correlation and a set of 13 features was chosen to the downstream analysis with five different machine learning algorithms.

Although each machine learning algorithm used a slightly different selection of quality features to distinguish true positive from false positive identifications, each algorithm outperformed Mascot in identification of false positives, indicating that machine learning algorithms with different spectral features bring clear added value to the spectra validation. When testing the classifiers with cross-validation of pY data, the false positive rate was reduced to 2.8% with a naïve Bayes algorithm using only three features (mz_{max}, *S, and* N_{noID}). This classifier was also the best at classifying the independent validation data giving AUC and PPV values of 93% and 91%, respectively. For this classifier, which was trained on pY data and validated with pS/pT data, this is a 51% decrease in false positives as compared to Mascot while retaining 94% of true positives. This shows that the classification is robust and very applicable to general phosphoproteomics validation. Since there were no statistical differences between the results of the different classifiers tested, phoMSVal has been implemented so that the user can choose the classification algorithm. The naïve Bayes classifier is used as a default.

Applying phoMSVal to the 2,662 MS/MS spectra including pY/pS/pT data, cross-validation with naïve Bayes resulted in AUC of 96% and PPV of 96%. This corresponds to a 70% reduction of the rate of false positives over Mascot's original peptide assignments while losing less than 4% of true positives. Our machine learning algorithm can almost ideally reconstruct the manual validation efforts of a dedicated expert mass spectrometrist, but in only a fraction of the time. Furthermore, phoMSVal is applicable to manage and validate spectra analyzed with other software than Mascot, such as SEQUEST and OMSSA [38], because the quality features are calculated directly from the spectra. We expect that phoMSVal will be a generally applicable tool that should significantly decrease the number of false positive identifications for many high-throughput phosphoproteomics datasets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the White lab for donating manually curated spectra to the data used in the case study. Funding: Financial support from Helsinki University Funds, Sigrid Jusélius Foundation, Biocentrum Helsinki, Academy of Finland (project 125826) and the Helsinki Biomedical Graduate School (HBGS) is gratefully acknowledged. Scott Carlson is supported by a Graduate Research Fellowship from the National Science Foundation. This work was partially supported by NCI Integrative Cancer Biology Program grant U54-CA11297 and by NCI grant R01-CA118705.

Abbreviations

ANN	artificial neural network
AUC	area under the ROC curve
PPV	positive predictive value
ROC	receiver operating characteristic
SILAC	Stable isotope labeling with amino acids in cell culture
SVM	support vector machine

References

1. Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, et al. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat Biotechnol. 2002; 20:301–305. [PubMed: 11875433]
2. Olsen JV, Blagoev B, Gnad F, Macek B, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell. 2006; 127:635–648. [PubMed: 17081983]
3. Moser K, White FM. Phosphoproteomic analysis of rat liver by high capacity IMAC and LC-MS/MS. J Proteome Res. 2006; 5:98–104. [PubMed: 16396499]
4. Larsen MR, Thingholm TE, Jensen ON, Roepstorff P, et al. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. Mol Cell Proteomics. 2005; 4:873–886. [PubMed: 15858219]
5. Rush J, Moritz A, Lee KA, Guo A, et al. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. Nat Biotechnol. 2005; 23:94–101. [PubMed: 15592455]
6. Ong S, Blagoev B, Kratchmarova I, Kristensen DB, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002; 1:376–386. [PubMed: 12118079]
7. Ross PL, Huang YN, Marchese JN, Williamson B, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004; 3:1154–1169. [PubMed: 15385600]
8. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20:3551–3567. [PubMed: 10612281]
9. Yates JR, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem. 1995; 67:1426–1436. [PubMed: 7741214]
10. Shilov IV, Seymour SL, Patel AA, Loboda A, et al. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol Cell Proteomics. 2007; 6:1638–1655. [PubMed: 17533153]
11. Peng J, Elias JE, Thoreen CC, Licklider LJ, et al. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res. 2003; 2:43–50. [PubMed: 12643542]
12. Elias JE, Gibbons FD, King OD, Roth FP, et al. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol. 2004; 22:214–219. [PubMed: 14730315]
13. Choi H, Ghosh D, Nesvizhskii AI. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res. 2008; 7:286–292. [PubMed: 18078310]
14. Chen Y, Zhang J, Xing G, Zhao Y. Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra. J Proteome Res. 2009; 8:3141–3147. [PubMed: 19368407]
15. Nichols AM, White FM. Manual validation of peptide sequence and sites of tyrosine phosphorylation from MS/MS spectra. Methods Mol Biol. 2009; 492:143–160. [PubMed: 19241031]
16. Koenig T, Menze BH, Kirchner M, Monigatti F, et al. Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. J Proteome Res. 2008; 7:3708–3717. [PubMed: 18707158]
17. Tanner S, Shu H, Frank A, Wang L, et al. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem. 2005; 77:4626–4639. [PubMed: 16013882]
18. Salmi J, Moulder R, Filén J, Nevalainen OS, et al. Quality classification of tandem mass spectrometry data. Bioinformatics. 2006; 22:400–406. [PubMed: 16352652]
19. Beausoleil SA, Villén J, Gerber SA, Rush J, et al. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol. 2006; 24:1285–1292. [PubMed: 16964243]

20. Ruttenberg BE, Pisitkun T, Knepper MA, Hoffert JD. PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. J Proteome Res. 2008; 7:3054–3059. [PubMed: 18543960]

21. Gnad F, Ren S, Cox J, Olsen JV, et al. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol. 2007; 8:R250. [PubMed: 18039369]

22. Gao J, Agrawal GK, Thelen JJ, Obradovic Z, et al. A New Machine Learning Approach for Protein Phosphorylation Site Prediction in Plants. Lect Notes Comput Sci 2009. 2009; 5462:18–29.

23. Lu B, Ruse C, Xu T, Park SK, et al. Automatic validation of phosphopeptide identifications from tandem mass spectra. Anal Chem. 2007; 79:1301–1310. [PubMed: 17297928]

24. Allmer J, Kuhlgert S, Hippler M. 2DB: a Proteomics database for storage, analysis, presentation, and retrieval of information from mass spectrometric experiments. BMC Bioinformatics. 2008; 9:302. [PubMed: 18605993]

25. Myers T, Law W, Eng JK, McIntosh M. Installation and use of the Computational Proteomics Analysis System (CPAS). Curr Protoc Bioinformatics. 2007; Chapter 13(Unit 13.5)

26. Deutsch EW, Lam H, Aebersold R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiol Genomics. 2008; 33:18–25. [PubMed: 18212004]

27. Nesvizhskii AI, Aebersold R. Interpretation of Shotgun Proteomic Data. Mol Cell Proteomics. 2005; 4:1419–1440. [PubMed: 16009968]

28. Grobei MA, Qeli E, Brunner E, Rehrauer H, et al. Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. Genome Res. 2009; 19:1786–1800. [PubMed: 19546170]

29. Zhang Y, Wolf-Yadlin A, Ross PL, Pappin DJ, et al. Time-resolved Mass Spectrometry of Tyrosine Phosphorylation Sites in the Epidermal Growth Factor Receptor Signaling Network Reveals Dynamic Modules. Mol Cell Proteomics. 2005; 4:1240–1250. [PubMed: 15951569]

30. Bern M, Goldberg D, McDonald WH, Yates JR. Automatic quality assessment of peptide tandem mass spectra. Bioinformatics. 2004; 20(Suppl 1):i49–54. [PubMed: 15262780]

31. Witten, IH.; Frank, E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; San Francisco: 2005.

32. Wong, L., editor. The Practical Bioinformatician. World Scientific; New Jersey: 2004.

33. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006; 8:861–874.

34. White FM. Quantitative phosphoproteomic analysis of signaling network dynamics. Curr Opin Biotechnol. 2008; 19:404–409. [PubMed: 18619541]

35. Macek B, Mann M, Olsen JV. Global and site-specific quantitative phosphoproteomics: principles and applications. Annu Rev Pharmacol Toxicol. 2009; 49:199–221. [PubMed: 18834307]

36. Tasaki S, Nagasaki M, Oyama M, Hata H, et al. Modeling and estimation of dynamic EGFR pathway by data assimilation approach using time series proteomic data. Genome Inform. 2006; 17:226–238. [PubMed: 17503395]

37. Kapp EA, Schütz F, Connolly LM, Chakel JA, et al. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics. 2005; 5:3475–3490. [PubMed: 16047398]

38. Geer LY, Markey SP, Kowalak JA, Wagner L, et al. Open mass spectrometry search algorithm. J Proteome Res. 2004; 5:958–64. [PubMed: 15473683]

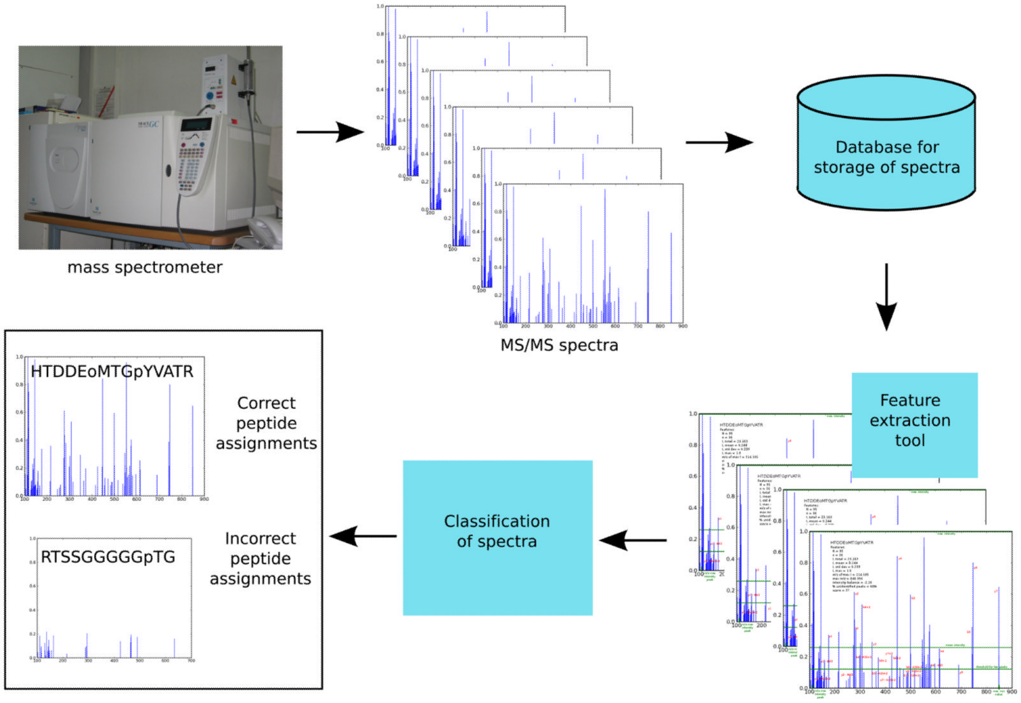
mass spectrometer
MS/MS spectra
Database for storage of spectra
Feature extraction tool
Classification of spectra
HTDDEoMTGpYVATR
Correct peptide assignments
RTSSGGGGGpTG
Incorrect peptide assignments

Figure 1.

An overview of phoMSVal. Data are obtained from a phosphospecific mass spectrometry experiment, the spectra are analyzed by Mascot and the data are stored in a database. For each spectrum, features are extracted and used as input into a classifier that separates spectra into two classes: correct or incorrect peptide assignments.

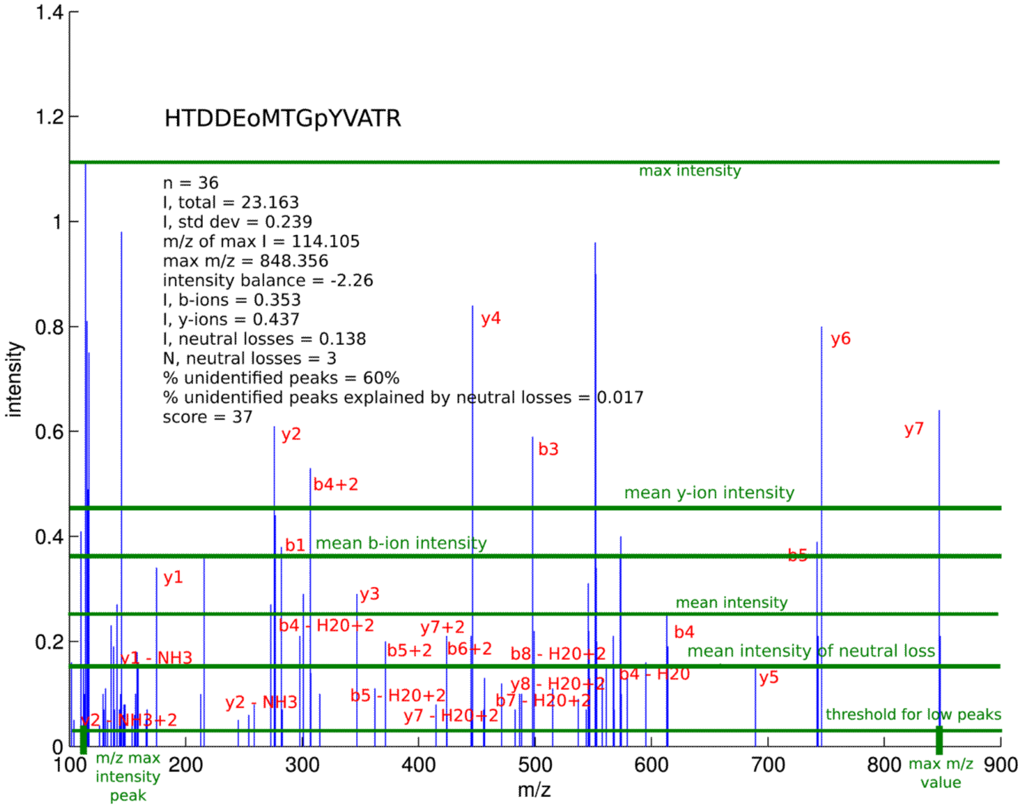

Figure 2.

MS/MS spectrum assigned to the tyrosine phosphorylated peptide from the activation loop of the p38 mitogen-activated protein kinase. Peak assignments are highlighted in red, and features extracted from the spectrum are shown in green. Phosphorylated residues are preceded by lower-case “p”, and oxidation of methionine to the sulfoxide is indicated by lower-case “o”.

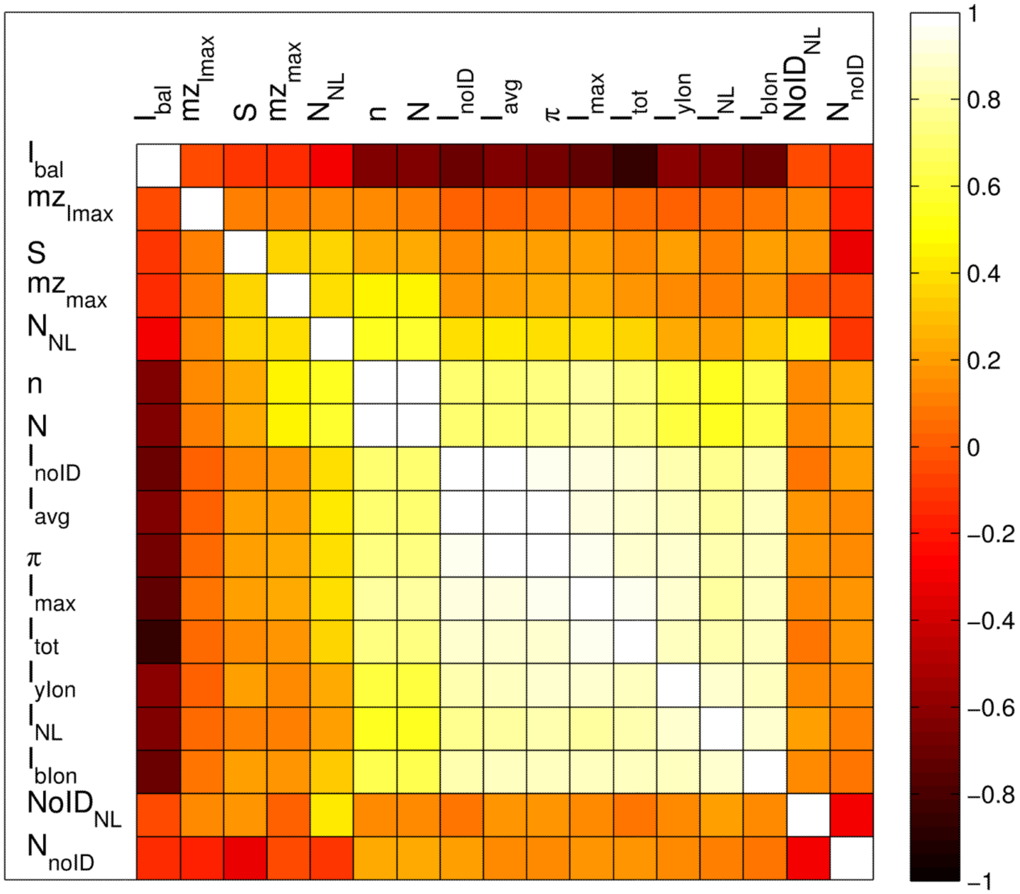

Figure 3.

Correlation of all 17 features extracted from phospho-MS/MS spectra. The features are grouped based on clustering of their correlations. The colors correspond to correlation values: Light and yellow denote strong positive correlation whereas black and dark colors correspond to strong negative correlation, and orange/red coloring weak correlation. The features *N* (number of peaks), I_{max} (maximum intensity), I_{avg} (mean intensity) and I_{noID} (mean intensity of unidentified peaks) were removed from further analysis due to high correlation (>95%).

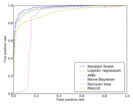

Figure 4.

ROC curves for random forest, logistic regression, artificial neural network, naïve Bayes and decision tree classifiers using cross-validation with training data. The plot of the original Mascot identifications in ROC space is also shown for comparison.

Table 1

The comparison of descriptive statistics for classifiers using all thirteen of the features.

Type of classifier	AUC, CV	PPV, CV	AUC, IV	PPV, IV
Random forest	0.977	0.967	0.890	0.892
Logistic regression	0.967	0.963	0.920	0.910
ANN	0.965	0.969	0.912	0.925
Naïve Bayes	0.950	0.964	0.715	0.892
Decision tree	0.835	0.959	0.781	0.889

Table 2

The features used in the study in order of their significance in classifying phosphorylated MS/MS data according to random forest classifier variable importance test.

Feature	**Decrease in acuracy**
Mascot score, S	0.760
Percent of unidentified peaks, N_{noID}	0.759
Mean intensity of b-ions, I_{bIons}	0.650
Maximum *m/z* value, mz_{max}	0.555
Number of low peaks, n	0.490
Mean intensity of fragment neutral loss peaks, I_{NL}	0.485
Intensity balance, I_{bal}	0.477
m/z value of maximum intensity peak, mz_{Imax}	0.475
Percent of unidentified peak intensity explained by neutral losses, $NoID_{nL}$	0.446
Mean intensity of y-ions, I_{yIons}	0.434
Total intensity, I_{tot}	0.423
Standard deviation of intensities, σ	0.422
Number of neutral losses, N_{NL}	0.368

Table 3

The descriptive statistics for best feature set for each classifier using cross-validation (CV) and independent validation set (IV) and features used in best classifiers.

Type of classifier	**AUC, CV**	**PPV, CV**	**AUC, IV**	**PPV, IV**	**features**
Naïve Bayes	0.966	0.972	0.928	0.913	mz_{max}, S, N_{noID}
ANN	0.970	0.965	0.923	0.903	n, σ, mz_{max}, I_{tot},, S, N_{noID}, I_{yIons},, I_{NL}
Logistic regression	0.968	0.964	0.922	0.907	n, mz_{Imax}, mz_{max}, I_{tot}, I_{bal}, S, N_{noID}, I_{yIons}, N_{NL}, $NoID_{nL}$
Random forest	0.978	0.965	0.880	0.887	n, σ, mz_{Imax}, mz_{max}, I_{tot}, I_{bal}, S, N_{noID}, I_{bIons}, N_{NL}
Decision tree	0.938	0.961	0.855	0.884	n, S, N_{noID}