Published in final edited form as:

J Proteome Res. 2009 February ; 8(2): 798–807. doi:10.1021/pr8005556.

Phosphorylation Site Mapping of Endogenous Proteins: A Combined MS and Bioinformatics Approach

Jeffrey M. Sundstrom¹, Christopher J. Sundstrom⁵, Scott A. Sundstrom⁵, Patrice E. Fort², Richard L.H. Rauscher^{3,4}, Thomas W. Gardner^{1,2}, and David A. Antonetti^{1,2}

¹Department of Cellular and Molecular Physiology, Penn State University College of Medicine, Hershey, PA 17033, USA

²Department of Ophthalmology, Penn State University College of Medicine, Hershey, PA 17033, USA

³Penn State Cancer Institute, Penn State University College of Medicine, Hershey, PA 17033, USA

⁴Department of Computer Science and Engineering, Penn State University, University Park, PA, 16802, USA

⁵Code Platinum Inc., Dover, NH 03820, USA

Abstract

We present a novel approach that combines MALDI-TOF profile analysis and bioinformatics-based inclusion criteria to comprehensively predict phosphorylation sites on a single protein of interest from limiting sample. It is technologically difficult to unambiguously identify phosphorylated residues, as many physiologically important phosphorylation sites are of too low abundance *in vivo* to be unambiguously assigned by mass spectrometry. Conversely, phosphorylation site prediction algorithms, while increasingly accurate, nevertheless overestimate the number of phosphorylation sites. In this study, we show that MODICAS, an MS data management and analysis tool, can be effectively merged with the bioinformatics attributes of residue conservation and phosphosite prediction to generate a short list of putative phosphorylation sites that can be subsequently verified by additional methodologies such as phosphospecific antibodies or mutational analysis. Therefore, the combination of MODICAS driven MS data analysis with bioinformatics-based filtering represents a substantial increase in the ability to putatively identify physiologically relevant phosphosites from limited starting material.

Keywords

Mass spectrometry; phosphorylation; post-translational modification; AKT1

Introduction

Current proteomics research emphasizes the identification and monitoring of post-translational modifications (PTMs) as many cellular events are controlled through the covalent modification of proteins¹. PTMs, such as phosphorylation and glycosylation, often regulate protein function by altering protein interactions, cellular distribution, degradation,

or enzymatic activity. As such, PTMs modulate both the processing and output of the signal transduction pathways that define the cellular response to ever changing environmental conditions. The functional state of signaling pathways is frequently regulated by dynamic protein phosphorylation and maladaptive responses of these pathways often characterize disease states at the molecular level. Therefore, identification of physiologically relevant PTMs, particularly phosphorylation, is central to understanding how signal transduction pathways are dynamically regulated in health and disease.

In order to understand how phosphorylation regulates cellular processes, researchers often attempt to comprehensively map phosphosites on a given protein of interest. The success of comprehensive phosphosite mapping has an enormous impact on the level of understanding that can be achieved. Traditionally, phosphosite maps are generated through the use of radiolabelled protein in combination with other techniques such as 2D-gel electrophoresis, phosphopeptide mapping, Edman degradation, and mutational analysis²⁻⁴. Although these methods are effective, they are time consuming, expensive, and require large amounts of target protein. In addition, the use of radiolabelled protein precludes the analysis of tissue obtained from animal models or clinical samples.

Recent methodologies involve the use of tandem mass spectrometry (tandem MS) to identify phosphosites by detecting the change in mass associated with the addition of a phosphate group. This approach allows for the unambiguous identification of the modified residue(s) and is also amenable to high-throughput experimentation². The power of this approach has been demonstrated by the identification of large numbers of novel phosphorylation sites present in complex protein mixtures obtained from tissue or cell lysates⁵⁻⁶. However, this approach generally fails to produce comprehensive phosphorylation maps on any given protein and instead sporadically identifies phosphosites on many proteins. This is due to several factors including the generally low stoichiometry of phosphorylation as well as the suppression in ionization efficiency induced by phosphorylation. These difficulties are compounded by the fact that only 10% of tandem MS spectra lead to the positive identification of modified sites⁷.

To circumvent these difficulties, proteins are frequently overexpressed in cell culture or phosphorylated *in vitro* and subsequently subjected to tandem MS analysis. This strategy often generates inaccurate maps of phosphosites that do not reflect the nature and position of the phosphosites *in vivo*³. This situation is further complicated as phosphorylation patterns on a given protein are likely dependent on the cells or tissue from which it is isolated⁸. Taken together, these complexities partially explain that while over 2000 modifications have been identified, only 450 have definitively been shown to occur *in vivo*⁴. This number represents less than 0.1% of the 100,000 phosphorylation sites currently thought to be present in the human proteome⁹. Therefore, approaches capable of accurately and efficiently generating phosphosite maps on individual proteins derived from specific tissues are needed.

Alternatively, phosphorylation sites can be identified by a wide range of prediction strategies. These phosphosite prediction scanners have become increasingly accurate as more raw data becomes available and more advanced motif identification procedures are implemented. KinasePhos, based on hidden Markov Models, was trained on phosphosite data from PhosphoBase and resulted in an algorithm that detects 18 different protein kinase motifs¹⁰. Although KinasePhos is more accurate than Scansite, none of these predictive algorithms are accurate enough to unambiguously identify phosphorylation sites.

In this study, we establish an alternative approach to putatively identify physiologically relevant phosphosites that exploits the synergy between MS data analysis coupled to

bioinformatics-based inclusion criteria consisting of S/T/Y conservation and phosphosite prediction analysis. Central to this approach was the development of MODICAS (modification assessment software). This software was designed to manage, compare, and mine MS Data in order to comprehensively map phosphosites on a single protein of interest. AKT1, a protein containing nine validated *in vivo* phosphorylation sites, was used to test our approach. MODICAS identified and filtered phosphopeptides from multiple observations of immunoprecipitated AKT. The resulting data set was subsequently filtered for S/T/Y residue conservation and phosphosite prediction. Assessment with phosphospecific antibodies confirmed that the merger of empirically derived MALDI-TOF profiles and bioinformatics resulted in the identification of six of the nine previously validated *in vivo* phosphorylation sites. In the adjoining paper, this method was applied to the tight junction protein occludin and resulted in the identification of 5 putative phosphosites, one of which was shown to be regulated by vascular endothelial growth factor through the development of a phospho-specific antibody. These data demonstrate the effectiveness of this combined approach to facilitate the identification of physiologically relevant and tissue specific phosphosites from limited amounts of material.

Methods

Sample Preparation and MS Data Acquisition

AKT1 was immunoprecipitated from rat retinal lysates and MALDI-TOF profiles were acquired. Retina was chosen as previous studies have shown that AKT1 is constitutively active in retina¹¹. Briefly, a single retina was isolated from a male Sprague-Dawley rat (Charles River Laboratories, Wilmington MA) and lysed in buffer consisting of 50 mM Hepes, 137 mM NaCl, 5 mM NaF, 1 mM β -glycerolphosphate, 1 mM EDTA, 1 mM EGTA, 0.1 mM NaVO₄, 1% NP-40, and 1 tablet of Protease Inhibitor Cocktail (Roche). AKT1 was immunoprecipitated (Santa Cruz) in triplicate from 400 μ g of total protein lysate, subjected to SDS-PAGE, and stained with Sypro Ruby (Molecular Probes). The location of AKT1 containing bands were determined by running a fraction of each reaction in parallel and immunoblotting with anti-AKT1. The bands were excised, destained twice (50% acetonitrile (AcN) containing 200mM NH₄CO₃ pH 8) for 45min @ 37°C and dried completely. Each sample was then reduced (2mM TCEP, 25mM NH₄CO₃ pH 8) for 15min @ 37°C with agitation and alkylated (20mM iodoacetamide, 25mM NH₄CO₃ pH 8) for 30min @ 37°C in the dark. The samples were then washed three times with 25mM NH₄CO₃ pH 8 and dried in a SpeedVac and rehydrated with 1.5X original gel slice volume of 0.02 μ g/ μ l of trypsin (sequencing grade modified trypsin, Promega) in 50% AcN, 40mM NH₄CO₃ pH 8, 0.1% w/v n-octylglucoside for 60min @ RT to allow for the concentrated trypsin to diffuse into the gel slice. An additional 50 μ l of 50% AcN, 40 mM NH₄CO₃ pH 8, 0.1% n-octylglucoside was added and the samples were incubated for 16–18hr @ 37°C with agitation. The supernatant containing AKT1 tryptic peptide fragments was removed, dried and resuspended in 200 μ l of water three times. After the last wash, the samples were resuspended in 20 μ l water and cleaned (10 $\times$ 3min binding time) with strong cation exchange ZipTip (ZipTipSCX; Eppendorf) columns according to the manufactures instructions. The samples were eluted directly onto the MALDI plate (57 mm $\times$ 57 mm stainless steel plate) in a volume of 1.0 μ l of freshly prepared elution solution containing 5% NH₄OH and 30% methanol. The sample was allowed to dry and 0.5 μ l of matrix was applied (alpha-cyano-4-hydroxycinnamic acid (CHCA). MALDI-TOF profiles were acquired in positive ion mode from an Applied Biosystems 4700 Proteomics Analyzer. The resulting peak lists were exported for further analysis.

Data Analysis – MODICAS

MODICAS was designed to facilitate the management and analysis of MS data from multiple experiments. MODICAS is readily expandable as this prototype was developed using Visual Basic (VB) classes within Microsoft Access. The overall database structure encompasses a single type of data unit (*PeptideGroup*) derived from either MS data peak lists (*DataImportMS*) or from *in silico* digests (*DataImportDigest*). After the data has been imported, the user applies functional modules (*DataSubtract* or *DataMatch*) in order to accomplish the desired task. Each functional module is discussed below and the overall data processing scheme conducted by current and planned future versions of MODICAS is shown in Figure 1.

PeptideGroup—The fundamental unit of data within MODICAS is the *PeptideGroup*, a data table derived from MS data peak lists, *in silico* protein digests with or without potential PTMs, or other sources. MODICAS groups all peptides/masses and their data attributes into a named *PeptideGroup*. As discussed below, functions are applied to one or more *PeptideGroups* resulting in derivative daughter *PeptideGroups* thus creating a hierarchical organization of data.

DataImportMS—This data module imports data from spreadsheets containing MS peak lists consisting of peptide masses and corresponding data attributes. This data is incorporated into a named *PeptideGroup* within MODICAS and can consist of sample data, gel-blank data, or lists of common contaminant peaks.

DataImportDigest—This data module imports data from spreadsheets containing theoretical MS peak lists derived from *in silico* protein digests and consists of peptide masses and corresponding data attributes. This user-defined reference look-up table can be adjusted for any PTM of interest and can also incorporate information regarding isotopic labeling. For this study, AKT1 was digested with trypsin *in silico* using ProteinProspector (www.prospector.ucsf.edu) with a maximum of two missed cuts, a maximum of one phosphorylation per peptide, and with variable phosphorylation of serine, threonine, or tyrosine residues. The results were incorporated into a spreadsheet and imported into MODICAS with the *DataImportDigest* function.

DataSubtract—This functional module creates a derivative daughter *PeptideGroup* by removing common data points from one of two *PeptideGroups*. This function is initially used to remove noise from a given MS data set and may be subsequently used to track uniquely observed peptides from one experiment to another.

DataMatch—This functional module compares MS data contained in a *PeptideGroup* or derivative daughter *PeptideGroup*, to *in silico* protein digests imported by *DataImportDigest*. This peptide mapping function identifies experimental masses and links them to their associated theoretical proteolytic peptides. As such this function identifies unmodified and modified peptides present in the experimental data set. Data from different samples may be subsequently combined in order to identify peptides observed in multiple experiments. The mass tolerance can be set in daltons or ppm. For this study, a mass tolerance of 100ppm was used.

Data Analysis - FindMod

Each set of MS data was submitted to FindMod, which returned a flat file containing identified peptides, both modified and unmodified. These tabular results were manually recorded back onto a spreadsheet containing the raw MS data in order to track the identified

peptide and note the modification state. This data was manually extracted to compile a putative phosphopeptide map for each experiment.

Residue Conservation

AKT1 sequence from *H. sapiens*, *M. musculus*, *R. norvegicus*, *G. gallus*, and *X. tropicalis* were aligned using the ClustalV algorithm. Tyrosine residues were considered to be conserved if identity was preserved from *R. norvegicus* to *X. tropicalis* while serines and threonines were allowed to be substituted for each other across these species. This evolutionary distance has been suggested to be appropriate for PTM mapping¹².

Phosphorylation Site Prediction

Protein sequence *R. norvegicus* AKT1 was subjected to phosphosite predication scanning by KinasePhos (www.kinasephos.mbc.nctu.edu.tw). KinasePhos was set to 95% specificity and the returned predictions included the predicted kinase group and an associated expectation value.

Immunoblotting

Retinas were isolated from male Sprague-Dawley rats (Charles River Laboratories, Wilmington MA) and lysed in buffer consisting of 50 mM Hepes, 137 mM NaCl, 5 mM NaF, 1 mM β -glycerolphosphate, 1 mM EDTA, 1 mM EGTA, 0.1 mM NaVO₄, 1% NP-40, and 1 tablet of Protease Inhibitor Cocktail (Roche). The lysates were subjected to SDS-PAGE and immunoblotted with α -AKT1 (1:200 dilution; Santa Cruz, sc-5298), α -pS129-AKT1 (1:2000 dilution; a generous gift from Maria Ruzzene), α -pT308-AKT1 (1:1000 dilution; Cell Signaling - #4056), α -Y326-AKT1 (1:1000 dilution; Cell Signaling - #2968), α -pT450-AKT1 (1:1000 dilution; Cell Signaling - #9267), and α -pS473-AKT1 (1:1000 dilution; Cell Signaling - #9271). ECL was performed with mouse or rabbit HRP conjugated secondary IgG (GE Healthcare/Amersham - #NA931).

Results and Discussion

Data Collection and Processing

The ability to identify phosphorylation sites from raw MS data using phosphopeptide mapping combined with bioinformatics-based inclusion criteria was assessed. For clarification, a generic diagram of the data analysis scheme is given in Figure 1 while the analysis of AKT1 data is depicted in Figure 2. AKT1 was immunoprecipitated from retinal lysates (400 ug total protein per reaction) in triplicate, subjected to SDS-PAGE, excised, digested with trypsin, and analyzed by MALDI-TOF MS. The peak list from each experiment was incorporated into a spreadsheet and the *DataImport* module was used to import this data into MODICAS. In addition, a peak list representative of background MS noise was generated from a slice of gel containing no sample and imported. The *DataSubtract* function was then used to remove non-specific peaks from each experimental mass list with a user-defined mass tolerance of 100 parts per million (ppm). To create a reference table containing expected masses indicative of AKT1 tryptic peptides, AKT1 was digested *in silico* by ProteinProspector (www.prospector.ucsf.edu) with a maximum of two missed cuts, a maximum of two phosphorylations per peptide, and with variable phosphorylation of serine, threonine, or tyrosine residues. These results were incorporated into a spreadsheet and imported into MODICAS with the *DataImportDigest* module. The *DataMatch* function was then used to identify the AKT1 peptides and phosphopeptides present in each sample. In this experiment, a tolerance of 100ppm was used to identify experimentally observed masses from each AKT1 sample that are also present in the AKT1

reference table. This process resulted in the generation of three independent *PeptideGroups* containing observed AKT1 peptides and phosphopeptides (Fig. 2, top).

In order to obtain a more reliable AKT1 phosphosite solution set, the data was combined and filtered (Fig. 2, bottom). MODICAS combined the data from each experiment resulting in a total of 40 (Sample1=23, Sample2=9, Sample3=8) data points representing observed masses that matched predicted phosphopeptide masses. These data points were then filtered by three different methods. First, MODICAS filtered the data for multiple observations with the requirement that each putative phosphopeptide be observed in at least two of the three biological replicates. This process resulted in the removal of thirteen of the original 40 observed data points. The remaining data was filtered for phosphosite prediction and S/T/Y residue conservation, which resulted in the removal of twelve additional data points. Following this process, 15 data points remained and these potential phosphopeptides were located in five discrete regions of interest (Fig. 3). Specifically, AKT1 phosphopeptide regions were observed at AKT1^{122..142}, AKT1^{145..168}, AKT1^{302..328}, AKT1^{378..386}, and AKT1^{466..480}. The combination of MS data with residue conservation and phosphosite prediction suggests that up to 13 phosphosites may be present within these five regions (Fig. 4). These data are based on 51.5% coverage of AKT1 residues in this study.

MODICAS and Bioinformatics - Identification of Known AKT1 Phosphosites

For purposes of this study, the operational definition of *in vivo* AKT1 phosphosites will be restricted to those residues identified by phosphosite specific antibodies or with the combined use of radiolabelled tryptic peptide mapping coupled with Edman degradation. A survey of the literature revealed that phosphorylation has been confirmed to occur *in vivo* on endogenously expressed AKT1 on nine different sites^{13–20}. Residues identified solely by *in vitro* phosphorylation or overexpression of AKT1 have been excluded from this analysis²¹.

In the present study, the combination of MODICAS driven MS data analysis and bioinformatics-based filtering correctly identified six of the nine previously identified AKT1 phosphosites (Fig. 5A). The two sites most critical to the regulation of AKT1 activity, T308 and S473¹⁴ were identified. These sites have previously been shown to be phosphorylated in the retina²². AKT1 phosphosites known to upregulate AKT1 activity, Y315 and Y326, were also identified^{15, 18}. Lastly, constitutive AKT1 phosphosites required for maximal AKT1 activity, S124 and S129^{13, 17}, were identified.

Western blotting was used in order to confirm the identity of AKT1 phosphosites in retina. Retinal lysates were subjected to SDS-PAGE and immunoblotted with α -AKT1, α -pS129-AKT1, α -pT308-AKT1, α -Y326-AKT1, α -pT450-AKT1, and α -pS473-AKT1 (Fig. 5B). Consistent with the MS and bioinformatics analysis, AKT1 is phosphorylated on S129, T308, Y326, and S473 in retinal lysates. As discussed in more detail below, T450 was not identified by the combined MS and bioinformatics approach but is phosphorylated in retinal lysates. Unfortunately, phosphosite specific antibodies were not available for pS124-AKT1 or pY315-AKT1.

MODICAS and Bioinformatics - AKT1 Phosphosites Not Observed

AKT1 phosphosites not identified from retinal lysates are noted in Figure 6. Phosphorylation of T34 is a negative regulator of AKT1 activity and is observed only after ceramide treatment of L6 myotubes²³. In the current study, although unmodified peptides inclusive of T34 were identified (Fig. 3), no phosphopeptides inclusive of T34 were observed. These results suggest that AKT1 is not phosphorylated at T34 under normal physiological conditions in the retina. Unfortunately, MS coverage of AKT1 in this set of experiments did not include coverage of T450, although phosphorylation of T450 has been observed by

multiple groups. The smallest tryptic peptide covering T450 is AKT1^{437..465} (RYFDEEFTAQMITITPPDQDDSMCEVDSEK), a peptide with a pI of 3.9, which suggests a low positive ion MALDI-TOF ionization potential. To circumvent this problem, chemical cleavage of methionine residues with cyanogen bromide might be useful in future experiments. Lastly, phosphorylation of Y474 has previously been observed *in vivo*¹⁶ but was not identified in this analysis. While phosphopeptides corresponding to AKT1^{466..480} were observed in the present study (Fig 3), Y474 was not predicted to be phosphorylated by KinasePhos and, therefore, may have been inappropriately discarded.

MODICAS and Bioinformatics - Novel AKT1 Phosphosites and False Positives

The results of our combined approach imply that up to seven additional phosphosites may be present on AKT1 in retinal lysates. The data putatively identifies three phosphosites contained within the AKT1^{145..168} phosphopeptide region (Fig. 4). It is likely that either T146 or Y152 are phosphorylated in retina as MS data analysis identified a monophosphorylated peptide of AKT1^{145..154} in multiple biological replicates. Moreover, the monophosphorylated AKT1^{159..168} peptide identified in multiple replicates contains T160 but no other S/T/Y are present suggesting that T160 is likely phosphorylated on retinal AKT1.

One of these putative novel sites, T146, is located in the linker region between the PH and catalytic domains. This region is the most disparate domain amongst the three AKT isoforms and thus, may allow for isoform specific regulation. Interestingly, recent therapeutic strategies incorporating allosteric modifiers of enzyme function have been targeted to this region²⁴. Although KinasePhos suggests PKA phosphorylates T146, another phosphosite scanner (Scansite) suggests that PKC- ϵ may be the responsible kinase. This is interesting as PKC- ϵ has recently been shown to modulate AKT function^{25, 26}. The remaining two putative sites identified within AKT1^{145..168}, Y152 and T160, are located in the N-terminal portion of the catalytic domain. Although compelling, the relevance of T146, Y152 or T160 to the *in vivo* phosphorylation state of AKT1 remains to be determined.

The data also putatively identifies a phosphosite contained within the AKT1^{378..386} phosphopeptide region. The combined MS and bioinformatics method suggests that S378 is phosphorylated on AKT1 in retinal lysates (Fig. 4). However, careful analysis of each biological replicate where this phosphopeptide was observed suggests that this peptide is phosphorylated on two residues. As such, it seems plausible that the conserved residue S381 is also phosphorylated as there are no other conserved S/T/Y residues. While phosphopeptides corresponding to AKT1^{378..386} were observed in the present study (Fig 3), S381 was not predicted to be phosphorylated by KinasePhos and, therefore, may have been inappropriately discarded. While either of these sites would be novel AKT1 phosphosites, the *in vivo* relevance of these sites remains to be confirmed by additional studies.

The remaining three putative phosphosites, S122, S126, T312, are in close proximity to known phosphosites and these known sites are likely responsible for the phosphopeptides observed in the MALDI-TOF profiles. Therefore, these sites are less likely to be newly identified sites and their identification possibly resulted from either the ambiguity regarding specific residues inherent in MS analysis or from the generally over predictive nature of phosphosite scanners²⁷.

Assessment of Methods

The ability of MS data analysis combined with bioinformatics to accurately identify AKT1 phosphosites was compared to both MS and bioinformatics in isolation. The same set of data used throughout this study was used for this comparison. MS data analysis was conducted

using MODICAS to assess a purely empirical approach to the identification of AKT1 phosphosites. This analysis resulted in 25 phosphosite identifications, six of which corresponded to confirmed AKT1 phosphosites (Fig. 6A). Similarly, residue conservation and phosphosite prediction was used to assess the ability of a purely bioinformatics-based approach to identify AKT1 phosphosites. This analysis resulted in 21 phosphosite identifications, seven of which correspond to confirmed AKT1 phosphosites (Fig. 6B). As discussed above, the combination of MS data analysis and bioinformatics identified a total of 13 phosphosites, six of which correspond to confirmed AKT1 phosphosites (Fig. 4 and Fig. 5).

The sensitivity (proportion of known phosphosites correctly identified ($Tp/(Tp+Fn)$) and specificity (proportion of nonphosphorylated residues correctly identified ($Tn/(Tn+Tp)$) for each method was calculated (where Tp = true positives, Fp = false positives, Tn = true negatives, and Fn = false negatives). For this analysis, identification of phosphosites that matched any of the nine previously verified *in vivo* AKT1 phosphosites (T34, S124, S129, T308, Y315, Y326, T450, S473, and Y474) were considered as true positives (Tp) while phosphosites identified outside of this group of nine sites were considered false positives (Fp). Identification of unmodified S/T/Y residues that matched any of the remaining 63 S/T/Y were considered as true negatives (Tn) while unmodified sites identified outside of this group of 63 S/T/Y were considered as false positives (Fn).

Either method alone or the combination of the methods resulted in the identification of approximately the same number of phosphosites as evidenced by the sensitivity associated with each method. Specifically, MS data analysis combined with bioinformatics identified six of the nine known AKT1 phosphosites, while MS data analysis or bioinformatics alone identified six or seven phosphosites, respectively (Fig 6C and 6D).

Implementation of either method in isolation overestimated the number of AKT1 phosphosites as evidenced by the lower specificity of either method alone (MS = 70%; bioinformatics = 78%) when compared to the combined approach (MS and bioinformatics = 89%). As such, the combination of MS data analysis with bioinformatics identified true negatives with considerably higher accuracy (Fig 6C and 6D).

The calculation of sensitivity and specificity above were done in the most conservative manner possible as each putatively novel phosphosite was considered to be a false positive data point. In contrast, the false discovery rate (false phosphosites/true phosphosites) was calculated based on the assumption that each of the putative novel phosphosites are true positives. To calculate the false discovery rate, a decoy data base was created by randomizing the AKT1 amino acid sequence. We chose to use randomized AKT1 so that the amino acid composition of the decoy would be equivalent to the actual sequence as it was particularly important to keep number of S/T/Y residues constant. The randomized AKT1 sequence was digested *in silico* with ProteinProspector and imported into MODICAS. The *DataMatch* function was then used to identify the random AKT1 phosphopeptides present in each sample and the resulting phosphopeptides were filtered for multiple observations and phosphosite prediction. This process identified a single phosphorylation site within the randomized AKT1 sequence which contained a total of 72 S/T/Y residues. As such, the discovery rate using the decoy database was 1/72 while the experimental discovery rate was 13/72. Therefore, the false discovery rate (decoy discovery rate/true discovery rate) was calculated to be 0.076, or 1 site out of every 13 identified (Fig 6D).

MALDI TOF-TOF Studies

MALDI TOF-TOF experiments on the phosphopeptides identified above as well as on AKT1 tryptic peptides separated by C18 peptide chromatography failed to yield reliable

spectra (data not shown). As such, tandem MS was unable to identify any of the known AKT1 phosphosites from these samples. Although unambiguous identification was not possible, treatment of samples with alkaline phosphatase after acquisition of initial spectra will be used in future experiments to verify that the masses of interest represent phosphorylated peptides. Inclusion of this information in the algorithm is predicted to enhance the predictive accuracy of the current strategy.

Comparison of MODICAS and FindMod

The design focus of MODICAS was to facilitate the mapping of covalent modifications from MS data obtained from endogenously expressed proteins. Although FindMod was developed for this purpose²⁸, the major advantage of MODICAS lies in the fact that it is a software tool created within a relational database scheme in order to facilitate data storage and analysis. The efficiency of both FindMod and MODICAS to generate phosphopeptide maps from the MALDI-TOF profiles used in this study was compared. Analysis revealed that manual tabulation of the FindMod flat files occurred over a period of weeks, whereas MODICAS driven MS data analysis was conducted in a several hours. Moreover, the functional algorithms within MODICAS allow for a more in depth analysis of data across multiple experiments and/or samples. As such, MODICAS was able to compile the results of the three independent experiments rapidly, a process FindMod is not capable of conducting. Future versions of MODICAS will automate the flow of data analysis and attempt to assign probability scores to putatively identified phosphosites.

Conclusions

The ability to comprehensively identify phosphorylation sites on a selected protein remains a hurdle in signal transduction biology as many physiologically important phosphorylation sites are of too low abundance *in vivo* to be unambiguously identified by mass spectrometry. Conversely, phosphorylation site prediction algorithms, while increasingly accurate, nevertheless overestimate the number of phosphorylation sites. In this study, we show that MS data can be effectively merged with the bioinformatics attributes of residue conservation and phosphosite prediction to generate a short list putative phosphorylation sites to be validated by other means, e.g. phospho-specific antibodies or site-directed mutagenesis. Further, MODICAS, an MS data management and analysis tool, was developed to facilitate this process. Therefore, the combination MODICAS driven MS data analysis with bioinformatics-based filtering represents substantial increase in the ability to putatively identify physiologically relevant phosphosites from limited starting material.

Notably, this approach was successful on MALDI-TOF profiles obtained from limited starting material and does not rely on the acquisition of tandem MS data or require radiolabelled protein. As such, this methodology is both robust and technologically simple. This novel approach should prove useful in several contexts as researchers will be more readily able to map modifications from samples derived from cell culture, animal models, or clinical samples. The improved ability to putatively identify phosphosites from limited samples, specifically clinical samples²⁹, is perhaps the most significant benefit. This approach is not limited to phosphosite mapping and may potentially be used to map any PTM detectable by MS provided that predictive scanners also exist. As such, the capacity of this methodology to map disease-associated PTMs directly from physiologically relevant samples may serve to better elucidate the signal transduction pathways altered in disease, identify novel disease biomarkers, and may ultimately identify novel therapeutic targets.

Acknowledgments

This research was supported by NIH grant NEI-012021 (DAA), from funding by the Juvenile Diabetes Research Foundation (DAA), the Pennsylvania Lions Sight Conservation and Eye Research Foundation (DAA) and a gift from Jack and Nancy Turner (TWG). We would like to thank Dr. Maria Ruzzene for the generous provision of α -pS129-AKT1 antibody. We would also like to thank Dr. Todd Fox for sample preparation and Dr. Bruce Stanley for sharing his expertise in MS analysis.

References

1. Weston AD, Hood L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* 2004;3(2):179–196. [PubMed: 15113093]
2. Reinders J, Sickmann A. State-of-the-art in phosphoproteomics. *Proteomics* 2005;5(16):4052–4061. [PubMed: 16196093]
3. Delom F, Chevet E. Phosphoprotein analysis: from proteins to proteomes. *Proteome Sci* 2006;4(1):15. [PubMed: 16854217]
4. Morandell S, Stasyk T, Grosstessner-Hain K, Roitinger E, Mechtler K, Bonn GK, Huber LA. Phosphoproteomics strategies for the functional analysis of signal transduction. *Proteomics*. 2006
5. Nousiainen M, Sillje HH, Sauer G, Nigg EA, Korner R. Phosphoproteome analysis of the human mitotic spindle. *Proc Natl Acad Sci U S A* 2006;103(14):5391–5396. [PubMed: 16565220]
6. Amanchy R, Kalume DE, Iwahori A, Zhong J, Pandey A. Phosphoproteome analysis of HeLa cells using stable isotope labeling with amino acids in cell culture (SILAC). *J Proteome Res* 2005;4(5):1661–1671. [PubMed: 16212419]
7. Savitski MM, Nielsen ML, Zubarev RA. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics* 2006;5(5):935–948. [PubMed: 16439352]
8. Graham ME, Anggono V, Bache N, Larsen MR, Craft GE, Robinson PJ. The in vivo phosphorylation sites of rat brain dynamin I. *J Biol Chem* 2007;282(20):14695–14707. [PubMed: 17376771]
9. Kalume DE, Molina H, Pandey A. Tackling the phosphoproteome: tools and strategies. *Curr Opin Chem Biol* 2003;7(1):64–69. [PubMed: 12547428]
10. Huang HD, Lee TY, Tzeng SW, Horng JT. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res* 2005;33(Web Server issue):W226–W229. [PubMed: 15980458]
11. Reiter CE, Sandirasegarane L, Wolpert EB, Klinger M, Simpson IA, Barber AJ, Antonetti DA, Kester M, Gardner TW. Characterization of insulin signaling in rat retina in vivo and ex vivo. *Am J Physiol Endocrinol Metab* 2003;285(4):E763–E774. [PubMed: 12799319]
12. Cooper GM, Brudno M, Green ED, Batzoglu S, Sidow A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res* 2003;13(5):813–820. [PubMed: 12727901]
13. Alessi DR, Andjelkovic M, Caudwell B, Cron P, Morrice N, Cohen P, Hemmings BA. Mechanism of activation of protein kinase B by insulin and IGF-1. *Embo J* 1996;15(23):6541–6551. [PubMed: 8978681]
14. Bellacosa A, Chan TO, Ahmed NN, Datta K, Malstrom S, Stokoe D, McCormick F, Feng J, Tsichlis P. Akt activation by growth factors is a multiple-step process: the role of the PH domain. *Oncogene* 1998;17(3):313–325. [PubMed: 9690513]
15. Chen R, Kim O, Yang J, Sato K, Eisenmann KM, McCarthy J, Chen H, Qiu Y. Regulation of Akt/PKB activation by tyrosine phosphorylation. *J Biol Chem* 2001;276(34):31858–31862. [PubMed: 11445557]
16. Conus NM, Hannan KM, Cristiano BE, Hemmings BA, Pearson RB. Direct identification of tyrosine 474 as a regulatory phosphorylation site for the Akt protein kinase. *J Biol Chem* 2002;277(41):38021–38028. [PubMed: 12149249]

17. Di Maira G, Salvi M, Arrigoni G, Marin O, Sarno S, Brustolon F, Pinna LA, Ruzzene M. Protein kinase CK2 phosphorylates and upregulates Akt/PKB. *Cell Death Differ* 2005;12(6):668–677. [PubMed: 15818404]
18. Jiang T, Qiu Y. Interaction between Src and a C-terminal proline-rich motif of Akt is required for Akt activation. *J Biol Chem* 2003;278(18):15789–15793. [PubMed: 12600984]
19. Jung HS, Kim DW, Jo YS, Chung HK, Song JH, Park JS, Park KC, Park SH, Hwang JH, Jo KW, Shong M. Regulation of protein kinase B tyrosine phosphorylation by thyroid-specific oncogenic RET/PTC kinases. *Mol Endocrinol* 2005;19(11):2748–2759. [PubMed: 15994200]
20. Shao Z, Bhattacharya K, Hsich E, Park L, Walters B, Germann U, Wang YM, Kyriakis J, Mohanlal R, Kuida K, Namchuk M, Salituro F, Yao YM, Hou WM, Chen X, Aronovitz M, Tsichlis PN, Bhattacharya S, Force T, Kilter H. c-Jun N-terminal kinases mediate reactivation of Akt and cardiomyocyte survival after hypoxic injury in vitro and in vivo. *Circ Res* 2006;98(1): 111–118. [PubMed: 16306447]
21. Li X, Lu Y, Jin W, Liang K, Mills GB, Fan Z. Autophosphorylation of Akt at threonine 72 and serine 246. A potential mechanism of regulation of Akt kinase activity. *J Biol Chem* 2006;281(19): 13837–13843. [PubMed: 16549426]
22. Reiter CE, Wu X, Sandirasegarane L, Nakamura M, Gilbert KA, Singh RS, Fort PE, Antonetti DA, Gardner TW. Diabetes reduces basal retinal insulin receptor signaling: reversal with systemic and local insulin. *Diabetes* 2006;55(4):1148–1156. [PubMed: 16567541]
23. Powell DJ, Hajduch E, Kular G, Hundal HS. Ceramide disables 3-phosphoinositide binding to the pleckstrin homology domain of protein kinase B (PKB)/Akt by a PKCzeta-dependent mechanism. *Mol Cell Biol* 2003;23(21):7794–7808. [PubMed: 14560023]
24. Kumar CC, Madison V. AKT crystal structure and AKT-specific inhibitors. *Oncogene* 2005;24(50):7493–7501. [PubMed: 16288296]
25. Zhou HZ, Karliner JS, Gray MO. Moderate alcohol consumption induces sustained cardiac protection by activating PKC-epsilon and Akt. *Am J Physiol Heart Circ Physiol* 2002;283(1):H165–H174. [PubMed: 12063287]
26. Li L, Sampat K, Hu N, Zakari J, Yuspa SH. Protein kinase C negatively regulates Akt activity and modifies UVC-induced apoptosis in mouse keratinocytes. *J Biol Chem* 2006;281(6):3237–3243. [PubMed: 16338928]
27. Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 2004;4(6):1633–1649. [PubMed: 15174133]
28. Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, Ou K, Sanchez JC, Bairoch A, Williams KL, Hochstrasser DF. High-throughput mass spectrometric discovery of protein post-translational modifications. *J Mol Biol* 1999;289(3):645–657. [PubMed: 10356335]
29. Hanash S. Disease proteomics. *Nature* 2003;422(6928):226–232. [PubMed: 12634796]

Figure 1. Combined MS and Bioinformatics Approach

MS data from multiple experiments are imported into MODICAS and non-specific peaks are filtered by the *DataSubtract* function. Observed peptides are identified using the *DataMatch* function which compares reference table masses to observed masses. The data is then filtered for multiple observations, residue conservation and phosphosite prediction. Large grey box indicates functions currently performed by MODICAS.

Figure 2. Analysis of AKT1 Phosphosites

MALDI-TOF profiles inclusive of the most abundant 250 peaks from each of three separate AKT1 immunoprecipitations and a gelblank were imported into MODICAS. Non-specific peaks were filtered by the *DataSubtract* function while peptides and phosphopeptides were identified using the *DataMatch* function. The data was then filtered by MODICAS for multiple observations and subsequently by bioinformatic constraints consisting of phosphosite prediction and S/T/Y conservation. Large grey box indicates functions currently performed by MODICAS.

Sample	Start	End	ThMass	ExpMass	dM	MC	Sequence	Modification	Putative Phosphosites
2	1	8	889.48	889.56	-83.64	0	(-)MNDVAIVK(E)		
1	1	14	1639.86	1639.84	13.00	1	(-)MNDVAIVKEGWLHK(R)		
2	1	14	1639.86	1639.91	-26.75	1	(-)MNDVAIVKEGWLHK(R)		
3	1	14	1639.86	1639.97	-67.70	1	(-)MNDVAIVKEGWLHK(R)		
2	9	15	925.50	925.48	17.87	1	(K)EGWLHKR(G)		
3	9	15	925.50	925.52	-25.73	1	(K)EGWLHKR(G)		
2	26	39	1678.88	1679.05	-98.09	1	(R)YFLLKNDGTFIGYK(E)		
1	49	64	1865.89	1865.80	45.00	0	(R)ESPLNFSVAQCQLMK(T)	1CAM-C	
2	77	86	1319.68	1319.71	-26.92	0	(R)CLQWTTVIER(T)	1Cys-am	
2	113	121	1184.49	1184.52	-28.32	0	(R)QEEETMDFR(S)		
1	122	142	2183.95	2184.08	-60.38	0	(R)SGSPSDNSGAEMEVALAKPK(H)	1PO4	S122, S124, S126, S128
2	122	142	2183.95	2184.02	-30.09	0	(R)SGSPSDNSGAEMEVALAKPK(H)	1PO4	S122, S124, S126, S128
1	143	158	1994.05	1994.00	24.40	2	(K)HRVTMNEFEYLKLLGK(G)	1Met-ox	
2	145	154	1273.61	1273.73	-90.27	0	(R)VTMNEFEYLK(L)		
2	145	154	1369.57	1369.52	43.25	0	(R)VTMNEFEYLK(L)	1Met-ox 1PO4	T146, Y152
3	145	154	1369.57	1369.62	-31.97	0	(R)VTMNEFEYLK(L)	1Met-ox 1PO4	T146, Y152
3	145	154	1353.58	1353.45	99.21	0	(R)VTMNEFEYLK(L)	1PO4	T146, Y152
1	145	163	2255.12	2255.04	32.50	2	(R)VTMNEFEYLKLLGKGTFGK(V)	1PO4	T146, Y152, T160
3	145	163	2255.12	2254.99	57.61	2	(R)VTMNEFEYLKLLGKGTFGK(V)	1PO4	T146, Y152, T160
2	155	163	920.56	920.52	37.40	1	(K)LLGKGTFGK(V)		
3	155	163	920.56	920.61	-61.26	1	(K)LLGKGTFGK(V)		
2	159	168	1141.64	1141.71	-62.62	1	(K)GTFGKVLV(K)	1PO4	T160
3	159	168	1141.64	1141.68	-35.36	1	(K)GTFGKVLV(K)	1PO4	T160
1	169	179	1333.66	1333.60	40.95	2	(K)DEATGRYYAMK(I)	1Met-ox	
1	171	179	1076.52	1076.43	83.81	1	(K)ATGRYYAMK(I)	1Met-ox	
1	171	179	1060.52	1060.54	-14.97	1	(K)ATGRYYAMK(I)		
3	171	179	1076.52	1076.55	-32.53	1	(K)ATGRYYAMK(I)	1Met-ox	
1	175	182	1045.58	1045.56	17.89	1	(R)YYAMKILK(K)	1Met-ox	
2	175	182	1045.58	1045.57	4.11	1	(R)YYAMKILK(K)	1Met-ox	
2	190	200	1284.62	1284.69	-53.32	0	(K)DEVAHTLTENR(V)		
1	190	206	1982.01	1982.01	-0.19	1	(K)DEVAHTLTENRVLQNSR(H)		
2	242	251	1264.64	1264.69	-40.24	2	(R)IYVSEDRAR(F)		
1	269	284	1963.07	1962.95	57.63	2	(K)NVVYRDLENLMLDK(D)		
3	274	284	1331.72	1331.68	36.22	1	(R)DLLENLMLDK(D)		
2	277	284	975.52	975.57	-57.61	0	(K)LENLMLDK(D)		
3	285	297	1446.74	1446.63	74.28	1	(K)DGHIKTFDGLCK(E)		
1	290	301	1323.70	1323.65	35.56	1	(K)ITDFGLCKEGIK(D)		
1	302	328	3223.30	3223.35	-15.53	1	(K)DGATMKTFCGTPEYLAPEVLEDNDYGR(A)	1Cys-am 2PO4	T308, T312, Y315, Y326
2	308	328	2549.00	2548.76	91.51	0	(K)TFCGTPEYLAPEVLEDNDYGR(A)	2PO4	T308, T312, Y315, Y326
3	308	328	2389.07	2389.28	-88.96	0	(K)TFCGTPEYLAPEVLEDNDYGR(A)		
2	368	377	1115.62	1115.67	-48.52	1	(R)FPRTLGPFAK(S)		
3	378	386	958.63	958.66	-36.89	1	(K)SLLSGLLK(D)		
1	378	386	1118.56	1118.50	56.15	1	(K)SLLSGLLK(D)	2PO4	S378
2	378	385	990.47	990.56	-91.65	0	(K)SLLSGLLK(K)	2PO4	S378
2	392	400	833.40	833.47	-80.86	0	(R)LGGSSEDAK(E)		
1	401	406	829.40	829.42	-21.70	0	(K)EIMQHR(F)	1Met-ox	
1	466	480	1826.73	1826.88	-84.47	0	(R)RPHFPQFSYSASGTA(-)	1 Met 2PO4	S473
3	466	480	1812.71	1812.84	-67.43	0	(R)RPHFPQFSYSASGTA(-)	2PO4	S473

Figure 3. Observed AKT1 tryptic peptides

The results of MS peptide mapping and bioinformatic filtering are shown. All modified and unmodified peptides are shown after clustering all three biological replicates. Each phosphopeptide region has been highlighted in gray. For each peptide, the following data attributes are listed: 'Sample' – biological replicate of observed peptide; 'Start' – location of n-terminus amino acid of peptide; 'End' – location of c-terminus amino acid of peptide; 'ThMass' – mass of *in silico* tryptic peptide; 'ExpMass' – observed peptide mass; 'dM' – delta mass, given in parts per million; 'MC' – number of missed cuts in peptide; 'Sequence' – primary amino acid sequence of observed peptides; 'Modification' – posttranslational modification present on peptide (CAM – carbamidomethyl cysteine, Cys-am – acrylamide modified cysteine, PO4 – phosphorylation of S/T/Y, Met-ox – oxidized methionine); 'Putative Phosphosites' – potential phosphosites within each phosphopeptide of AKT1 are enumerated.

4a

rAKT1 Region of Interest	Predicted Kinase - KinasePhos	Known Kinase
122..142 <u>S</u> G <u>S</u> P <u>S</u> D <u>N</u> S <u>G</u> A <u>E</u> E <u>M</u> E <u>V</u> A <u>L</u> A <u>K</u> P <u>K</u>		
S122	IKK	-
S124 *	cdc2, IKK	Unknown Kinase
S126	ATM	-
S129 *	CKII, IKK	Unknown Kinase
145..168 <u>V</u> T <u>M</u> N <u>E</u> F <u>E</u> <u>Y</u> L <u>K</u> L <u>L</u> G <u>K</u> G <u>T</u> F <u>G</u> K <u>V</u> L <u>V</u> K		
T146	PKA	-
Y152	EGFR, Jak	-
T160	other_MDD	-
308..328 <u>I</u> F <u>C</u> G <u>T</u> P <u>E</u> <u>Y</u> L <u>A</u> P <u>E</u> V <u>L</u> E <u>D</u> N <u>D</u> <u>Y</u> G <u>R</u>		
T308 *	other_MDD	PDK1
T312	other_MDD	Unknown Kinase
Y315 *	EGFR, Src, Jak, Syc, INSR, other_MDD	-
Y326 *	EGFR, Src, Jak, INSR, other_MDD	Unknown Kinase
378..386 <u>S</u> L <u>L</u> <u>S</u> G <u>L</u> L <u>K</u> K		
S378	CKI, IKK	-
466..480 <u>R</u> P <u>H</u> F <u>P</u> Q <u>F</u> S <u>Y</u> S <u>A</u> S <u>G</u> T <u>A</u>		
S473 *	PKA, other_MDD	PDK1 Regulated

Figure 4.

A. AKT1 Phosphopeptide Regions of Interest and Phosphosite Analysis. The results of MS peptide mapping and bioinformatic filtering are shown as five discrete regions of interest. For each region of interest, the amino acid location is given for the first and last residue. Within each region of interest, conserved S/T/Y residues are underlined. Residues that are conserved and predicted are listed separately under each region of interest. For each of these residues the corresponding predicted and/or known kinase is given. Predicted kinases are derived from KinasePhos 2.0 and represent clusters of kinases based on maximal dependence decomposition (MDD). IKK, cdc2, ATM, CKII, PKA, EGFR, Jak, Src, Syc, INSR, and CKI represent such kinase clusters. In addition, certain sites are designated as

‘other_MDD’ as this cluster represents multiple kinases. An ‘*’ indicates sites previously been shown to be phosphorylated on AKT1 *in vivo*.

B. Summary of Combined MS and Bioinformatics Approach. Distribution of AKT1 phosphosites identified from retinal immunoprecipitates. MALDI-TOF profiles of tryptic AKT1 was analyzed in triplicate by MODICAS and then filtered for residue conservation and phosphopeptide prediction. AKT1 sequence from *H. sapiens*, *M. musculus*, *R. norvegicus*, *X. tropicalis*, and *G. gallus* was aligned using ClustalV algorithm. All sites identified by the combined MS and bioinformatics approach are highlighted. Red - sites identified by MS and bioinformatics and previously identified *in vivo*. Yellow - sites not identified by MS and bioinformatics but previously identified *in vivo*. Green – potential novel phosphosites identified by MS and bioinformatics. AKT1 domains: PH Domain – green dashed line; Catalytic Domain – blue line; Regulatory Domain – red dashed line.

A.

B.

Figure 5. Confirmation of AKT1 Phosphosites

(A) Each of nine previously confirmed *in vivo* AKT1 phosphosites are displayed on a schematic AKT1. Residues that were identified in this study have been marked with an '*'. (B) Retinal lysates were subjected to SDS-PAGE and immunoblotted with α -AKT1, α -pS129-AKT1, α -pT308-AKT1, α -Y326-AKT1, α -pT450-AKT1, and α -pS473-AKT1. The primary antibodies are listed above their respective bands.

A.

B.

C.

D.

Method	Sensitivity	Specificity
MS	6/9	0.70
Bioinformatics	7/9	0.78
MS and Bioinformatics	6/9	0.89

Method	False Discovery Rate
MS and Bioinformatics	(1/72) / (13/72) 0.076

Figure 6. Assessment of Combined MODICAS and Bioinformatics Approach

(A) MS data analysis of AKT1 immunoprecipitates as conducted in Figure 2. (B) Bioinformatics - AKT1 data analyzed for phosphosites with a purely bioinformatics approach including S/T/Y residue conservation and phosphosite prediction criteria. (C) Venn diagram comparing the solution set of each of the three different phosphosite identification strategies versus the set of previously confirmed *in vivo* phosphosites. Upper circle – represents set of sites identified by MS, Bioinformatics, or MS and Bioinformatics. Lower circle – represents set of nine previously confirmed *in vivo* phosphosites. True positives (Tp) – represented by the intersection of the two sets being compared. False positives (Fp) – represented by the upper region of each diagram. False negatives (Fn) – represented by the lower region of each diagram. (D) Comparison of the sensitivity and specificity of MS data analysis, bioinformatics, and the combined MS and bioinformatics approach. The sensitivity (proportion of known phosphosites correctly identified ($Tp/(Tp + Fn)$)) and specificity (proportion of nonphosphorylated residues correctly identified ($Tn/(Tn + Fp)$)) for each method was calculated (where Tp = true positives, Fp = false positives, Tn = true negatives, and Fn = false negatives). For this analysis, the nine previously verified *in vivo* AKT1 phosphosites (T34, S124, S129, T308, Y315, Y326, T450, S473, and Y474) constituted the set of true positives (Tp). Each of the remaining 63 S/T/Y residues were

classified as unmodified residues and defined as true negatives (Tn). False discovery rate (decoy discovery rate/true discovery rate).