Journal of Proteomics & Bioinformatics - Open Access
Research Article JPB/Vol.2/March 2009

Improving Phosphopeptide/protein Identification using a New Data Mining Framework for MS/MS Spectra Preprocessing

Fabio R. Cerqueira[1,*], Sandra Morandell[2], Stefan Ascher[2], Karl Mechtler[3], Lukas A. Huber[2], Bernhard Pfeifer[1], Armin Graber[4], Bernhard Tilg[1], Christian Baumgartner[1*]

[1]Research Group for Clinical Bioinformatics, Institute of Biomedical Engineering, University for Health Sciences, Medical Informatics and Technology, Eduard Wallnoefer Zentrum 1, 6060, Hall in Tirol, Austria

[2]Biocenter, Division of Cell Biology, Medical University of Innsbruck, Fritz-Pregl Str. 3, 6020, Innsbruck, Austria

[3]Research Institute of Molecular Pathology (IMP), Dr Bohr-Gasse 7 and Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), Dr Bohr Gasse 3, A1030 Vienna, Austria

[4]Institute for Bioinformatics, University for Health Sciences, Medical Informatics and Technology, Eduard Wallnoefer Zentrum 1, 6060, Hall in Tirol, Austria

*Corresponding authors: Fabio R. Cerqueira, Research Group for Clinical Bioinformatics, Institute of Biomedical Engineering, University for Health Sciences, Medical Informatics and Technology, Eduard Wallnoefer Zentrum 1, 6060, Hall in Tirol, Austria, Tel: +43 50 8648 3827; Fax: +43 50 8648 673827; E-mail: fabio.cerqueira@umit.at

Christian Baumgartner, Research Group for Clinical Bioinformatics, Institute of Biomedical Engineering, University for Health Sciences, Medical Informatics and Technology, Eduard Wallnoefer Zentrum 1, 6060, Hall in Tirol, Austria, E-mail: christian.baumgartner@umit.at

Received January 27, 2009; Accepted March 21, 2009; Published March 21, 2009

Citation: Cerqueira FR, Morandell S, Ascher S, Mechtler K, Huber LA, et al. (2009) Improving Phosphopeptide/protein Identification using a New Data Mining Framework for MS/MS Spectra Preprocessing. J Proteomics Bioinform 2: 150-164. doi:10.4172/jpb.1000072

Abstract

Phosphopeptide/protein identification using tandem mass spectrometry (MS/MS) is a challenging issue in proteomics research. In particular, phosphopeptides typically exhibit low intensity peaks of *b* and *y* ions in spectra when serine or threonine is phosphorylated. Consequently, the existing algorithms for peptide and protein identification generate a high false discovery rate when coping with phosphopeptide spectra. In order to increase the number of correct phosphopeptide identifications using database search, a new data mining approach for spectra preprocessing is proposed. A support vector machine classifier is used to calculate the probability of each peak representing a *b* or *y* ion. Next, low-probability peaks are removed from spectra, while remaining peaks have their intensities enhanced. As a result, a huge increase in signal-to-noise ratio is provided and the chances of detecting important peaks are significantly advanced. Experiments using MASCOT and SEQUEST along with Peptide/ProteinProphet and a decoy database approach showed a significant improvement in the sensitivity of phosphopeptide identification without compromising specificity, demonstrating that our new strategy for MS/MS spectra preprocessing is a powerful proteomics tool for enhancing phosphopeptide identifications.

Key words: Data mining; Tandem mass spectrometry; Spectra preprocessing; Phosphoproteomics; Peptide/protein identification

Introduction

Protein phosphorylation is a key post-translational modification for many cellular processes in all living organisms (Morandell et al., 2006). Moreover, abnormal phosphorylation can be responsible for many serious diseases such as diabetes and cancer (Kocher et al., 2006). Therefore, a complete understanding of many biological processes and signaling pathways depends on a profound analysis of involved phosphorylated proteins.

Liquid chromatography (LC) coupled with tandem mass spectrometry (MS/MS) has been the method of choice for protein identification in complex mixtures (Steen and Mann, 2004). On the other hand, such a high throughput approach can generate thousands of MS/MS spectra in a single run, turning the manual interpretation into an infeasible practice (Kappet al., 2005). By that means, computational tools for spectra interpretation became vital for large-scale MS-based proteomics. However, the most popular algorithms for peptide and protein identification produce a high false discovery rate (FDR) (Kapp et al., 2005). This is mainly due to the presence of background signals in spectra, i.e., peaks representing isotopes, internal fragments, electronic noise, chemical noise, and ions originated from unknown fragmentation pathways (Gentzel et al., 2003; Mujezinovic et al., 2006). In the case of phosphopeptides, there is an additional source of complexity. Phosphopeptide fragmentation in low energy dissociation tends to occur in phosphate groups when serine or threonine is phosphorylated, which is normally the case. Hence, it is common to see a very prominent peak in the central region of the spectrum (Figure 1), corresponding to neutral losses of HPO_3 or H_3PO_4 groups from the precursor ion (Kocher et al., 2006). Therefore, other necessary fragmentations to form *b* and *y* ions are scarce. The resulting low signals of such ions are then easily confounded with noise peaks (Figure 1). For this reason, false positive (FP) identifications occur even more frequently for phosphopeptide spectra interpretation.

In order to avoid misinterpretations, various algorithms have been proposed to process spectra before ion search. One approach focuses on the deconvolution of multiply charged peaks and deisotoping (Ferrige et al., 1991; Reinhold and Reinhold, 1992; Zhang and Marshall, 1998). More recently, other cleaning steps have been included. Gentzel et al., (2003), for example, considered peak centroiding, joining of redundant spectra and automatic calibration. However, no special procedure thereby was undertaken for noise elimination. Mujezinovic et al., (2006) propose a sophisticated noise removal procedure, but, as described by the authors, this approach is only robust for mild inaccuracies in spectra. The main idea in the work of Jaitly et al.,(2004) was to elucidate peaks of interest instead of background peaks. A maximum likelihood estimate is used to classify peaks corresponding to fragment ions, achieving high sensitivity and specificity. Nevertheless, no experiment is presented in this work concerning the impact of peak classification on spectra interpretation. All cited studies provided an important enhancement on MS/MS spectra preprocessing methods. However, to the best of our knowledge, the particular issue of low quality of phosphopeptide spectra

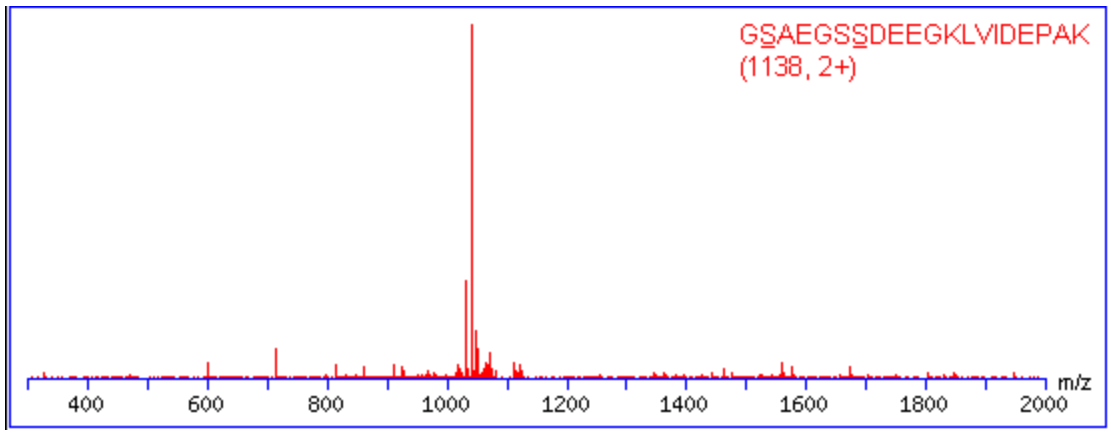

Figure 1: Typical phosphopeptide MS/MS spectrum. The reported sequence indicates phosphorylation in the first and third serines. Notice the intensity of the central peak at *m/z* = 1040 compared to intensities of other peaks. This prominent peak corresponds to the neutral loss of two H_3PO_4 groups (−196 Da, offset of −98).

remains unaddressed. Recent works (Ishihama et al., 2007; Imanishi et al., 2007; Lu et al., 2007; Hoffert et al., 2007) have been proposed for a more efficient validation of phosphopeptide identifications. These methods act as good filters to separate correct from incorrect outcomes. Nonetheless, no effort is presented to improve the quality of phosphopeptide spectra in order to decrease the number of incorrect sequence assignments.

In this paper, we propose a new data mining approach for advancing phosphopeptide/protein identification in standard database (DB) search tools by introducing a novel preprocessing modality for spectra before ion search. The method trains a support vector machine classifier for determining the useful peaks in a spectrum concerning the procedure of assigning a peptide sequence. An important characteristic of the proposed method is that, instead of having an unique training set (TS), a particular TS is dynamically constructed for each dataset in analysis so that a specialized classifier can be obtained for each run. This step provides high flexibility concerning variations in spectra caused by different acquisition or sample preparation methods as well as the mass spectrometer type used. A 10-fold cross validation showed a classifier accuracy of around 80%. In our experiments, we verified that about 84% of peaks in the original spectra could be eliminated after preprocessing (after applying the learned classifier), yet the identifications could be improved. In a decoy DB analysis on SEQUEST (Eng et al., 1994) results, for instance, we could observe increases in the number of correct assignments varying from 33% to 60%. The same analysis on MASCOT (Perkins et al., 1999) results showed enhancements between 20% and 35% when employing the identity threshold. The statistical models provided by Peptide/ProteinProphet (Keller et al., 2002; Nesvizhskii et al., 2003) confirmed the power of our method. The sensitivity curves plotted for the preprocessed datasets dominated the sensitivity curves for original data in all cases. The error curves for treated spectra, on the other hand, were mostly dominated by the error curves of noisy data. Examining the proteins infered by ProteinProphet, a higher number of matches and better coverage could be normally observed after the application of our preprocessing procedures.

Materials and Methods

Sample Preparation and LC/MS

In two independent experiments, samples were prepared according to the following protocol. Cytosol of Mek1$^{-/-}$ MEFs was pretreated in a KESTREL-like approach (Knebel et al., 2001), followed by in vitro kinase assay using Chemical Genetics tools (Shah et al., 1997). Samples were incubated with or without an ATP-binding pocket mutant of constitutive active GST-Mek1 in the presence of a N^6-modified ATP-analog. Proteins were precipitated according to Wessel and Flugge, (1984) and a proteolytic digestion with trypsin was performed as described by de la Fuente van Bentem et al., (2006). For the enrichment of phosphorylated peptides, the samples were methylester-modified with normal (light) or deuterated (heavy) methanol (Ficarro et al., 2002), and the IMAC protocol was performed (de la Fuente van Bentem et al., 2006) using two different elution steps with 50mM and 125mM Na_2HPO_4, respectively. In a third experiment, samples were prepared as above, except for the in vitro kinase assay.

Eluted peptides were separated by reversed phase high performance liquid chromatography (RP-HPLC) coupled to a LTQ FT mass spectrometer (hybrid linear ion trap / Fourier transform ion cyclotron resonance (Thermo Electron, Bremen)) using a multistage activation method as described in Schroeder et al., (2004).

MS/MS Data

For each experiment described above, we took data files in Xcalibur raw format corresponding to the samples with GST-Mek1, methylester-modified with normal methanol, eluted with 125mM Na_2HPO_4. A total of three raw files were then converted to dta files, resulting in 24405 (SET1), 23668 (SET2) and 18996 (SET3) spectra, respectively. SEQUEST (Bioworks v3.3, Thermo Electron) was run on these three datasets so that we could filter high-quality identifications for constructing the TSs. In order to evaluate the preprocessing procedures, we randomly picked 10000 files (spectra) from each dataset above (excluding spectra used to build the TSs). After applying our approach on these selected spectra, three additional sets of spectra were generated, resulting in six datasets for the analysis described in Section Results and Discussion, three of them containing original 10000 spectra each, and the others containing the respective preprocessed versions. These datasets are defined here simply as SET1_O, SET2_O, SET3_O, SET1_P, SET2_P and SET3_P, respectively, where "_O" and "_P" indicate original and preprocessed data.

In order to perform experiments using MASCOT (v2.2,

Matrix Science), each set was converted into the MASCOT generic format mgf. All searches in MASCOT and SEQUEST were performed against the mouse IPI database (v3.18) (Kersey et al., 2004). The search parameters were set the same for all runs. Enzyme: trypsin; missed cleavages: up to 3; fixed modifications: carbamidomethyl (C), methyl (C-term), Methyl (DE); variable modifications: oxidation (M), phosphorylation (ST), phosphorylation (Y); peptide charges: +1, +2 and +3; protein mass: unrestricted; mass values: monoisotopic; peptide mass tolerance: $\pm$10 ppm; fragment mass tolerance: $\pm$0.6 Da.

Data Mining Framework for Spectra Preprocessing

Figure 2a illustrates the general overview of our new framework for spectra preprocessing. Initially, the original spectra are used for the TS construction (details in Figure 2b). Next, a support vector machine classifier is trained using the constructed TS. Finally, the resulting model is used to classify peaks and adjust intensities (details in Figure 2c) in the original spectra, giving rise to more easily interpretable spectra. The following sections provide detailed descriptions of each step.

Applying Filters for High-quality Identifications

In the proposed framework, TSs are constructed on-the-fly, i.e., one for each dataset in analysis. In order to obtain peaks for representing instances in the TS, a set of MS/MS spectra with respective correct sequences is necessary. For this purpose, some standard database search tool is run a priori on the data. In our experiments, SEQUEST was chosen to provide interpretations of spectra. Next, strict filter constraints are applied on the obtained identifications for minimizing the risk of selecting incorrect sequence assignments (for minimizing noise in the TS). These constraints are as follows:

i. For each spectrum consider only the top hit (as tools normally provide other alternative sequences).

ii. For each spectrum pick the best hit when searching with different charge states (this is common when the exact precursor charge cannot be accurately distinguished).

iii. Pick the best hit among redundant answers (different spectra assigned to the same sequence. It is to assure a

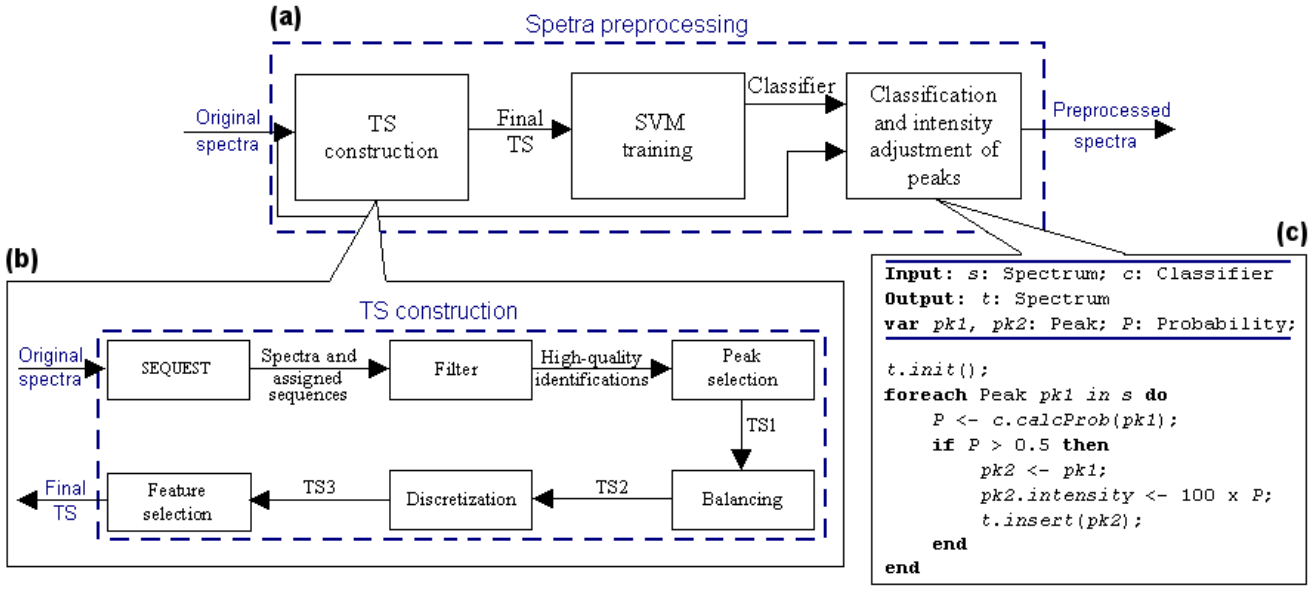

Figure 2: Framework for spectra preprocessing. (a) General overview: The TS is constructed as illustrated in (b). The obtained TS is then used to train a support vector machine classifier. In the last stage, the resulting model is applied to the original spectra for classifying peaks and adjusting intensities. (b) Dynamic TS construction details: A standard DB search engine for peptide identification (SEQUEST in our case) is run a priori to provide spectra interpretations. Next, filters are applied to separate high-quality identifications. From the resulting filtered set, peaks are taken to form the first TS version. In order to optimize the classifier's accuracy, three additional steps are performed: Balancing of TS, discretization and feature selection, producing a final TS for the model construction. (c) Pseudo-code for peak classification and intensity adjustment: The classifier is used to assign a probability *P* to each peak. Low-probability peaks are removed from spectra, while the remaining peaks have their intensities adjusted according to *P*.

high variability among examples in the TS).

iv. Consider only spectra for which the second-ranked hit has evaluation at least 25% worse than the top hit (Kapp et al., 2005) (in SEQUEST this is measured by ΔCn).

v. For selected identifications so far keep only ones for which the reported protein has at least one different assigned peptide sequence in another hit (it provides a higher confidence in the protein identification (Elias et al., 2005)).

vi. For selected identifications so far keep the best 10%.

vii. For each kept spectrum and its assigned sequence calculate the theoretical *m/z* values for *b* as well as *y* ions and match them to the observed peaks. If at least 70% of both expected series are covered by identified peaks (no matter which charge), the result is maintained.

The application of the described constraints leads to a very reduced set of spectra. The idea is to achieve a high precision, avoiding noise in the TS. Experiments demonstrated that about 50 spectra are enough to provide a TS containing more than 1000 peaks (TSs can be found in supplementary data - appendix Addendum: Web Supplement). We assume that peaks have similar patterns among spectra in the dataset.Therefore, even learning with a reduced subset, we would be able to accurately classify peaks in the remaining spectra. The validity of this assumption is demonstrated in our statistical evaluations described in Section Statistical Analysis , since the identifications for preprocessed spectra are improved. After applying the constraints, if the number of selected spectra is not enough to provide a significant TS, we suggest changing the percentage in item vi, as itens iv and vii already provide strong cutoffs. It is also a good alternative to test changing the minimum ΔCn (item iv) or even the percentage of *b* and *y* ions found (item vii) until getting a combination of values leading to a suitable TS. However, the values stated above are normally apropriate. In the case of SET3, for instance, only item vi was altered, i.e., we selected the best 20% results to obtain enough spectra. For the other datasets, the restrictions were applied exactly as proposed. A total of 68, 49 and 47 spectra were selected from SET1, SET2 and SET3, respectively, to form each training set.

Selecting Peaks

Once having selected spectra and their assigned sequences, the next step is to separate peaks in spectra to form the TS. For each spectrum, peaks are labeled *golden* if they match to the theoretical *b* and *y m/z* values (item vii), but only ions of charges +1 and +2 are considered. Any other peak is labeled as background. In this manner, signals representing internal fragments, isotopes, electronic and chemical noise, the precursor and its derivatives, neutral losses, and *b*/*y* ions of higher charges are taken as background. Our assumption is that leaving only peaks representing $b^{+1/+2}$ and $y^{+1/+2}$ ions in spectra is the cleanest and simplest way to minimize misinterpretations. Besides, the *m/z* values for ions of higher charges are difficult to handle due to resolution issues. After separating golden and background peaks, each one has its features extracted from spectra (Section Feature Selection) and is used to compose the first TS version.

Balancing of TS

In many real-world classification problems, the majority of instances in a TS is from one of the classes. In binary classification, the minority class is normally the class of interest. This situation occurs in our case. The first TS version obtained from SET1, for example, contained 40184 peaks of the class "background" and 1022 peaks of the class "golden". Such an imbalanced TS often causes machine learning algorithms to perform poorly on the minority class, i.e., the rare instances are often treated as noise (Dehmeshki et al., 2003). In order to make both classes have the same weight for the model construction, the TS is balanced using a supervised instance resampling algorithm. This algorithm produces a random subsample of instances in the class background, applying a sampling with replacement approach (Hsia, 2005), resulting in the same number of examples in both classes. In this stage, the second TS version is established.

Feature Discretization

The third TS variant is generated from a discretization of the given numeric features (Section Feature Selection). According to Witten and Frank, (2005), learning methods that are able to handle numeric features often produce better results and work faster when features are discretized beforehand. Experiments confirmed that the classification accuracy is improved after discretization (supplementary data). In our approach, we employ a supervised discretization algorithm that uses an entropy minimization heuristic for discretizing the range of a continuous-valued feature into multiple inter-

vals (Fayyad and Irani, 1993; Witten and Frank, 2005).

Feature Selection

In order to accomplish our classification, we had to select important features that clearly distinguish golden peaks from background signals. For the employed mass spectrometer (Section Sample Preparation and LC/MS), peaks of *b* and *y* ions are normally accompanied by signals with lower *m/z* values representing neutral losses such as water, ammonia, two molecules of water, and water+ammonia (Frank and Pevzner, 2005). The occurrence of *a* ions is common as well and these ions can undergo a neutral loss of water or ammonia (Frank and Pevzner, 2005). Isotope variants (more commonly +1 Da and +2 Da) are also very frequent and can present the same cited *satellite peaks*. Putting together all these information and initially considering +1 charged ions, the first set of features of a given *target* (peak being classified) had its composition based on the presence of the described satellite peaks at the positions shown in Figure 3a. However, in order to increase the classifier's discriminatory power, we added features regarding also background peaks, as demonstrated in Figures 3b and 3c. If we consider a signal representing loss of water as the reference, for example, satellite peaks are expected to be found at the positions shown in Figure 3b. Hence, it is appropriate to include in the set of features also the occurrence of peaks at the positions -25, -10, -9, -8, $+3$, $+18$, $+19$ and $+20$ in order to improve the distinction between signals representing loss of water and golden peaks. On the other hand, when taking as reference a peak representing a loss of CO (*a* ion), its expected vicinity is composed as stated in Figure 3c. Consequently, the inclusion of the positions -8, -7, -6, -5, $+10$, $+11$, $+12$, $+13$, $+28$, $+29$, $+30$ in the set of features increases the dissimilarities between peaks representing *a* ions and golden peaks. This reasoning was applied for the 17 satellite peaks shown in figure 3a, giving rise to 76 distinct features. Since +2 charged ions are also considered, peak positions are divided by two, ending up with 128 features (the division leads to many repeated positions). Finally, since noise can be very abundant in spectra, random signals might occupy the described positions, being thus confounded with satellite peaks. As a result, instead of considering the occurrence of peaks at the proposed positions, the set of features was formed by the intensities of such peaks normalized to (divided by) the target's intensity. Therefore, we always start with 128 features (128 normalized intensities of satellite peaks) for each dataset being operated. However, in order to remove irrelevant features, the final TS is generated using a correlation-based feature selection heuristic (Plant et al., 2006). This method assigns higher scores to feature subsets whose features show high correlation with the class and low correlation among each other.

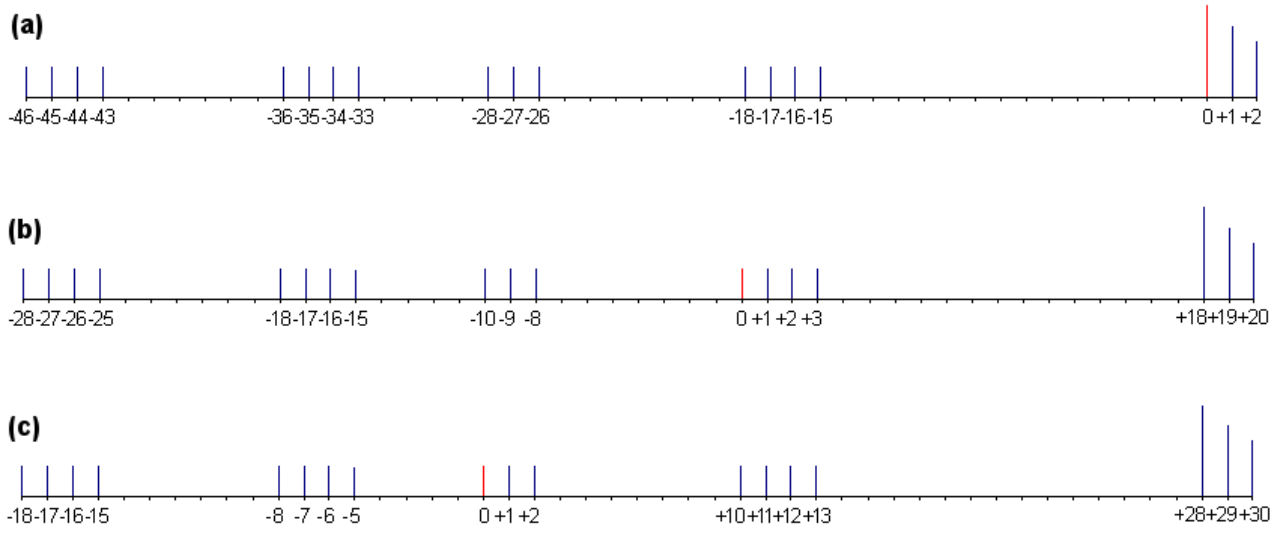

Figure 3: Simplified scheme representing the vicinity of common peaks in spectrum for +1 charged ions. The peaks are shown in relative positions, indicating offset in Dalton. (a) Golden peak (red) and peaks appearing frequently in its vicinity (blue). (b) Changing the reference to a peak representing loss of water (red). (c) Changing the reference to a peak representing loss of CO (red).

Learning with Support Vector Machines (SVMs)

SVMs (Cristianini and Shawe-Taylor, 2000) constitute a very powerful classification method with outstanding performance applied to many biomedical problems (Noble, 2006). Given instances x_i, $i = 1, ..., l$ with labels $y_i \in \{1, -1\}$, the training of SVMs is the task of solving the following quadratic optimization problem:

$$\min_{\alpha} f(\alpha) = \frac{1}{2}\alpha^T Q\alpha - e^T\alpha$$
$$\text{subject to } 0 \le \alpha_i \le C,\ i = 1, ..., l, \qquad (1)$$
$$y^T\ \alpha = 0$$

where e is the vector of all ones, C is the upper bound of all variables, Q is an $l \times l$ symmetric matrix with $Q_{ij} = y_i y_j K(x_i, x_j)$, $K(x_i, x_j)$ is the kernel function, and the optimal weights α_i are searched for defining a hyperplane that separates the two classes so that unknown instances can be accurately classified. In our case, a pair (x_i, y_i) represents a peak in the training set, where x_i is the vector of features (normalized/discretized intensities) and y_i denotes whether the peak is golden (1) or background (-1).

The size of the matrix Q is proportional to the square of the number of training examples, which can lead to huge storage requirements. Consequently, decomposition methods are applied to modify only a subset of the vector α in each iteration. In this context, we chose a powerful method termed LIBSVM (Fan et al., 2005) to implement in our framework, as this algorithm presents faster convergence compared to previously proposed approaches, without compromising the result quality. Another advantage of LIBSVM is that it is able to provide a probability estimate to determine the confidence level of a classification instead of predicting "hard" class labels (Wu et al., 2004). This is very suitable to our goals, as can be seen in the next section.

We selected the radial basis function (Cristianini and Shawe-Taylor, 2000) as the kernel function with $\gamma = 1/k$, where k is the number of features. Using 10-fold cross validation, the performance of LIBSVM for golden peaks classification in the final TSs gathered from SET1, SET2 and SET3 is summarized in Table 1.

Data set	Sensitivity	Specificity	Accuracy	AUC
SET1	75.4	84.1	80.58	0.873
SET2	75	80	78	0.835
SET3	77.6	81.6	80.05	0.863

AUC is area under the ROC curve.

Table 1: Performance of LIBSVM for golden peaks classification using 10-fold cross validation in the final TSs built from SET1, SET2 and SET3.

Classifying Peaks and Adjusting Intensities

The SVM classifier is used in the proposed framework to calculate the probability P of a peak designated to be golden. The peak is only kept in spectrum if $P > 0.5$. In this case, its intensity is set to $P \times 100$ (algorithm in Figure 2c). As a consequence, only the most significant peaks for peptide sequence identification are maintained in order to avoid FP matches caused by additional mass combinations of background peaks. The adjustment of intensities is also important to decrease the number of wrong answers, since many popular tools for peptide and protein identification consider only the most intense peaks in the ion search procedure. The SVM runs took on average 1.02*ms*, 0.49*ms* and 0.37*ms* (the resulting TS from SET1 had a higher number of features) per peak for generating the datasets SET1_P, SET2_P and SET3_P, respectively.

Results and Discussion

In this paper we applied a new data preprocessing method for MS/MS spectra with the objective of enhancing the rate of correct phosphopeptide/protein identifications in standard DB search tools. Experiments using the proposed approach were performed on three selected datasets (SET1-3) generated from phospho-enriched samples, with results achieving a significant positive effect on the quality of phosphopeptide assignments. Since no reference data was available during this work, the experiments were based on a comparative analysis between results before and after spectra preprocessing.

First, a manual inspection of MASCOT and SEQUEST results was performed. An important observation was the huge increase in signal-to-noise ratio (S/N) for treated spectra as illustrated in Figure 4 (another example can be seen in Figure 2S of supplementary data). Moreover, in cases in which original and preprocessed spectra led to the same peptide sequence, a higher number of ions in the preprocessed spectra was frequently observed (Figures 4 and 5). Even in cases in which fewer ions were found, higher scores were normally generated as a consequence of a better S/N and intensity adjustment.

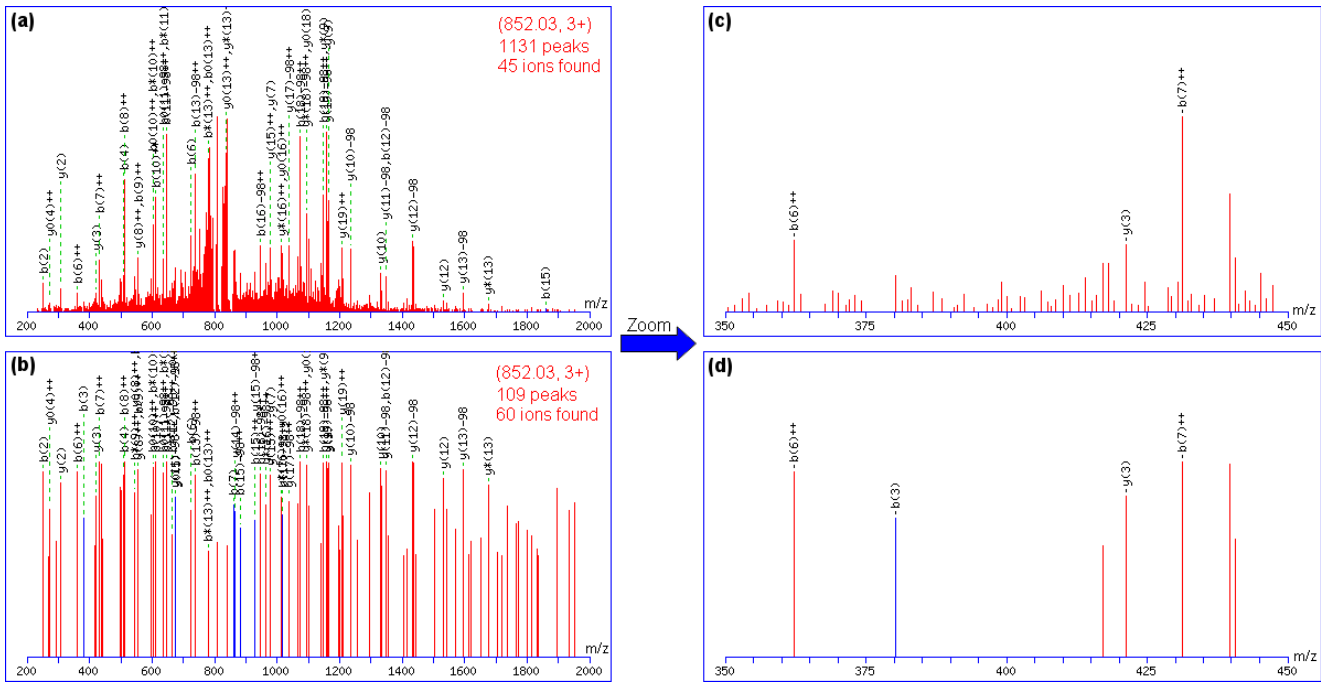

Figure 4: Comparison of MASCOT results for a spectrum before and after preprocessing, supporting the same peptide sequence HNQDSQHCSLSGDEEDELFK. (a) Original spectrum (SET1_O). (b) Preprocessed spectrum (SET1_P). (c) Detail of original spectrum (*m/z*: 350 to 450). (d) Detail of preprocessed spectrum (*m/z*: 350 to 450). In (b), it can be seen a huge decrease in the number of peaks, yet a higher number of ions was found. In (a) only 3.98% of peaks are annotated, while in (b) this value raised to 55.05%. The treated spectrum shows that MASCOT was able to detect 7 new golden peaks (in blue) that were hidden in the background before preprocessing. Nevertheless, *b*(15) got lost. In (c) and (d) the efficiency of the cleaning and the intensity adjustment steps for the zoomed range is demonstrated (another range can be seen in Figure 1S of supplementary data). See supplementary data for the complete result report and for zooming other regions of the figure.

We have observed that peaks representing neutral losses of water and ammonia were regularly kept in spectra. It is probably due to the fact that these ions have similar properties when compared to *b* and *y* ions, such as the presence of isotopes and further water losses, suggesting similarities in their vicinities as well. However, since standard DB search algorithms expect the occurrence of water and ammonia losses, such misclassifications did not prevent our method from yielding to important improvements (Section Statistical Analysis). Reviewing Figure 4b, we can also observe some nonannotated peaks. This is either due to their lower intensities (i.e., they have *P* close to 0.5) or due to the fact that they represent ions (precursor and its derivatives, internal fragments, etc) with properties similar to golden peaks, alike water and ammonia losses. But, even with some misclassifications, the increase in S/N was remarkable.

Another kind of manual inspection we performed was the analysis of the number of matches and the coverage for the same proteins identified before and after preprocessing. This investigation strongly indicated that many spectra could be correctly interpreted after preprocessing. An example from SEQUEST results is shown in Figure 5. For the given protein identified in SET1_O and SET1_P, besides observing a higher number of matches and better coverage for the second dataset, we can also notice that the quality parameters were mostly better for the peptides reported in the preprocessed case, turning this protein into a significant identification with a much better position in SEQUEST ranking for SET1_P. For an extended manual inspection of the peptides and proteins identified in common between original and preprocessed data, see the complete result report of MASCOT and SEQUEST for all datasets in the supplementary data.

Statistical Analysis

In order to provide a comprehensive analysis of the over-

(a)

		Reference				P (pro)	Sf	Score	Coverage	MW	Accession	Peptide (Hits)	
		Scan(s)	Peptide	MH+	ΔM	z	P (pep)	Sf	XC	ΔCn	Sp	RSp	Ions	Count
☐	⊟	Isoform 1 of PITSLRE serine/threonine-protein kinase CDC2L1					1.0	1.34	18.2		91457.7	IPI00110050.2	4 (2 2 0 0 0)	
		10717,	K.RGTS#PRPPEGGLGYSQLGDDDLK.E	2565.23	0.02	3	1.0	0.61	4.296	0.011	341.1	1	30/13	1
		10717,	K.RGT#SPRPPEGGLGYSQLGDDDLK.E	2565.23	0.02	3	1.0	0.58	4.248	0.486	320.9	2	29/13	1
		10744,	K.RGTS#PRPPEGGLGYSQLGDDDLK.E	2565.23	0.02	3	1.0	0.70	4.351	0.029	439.0	1	35/13	1
		10744,	K.RGT#SPRPPEGGLGYSQLGDDDLK.E	2565.23	0.02	3	1.0	0.64	4.224	0.482	416.7	2	34/13	1

(b)

		Reference				P (pro)	Sf	Score	Coverage	MW	Accession	Peptide (Hits)	
		Scan(s)	Peptide	MH+	ΔM	z	P (pep)	Sf	XC	ΔCn	Sp	RSp	Ions	Count
☐	⊟	Isoform 1 of PITSLRE serine/threonine-protein kinase CDC2L1					2.4e-00	1.9	32.3		91457.7	IPI00110050.2	7 (2 2 1 2 0)	
		229,	R.DHRMEIT#IRNS#PYRR.E	2145.98	-0.02	3	0.8	0.0	1.457	0.151	98.5	90	13/11	1
		8265,	-.M*GDEKDS#WKVKTLDEILQEK.K	2586.28	-0.02	3	0.7	0.0	2.138	0.089	200.7	2	17/11	1
		10717,	K.RGTS#PRPPEGGLGYSQLGDDDLK.E	2565.23	0.02	3	2.4e-00	0.9	5.548	0.013	1189.9	1	38/13	1
		10717,	K.RGT#SPRPPEGGLGYSQLGDDDLK.E	2565.23	0.02	3	2.4e-00	0.9	5.478	0.511	1182.4	2	38/13	1
		10744,	K.RGTS#PRPPEGGLGYSQLGDDDLK.E	2565.23	0.02	3	3.3e-00	0.8	5.708	0.013	699.6	1	35/13	1
		10744,	K.RGT#SPRPPEGGLGYSQLGDDDLK.E	2565.23	0.02	3	3.3e-00	0.8	5.635	0.466	694.9	2	35/13	1
		14686,	K.SLMETM*KQPFLPGEVKTLMIQLLSGVK.H	3076.69	0.03	3	0.8	0.0	0.715	0.302	76.4	4	10/10	1

Figure 5: Comparison of SEQUEST results for a protein identified before and after preprocessing. (a) SET1 O. (b) SET1 P. In (b), three new sequences were included, resulting in a higher number of matches and better coverage. Comparing spectra with the same scan number in (a) and (b), the quality values were clearly more favorable after preprocessing. The symbol "#" means phosphorylation, while "*" denotes oxidation of methionine.

all benefit of our method, statistical measurements from SEQUEST, MASCOT and TransProteomic Pipeline (Keller et al., 2005) were considered. In a first analysis, we estimated the number of correct assignments made by MASCOT and SEQUEST. This calculation was performed using a decoy DB approach (Elias et al., 2005; Balgley et al., 2007). In this method, the search is repeated against a database with randomized or reversed sequences (decoy DB). As a consequence of this search, the number of positive matches can be used to estimate the FDR. MASCOT employs a decoy DB composed by randomized sequences. Here, the FDR calculation is based on the number of significant hits in normal and decoy DBs. Note that, in MASCOT, a peptide identification is regarded as significant if its score is above the homology or identity thresholds, the latter being the strongest quality measure (Perkins et al., 1999). Table 2 shows the comparative analysis after MASCOT search where a significantly improved sensitivity (= rate of correct hits obtained) and specificity (= rate of incorrect hits below the employed threshold) could be in general demonstrated. When using the identity threshold, the number of estimated correct hits increased 20% (from 175 to 210), 35% (from 48 to 65) and 33% (from 94 to 125), respectively, after spectra preprocessing. For the homology threshold, though in a lower percentage, all treated datasets presented improvements in the number of correct hits as well. Note yet that, except for one case (from 3.7% to 4.46%), the FDR was systematically diminished.

	# of hits in Mouse DB	# of hits in Decoy DB	FDR
SET1_O	176	1	0.57 %
	378	14	3.7 %
SET1_P	210	0	0 %
	404	18	4.46 %
SET2_O	48	0	0 %
	170	8	4.71 %
SET2_P	65	0	0 %
	186	7	3.76 %
SET3_O	99	5	5.05 %
	288	16	5.56 %
SET3_P	126	1	0.79 %
	305	10	3.28 %

Table 2: Analysis of significant hits in MASCOT results using a decoy DB approach. In the first line of each case the values are based on the identity threshold, while in the second line the homology threshold is the reference.

Since SEQUEST does not provide any significance threshold for peptide identifications, we employed several score thresholds for calculating the FDR. It should be noted that such experiments have to be done separately for each result because a particular threshold used in a dataset does not necessarily provide the same FDR in another dataset. For our experiments, a decoy DB was generated by reversing the peptide sequences contained in the original

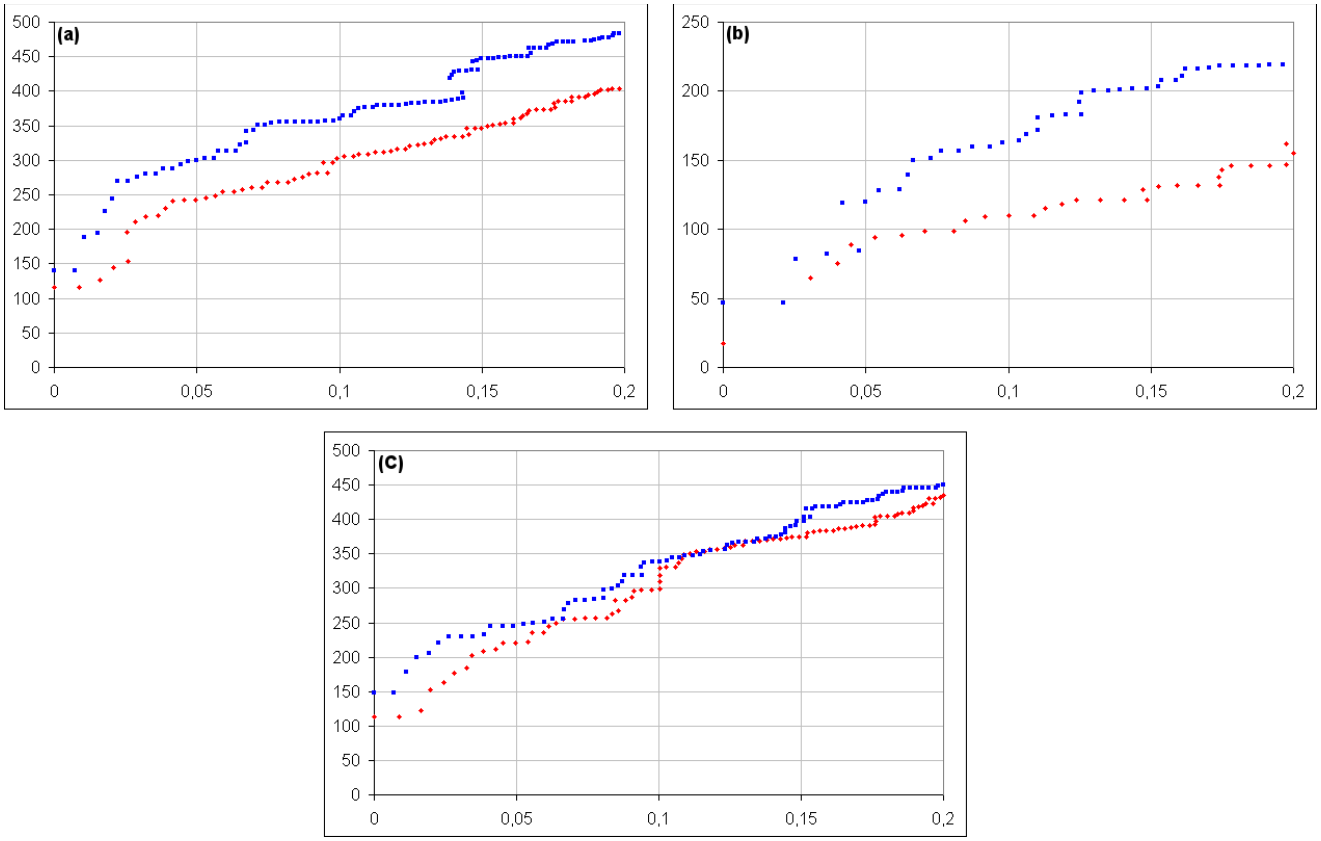

Figure 6: Decoy DB analysis for SEQUEST results. (a) SET1_O/P. (b) SET2_O/P. (c) SET3_O/P. Each data point corresponds to a different Xcorr threshold. The *x* axis indicates the estimated FDR (calculated by counting the hits above the threshold for the decoy DB), while the *y* axis shows the number of hits above the threshold when using the mouse DB. The red curves correspond to the original datasets and the blue curves relate to the treated datasets. Taking the same FDRs in original and preprocessed data, a higher number of hits comming from the mouse DB was in general observed for treated spectra.

DB (Elias et al., 2005; Balgley et al., 2007). A broad range of Xcorr thresholds was used for each dataset in order to obtain FDRs varying from 0 to 0.2. Figure 6 shows for the same FDRs that preprocessed datasets lead normally to a higher number of correct assignments. Even having some regions in Figure 6b and 6c in which the curves coincide, it is clear the dominance of the blue curves in all cases. For some FDR intervals, it can be seen an augmentation in the number of correct identifications of up to 40%, 60%, 33% in the datasets, respectively, after using our approach.

To further verify our studies, we used also Trans-Proteomic Pipeline (v3.2). This tool is a pipeline for analysis of LC-MS/MS proteomics data, containing modules for validation of peptide identifications (PeptideProphet) and protein inference (ProteinProphet). PeptideProphet was used to evaluate MASCOT and SEQUEST results. PeptideProphet uses the expectation maximization algorithm to compute probabilities of peptide assignments. Applying its statistical model, PeptideProphet provides the sensitivity/FDR tradeoff as ilustrated in Figure 7. The plots show that preprocessed datasets presented a better sensitivity, yet same or lower error in general for both MASCOT and SEQUEST when compared to the original data. Particularly in Figure 7b, the difference for treated spectra is very prominent.

In Figure 8, Venn Diagrams are shown. The diagrams contain the number of identifications reported by PeptideProphet with probability 0.9 or higher. This threshold is appropriate as the error rates for this value were low (varied from 1% to 3%. See supplementary data). It can be noted that the number of exclusive identifications for preprocessed data was always higher than for initial datasets.

In Figure 9, our analyses are moved to the protein level by means of ProteinProphet. This program uses PeptideProphet results to identify proteins and to associate probabilities to these identifications. As can be seen in Fig-

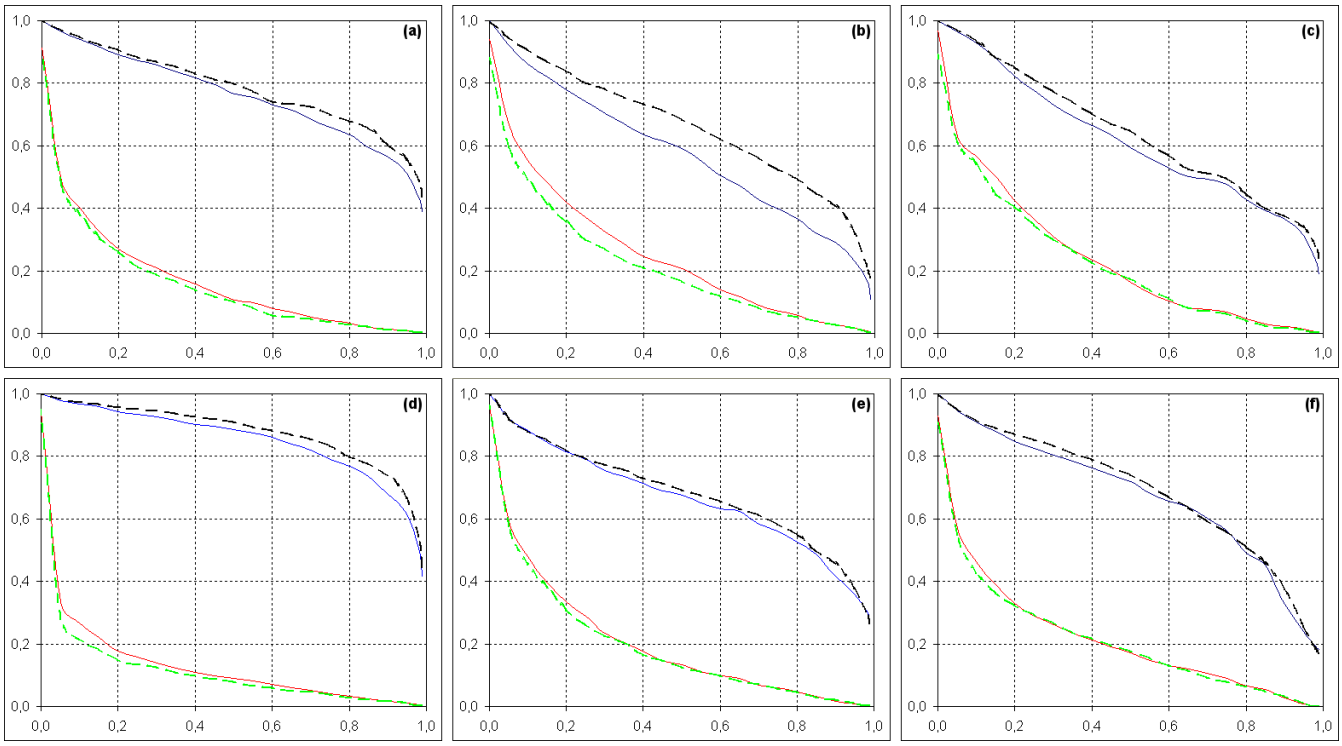

Figure 7: Sensitivity and FDR curves provided by PeptideProphet. (a-c) SET1_O/P, SET2_O/P, and SET3_O/P, respectively, using MASCOT. (d-f) SET1_O/P, SET2_O/P, and SET3_O/P, respectively, using SEQUEST. The *x* axis shows the different probability thresholds, while the *y* axis designates sensitivity and error. The blue and the red curves represent sensitivity and FDR, respectively, of untreated data, while the black and the green dashed curves describe sensitivity and FDR, respectively, of preprocessed spectra. The dominance tendency of the black curves upon the blue ones as well as the red curves upon the green ones indicates that the sensitivity was improved, yet the error decreased in the datasets obtained after application of the proposed method.

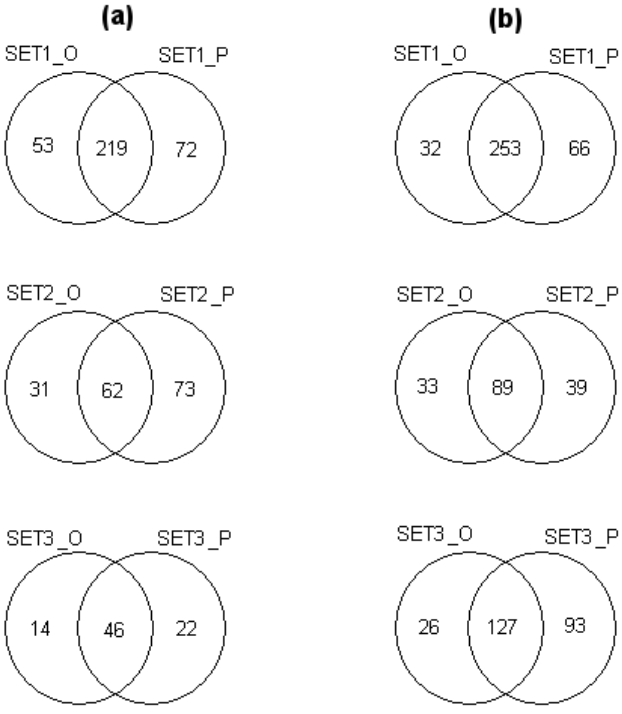

Figure 8: Venn Diagrams of peptide identifications reported by PeptideProphet with Prob≥0.9. (a) Using MASCOT results. (b) Using SEQUEST results. The number of exclusive identifications is always higher in treated datasets for both tools.

ure 9, ProteinProphet provides also the sensitivity and FDR curves, which open again the possibility to compare the different scenarios for the tested datasets. The Figure reveals once more the improvements in sensitivity and specificity for the datasets produced by our method. In the protein level, the enhancements are even more obvious. The area between black and green curves (related to treated data) is clearly higher than the area between blue and red curves (regarding the unprocessed data) for all cases in both MASCOT and SEQUEST runs.

Finally, ProteinProphet was used also to compare the same proteins identified in original and preprocessed data. For each protein identified in both cases, we computed the num-

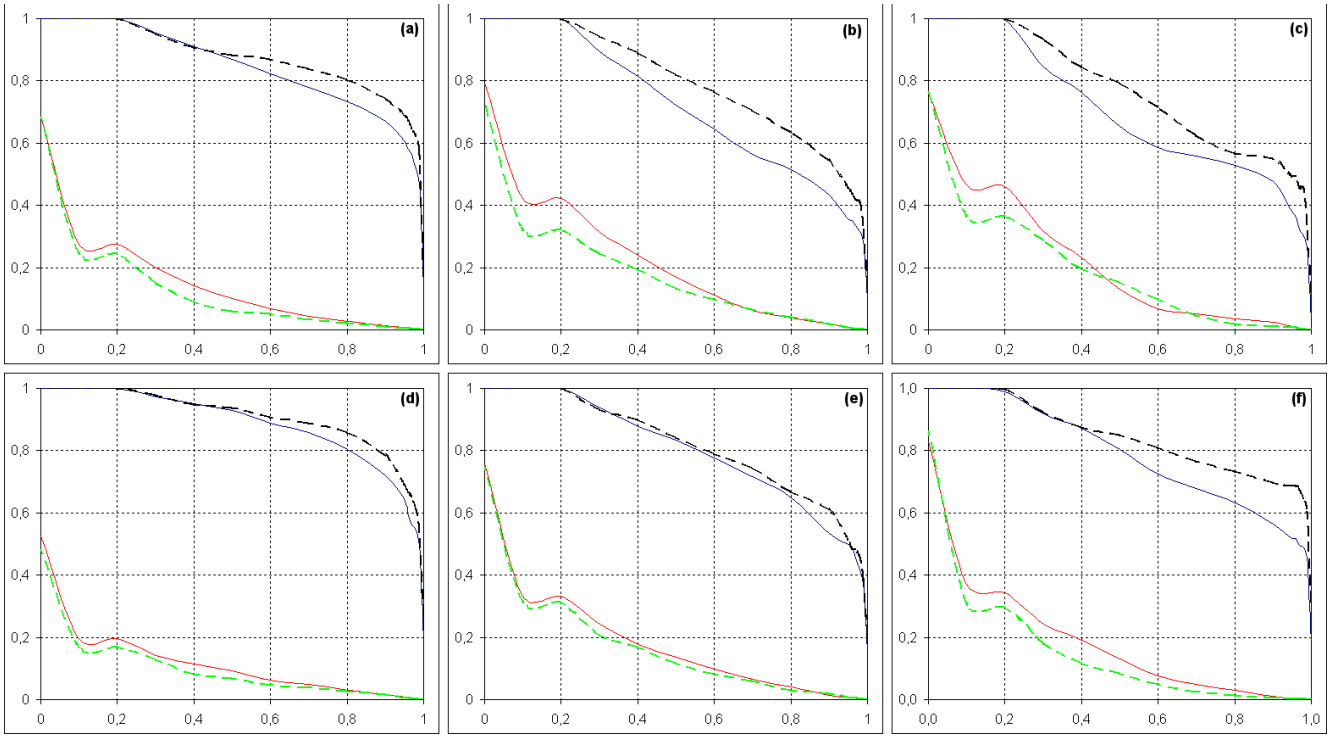

Figure 9: Sensitivity and FDR curves provided by ProteinProphet. (a-c) SET1_O/P, SET2_O/P, and SET3_O/P, respectively, using MASCOT. (d-f) SET1_O/P, SET2_O/P, and SET3_O/P, respectively, using SEQUEST. The description of these plots is the same provided in Figure 7, with the difference that here the analysis is in the protein level. Once more the preprocessed datasets demonstrated a better sensitivity and specificity (lower FDR). The enhancements in the protein level are even more prominent than in the peptide level (Figure 7).

ber of peptide matches and sequence coverage in order to verify if both measurements get improved after preprocessing. Figure 10 shows this comparison. It can be seen that the occurrence of proteins with more matches or higher coverage is much more frequent in treated than in original datasets.

Storage and Runtime Requirements

The experiments described in previous sections were performed using a stand-alone PC: Pentium(R) 4 CPU 3.20GHz, 1GB RAM, under Windows XP. A significant decrease in storage requirements was achieved, as the number of peaks in treated spectra could be reduced on average by 84%. The running times of MASCOT and SEQUEST for preprocessed data, in turn, could be shortened by approximately 28%. However, additional processing time is necessary for TS construction and dataset treatment (Section Classifying Peaks and Adjusting Intensities). On the other hand, it should be noted that curators take days or even weeks for separating correct from incorrect identifications manually, which means that the additional time caused by our data mining method is therefore negligible. Furthermore, the overall processing time can be significantly reduced using a powerful computer cluster.

Conclusion

Existing computational tools for MS/MS spectra interpretation produce a high number of FP identifications. This is due to the presence of background signals in spectra, which introduces multiple mass combinations. Phosphopeptide spectra, produced by low energy dissociation, demonstrate an additional source of complexity because the phosphopeptide fragmentation is frequently poor, giving rise to low intensity peaks of *b* and *y* ions. Consequently, phosphopeptide spectra interpretation presents an even higher false discovery rate.

In order to decrease the number of FP identifications for phosphopeptides, we have developed a new data mining approach for spectra preprocessing. With a dynamic TS construction, our method trains a SVM algorithm to classify

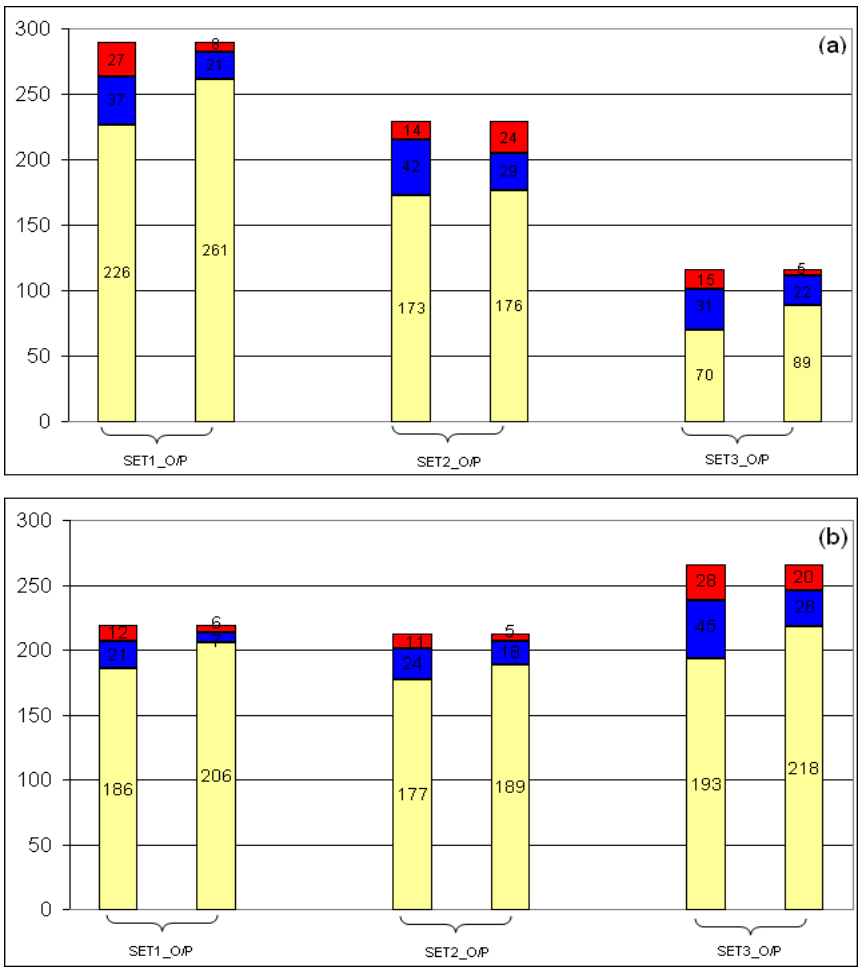

Figure 10: Analysis of ProteinProphet concerning the number of proteins with more matches / better coverage among the same proteins identified before and after the application of the proposed method. The red region contains the number of proteins having more matches/higher coverage in original dataset. The blue region shows the number of proteins having more matches/higher coverage in preprocessed dataset. Finally, the yellow part holds the number of proteins in which no difference arose between data before and after preprocessing (the number of matches or the coverage did not change). In each comparison, the left bar shows the analysis related to the number of matches, while the right bar regards the coverage. The number of proteins with more matches or better coverage is normally much higher after preprocessing.

(cleaning step) and adjust intensities of peaks in MS/MS spectra. The cleaning step facilitates an elimination of up to 84% of peaks, leading to a huge increase in S/N, while the intensity adjustment step improves the chances of detecting important signals. Since we could not find any previously proposed methods for phosphopeptide spectra preprocessing, no comparisons could be performed between our procedures and other approaches. Nonetheless, the presented statistical evaluations demonstrate that our method improves sensitivity of phosphopeptide identification significantly without compromising specificity. A decoy DB analysis using MASCOT and SEQUEST searches showed that the number of correct hits could be significantly enhanced in all preprocessed datasets. Likewise, Peptide/ProteinProphet confirmed this benefit for both tools. In summary, our experiments demonstrate that our data mining framework for preprocessing MS/MS spectra is a powerful tool for enhancing phosphopeptide/protein identifications in standard DB search tools.

Acknowledgment

This work is supported by the Austrian Genome Program (GEN-AU), project Bioinformatics Integration Network (BIN II). The work in the Huber laboratory is supported by the Austrian Proteomics Platform (APP) within the GEN-AU, Vienna, Austria.

Addendum: Web Supplement

Supplementary resources are available at the website: http://biomed.umit.at/page.cfm?pageid=515. It contains supplementary figures; SEQUEST, MASCOT and TransProteomic Pipeline results; and training sets. The software is available on request from the authors.

References

1. Balgley BM, Laudeman T, Yang L, Song T, Lee CS (2007) Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. Mol Cell Proteomics 6: 1599-1608. » CrossRef » Pubmed » Google Scholar

2. Cristianini N, Shawe TJ (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press Cambridge. » Google Scholar

3. de la Fuente van Bentem S, Anrather D, Roitinger EDA, et al. (2006) Phosphoproteomics reveals extensive *in vivo* phosphorylation of Arabidopsis proteins involved in RNA metabolism. Nucleic Acids Res 34: 3267-3278. » CrossRef » Pubmed » Google Scholar

4. Dehmeshki J, Karakoy M, Casique MV (2003) A rule-based scheme for filtering examples from majority class in an imbalanced training set. In: MLDM 2003: pp215-223. » CrossRef » Google Scholar

5. Elias JE, Haas W, Faherty BK, Gygi SP (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods 2: 667-675. » CrossRef » Pubmed » Google Scholar

6. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5: 976-989. » CrossRef » Google Scholar

7. Fan R, Chen P, Lin C (2005) Working set selection using second order information for training support vector machines. J Machine Learning Research 6: 1889-1918. » CrossRef » Google Scholar

8. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuousvalued attributes for classification learning. In: Proceedings of the international joint conference on uncertainty in AI 1993: pp1022-1027. » CrossRef » Google Scholar

9. Ferrige AG, Seddon MJ, Jarvis S, Skilling J, Aplin R (1991) Maximum entropy deconvolution in electrospray mass spectrometry. Rapid Commun Mass Spectrom 5: 374-377. » CrossRef » Google Scholar

10. Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, et al. (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat Biotechnol 20: 301-305. » CrossRef » Pubmed » Google Scholar

11. Frank A, Pevzner P (2005) PepNovo: *De novo* peptide sequencing via probabilistic network modeling. Anal Chem 77: 964-973. » CrossRef » Pubmed » Google Scholar

12. Gentzel M, Kocher T, Ponnusamy S, Wilm M (2003) Preprocessing of tandem mass spectrometry data to support automatic protein identification. Proteomics 3: 1597-1610. » CrossRef » Pubmed » Google Scholar

13. Hoffert JD, Wang G, Pisitkun T, Shen RF, Knepper MA (2007) An automated platform for analysis of phosphoproteomic datasets: application to kidney collecting duct phosphoproteins. J Proteome Res 9: 3501-3508. » CrossRef » Pubmed » Google Scholar

14. Hsia J (2005) Encyclopedia of biostatistics, 2nd Edition. Wiley, Chichester, Ch. Sampling with and without replacement.

15. Imanishi SY, Kochin V, Ferraris SE, Thonel A, Pallari HM, et al. (2007) Reference-facilitated phosphoproteomics: Fast and reliable phosphopeptide validation by μLC-ESI-Q-TOF MS/MS. Mol Cell Proteomics 6: 1380-1391. » CrossRef » Pubmed » Google Scholar

16. Ishihama Y, Wei FY, Aoshima K, Sato T, Kuromitsu J, et al. (2007) Enhancement of the efficiency of phosphoproteomic identification by removing phosphates after phosphopeptide enrichment. J Proteome Res 3: 1139-1144. » CrossRef » Pubmed » Google Scholar

17. Jaitly D, Belanger RP, Faubert D, Thibault P, Kearney P (2004) MSMS peak identification and its applications. In: ISMB/ECCB 2004.

18. Kapp EA, Schtz F, Connolly LM, Chakel JA, Meza JE, et al. (2005) An evaluation, comparison, and accurate benchmarking of several publicy available MS/MS search algorithms: Sensitivity and specificity analysis. Proteomics 5: 3475-3490. » CrossRef » Pubmed » Google Scholar

19. Keller A, Eng J, Zhang N, Li X, Aebersold R (2005) A uniform proteomics MS/MS analysis platform utilizing

open XML file formats. Mol Syst Biol 1-8. »CrossRef » Pubmed » Google Scholar

20. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74: 5383-5392. »CrossRef » Pubmed » Google Scholar

21. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, et al. (2004) The international protein index: An integrated database for proteomics experiments. Proteomics 4: 1985-1988.» CrossRef » Pubmed » Google Scholar

22. Knebel A, Morrice N, Cohen P (2001) A novel method to identify protein kinase substrates: eEF2 kinase is phosphorylated and inhibited by SAPK4/p38δ. EMBO J 20: 4360-4369. »CrossRef » Pubmed » Google Scholar

23. Kocher T, Savitski MM, Nielsen ML, Zubarev RA (2006) PhosTShunter: A fast and reliable tool to detect phosphorylated peptides in liquid chromatography fourier transform tandem mass spectrometry data sets. J Proteome Res 5: 659-668.» CrossRef » Pubmed » Google Scholar

24. Lu B, Ruse C, Xu T, Park SK, Yates J (2007) Automatic validation of phosphopeptide identifications from tandem mass spectra. Anal Chem 4: 1301-1310. »CrossRef » Pubmed » Google Scholar

25. Morandell S, Stasyk T, Grosstessner HK, Roitinger E, Mechtler K, et al. (2006) Phosphoproteomics strategies for the functional analysis of signal transduction. Proteomics 6: 4047-4056» CrossRef » Pubmed » Google Scholar

26. Mujezinovic N, Raidl G, Hutchins JRA, Peters J, Mechtler K, et al. (2006) Cleaning of raw peptide MS/MS spectra: Improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. Proteomics 6: 5117- 5131.» CrossRef » Pubmed » Google Scholar

27. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75: 46460-4658. »CrossRef » Pubmed » Google Scholar

28. Noble WS (2006) What is a support vector machine. Nat Biotechnol 24: 1565-1567. »CrossRef » Pubmed » Google Scholar

29. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551-3567.» CrossRef » Pubmed » Google Scholar

30. Plant C, Osl M, Bernhard T, Baumgartner C (2006) Feature selection on high throughput SELDI-TOF mass-spectrometry data for identifying candidates in ovarian and prostate cancer. In: IEEE ICDM 2006 workshop on data mining in bioinformatics (DMB 2006). p. not shown.

31. Reinhold BB, Reinhold VN (1992) Electrospray ionization mass spectrometry: Deconvolution by an entropy based algorithm. J Am Soc Mass Spectrom 3: 207-215. »CrossRef » Google Scholar

32. Schroeder MJ, Shabanowitz J, Schwartz JC, Hunt DF, Coon JJ (2004) A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. Anal Chem 76: 3590-3598. »CrossRef » Pubmed » Google Scholar

33. Shah K, Liu Y, Deirmengian CMSK (1997) Engineering unnatural nucleotide specificity for Rous sarcoma virus tyrosine kinase to uniquely label its direct substrates. Proc Natl Acad Sci USA 94: 3565-3570. »CrossRef » Pubmed » Google Scholar

34. Steen H, Mann M (2004) The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol 5: 699-711. »CrossRef » Pubmed » Google Scholar

35. Wessel D, Flugge UI (1984) A method for the quantitative reco very of protein in dilute solution in the presence of detergents and lipids. Anal Biochem 138: 141-143. »CrossRef » Pubmed » Google Scholar

36. Witten IH, Frank E (2005) Data mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco.

37. Wu TF, Lin CJ, Weng RC (2004) Probability estimates for multi-class classication by pairwise coupling. JMLR 99: 975-1005. »CrossRef » Google Scholar

38. Zhang Z, Marshall AG (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. J Am Soc Mass Spectrom 9: 225-233» CrossRef » Pubmed » Google Scholar