A scoring model for phosphopeptide site localization and its impact on the question of whether to use MSA

Juliana de S. da G. Fischer, Marlon D.M. dos Santos, Fabricio K. Marchini, Valmir C. Barbosa, Paulo C. Carvalho, Nilson I.T. Zanchin

Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil
Functional Genomics Laboratory, Carlos Chagas Institute, Fiocruz, Paraná, Brazil
Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 26 November 2014
Accepted 16 January 2015
Available online 23 January 2015

Keywords:
Phosphoproteomics
Phosphopeptide enrichment
HAP chromatography
Multistage activation (MSA)

ABSTRACT

The production of structurally significant product ions during the dissociation of phosphopeptides is a key to the successful determination of phosphorylation sites. These diagnostic ions can be generated using the widely adopted MS/MS approach, MS3 (Data Dependent Neutral Loss — DDNL), or by multistage activation (MSA). The main purpose of this work is to introduce a false-localization rate (FLR) probabilistic model to enable unbiased phosphoproteomics studies. Briefly, our algorithm infers a probabilistic function from the distribution of the identified phosphopeptides' XCorr Delta scores (XD-Scores) in the current experiment. Our module infers p-values by relying on Gaussian mixture models and a logistic function. We demonstrate the usefulness of our probabilistic model by revisiting the "to MSA, or not to MSA" dilemma. For this, we use human leukemia-derived cells (K562) as a study model and enriched for phosphopeptides using the hydroxyapatite (HAP) chromatography. The aliquots were analyzed with and without MSA on an Orbitrap-XL. Our XD-Scoring analysis revealed that the MS/MS approach provides more identifications because of its faster scan rate, but that for the same given scan rate higher-confidence spectra can be achieved with MSA. Our software is integrated into the PatternLab for proteomics freely available for academic community at http://www.patternlabforproteomics.org.

Biological significance
Assigning statistical confidence to phosphorylation sites is necessary for proper phosphoproteomic assessment. Here we present a rigorous statistical model, based on Gaussian mixture models and a logistic function, which overcomes shortcomings of previous tools. The algorithm described herein is made readily available to the scientific community by integrating it into the widely adopted PatternLab for proteomics.

This article is part of a Special Issue entitled: Computational Proteomics.

© 2015 Elsevier B.V. All rights reserved.

This article is part of a Special Issue entitled: Computational Proteomics.

Corresponding authors at: Rua Prof. Algacyr Munhoz Madder 3775, Cidade Industrial de Curitiba, CEP: 81350-010, Brazil.
E-mail addresses: paulo@pcarvalho.com (P.C. Carvalho), nzanchin@gmail.com (N.I.T. Zanchin).
1. Introduction

Protein phosphorylation is a key regulatory mechanism in biological systems, being a central regulator of intracellular processes [1,2]. Reversible protein phosphorylation is involved in the regulation of various processes, such as cellular signaling and communication, cell homeostasis, differentiation, proliferation and survival, protein degradation, and metabolism, to name a few [3]. Consequently, much effort has been directed toward developing tools and methods to effectively unravel the phosphoproteome of complex protein mixtures [4], striving to ultimately obtain a better understanding of the biological system at hand [5,6].

Large-scale phosphopeptide identification experiments are usually carried out by using mass spectrometry to dissociate peptides and to match their experimental tandem mass spectra to those theoretically generated from a sequence database [7]. This dissociation procedure happens by isolating a precursor ion of a given m/z and then colliding it with neutral atoms or molecules in gas phase to ultimately obtain a mass spectrum of the precursor’s product ions. Collision-induced dissociation (CID) [8] and Higher-energy collisional dissociation (HCD) [9] are examples of strategies for dissociating peptide ions (hereafter we refer to this procedure simply as MS2). Likewise, the sequential isolation and fragmentation (e.g., MS3) of these product ions can be accomplished by trapping analyzers multiple times in a process known as MSn. When analyzing phosphopeptides, the corresponding precursor ions commonly undergo neutral losses of an H3PO4, or ~98 Da during the procedure. As such, performing MS3 data-dependent neutral-loss (DDNL) analyses on these precursor ions becomes an interesting strategy to further increase confidence in phosphopeptide and phosphorylation-site identification. Yet, partial fragmentation of the precursor takes place during its first isolation, leading to a decrease in the signal-to-noise ratio in the MS3 spectra.

Multistage activation (MSA) emerged as a fragmentation strategy very suitable for large-scale phosphoproteomic experiments [10], especially for Orbitrap-XL instruments. MSA, sometimes described as a pseudo-MS2, allows for the neutral-loss ion originating from the precursor to undergo collision activation while still in the ion trap. This makes MSA spectra significantly faster to acquire than MS3, since there is no need for another isolation step. Thus, a single MSA mass spectrum contains product ions from both the precursor and its neutral-loss product. On the other hand, MSA spectra take more time to acquire than the standard MS2. Even though MS2 contains fewer reporter ions, site determination can still be satisfactorily accomplished in many cases. As a matter of fact, it can be argued that even though the success rate for site determination in MS2 spectra is lower than that of MSA, this could be compensated for by the higher number of MS2 spectra. In our view, a consensus has yet to emerge in the scientific community regarding when to MSA, or not to MSA. Even though previous publications have confronted these strategies [11], the literature continues to provide examples of high-quality phosphoproteomic studies that do [12] and do not use [13] MSA for large-scale studies. We must point out that for last generation instruments, HCD has significantly improved and is routinely used in the analysis of phosphopeptides.

In this work we introduce a scoring system for assessing phosphoproteomic data. We exemplify its usefulness by performing an unbiased comparison between the MSA and MS2 dissociation strategies, and extend this discussion from where previous works left off [11]. In general, peptide spectrum matching (PSM) search tools such as SEQUEST [7] and Mascot [14] do not provide a score for proper post-translational modification site assignment. Notwithstanding this, several scoring systems have been developed to assign confidences to phosphorylated sites, all employing local heuristics to check for the presence and intensity order of diagnostic-ion fragments in tandem mass spectra [15,16].

Stemming from a different paradigm, we bring attention to the widely adopted Mascot Delta score or MD-Score [17]. Although simple in its essence, the MD-Scoring system has been demonstrated to be extremely efficient when benchmarked on a set of 180 synthetic peptides against competing scoring systems. Briefly, in a mass spectrum whose top scoring candidate is a phosphopeptide, the MD-Score is calculated as the difference between the top two Mascot ion scores of alternative phosphorylation sites in the same peptide sequence. The “delta” concept behind the MD-Score has been found to be so efficient that other groups extended it to different search engines and demonstrated how to benefit from the systematic, simultaneous use of more than one such engine [18]. A drawback of the MD-Score is its exclusive applicability to Mascot [14], a commercially available search engine, which thus restricts the broader proteomics community from adopting it. By capitalizing on its simplicity and efficiency, we introduce the XD-Score, which has roots on the same premises but uses the Cross Correlation (XCorr), itself employed by widely adopted search engines such as SEQUEST [7], ProLuCID [19], and the open-source and freely available Comet [20]. We make the XD-Scoring system freely available to academic use as a new module integrated into the PatternLab for proteomics software suite [21,22]. PatternLab is a one-stop shop for analyzing proteomic data. It is fully GUI-driven and very user-friendly, so including a natively integrated scoring system into it has the potential of encouraging the scoring system’s widespread adoption.

One novelty of our XD-Scoring system is its freedom from previously established, fixed-cutoff scores, unlike MD-Scoring. Quoting the evaluators of MD-Scoring, the system needs to be tuned for every given setup: “we make the physical phosphopeptide collection available to the community so that any laboratory can determine and implement MD-score characteristics for their particular analytical set up” [17]. To overcome this limitation, we propose to sophisticate the strategy in order to make it to adjust its cutoff scores dynamically to reflect p-values that are inherent to the nature of the experiment at hand. This depends on many factors, including the mass spectrometer being used, database size and sample complexity, so the proposed automation relieves the user from the burden of having to make adjustments based on phosphopeptide libraries for each instrumentation setup; such libraries are made available, and there are many advices on how to use them [23]. To accomplish this, the new module relies on the distribution of XCorr scores to establish probabilistic cutoffs.
We believe our framework can be made to function with the MD-Scoring system as well, which is bound to favor an unbiased comparison between the different dissociation and enrichment strategies. Our scoring system is exemplified in this work using the human leukemia-derived cell lineage K562 [24].

Another challenge when dealing with phosphoproteomic studies is that phosphorylated proteins are generally substoichiometric, thus making it imperative to enrich for phosphorylated peptides prior to large-scale comprehensive analyses. In this regard, there are a number of existing methods for enriching for phosphorylated peptides; for an in-depth understanding of the methods related to phosphoanalyses we recommend reviews that address widely adopted strategies, such as Immobilized Metal Affinity Chromatography, Titanium Dioxide [25–27], and even more general fractionation strategies [28]. In this work, we relied on the hydroxyapatite (HAP) chromatography for enriching for phosphopeptides. HAP chromatography uses a naturally occurring metal salt (Ca_{10}(PO_{4})_{3}(OH)) and crystallizes as a hexagonal, close-packed structure, positioning calcium ions at the surface in a triangle a few angstroms apart [29]. HAP chromatography has been previously used for analytical protein separations [30,31]. More recently, Bio-Rad Laboratories introduced macroporous ceramic HAP (CHT), thus improving mechanical stability, and increased surface area 20-fold over microcrystalline HAP, thus yielding significant gains for using this material in high performance liquid chromatography (HPLC) systems [32,33]. Two recent publications have demonstrated the effectiveness of HAP chromatography in protein standards using commercially available spherical ceramic macroporous hydroxyapatite particles [34] and monolith-imbedded hydroxyapatite nanoparticles [35].

2. Experimental section

2.1. Materials

Leukemia-derived cells of the K562 lineage were kindly donated by the Laboratory of Tumoral Immunology of the Federal University of Rio de Janeiro, Brazil. RPMI 1640 medium, penicillin, streptomycin, and fetal bovine serum (FBS) were purchased from Gibco. Qubit® Protein Assay Kit and RapiGest SF acid-labile surfactant were purchased from Invitrogen (Carlsbad, CA) and Waters Corp. (Milford, MA), respectively. Hydroxyapatite resin was acquired from Bio-Rad (cat. # 158-2000). PhosSTOP Phosphatase Inhibitor cocktail tablet was purchased from Roche (cat. # 4906837001). All other laboratory reagents were acquired from Sigma-Aldrich (St. Louis, MO), unless specified otherwise.

2.2. Cell culture and sample preparation

K562 cells were cultivated in RPMI-1640 complete media supplemented with 10% heat inactivated fetal bovine serum (FBS), 2 mM l-glutamine, 100 U/mL penicillin, 100 μg/mL streptomycin, under 5% CO₂ and 95% humidified atmosphere at 37 °C in a CO₂ incubator. Cells were cultured and maintained in logarithmic growth phase until the number of cells reached 5 × 10⁶ cells/mL. The medium from all cultures was discarded and the cells were washed once with cold phosphate buffer saline (PBS).

Proteins were extracted from the cell pellets using RapiGest surfactant (0.2% final) conjointly with PhosSTOP inhibitor cocktail (1/10). Then, proteins were sonicated for 2 h in a cold water bath to break DNA strands. Subsequently, samples were centrifuged during 30 min at 20,000 × g to remove insoluble material. Afterwards, the total protein concentration was determined with the Qubit® Protein Assay Kit together with the manufacturer’s instructions. Ten milligrams of proteins were reduced with 10 mM (final) of dithiothreitol at 37 °C and, after cooled to room temperature, the proteins were alkylated with 30 mM (final) of iodoacetamide for an additional 30 min at room temperature. Finally, the proteins were digested with sequencing grade modified trypsin (E/S-1/100) during 18 h at 37 °C.

Following digestion, the reaction was acidified with 100% trifluoroacetic acid (0.4% final) to stop proteolysis and incubated at 37 °C for an additional 60 min to enhance the hydrolysis of RapiGest. Subsequently, the sample was centrifuged during 30 min at max speed (20,000 × g, 5 °C) to remove insoluble material and the soluble peptide mixture was collected and separated into six equal parts to continue with the following enrichment methods in duplicate technical replicates. The pH of the peptides mixture was measured with a pH string and showed to be approximately 3.0.

2.3. Phosphopeptide enrichment using HAP chromatography

This section was done according to Fonslow et al. with some modifications [33]. First, the pH of the digestion peptides was adjusted to approximately 7.5 with sodium hydroxide. Subsequently, acetonitrile (100%) was added to the sample to be 40% final (v/v). The HAP resin (30 mg) was activated with 500 μL of methanol and equilibrated with 1 mL of 20 mM Tris pH 8.0, 40% ACN. After preparation of the HAP resin, the peptides were incubated with 100 μL of the solution containing HAP (Tris pH 8.0, 40% ACN) for 1 h with end-over-end rotation at room temperature. Afterwards, the sample was spun down at 1000 × g and all liquid was removed. Then the resin containing the phosphopeptides was washed three times with 20 mM Tris, pH 8.0; 60% ACN (1 mL each wash). Finally, the resin was incubated with 500 μL of elution buffer A (20 mM K₂PO₄, pH 8.0) for 10 min with end-over-end rotation, followed by the incubation with the elution buffer B (100 mM of K₃HPO₄, pH 8.0). Each of the eluted samples was cleaned up with stage-tip and used for performing an independent LC/MSMS analysis.

2.4. LC–MS/MS acquisition

The dried phosphopeptides were subjected to LC-MS/MS analysis with a Thermo Scientific Easy-nLC 1000 ultra-high performance liquid chromatography (UPLC) system coupled with a LTQ-Orbitrap XL ETD mass spectrometer (Mass Spectrometry Facility-RTP024H PDTIS/Carlos Chagas Institute-Fiocruz Parâana), as follows. The peptide mixtures were loaded onto a column (75 mm i.d., 15 cm long) packed in house with a 3.2 μm ReproSil-Pur C18-AQ resin (Dr. Maisch) with a flow of 500 nL/min and subsequently eluted with a flow of 250 nL/min from 5% to 40% ACN in 0.5% formic acid and 0.5% DMSO, in a 120 min gradient [36]. The mass spectrometer was set in data-dependent
mode to automatically switch between MS and MS/MS (MS2) acquisition. Survey full scan MS spectra (from m/z 300–2000) were acquired in the Orbitrap analyzer with the resolution R = 60,000 at m/z 400 (after accumulation to a target value of 1,000,000 in the linear trap). The ten most intense ions were sequentially isolated and fragmented in the linear ion trap using collisional induced dissociation at a target value of 10,000. Previous target ions selected for MS/MS were dynamically excluded for 90 s. Total cycle time was approximately 3 s. The general mass spectrometric conditions were: spray voltage, 2.4 kV; no sheath and auxiliary gas flow; ion transfer tube temperature 175 °C; collision gas pressure, 1.3 mTorr; normalized energy collision energy using wide-band activation mode; 35% for MS2. Ion selection thresholds were: 250 counts for MS2. An activation q = 0.25 and activation time of 30 ms were applied in MS2 acquisitions. Technical replicates were performed with, and without, multistage activation.

2.5. Raw MS data analysis

The reference proteome set of Homo sapiens, composed of 68,625 sequences, was downloaded from the UniProt consortium on the 4th of July, 2014. PatternLab was used for generating a target-decoy database by grouping subset sequences, adding the 4th of July, 2014. PatternLab was used for generating a Bayesian discriminator. The identifications were sorted in non-decreasing order according to the discriminator score. A cutoff score was established to accept a false-discovery rate (FDR) of 1% at the peptide level based on the number of results were post-processed to only accept PSMs with less tryptic status or charge state. Additionally, a minimum label decoys. This procedure was independently performed on each data subset, resulting in an FDR that was independent of tryptic status or charge state. Additionally, a minimum sequence length of six amino-acid residues was required. Results were post-processed to only accept PSMs with less than 6 ppm from the global identification average. One-hit wonders (i.e., proteins identified with only one mass spectrum) having an XCorr of less than 2 were discarded. This last filter led to FDRs, now at the protein level, to be lower than 1% for all search results. For each mass spectrum, the search engine reported the top 50 scoring candidates to enable an effective XD-Score calculation.

2.7. Assessment of a global false-localization rate using PatternLab’s XD-Score module

The XD-Scoring module for false localization rate takes as input the sqt file (i.e., Comet’s output), which includes the top 50 identification candidates for each tandem mass spectrum. The module will preserve only results of tandem mass spectra whose top-scoring identification is a phosphopeptide. As a second step, all results having no difference in XCorr between the top-ranking candidate and the next (i.e., DeltaCN = 0) are eliminated. We use the scores from the remaining results to set up a logistic function that enables the estimation of p-values, as follows.

Let X be a random variable representing the score of some randomly chosen identification, and let $L = \ln(X)$. Our data are distributed in a way that is heavily skewed toward the lower values, so we use L instead of X. We wish to establish a p-value p to support the decision of whether a particular L = l refers to the positive (i.e., confident identification) or the negative (i.e., unconfident identification) class. We expect very low values of l to yield a p-value near 1. This suggests the use of a complementary logistic function, that is,

$$p = 1 - \frac{1}{1 + e^{-\gamma(l - \theta)}} = 1 - \frac{1}{1 + e^\gamma \theta}$$

for suitable choices of $\theta$ (the value of l for which $p = 0.5$) and $\gamma > 0$ (which regulates the function’s steepness at $l = \theta$, given by $-\gamma/4$). One difficulty, though, is that establishing $\theta$ and $\gamma$ by logistic regression is infeasible in our case, since all our data are unlabeled (i.e., we do not know, a priori, whether the phosphopeptide identification is correct or not). We then approach parameter determination through the following heuristic.

First, we resort to a representation of the data distribution by a mixture of two Gaussians, whose parameters can be determined in an unsupervised manner (in our algorithm, by expectation maximization), and thereby do away with the need for labeled data. We then use these two Gaussians in the determination of $\theta$. Denoting the integrals of the positive-class Gaussian by $P^+$ and those of the negative-class Gaussian by $P^-$, we achieve this by requiring that, for $l = \theta$, the two "misclassification" probabilities be the same. That is,

$$\Pr(L \leq \theta \mid \text{positive class}) = \Pr(L \geq \theta \mid \text{negative class})$$

or, equivalently,

$$P^-(L \leq \theta) = P^+(L \geq \theta)$$.

What is left to determine is the value of $\gamma$, and to this end we resort to some widely accepted pair $(\bar{L}, \bar{p})$. In other words, we choose $\gamma$ such that

$$\gamma = \frac{\ln((1-p)/p)}{L-\theta}.$$
To determine values for this pair, we refer to conservative values as seen by the scientific community, such as identifications with an XCorr of 3.0 and a DeltaCN of 0.1, which are considered by many to be very good. In this case, a worst-case scenario — of two isobaric phosphopeptide sequence candidates (i.e., candidates differing only in site localization) ranking as the top two sequences — would yield an XD-Score of 0.3. This allows us to determine a pair, (In 0.3, 0.01), that should be acceptable for a wide range of experiments and robust enough to support different mass spectrometers and database configurations.

We note, finally, that owing to our choice of a complementary logistic function to calculate \( p \) from \( l \), it is straightforward to determine \( p \) directly from a value \( x \) of the original score \( X \). That is, using \( x = e^l \) we have

\[
p = \frac{1}{1 + (xe^{-\theta})^\gamma}.
\]

3. Results and discussion

3.1. To MSA, or not to MSA, that is the question

The K562 tryptic peptides were enriched for phosphopeptides by HAP chromatography and then analyzed by tandem mass spectrometry with and, without MSA, according to Experimental section. Fig. 1 presents Venn diagrams summarizing the comparison of the sequences identified with these methods.

The results in Fig. 1 show that HAP-MS2 provided more phosphopeptide identifications than HAP-MSA for this scenario. We employed our XD-Scoring system to attribute \( p \)-values to the phosphopeptide site assignments. Fig. 2 shows the graphical user interface (GUI) of our software while performing the HAP-MSA analysis.

Fig. 3 shows the number of spectra satisfying a \( p \)-value cutoff. Our assessment of Fig. 3 is that, for the experiment at hand, in general HAP-MS2 provided more confident phosphopeptide site attributions than HAP-MSA. Yet, these results leave us wondering whether HAP-MS2 was more efficient mainly because of its faster scan rate. We note that 12,713 mass spectra were acquired for the HAP-MS2 dataset against only 9918 for HAP-MSA. In this regard, we prepared a computer script that randomly removes mass spectra from the HAP-MS2 file until it has the same number of spectra as HAP-MSA. We call this artificially generated proof-of-concept condition HAP-MS2 reduced, or simply HAP-MS2R. Redoing the analysis (search, filtering with SEPro, and \( p \)-value attribution to phosphosites) allowed us to investigate the hypothetical condition in which HAP-MS2 would have operated at the same scan rate as HAP-MSA. The resulting phosphopeptide comparison is presented in Fig. 4. Fig. 5 then follows with the number of phosphopeptides for a given \( p \)-value in the new condition done in triplicates.

These new results suggest that, in the proof-of-concept condition, HAP-MSA was significantly more efficient than HAP-MS2R, thus demonstrating the usefulness of MSA is generating more informative spectra and thereby increasing, in general, the confidence in site determination. By roughly contrasting Figs. 3 and 5, we note that, while HAP-MS2 provides some 50 spectra in excess of those of HAP-MSA under the 0.01 \( p \)-value, comparing HAP-MS2R versus HAP-MSA reveals a much stronger advantage in favor of HAP-MSA: almost 150 high-quality spectra are provided in excess of those of HAP-MS2R for the same \( p \)-value. In all, this suggests that, if MSA were as fast as MS2, it might be always the method of choice for analyzing large-scale phosphoproteomic experiments. As newer mass spectrometers with faster scan rates become available, we believe the appeal of using MSA will increase even more. As for very simple protein mixtures, our results suggest that, in general, MSA should fare better than MS2. However, one should mind the caveat that defining what is meant by a “simple mixture” given the instrumentation setup at hand remains elusive.

3.2. Considerations on the XD-Scoring system

We note that there are other implications when analyzing highly complex mixtures such as the ones of this work. As recently demonstrated, identification scores can degrade due to contaminating nearly-isobaric ions that are co-isolated and

---

**Fig. 1** – Area-proportional Venn diagrams of the numbers of peptide sequences identified from K562 using HAP chromatography analyzed by MS/MS (HAP-MS2) and by MSA (HAP-MSA). Panel A refers to all identified peptides (phosphorylated or not after enrichment), and panel B to phosphopeptides only. All identifications are under an FDR of 1% as provided by SEPro.
fragmented together with the target ions [38]. Indeed, reports have demonstrated that chimera mass spectra can easily range from 10% in yeast [39] to 50% or even more in complex samples like the K562 cells analyzed here [40]. In this respect, we believe that our method can provide an unbiased approach by using p-values obtained from scoring distributions, which in

Fig. 2 – XD-Scoring software graphical user interface (GUI) for the HAP-MSA analysis. The topmost textbox allows the user to specify a search engine result in the “.sqt” format [30]. Then the user can optionally specify a precursor charge state and choose to consider only peptides having a given number of phosphorylation sites when generating the statistical model. By pressing the “Calculate button”, a list of the XD-Scores for all search results satisfying the user’s criterion appears in the textbox below (left side of the GUI). Clicking on the Generate GMM button calculates a Gaussian mixture model and displays the resulting Gaussians (top plot, in which the x-axis refers to the natural logarithm of the XD-Score and the y-axis to the number of identifications). The logistic function generated using the GMM is displayed in the bottom plot by the thick purple curve, the green curve is the cumulative distribution of the green Gaussian and the red curve is the complementary distribution of the red Gaussian; once again, the x-axis refers to the natural logarithm of the XD-Score while the y-axis provides the corresponding p-value. Finally, the user can specify a SEPro file containing the confident peptide identifications and enabling the program to output a table associating a p-value to each site attribution. The tables generated, for the work at hand, are provided as part of the supplementary material available at the project’s website.

Fig. 3 – Number of phosphopeptide identifications satisfying a given p-value. The x-axis represents the p-values obtained from our logistic function and the y-axis represents the numbers of phosphopeptide identifications, provided by SEPro, satisfying that p-value. The plots on the right result from zooming in on the region of low p-values in the plots on the left.
turn can adapt to sample complexity and even to instrumentation setup. We thus believe to also be contributing a step toward overcoming the limitations pointed out by the original MD-Score manuscript, in which scores were generated using a standard library and thus required adaptation for different mass spectrometers and setups.

3.3. Final considerations

Here we introduced the XD-Scoring framework and exemplified its usefulness by comparing MSA with standard CID MS2 dissociation. Although the purpose of our example was to determine which dissociation strategy method is more effective, the conclusion is that this may vary with the complexity of the sample at hand. Yet, determining such complexity a priori might not be trivial. The XD-Scoring system stems from the well-established MD-Scoring system and extends it by attributing p-values based on distributions raised from experiment-wide scores. It is integrated into the freely available PatternLab for proteomics.

3.4. Data and software availability

Our mass spectrometry raw files, search results, list of proteins unique to each condition, and PatternLab intermediate files are all made available at http://proteomics.fiocruz.br/software/xdscore/.

The XDScore analyzer is part of the PatternLab for proteomics software suite [21,22]. The software was programed in C# and requires .NET 4.5 and a computer running Windows 7 or later.

Author contributions

NZ planned and designed the experiments. JSGF and MDMS performed the experiments. PCC and VCB analyzed the data and idealized the XD-Score framework. PCC programed the software. FKM aided during sample preparation. PCC, VCB, and JSGF wrote the paper. All authors reviewed the paper.

Acknowledgments

The authors thank the Program for Technological Development in Tools for Health, PDTIS-FIOCRUZ, for use of its facilities (RPT02H PDTIS/Carlos Chagas Institute-Fiocruz, Paraná). The authors thank Dr. Michel Batista for running the samples in the mass spectrometer. The authors thank Dr. Vivian M. B. D. Rumjanek, head of the Laboratory of Tumoral Immunology of the Federal University of Rio de Janeiro, for kindly donating the K562 cell lineage. This work was funded by PNPD-CAPES, Universal-CNPq grant number 21465213.5.0000.5248, FAPERJ BBP, Programa Estratégico de Apoio à Pesquisa em Saúde (PAPES), and Fundação Oswaldo Cruz.

Fig. 4 – Area-proportional Venn diagrams for the phosphopeptide identifications originating from HAP-MSA and HAP-MSR (HAP-MS2 reduced).

Fig. 5 – Number of phosphopeptide identifications satisfying a given p-value. The x-axis represents the p-values obtained from our logistic function and the y-axis represents the number of identifications satisfying that p-value. HAP-MS2R-1, HAP-MS2R-2, and HAP-MS2R-3 stand for the technical replicates of the artificially reduced MS2 analysis. The plots on the right result from zooming in on the region of low p-values in the plots on the left.
REFERENCES


