ARTICLE
Journal of Cellular Biochemistry 116:370–379 (2015)

Journal of Cellular Biochemistry

Exploring the Sequence Context of Phosphorylatable Amino Acids: The Contribution of the Upgraded MAPRes Tool

Zeeshan Iqbal,[1] Daniel C. Hoessli,[1,2] Wajahat M. Qazi,[3] Munir Ahmad,[4] Abdul Rauf Shakoori,[5*] and Nasir-ud-Din[1,6**]

[1]*Institute of Molecular Sciences and Bioinformatics, Lahore, Pakistan*
[2]*Panjwani Centre for Molecular Medicine and Drug Research, ICCBS, University of Karachi, Karachi, Pakistan*
[3]*Cybernetics Research Lab, Department of Computer Science, GC University, Lahore, Pakistan*
[4]*National College of Business Administration and Economics, Lahore, Pakistan*
[5]*School of Biological Sciences, University of the Punjab, Lahore, Pakistan*
[6]*H.E.J. Research Institute of Chemistry, ICCBS, University of Karachi, Karachi, Pakistan*

ABSTRACT

Several models that predict where post-translational modifications are likely to occur and formulate the corresponding association rules are available to analyze the functional potential of a protein sequence, but an algorithm incorporating the functional groups of the involved amino acids in the sequence analyses process is not yet available. In its previous version, MAPRes was utilized to investigate the influence of the surrounding amino acids of post- translationally and co-translationally modifiable sites. The MAPRes has been upgraded to take into account the different biophysical and biochemical properties of the amino acids that have the potential to influence different post- translational modifications (PTMs). In the present study, the upgraded version of MAPRes was implemented on phosphorylated Ser/Thr/Tyr data by considering the polarity and charge of the surrounding amino acids. The patterns mined by MAPRes incorporating structural information on polarity and charge of amino acids suggest distinct structure-function relationships for phosphorylated serines in a multifunctional protein such as the insulin-receptor substrate-1 (IRS-1) protein. The new version of MAPRes is freely available at http://www.imsb.edu.pk/Database.htm. J. Cell. Biochem. 116: 370–379, 2015.

KEY WORDS: POST- TRANSLATIONAL MODIFICATIONS; ASSOCIATION RULE MINING; PHOSPHORYLATION; POLARITY AND CHARGE; INSULIN-RECEPTOR SUBSTRATE PROTEIN 1; MAPRes; PHOSPHORYTABLE ANMINO ACIDS

Most proteins are multifunctional, both in vivo and in vitro [Jeffery, 1999], and it is important to elucidate how such multi-functionality arises [Hegyi and Gerstein, 1999], particularly in signaling proteins that undergo functional switches. In vivo, many proteins can be confined to minute niches to perform a variety of functions and prompted to perform yet other functions following slight modifications of their structure or conformation [Hegyi and Gerstein, 1999; Baker and Sali, 2001; Li et al., 2010]. To assess the protein's multi- functionality and to establish its structure-function relationship remains at present very difficult tasks. While such assessments can be made for blood proteins in vitro [Sennels et al., 2007] when isolated, purified and characterized by standard

Abbreviations: DIP, data inconsistency and preprocessing; ES, encoding scheme; IRS-1, insulin-receptor substrate-1; L-group, non-polar aliphatic; MAPRes, mining association patterns among preferred amino acid residues in the vicinity of amino acids targeted for post-translational modifications; N-group, negatively charged; NSR, negative sites retrieval; P-group, positively charged; PTMs, post-translational modifications; R-group, aromatic; RC, rule comparison; U-group, polar uncharged.

Conflict of interest: The authors have no conflict of interest to declare.

*Correspondence to: Dr. Abdul Rauf Shakoori, Distinguished National Professor and Director, School of Biological Sciences, University of the Punjab, Quaid-i-Azam Campus, Lahore 54590, Pakistan. E-mail: arshaksbs@yahoo.com, arshakoori.sbs@pu.edu.pk

**Correspondence to: Nasir-ud-Din, Institute of Molecular Sciences and Bioinformatics, Lahore, Pakistan. E-mail: prof_nasir@yahoo.com, chairman@imsb.edu.pk

Manuscript Received: 23 July 2014; Manuscript Accepted: 19 September 2014

Accepted manuscript online in Wiley Online Library (wileyonlinelibrary.com): 25 September 2014

DOI 10.1002/jcb.24983 • © 2014 Wiley Periodicals, Inc.

procedures [Moritz et al., 1994; Orengo et al., 1997; Takahashi et al., 2000], the vast number of multifunctional proteins performing concomitantly in a cell and exhibiting different structure-function relationships remain impossible to study [Hegyi and Gerstein, 1999]. Indeed, one needs to know how an isolated protein could change structurally and functionally when interacting with several other proteins, as it does in vivo. Furthermore, it is well known that in a protein, all the amino acids with substitution potential such as Ser, Thr, and Tyr, are not systematically modified, and the selection of the ones to be modified occurs on the basis of amino acid sequence, polarity and charge, which define a specific sequence context. Thus, classifications of the amino acids on the basis of physical, chemical, and structural properties are required to understand the environment of modifiable residues [Yaffe et al., 2001; Li et al., 2010].

In vivo, many cooperating proteins smoothly execute signaling and numerous other tasks, indicating that functional cooperation of proteins is a routine and necessary process in cells, and is likely to be perturbed in countless pathological situations [Li et al., 2010]. Multifunctionality is imparted to proteins when amino acids are post-translationally modified on hydroxyl-, amino- and carboxyl-groups. Substitutions on such groups are decisive chemical reactions in modulating protein functions. Protein proteolytic processing is also important in generating new functionalities as in the complement and coagulation cascades. The modification of the hydroxyl group on Ser, Thr, and Tyr by phosphorylation/dephosphorylation provides many opportunities for proteins to perform distinct functions. Phosphorylation of Ser473 and Thr308 in the Akt/PKB kinase at the membrane leads to its activation and generation of further intra-cytoplasmic and nuclear enzymatic and signaling activities [Kunkel et al., 2005]. Another example is that of cytochrome c (bovine heart) that also requires dephosphorylation of Tyr 48 and/or 97 to interact with the Apaf complex and activate caspases [Hüttemann et al., 2011]. Therefore, substitution of hydroxyl groups by phosphate creates functional switches governing metabolic adaptations and signaling, involving nearly sixteen hundred kinases [Cao and Chen, 2009]. Similarly, glycosylation of Ser and Thr by the sugars *O*-GlcNAc and *O*-GalNAc provides memory, stability and protection against proteolysis, conformation and stereo-electronic specificity [Slawson and Hart, 2003; Rexach et al., 2008; Martinez-Fleites et al., 2010].

Association rule mining between two or more objects in a large amount of data has been extensively applied on variety of databases to identify hidden knowledge [Agrawal et al., 1993; Agrawal and Srikant, 1994; Savasere et al., 1995]. The MAPRes algorithm has capacity to mine significantly preferred sites and association patterns/rules for neighboring amino acids around modified residues at different support levels and confidence levels. An association rule (e.g., $<G,2> = S$, i.e., Gly at position 2 around modified Ser) is an expression for significant residues at certain positions around modified residues. The total number of records in a dataset that have a certain pattern is the support level of that pattern while the ratio of the support value to the total number of records in a dataset is the confidence level of the pattern. MAPRes has been implemented on different modified protein datasets [Ahmad et al., 2008a; Ahmad et al., 2009; Iqbal et al., 2013].

In new version of MAPRes, biophysical and biochemical properties of the surrounding amino acids of modified residues has been taking into account through classification of the amino acids. The classification of amino acids is important to study different functional and metabolic aspects that are directly associated with their functional group or side chain structure. It has been reported that in the case of tissue injury involving different infections, asthma and allergic reactions, hypobromous acid is generated by the activated eosinophils. This molecule affects the protein folding by oxidation of aromatic amino acids and consequently inhibits protein function [Hawkins and Davies, 2005]. The polar and charge amino acids play a critical role to control different functional activities of proteins such as in human multidrug resistance protein 1 (MDR-1) and these amino acids have been shown to be involved in conferring resistance to different anticancer drugs. The polar and charged amino acids are indeed involved in controlling the substrate binding affinity to human MDR-1 [Zhang et al., 2003]. The new version of MAPRes has the capacity to analyze primary sequence of the proteins according to specific properties of the amino acids such as hydrophobicity, hydrophilicity, polarity, charge, etc. In this study, the previous version of MAPRes [Ahmad et al., 2008b] has been upgraded to analyze the sequence of amino acids surrounding modified residues, taking the charge and polarity of surrounding residues into account.

MATERIALS AND METHODS

MAPRes has been upgraded by the addition of four new modules that provide the opportunity to mine association patterns and estimate which are the significantly preferred residues around the modified sites on the basis of biochemical and biophysical properties of the amino acids. The estimation of polarity and charge of the neighboring amino acids is illustrated by mining sequence patterns for phosphorylated Ser/Thr/Tyr residues. Two types of analyses were performed by using the upgraded version of MAPRes. Firstly, the estimation of preferred amino acids and association patterns were determined according to the polarity and charge of the neighboring amino acids of phosphorylated and non-phosphorylated Ser/Thr/Tyr residues. Secondly, the general dataset was considered for mining significantly preferred sites and association patterns.

UPGRADING MAPRES

In the upgraded version of MAPRes, different new modules such as data inconsistency and pre-processing (DIP), Negative Sites Retrieval (NSR), Encoding Scheme (ES), and Rule Comparison (RC) have been integrated. The DIP module checks for inconsistency in the assembled data. This module performs: (i) Sequence Consistency Analysis; (ii) Modified Site Position Profile Analysis; and (iii) Duplicate Entry Analysis. The NSR module is very useful to investigate non-modified sites, as it efficiently segregates modified and non-modified sites and minimizes the large amount of non-modified data relative to modified sites. The ES module is the most essential update that supports analysis of primary sequences and mines association patterns for surrounding amino acids on the basis of their biophysical and biochemical properties. The RC module

was designed to compare the rules mined by MAPRes with those obtained with existing computational prediction models (See Manual for the description of these modules that is attached in the downloaded file of MAPRes in above described link).

DATASET PREPARATION

In this paper, updated MAPRes was applied on *O*-phosphorylation sites collected from Phospho.ELM version 9.0 [Dinkel et al., 2011] (http://phospho.elm.eu.org/), containing 8,718 proteins that have 31,754 entries for Ser, 7,449 for Thr, 3,371 for Tyr, and 1 for His as shown in Table I. The downloaded file from Phospho.ELM was converted into an XML file after removing the trivial information. The single entry of His was removed manually from the database, and the resulting XML data repository was processed for data inconsistency. Duplicate entries, incorrect sequences, and wrong position of the modified residues were found out by the DIP module and corrected to make the dataset utilizable for MAPRes. The NSR module was used to retrieve the non-phosphorylated Ser/Thr/Tyr. These non-modified sites were in large numbers as compared to the modified residues, which could generate uncertainties in the results mined by MAPRes. The NSR module also has the capacity to reduce the dataset of non-modified sites with respect to modified sites by ratio 1:1 and 1:2 (one-one and one-two ratio for modified and non- modified sites, respectively) (See Section 4.3 of MAPRes Manual for preparation of the non-modified dataset).

CLASSIFICATION OF AMINO ACIDS BASED ON POLARITY AND CHARGE

The resulting XML data was encoded by the ES module into five groups according to polarity and charge of the amino acids. This classification method distributes all 20 standard amino acids into five distinct groups such that the amino acids with non-polar aliphatic side chain (L-group) were marked 'L', aromatic (R-group) residues 'R', polar uncharged (U-group) 'U', negatively charged (N-group) 'N', and positively charged (P-group) 'P.'

MAPRES METHODOLOGY AND APPLICATION

The basic working principle of MAPRes is association pattern mining among the modified residues and significantly preferred amino acids in their vicinity. MAPRes first generates peptides of 21 amino acids in length (modified residues at 0 position and 10 amino acids both downstream and upstream) and determines the nature and number of significantly preferred amino acids in the vicinity of modified residues. Next, these significantly preferred amino acids are considered for the development of association between modified and significantly preferred amino acids [Ahmad et al., 2008b]. Herein, the updated version of the MAPRes was used for the preference estimation and association rule mining for amino acids surrounding modified and non-modified residues, not only for the primary sequence dataset, but also for the classified dataset incorporating to the polarity and charge of the amino acids. All of those Ser/Thr/Tyr residues which had not yet been reported for phosphorylation were considered as non-modified sites. The association patterns mined by MAPRes in this study were obtained at different support levels in different datasets (Table I).

RESULTS

MAPRes was implemented for the development of correlations between amino acids next to phosphorylated and non-phosphorylated residues. The analyses were performed on charge-specific sequence environments and general primary sequences of phosphorylated Ser/Thr/Tyr residues.

PREFERENCE ESTIMATION FOR CHARGE-SPECIFIC DATASETS

Preference estimation based on charge and polarity around phosphorylated and non-phosphorylated Ser/Thr/Tyr was done in two steps. Firstly, the frequency of occurrence for neighboring amino acids was estimated and secondly significantly preferred amino acids/sites were identified around phosphorylated (Fig. 1) and non-phosphorylated (Supplementary Fig. S1) Ser/Thr/Tyr. The observed frequencies of amino acids at each position (-10 to $+10$) around phosphorylated and non-phosphorylated Ser/Thr/Tyr was assessed. The frequency estimation around phosphorylated Ser, Thr, and Tyr indicated that L-group amino acids have highest frequency at various positions. Among other positions, the $+1$ position around phosphorylated Ser and Thr is significantly occupied by L-group amino acids (Fig. 1a, b). In contrast, phosphorylated Tyr has shown highest frequency for L-group amino acids at $+3$ position instead of $+1$ (Fig. 1c). The frequency graph for non- phosphorylated in charge and polarity based dataset observed that L-group amino acids have highest frequency at all positions (Supplementary Fig. S1).

The charge and polarity-based s-preferred estimation specified 32, 25, and 27 sites around phosphorylated Ser, Thr, and Tyr, respectively (Table I). MAPRes identified 26 out of 32 s-preferred sites for charged amino acids, precisely 14 for P-group amino acids, and 12 for N-group amino acids while other residues were U-group amino acids (Table II). Similarly, in case of phosphorylated Thr, P-group amino acids were found at 10 out of 25 s-preferred sites. Both, N- group and U-group amino acids were found at 6 s-preferred sites

TABLE I. Data Statistics of Phospho.Elm Data and Association Patterns Mined by MAPRes

	Ser	Thr	Tyr	Total
Number of modified sites	31,754	7,449	3,371	42,574
Number of non-modified sites	3,275,968	435,443	78,213	3,789,624
Significantly preferred sites (Phosphorylated)	32	25	27	84
Significantly preferred sites (non-phosphorylated)	28	19	18	65
Association patterns mined for phosphorylation	23	56	30	80
Association patterns for non-Pposphorylation	240	101	19	360

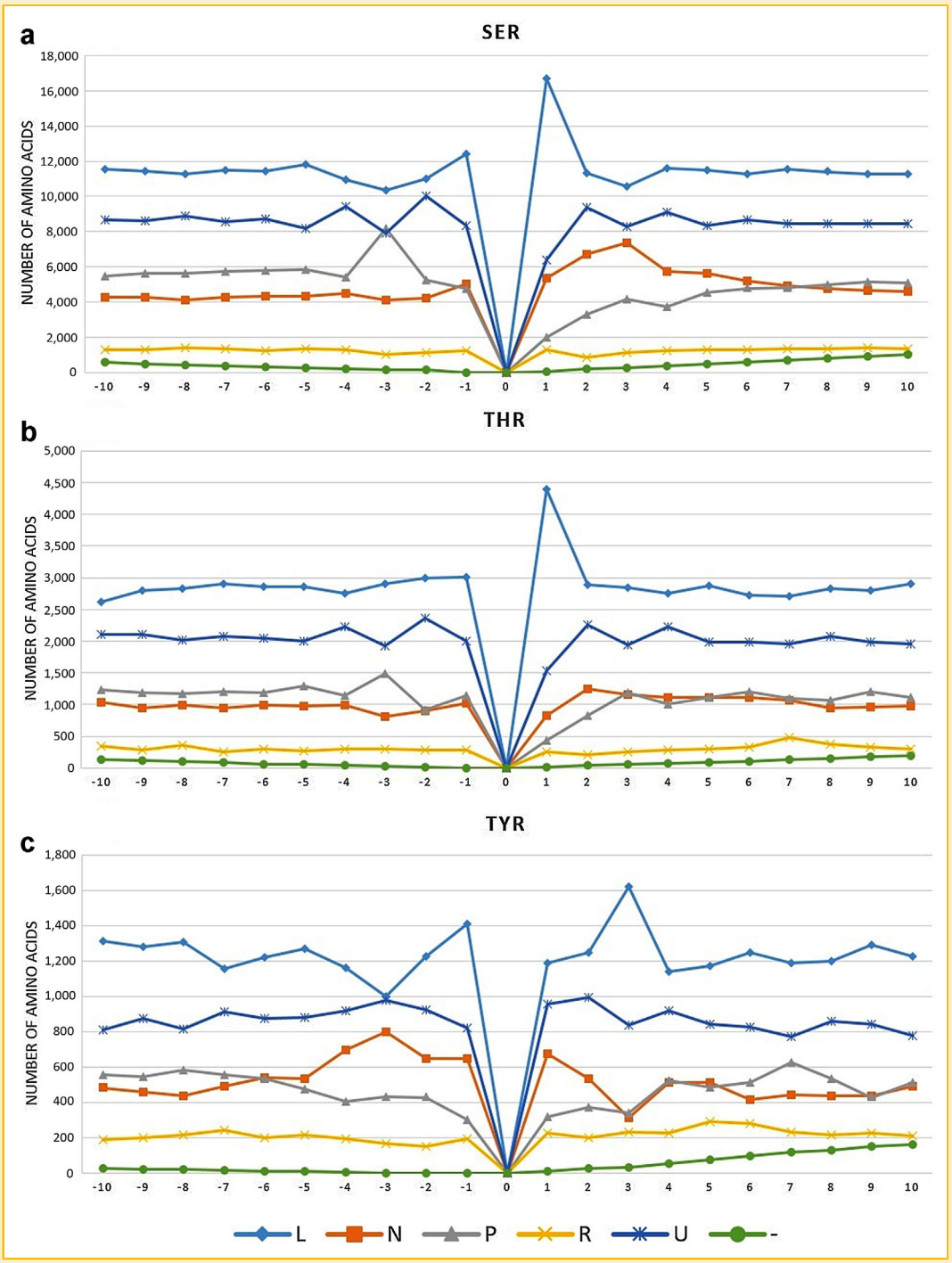

Fig. 1. The frequency diagram of surrounding amino acids of (a) phosphorylated Ser, (b) phosphorylated Thr, and (c) phosphorylated Tyr, classified as L (non-polar aliphatic), N (negatively charged), P (positively charged), R (aromatic) and U (polar uncharged), on the basis of defined classification of the amino acids.

around phosphorylated Thr. MAPRes showed N-group amino acids at 13 out of total 27 s- preferred sites in preference estimation around phosphorylated Tyr, and P-group amino acids were found selectively at 7 s-preferred sites.

The analyses performed for non-phosphorylated residues, MAPRes found 65 s-preferred sites in total (Table II). Among these s-preferred sites 28 were for Ser, 19 for Thr, and 18 for Tyr. U-group amino acids were found s-preferred at 20 and 17 different positions in vicinity of non-phosphorylated Ser and Thr, respectively. In contrast, non-phosphorylated Tyr has highest preference for R-group amino acids at 15 out of 18 s-preferred sites.

TABLE II. Significantly Preferred Positions Mined by MAPRes for Phosphorylated and Non- Phosphorylated Ser, Thr, and Tyr on the Basis of Polarity and Charge of the Surrounding Amino Acids

	Phosphorylated			Non-phosphorylated		
Classified	Ser	Thr	Tyr	Ser	Thr	Tyr
L	1,	−2, −1, 1,	−1, 3,	–	1,	–
N	−4,−1,1,2,3,4, 5,6,7,8,9,10,	2,3,4,5,6,7,	−10,−7,−6,−5,−4, −3,−2,−1,1,2,4,5,10,	–	–	–
P	−10,−9,−8,−7, −6,−5,−4,−3, −2,6,7,8,9,10,	−10,−9,−8,−7, −6,−5,−3,3, 6,9,	−10,−9,−8,−7,−6,7,8,	−7,−6,−5,−3,−1,5,7,8,	−1,	−6,1,5,
R	–	–	5,6,	–	–	−10,−9,−8,−7,−6, −4,−3,−2,−1,1, 4,7,8,9,10,
U	−8,−4,−2,2,4,	−10,−9,−4, −2,2,4,	−3,1,2,	−10,−9,−8,−7,−6, −5,−4,−3,−2,−1,1,2, 3,4,5,6,7,8,9,10,	−10,−9,−8,−6,−5, −4,−3,−2,2,3,4, 5,6,7,8,9,10,	–

PREFERENCE ESTIMATION FOR GENERAL DATASETS

Statistically preferred amino acids (without considering polarity and charge) of surrounding amino acids of phosphorylated Ser/Thr/Tyr were found by MAPRes. The frequency diagram for phosphorylated Ser and Thr showed that the Pro at +1 has highest frequency (Supplementary Fig. S2). Other amino acids which found at high frequency around phosphorylated Ser were Arg (at position −3), Ser (at position +2 and −2), and Glu (at position +3). The frequency diagram for phosphorylated Thr indicated that Pro (at position +2) and Ser (at position −1 and +3) had the highest frequency (Supplementary Fig. S2) and around phosphorylated Tyr, Glu, Asp, Pro, and Ser most frequently occurred at residues −3, −2, +3 and Ser at +1 position (Supplementary Fig. S2). After calculation of the frequency of each amino acid around phosphorylated residues, MAPRes estimated the s-preferred sites and found 309 s- preferred sites for phosphorylated Ser/Thr/Tyr (Supplementary Table SI). Among these s-preferred sites, 115 were for phosphorylated Ser, 98 were for phosphorylated Thr, and 96 for phosphorylated Tyr. The s-preferred site estimation showed that Ser, Pro, Arg, Asp, Gly, and Glu have highest preference in the vicinity of phosphorylated Ser/Thr/Tyr (Supplementary Table SI). Similarly, for the preference estimation for non-phosphorylated Ser/Thr/Tyr, the MAPRes found a total of 94 s-preferred amino acids.

ASSOCIATION PATTERNS FOR CHARGE-SPECIFIC DATASET ANALYSIS

MAPRes analyzed the charge-specific dataset and extracted 109 association patterns for phosphorylated Ser/Thr/Tyr. Among those, 23 association patterns were mined for phosphorylated Ser at support levels of 5%, 10%, 15% up to 50% (Table III). It was observed that L-group amino acids at +1 position had the highest support level (up to 50%) in the vicinity of phosphorylated Ser. The same pattern was also found around phosphorylated Thr but with a slightly higher support level (up to 60%). In addition to the +1 position, the −1 and −2 positions were also found preferred for L-group amino acids at high support level (up to 40%) around phosphorylated Thr. For phosphorylated Tyr, the L-group amino acids were also found at position −1 with a 40% support level. The U-group amino acids were significantly preferred at −2, −4, −8, +2, and +4 positions around phosphorylated Ser with 20% and 25% support level, but the −2 position was also found at a 30% support level. A similar trend was observed for phosphorylated Thr as U-group amino acids were preferred at −2, 2, and 4 position with 30% support level. Furthermore, MAPRes identified L-group amino acids at +3 and −1 position with 50% and 25%–40% support levels respectively in vicinity of phosphorylated Tyr. Around the phosphorylated Tyr, the L-group amino acids were found by MAPRes at −3, +1, and +4 position with 25% support level while at +2 position with 30% support level. The N-group amino acids were also found around phosphorylated Tyr at −3 position with 25% support level (Table III).

MAPRes also mined several association patterns for non-phosphorylated Ser/Thr/Tyr at different support levels. The U-group amino acids were mined by MAPRes at 30% support level for non-phosphorylated Ser and Thr (Supplementary Table SII). The L-group amino acids around non-phosphorylated Thr were found at even higher percent (40%) support level (Supplementary Table SII). Most interestingly, the P-group amino acids were identified at +1, +5, and −6 positions around non-phosphorylated Tyr with the highest support level (Supplementary Table SII).

ASSOCIATION PATTERNS FOR GENERAL DATASET ANALYSIS

The results concerning the environment of phosphorylated Ser/Thr/Tyr in the general dataset showed 188 unique association patterns (Supplementary Table SIII). The updated version of the MAPRes mined these association patterns at varying support levels, ranging from 5% to 40%. Out of these total association patterns/rules, 35 were for Ser, 50 for Thr, and 103 for Tyr within a variable range of confidence levels. The most significantly preferred residue in the vicinity of phosphorylated Ser was Pro at +1 position with maximum of 30% support level. A similar trend was observed also in the vicinity of the phosphorylated Thr as well with 10–40% support level. Moreover, the other association patterns for phosphorylated Ser mined with good confidence level were <Glu, +3> (i.e., Glu at position +3), <Arg, −3>, <Ser, +2, and −2> (Supplementary Table SIII).

TABLE III. The Association Rules Mined by MAPRes for Phosphorylated Ser/Thr/Tyr on the Basis of Polarity and Charge

Association rules	Confidence level	Support level	Association rules	Confidence level	Support level
Phosphorylated Ser			Phosphorylated Thr		
<N,1><N,2><N,3>	100	5	<L,−1><L,1><U,2>	100	5
<U,−2><L,1><U,4>	79.76	5	<L,−1><L,1><U,4>	100	5
<U,−4><L,1><U,2>	79.55	5	<L,−1><L,1><U,8>	100	5
<U,−4><L,1><U,4>	79.86	5	<L,1><U,2><U,4>	20.30	5
<U,−4><U,−2><L,1>	82.04	5	<L,1><U,2><U,8>	100	5
<U,−8><U,−4><L,1>	100	5	<L,1><U,4><U,8>	100	5
<N,1><N,3>	100	10	<L,−2><L,1><P,3>	100	5
<P,−3><L,1>	82.32	10	<L,−2><L,1><P,6>	100	5
<U,−4><U,−2>	83.47	10	<L,−2><L,1><P,9>	100	5
<L,1><U,2>	79.19	10, 15	<L,−2><L,1><U,2>	100	5
<L,1><U,4>	79.36	10, 15	<L,−2><L,1><U,4>	100	5
<U,−2><L,1>	80.99	10, 15	<L,−2><L,−1><U,4>	100	5
<U,−4><L,1>	79.98	15	<L,−2><L,1><U,8>	100	5
<U,−8><L,1>	100	10, 15	<U,−2><L,−1><L,1>	100	5
<N,2>	80.38	20	<U,−2><L,1><U,2>	19.71	5
<N,3>	87.66	20	<U,−2><L,1><U,4>	20.24	5
<P,−3>	84.91	20, 25	<U,−2><L,1><U,8>	100	5
<U,2>	74.32	20, 25	<U,−4><L,−1><L,1>	100	5
<U,4>	74.32	20, 25	<U,−4><L,1><U,2>	20.45	5
<U,−4>	81.79	20, 25	<U,−4><L,1><U,4>	20.14	5
<U,−8>	100	20, 25	<U,−4><L,1><U,8>	100	5
<U,−2>	82.19	20, 25, 30	<U,−4><L,−2><L,1>	100	5
<L,1>	79.77	20, 25, 30, 35, 40, 45, 50	<U,−4><L,−2><L,−1>	100	5
Phosphorylated Tyr			<U,−4><U,−2><L,1>	17.96	5
<L,−1><L,3><U,4>	100	5	<U,−7><L,−1><L,1>	100	5
<L,−1><N,1><L,3>	100	5	<U,−7><L,1><U,2>	100	5
<L,−1><U,1><L,3>	100	5	<U,−7><L,1><U,4>	100	5
<L,−1><U,2><L,3>	100	5	<U,−7><L,1><U,8>	100	5
<N,−3><L,− 1><L,3>	100	5	<U,−7><L,−2><L,1>	100	5
<N,−4><L,−1><L,3>	100	5	<U,−7><U,−2><L,1>	100	5
<U,2><L,3><U,4>	100	5	<U,−7><U,−4><L,1>	100	5
<U,−3><L,−1><L,3>	100	5	<U,−9><L,−1><L,1>	100	5
<L,−1><U,1>	100	10	<U,−9><L,1><U,2>	100	5
<L,−1><U,2>	34.68	10	<U,−9><L,1><U,4>	100	5
<L,−1><U,4>	33.55	10	<U,−9><L,1><U,8>	100	5
<L,3><U,4>	100	10	<U,−9><L,−2><L,1>	100	5
<N,1><L,3>	100	10	<U,−9><U,−2><L,1>	100	5
<N,−1><L,3>	100	10	<U,−9><U,−4><L,1>	100	5
<N,−2><L,3>	100	10	<U,−9><U,−7><L,1>	100	5
<N,−3><L,−1>	100	10	<L,−2><L,−1><L,1>	100	10
<N,−4><L,3>	100	10	<L,1><U,2>	20.81	15
<U,1><L,3>	100	10	<L,1><U,4>	20.64	15
<U,−3><L,−1>	100	10	<L,1><U,8>	100	15
<U,−3><L,3>	100	10	<L,−2><L,−1>	100	15
<N,−3><L,3>	100	15	<U,−2><L,1>	19.01	15
<U,2><L,3>	100	15	<U,−4><L,1>	20.02	15
<L,−1><L,3>	100	20	<U,−7><L,1>	100	15
<N,−3>	100	25	<U,−9><L,1>	100	15
<U,1>	100	25	<L,−1><L,1>	100	25
<U,−3>	100	25	<L,−2><L,1>	100	15, 20, 25
<U,4>	8.00	25	<U,2>	17.06	30
<U,2>	8.62	30	<U,−2>	17.81	30
<L,−1>	34.10	25, 30, 35, 40	<U,4>	17.68	30
<L,3>	100	50	<L,−1>	65.906	30, 35, 40
			<L,−2>	100	30, 35, 40
			<L,1>	20.23	30, 35, 40, ... 60

MAPRes searched and mined 107 association patterns by analyzing amino acids surrounding non-phosphorylated Ser/Thr/Tyr in the general dataset (Supplementary Table SIV). MAPRes determined 87 association rules for non-phosphorylated Ser, 14 for Thr, and only 6 for Tyr. It was determined that Ser itself was most significantly preferred at most of the positions around non-phosphorylated S at 5% support level, and varying confidence level ranging from 96.58 to 100 (Supplementary Table SIV).

Similarly, out of 14 association patterns/rules for non-phosphorylated Thr, five were found for Thr itself at 5% support level. The most frequent residue around non-phosphorylated Tyr was Phe at −7 and +9 position with 5% support level (Supplementary Table SIV).

VALIDATION OF ASSOCIATION RULES

The association rules mined by the upgraded MAPRes for phosphorylated and non-phosphorylated residues were investigated by existing phosphorylation prediction models as well. Prediction models based on polarity and charge of amino acids not being available, direct comparison of polarity and charge-based analysis with existing general computational models was not possible. However, for indirect comparison, the amino acids of the predicted dataset classified as previously defined was then searched for all mined association patterns established on the basis of polarity and charge of vicinal amino acids and showed a high level of conformity (Table IV). To perform this task, 30 proteins were selected randomly from the uniprot database without any prior knowledge of their phosphorylation status. These selected proteins were used to identify the phosphorylatable residues by utilizing DISPHOS and the NetPhos 2.0 prediction servers. DISPHOS predicted 293 sites for phosphorylated Ser, 98 sites for phosphorylated Thr, and 61 sites for phosphorylated Tyr. NetPhos 2.0 predicted 281, 95, and 54 sites for phosphorylated Ser, Thr and Tyr, respectively (Table IV). After the identification of potential predicted sites, 21 amino acid-long peptides were constructed for all Ser/Thr/Tyr from the selected 30 protein in the dataset. The association patterns mined by MAPRes explored in the selected peptides and calculated peptides that contained one or more association rules. MAPRes rules for phosphorylated Ser showed 90% consistency with DISPHOS, and 93% with NetPhos 2.0. For phosphorylated Thr, the consistency was 87% with DISPHOS and 85% with NetPhos 2.0. The comparative results for phosphorylated Tyr showed that 95% of the DISPHOS predictions and 85% of the NetPhos 2.0 ones were consistent with association patterns mined by MAPRes. The rules generated by MAPRes were also investigated in the vicinity of predicted and non-predicted sites of DISPHOS and NetPhos 2.0. Comparison with DISPHOS and NetPhos 2.0 results closely agreed with the results mined by MAPRes (Table IV).

IMPLEMENTATION OF UPGRADED MAPRES ON PHOSPHORYLATED SER IN HUMAN IRS-1 PROTEINS

To test the new version of MAPRes, we compared two different sets of Ser in the insulin-receptor protein 1 (IRS-1) protein, which critically modulates signals received by the insulin receptor [Fröjdö et al., 2009]. We have taken advantage of the study by Yi and others who performed a global assessment of phosphorylation occurring in vivo in the IRS-1 of human striated muscle cells [Yi et al., 2007]. In normal human subjects responding to insulin, the status of Ser in striated muscle IRS-1 was determined by mass spectrometry on IRS-1 peptides retrieved from the biopsy material. Using the upgraded MAPRes, we have compared the Ser that were found to be phosphorylated with 20 non-phosphorylated ones taken at random in the tryptic peptides retrieved from the analysis. The results are shown in Figure 2 in form of sequence logos document the charged and polar residues found in the environment of phosphorylated and non-phosphorylated Ser. While non-polar aliphatic residues predominate in the −1 to −5, and +1 to +5 positions of non-phosphorylated Ser, the environment of phosphorylated Ser shows predominance of P (basic polar) residues at position −3 and uncharged (U) polar residues at position −1. Overall, positions −2 to −5 exhibit more abundant positively charged residues in phosphorylated Ser than in non-phosphorylated Ser of IRS-1. The differences are less conspicuous in the +1 to +5 positions, where U is exchanged for L at +4 and +5 in Ser of phosphorylated IRS-1.

These distinctive properties of phosphorylated versus non-phosphorylated Ser suggest that incorporating polarity and charge in the analysis of the sequence environment could highlight distinctive functional properties associated with Ser in IRS-1. Phosphorylated Ser in IRS-1 have indeed been shown to be inhibitory or activating with regard to the downstream signaling of the insulin receptor [reviewed in Fröjdö et al., 2009]. The comparison of the rules mined by MAPRes with the peptides of these inhibitory and activating P- Ser suggests that distinctive properties can be assigned to each (Supplementary Table SV). The inhibitory and activating activities of such phosphorylated Ser residues is probably related to distinct signaling proteins associating with them, causing negative or positive modulation of the insulin signal [Fröjdö et al., 2009].

TABLE IV. Validation of the Rules Mined by Mapres With Existing Prediction Models for Phosphorylated Ser/Thr/Tyr

		DISPHOS			NetPhos 2.0		
	Number of rules mined by MAPRes	Total number of predicted sites	Number of peptides in which rules were found	Conformity with MAPRes rules (In %)	Total number of predicted sites	Number of peptides in which rules were found	Conformity with MAPRes rules (In %)
Ser	23	293	264	90	281	261	93
Thr	56	98	85	87	95	80	85
Tyr	30	61	58	95	54	46	85

To validate the rules mined by MAPRes with DISPHOS and NetPhos 2.0, 30 protein sequences were selected without any prior knowledge of phosphorylation. The prediction models were used for the prediction of phosphorylated sites in selected proteins. The total number of sites that were predicted by prediction models is in 3rd and 6th columns. The peptides of length 21 amino acids (10 amino acids on each side of predicted site) formulated computationally and patterns mined by MAPRes were searched in these peptides. In 5th and 7th columns the percentages of consistency of MAPRes patterns and surrounding environment of predicted sites are described.

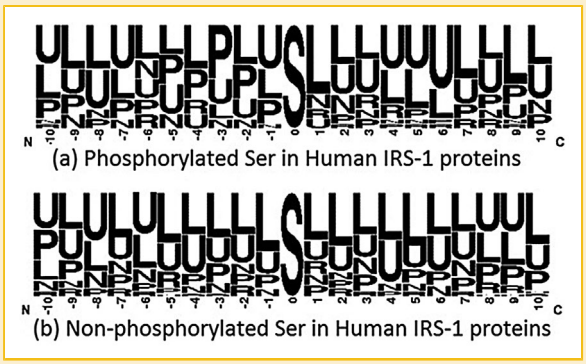

Fig. 2. Sequence context comparison of phosphorylated and non-phosphorylated Ser in human IRS-1. IRS-1 protein peptides extracted from needle biopsies of vastus lateralis muscle obtained from healthy donors following insulin stimulation, were analyzed by capillary high performance liquid chromatography tandem mass spectrometry (Yi et al., 2007). The sequence logo of the classified amino acids at each position around phosphorylated Ser (a), and nonphosphorylated Ser (b), are shown. The size of the letter corresponds to the frequency of L, P, N, and U class of amino acid. These frequency logos were generated utilizing Weblogo (http:/weblogo.berkeley.edu/).

DISCUSSION

A pure crystalline protein in vitro with a known amino acid sequence, polarity, charge, and conformation will not perform more than one function. However, proteins in vivo are always in contact with other proteins of different charge, polarity, and conformation. These interactions induce a controlled modulation of protein structure and consequently permit functional switches. Another way to induce multi-functionality is by substituting a functional group on amino acids with phosphate, acetyl, methyl, or sugar residues. For instance, the substitution with phosphate or glycosylation with N-glucosamine or N-acetyl glucosamine are important in that they may induce important multiple functions such as recognition, complex formation, or control of synthesis and degradation. In this study, we have developed rules that describe and may explain the influence of polarity and charge on the substitution of functional groups that result in protein multi-functionality.

The substituent groups commonly involved in protein modifications are phosphate, acetate, methyl, and quite abundantly sugar residues in glycosidic linkage [Rucker and McGee, 1993; Iakoucheva et al., 2004]. In a number of cases, the charged substituent group such as phosphate and acetate are expected to be influenced by the polarity and charge of acceptor proteins. It is considered that the polar and charged amino acids of substituent-accepting protein will have sufficient effect on site selection for transferring substituent polar and charged groups [Kitchen et al., 2008]. The data available suggests that the transferring enzymes (kinases and transferases) are specific for the accepting amino acid residues [Kitchen et al., 2008].

The assessment of the modification potential of a protein and its functional implications are difficult to study using wet lab techniques [Beausoleil et al., 2006]. In silico studies are therefore useful to identify the influence of surrounding amino acids on the modification potential of phosphorylatable Ser/Thr/Tyr residues. The sequence pattern of amino acids in the primary sequence around phosphorylated Ser/Thr/Tyr has been described earlier [Zhu et al., 2005; Qazi et al., 2006; Ahmad et al.,2008a,b, 2009], but further examination of the biophysical and biochemical properties of surrounding amino acids are critically important to define how amino acids are selected for modification. For instance, MAPRes was utilized to assess the properties (biophysical and biochemical) of amino acids surrounding acetylation sites. The results obtained from such studies stressed the importance of the nature of surrounding residues as well as their properties [Qazi et al., 2006; Ahmad et al., 2008a; Ahmad et al., 2009; Iqbal et al., 2013].

The upgraded version of MAPRes suggested various s-preferred positions and association patterns for charged (positively and negatively) and polar amino acids around phosphorylated Ser/Thr/Tyr. Several earlier studies described the occurrence and role of the different properties of amino acids surrounding modified residues [Iakoucheva et al., 2004]. It has been described by [2001] that the abundance of polar and non-polar amino acids among the nearest surrounding residues, is the signature for phosphorylated Ser. Similar positions are also proposed by MAPRes, so that L-group (at +1 position) and U-group (at +2, +4, −2, and −4 positions) amino acids are favored around Ser. Furthermore, N-group amino acids were reported for phosphorylated Tyr at various positions [Tatárová et al., 2012], and the occurrence of L-group and U-group residues at several positions around Tyr (<L, 3, −1>, <U, −1, −3, 2>) were also described in this study. Thus, the association patterns mined by MAPRes are highly consistent with previous results and add novel important characteristics to the environment of phosphorylated residues. MAPRes also found significantly preferred positions and association patterns for non-phosphorylated sites. It was significantly observed that non-phosphorylated Tyr has only P-group and R-group amino acids in their vicinity.

The association patterns mined by MAPRes were also compared with those results determined by explicit experiments. For instance, the substrate specificity of different families of kinases based on sequence around a phosphorylated site has been discussed as, for instance, Pro at +1 position is an absolute requirement for Proline-directed kinases and Ile at +1 for basophilic kinases [Zhu et al., 2005]. The Ile and Pro both are non-polar amino acids and their preference revealed the requirement of a specific physico-chemical environment around phosphorylated residues. In our analyses, a pattern <L, 1> => S (i.e., non-polar at +1 position around phosphorylated Ser) was mined by MAPRes at a maximum support level of 50% (Table III), while in previous analyses (without consideration of polarity and charge), the <Pro, +1> => Ser (Pro at +1 position, a non-polar amino acids around Ser) was found at only 10% support level. A pattern <P, 1><S, 4> => S in our previous analyses was mined only at 5% support, while in the present studies, a pattern <L, 1><U, 4> was mined at a maximum of 15% support level (Table III). Another important pattern for phosphorylated Thr, was Pro at +1 position (found in previous analyses at 35% support) was attributed a 55% support level in this study (Table III). These increased support levels, and the literature [Zhu et al., 2005] suggested that polarity and charge

of the amino acids is critical at certain position around phosphorylated residues [Zhu et al., 2005; Qazi et al., 2006; Ahmad et al., 2008a,b, 2009; Iqbal et al., 2013].

CONCLUSIONS

Association rules mined by upgraded version of MAPRes on the basis of polarity and charge of the amino is an interesting approach for the detection of useful motifs that can have important role for the regulation of phosphorylation on Ser/Thr/Tyr. The results generated by MAPRes were in line with existing computational models and literature. However, the approach of MAPRes is the development of functional and structural correlations between modified residues and surrounding amino acids, and it cannot be considered as the prediction model. The development of the correlation between modified sites and specific properties of the surrounding amino acids is the main reason for the upgradation in MAPRes and our application of the upgraded MAPRes to the analysis of IRS-1 datasets strongly suggests that such corrections may be found.

ACKNOWLEDGEMENT

Nasir-ud-Din acknowledges financial support for this research effort from Pakistan Academy of Sciences (PAS) and EMRO-COMSTECH.

REFERENCES

Agrawal R, Srikant R. 1994. Fast algorithms for mining association rules in large databases. Proc. 20th International Conference on Very Large Data Bases (VLDB94). Santiago de Chile, Chile, September 12–15. 1994, 487–499.

Agrawal R, Imielinski T, Swami AN. 1993. Mining association rules between sets of items in large databases. Proc. ACM SIGMOD Conference on Management of Data, D.C., Washington, May 26–28, 1993, 207–216.

Ahmad I, Hoessli DC, Qazi WM, Khurshid A, Mehmood A, Walker-Nasir E, Ahmad M, Shakoori AR. 2008a. MAPRes: An efficient method to analyze protein sequence around post-translational modification sites. J Cell Biochem 104:1220–1231.

Ahmad I, Mehmood A, Khurshid A, Qazi WM, Hoessli DC, Walker-Nasir E, Shakoori AR, Nasir-ud-Din. 2009. Phosphoproteome sequence analysis and significance: Mining association patterns around phoshorylation sites utilizing MAPRes. J Cell Biochem 108:64–74.

Ahmad I, Qazi WM, Khurshid A, Ahmad M, Hoessli DC, Khawaja I, Choudhary MI, Shakoori AR, Nasir-ud-Din. 2008b. MAPRes: Mining association patterns among preferred amino acid residues in the vicinity of amino acids targeted for post-translational modifications. Proteomics 8:1954–1958.

Baker D, Sali A. 2001. Protein structure prediction and structural genomics. Science 294:93–96.

Beausoleil SA, Villén J, Gerber SA, Rush J, Gygi SP. 2006. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 10:1285–1292.

Blom N, Gammeltoft S, Brunak S. 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294:1351–1362.

Cao X, Chen Y. 2009. Mitochondria and calcium signaling in embryonic development. Semin Cell Dev Biol 20:337-345.

Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F. 2011. Phospho. ELM: A database of phosphorylation sites-update 2011. Nucleic Acid Res 39:261–267.

Fröjdö S, Vidal H, Pirola L. 2009. Alterations of insulin signaling in type 2 diabetes: A review of the current evidence from humans. Biochim Biophys Acta 1792:83–92.

Hawkins CL, Davies MJ. 2005. The role of aromatic amino acid oxidation, protein unfolding, and aggregation in the hypobromous acid-induced inactivation of trypsin inhibitor and lysozyme. Chem Res Toxicol 18: 1669–1677.

Hegyi H, Gerstein M. 1999. The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. J Mol Biol 288:147–164.

Hjerrild M, Stensballe A, Rasmussen TE, Kofoed CB, Blom N, Sicheritz-Ponten T, Larsen MR, Brunak S, Jensen ON, Gammeltoft S. 2004. Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. J Proteome Res 3: 426–433.

Hüttemann M, Pecina P, Rainbolt M, Sanderson TH, Kagan VE, Samavati L, Doan JW, Lee I. 2011. The multiple functions of cytochrome c and their regulation in life and death decisions of the mammalian cell: From respiration to apoptosis. Mitochondrion 3:369–381.

Iakoucheva LM, Radivojac P, Brown CJ, O′Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acid Res 32:1037–1049.

Ingrell CR, Miller ML, Jensen ON, Blom N. 2007. NetPhosYeast: Prediction of protein phosphorylation sites in yeast. Bioinformatics 23:895–897.

Iqbal Z, Hoessli DC, Kaleem A, Munir J, Saleem M, Afzal I, Shakoori AR, Nasir-Ud-Din. 2013. Influence of the sequence environment and properties of neighboring amino acids on amino-acetylation: Relevance for structure-function analysis. J Cell Biochem 4:874–887.

Jeffery CJ. 1999. Moonlighting proteins. Trends Biochem Sci 24:8–11.

Kitchen J, Saunders RE, Warwicker J. 2008. Charge environments around phosphorylation sites in proteins. BMC Struct Biol 8:19.

Kunkel MT, Ni Q, Tsien RY, Zhang J, Newton AC. 2005. Spatio-temporal dynamics of protein kinase B/Akt signaling revealed by a genetically encoded fluorescent reporter. J Biol Chem 7:5581–5587.

Li S, Iakoucheva LM, Mooney SD, Radivojac P. 2010. Loss of post-translational modification sites in disease. Pac Symp Biocomput 15: 337–347.

Martinez-Fleites C, He Y, Davies GJ. 2010. Structural analyses of enzymes involved in the O-GlcNAc modification. Biochim Biophys Acta 1800:122–133.

Miller ML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I. 2008. NetPhosBac – A predictor for Ser/Thr phosphorylation sites in bacterial proteins. Proteomics 9:116–125.

Moritz RL, Reid GE, Ward LD, Simpson RJ. 1994. Capillary HPLC: A method for protein isolation and peptide mapping. Method 6:213–226.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH – A hierarchic classification of protein domain structures. Structure 5:1093–1109.

Qazi WM, Ahmed M, Hoessli DC, Ahmad I, Khawaja I, Wajahat T, Kaleem A, Walker-Nasir E, Rahman N, Shakoori AR, Nasir-Ud-Din. 2006. Consensus sequences as targets for phosphorylation of amino acids in phosphoproteins: Statistical computing analysis. Pak J Zool 38:55–63.

Rexach JE, Clark PM, Hsieh-Wilson LC. 2008. Chemical approaches to understanding O-GlcNAc glycosylation in the brain. Nat Chem Biol 4: 97–106.

Rucker RB, McGee C. 1993. Chemical modifications of proteins in vivo: Selected examples important to cellular regulation. J Nutr 123:977–990.

Savasere A, Omiecinski E, Navathe S. 1995. An efficient algorithm for mining association rules in large databases. Proc. of the 21st VLDB Conference, Zurich, Switzerland, 1995, 432–444.

Sennels L, Salek M, Lomas L, Boschetti E, Righetti PG, Rappsilber J. 2007. Proteomic analysis of human blood serum using peptide library beads. J Proteome Res 6:4055–4062.

Slawson C, Hart GW. 2003. Dynamic interplay between O- GlcNAc and O-phosphate: The sweet side of protein regulation. Curr Opin Struct Biol 13:631–636.

Takahashi H, Nakanishi T, Kami K, Arata Y, Shimada I. 2000. A novel NMR method for determining the interfaces of large protein–protein complexes. Nat Struct Biol 7:220-223.

Tatárová Z, Brábek J, Rösel D, Novotný M. 2012. SH3 domain tyrosine phosphorylation–Sites, role and evolution. PLoS One 7:e 36310.

Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, Cantley LC. 2001. A motif-based profile scanning approach for genome wide prediction of signaling pathways. Nat Biotechnol 19:348–353.

Yi Z, Langlais P, De Filippis EA, Luo M, Flynn CR, Schroeder S, Weintraub ST, Mapes R, Mandarino LJ. 2007. Global assessment of regulation of phosphorylation of insulin receptor substrate-1 by insulin in vivo in human muscle. Diabetes 6:1508–1516.

Zhang DW, Gu HM, Situ D, Haimeur A, Cole SP, Deeley RG. 2003. Functional importance of polar and charged amino acid residues in transmembrane helix 14 of multidrug resistance protein 1 (MRP1/ABCC1): Identification of an aspartate residue critical for conversion from a high to low affinity substrate binding state. J Biol Chem 278:46052–46063.

Zhu G, Fujii K, Belkina N, Liu Y, James M, Herrero J, Shaw S. 2005. Exceptional disfavor for proline at the P +1 position among AGC and CAMK kinases establishes reciprocal specificity between them and the proline-directed kinases. J Biol Chem 280:10743–10748.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.