Journal of Cellular Biochemistry 104:1220–1231 (2008)

MAPRes: An Efficient Method to Analyze Protein Sequence Around Post-Translational Modification Sites

Ishtiaq Ahmad,[1] Daniel C. Hoessli,[2] Wajahat M. Qazi,[1] Ahmed Khurshid,[1] Abid Mehmood,[1] Evelyne Walker-Nasir,[1] Munir Ahmad,[3] Abdul R. Shakoori,[4] and Nasir-ud-Din[1,5]*

[1]Institute of Molecular Sciences and Bioinformatics, Lahore, Pakistan
[2]Department of Pathology and Immunology, Centre Médical Universitaire, Geneva, Switzerland
[3]National College of Business Administration & Economics, Lahore, Pakistan
[4]School of Biological Sciences, University of the Punjab, New Campus, Lahore 54590, Pakistan
[5]HEJ Research Institute of Chemistry, University of Karachi, Karachi, Pakistan

Abstract Functional switches are often regulated by dynamic protein modifications. Assessing protein functions, in vivo, and their functional switches remains still a great challenge in this age of development. An alternative methodology based on in silico procedures may facilitate assessing the multifunctionality of proteins and, in addition, allow predicting functions of those proteins that exhibit their functionality through transitory modifications. Extensive research is ongoing to predict the sequence of protein modification sites and analyze their dynamic nature. This study reports the analysis performed on phosphorylation, Phospho.ELM (version 3.0) and glycosylation, OGlycBase (version 6.0) data for mining association patterns utilizing a newly developed algorithm, MAPRes. This method, MAPRes (Mining Association Patterns among preferred amino acid residues in the vicinity of amino acids targeted for post-translational modifications), is based on mining association among significantly preferred amino acids of neighboring sequence environment and modification sites themselves. Association patterns arrived at by association pattern/rule mining were in significant conformity with the results of different approaches. However, attempts to analyze substrate sequence environment of phosphorylation sites catalyzed for Tyr kinases and the sequence data for *O*-GlcNAc modification were not successful, due to the limited data available. Using the MAPRes algorithm for developing an association among PTM site with its vicinal amino acids is a valid method with many potential uses: this is indeed the first method ever to apply the association pattern mining technique to protein post-translational modification data. J. Cell. Biochem. 104: 1220–1231, 2008.

Key words: post-translational modifications of proteins (PTMs); phosphorylation; glycosylation; association pattern mining; vicinal sequence analysis of PTM sites

The information flow from Gene (DNA) to mRNA and protein is made possible by the coordination of complex biological processes. However, the functional versatility of proteins is in a large part ensured by various types of co- and post-translational modifications (PTMs). Protein PTMs such as phosphorylation, acetylation, glycosylation and methylation are involved in regulating various cellular activities. Almost every type of cellular function involves protein(s) and most often modulation of one function to another is regulated by one or more PTM type such as; enzyme activation [Konstantinopoulos et al., 2007], interaction of protein with protein [Baisse et al., 2007], of protein with membrane [Macher and Yen, 2007] and of protein with various matrices [Reviewed in Hynes, 2007]. The associative potential and functions of proteins depend on their precise 3D structure and the presence of specific modifications [Bork et al., 1998; Attwood, 2000]. Every one of the PTMs described to date is a result of complex biochemical steps and encompasses such a wide range of chemical properties that cannot be characterized by a single biochemical technique. Additionally, determination and regulation of PTMs in vitro hardly follow the same rules as in vivo, where the chemical environment is considerably more complex. Wet-lab approaches to face such challenges could be conveniently

*Correspondence to: Nasir-ud-Din, Institute of Molecular Sciences and Bioinformatics, 28 Nisbet Road, Lahore, Pakistan. E-mail: professor_nasir@yahoo.com

Received 26 November 2007; Accepted 17 December 2007
DOI 10.1002/jcb.21699

complemented by informatics and computational tools. The computational tools for predicting and the databases for searching possible patterns of PTM site(s) will eventually be helpful in formulating the basis of experimental investigations for protein(s) under study and avoid wastage of time. Such computational methods need to be highly specific and generate a minimal number of false positive predictions. Computational methods present a number of prominent achievements in predicting the PTMs, by applying different searching algorithms for matching the unique sequence motif(s) from databases [Yaffe et al., 2001; Falquet et al., 2002; Obenauer et al., 2003], hidden markov models (HMM) [Huang et al., 2005; Senawongse et al., 2005], data mining [Oyama et al., 2002; Creighton and Hanash, 2003; Ji and Tan, 2004; Georgii et al., 2005] and artificial neural networks (ANN) [Hansen et al., 1995; Blom et al., 1999; Iakoucaheva et al., 2004; Julenius et al., 2005] with different levels of accuracy.

This study utilizes MAPRes [Ahmad et al., unpublished work] to analyze the neighboring sequence patterns preferred for phosphorylation and for GalNAc modification sites and provides the comparison and validation of the association patterns mining results by MAPRes algorithm with those of the other available methods. MAPRes [Ahmad et al., unpublished work] mines association patterns among modification site(s) and statistically preferred amino acid residues in their neighboring environment. Our analysis data suggest that specific modification of a protein is often associated with a general type(s) of sequence pattern(s) around modification sites, and along with specific amino acids related to enzyme binding preferences that catalyze the process of modification.

MATERIALS AND METHODS

MAPRes was used to analyze phosphorylation sites from Phospho.ELM [version 3.0 by Diella et al., 2004] database and *O*-GalNAc glycosylation sites from OGlycBase 6.0 (www.cbs.dtu.dk/database/oglycbase/oglyc.base.html). The detailed methodology of MAPRes can be downloaded freely at URL www.imsb.edu.pk/mapres.

Data Preprocessing

Phospho.ELM contained information about 4,252 phosphorylation sites from 1,663 proteins. To ensure the consistency of information drawn from Phospho.ELM, ambiguities in the data were removed. This resulted in removal of redundant entries, misinformation between phosphorylated site and sequence, length of sequence, etc. This data processing resulted in removal of 18 proteins. Another entry for His (N-linked) phosphorylation was also removed. Thus, the data cleaning process resulted in 4,026 *O*-linked phosphorylation sites at Ser, Thr and Tyr acceptors drawn from 1,645 proteins (Table I).

OGlycBase is a database that contains information about glycosylated sites in glycoproteins. The glycosylation information in this database includes different glycosylating oligosaccharides on Ser, Thr, Asn and Asp. Only the proteins with oligosaccharides starting from GalNAc (Mucin type glycosylation) were selected for analysis. The information consistency and redundancy was also checked for all GalNAc modification sites. All ambiguous and redundant entries were removed including the sites in tandem repeats. This refining of the data resulted in 115 proteins with total 606 GalNAc modification sites (Table I).

Dataset Creation and Grouping

MAPRes requires protein and peptide datasets for each modification sites in order to perform analysis. Therefore in these analyses, peptide datasets and protein datasets were automatically generated and grouped accordingly by the MAPRes. Size of peptides in these analyses was set to 21 amino acids (taking modification site at zero position with 10 amino

TABLE I. Statistics of the Data Analyzed by MAPRes

Modification type	Amino acid	Number of peptides	Number of proteins	Total number of proteins
Phosphorylation	S	2,734	1,221	1,645
	T	634	452	
	Y	838	361	
O-GalNAc modification	S	170	54	115
	T	606	99	

acids on right and 10 on left side of modification site). Statistical details of all datasets generated from Phospho.ELM and OGlycBase are given in Table I.

Preference Estimation and Mining Association Patterns

In the data analyses of both modification types, MAPRes was applied to perform preference estimation followed by association patterns mining. Preference estimation was performed using peptide datasets and protein datasets generated from phosphorylation sites (Ser, Thr and Tyr) and *O*-GalNAc modification sites (Ser, Thr).

The preferred sites corresponding to Ser, Thr and Tyr in phosphorylation data analysis were used to mine association rules with different support values set to 5%, 10%, and 15%. Similarly, the association patterns mining was performed for *O*-GalNAc modification sites data for Ser and Thr on different support levels.

Validity of Association Rules

The association patterns, mined by MAPRes from phosphorylation and glycosylation data were compared with the results of earlier studies and existing computational methods. Firstly, the patterns mined by MAPRes were compared with the results reported earlier through a detailed literature search for any such results. Secondly, to find the level of conformity of the patterns mined by MAPRes with the results of existing computational methods we selected 30 proteins randomly belonging to different categories without any prior information about their specific phosphorylation and glycosylation sites from SwissProt [Boeckmann et al., 2003].

Comparison of Association Patterns Mined by MAPRes for Phosphorylation

Comparison of association patterns mined for phosphorylation sites using the sequence data of 30 randomly selected proteins was performed by first predicting the potential phosphorylation sites utilizing NetPhos2.0 [Blom et al., 1999], and DISPHOS 1.3 [Iakoucaheva et al., 2004]. The peptides with positive predictions by the two methods were stored as separate datasets. Then the peptides that contained the association patterns mined by MAPRes were searched and counted.

The data sources of the two methods NetPhos 2.0 and DISPHOS 1.3 were two different versions of Phosphobase and the data analyzed by MAPRes was also the same Phospho.ELM [Diella et al., 2004] utilized by DISPHOS 1.3. Therefore two other comparisons were also performed for the phosphorylation patterns, mined by MAPRes that utilized different data sources. We scanned all 30 protein sequences for phosphorylation motifs through *Scansite 2.0* [Obenauer et al., 2003] and then *Prosite* [Falquet et al., 2002] available on the server *Motif Scan* of *MyHits* (http://myhits.isb-sib.ch/cgi-bin/motif_scan). All the peptides with phosphorylation motifs predicted by Scansite and *Prosite* were stored as separate datasets. Again the peptides that contained the association rules were searched and counted.

Comparison of Association Patterns for *O*-GalNAc Modification

Two prediction methods, NetOGlyc, version 3.1 [Julenius et al., 2005] and Oglyc [Li et al., 2006], for predicting the potential of GalNAc modification on Ser/Thr were utilized to find the level of validation of the association patterns mined by MAPRes. The method NetOGlyc, version 3.1 [Julenius et al., 2005] had the same training data source (i.e., OGlycBase) analyzed by MAPRes but the prediction method Oglyc [Li et al., 2006] utilized the training data from SwissProt [Boeckmann et al., 2003]. Then peptides with positive prediction for *O*-GalNAc modification by the two methods were stored as separate datasets and the peptides among them containing association patterns mined by MAPRes were searched and counted.

RESULTS

The association analysis of both phosphorylation and glycosylation data by MAPRes resulted in various association patterns for phosphorylation (Table II) and *O*-GalNAc modification (Table III).

Phospho.ELM Analysis Results

The total rules generated for Ser, Thr and Tyr, without filtering these rules against any confidence level, were 117. But there were some identical patterns/rules mined at different support levels. Therefore, these identical patterns mined at higher support value become a

TABLE II. Association Patterns Mined by MAPRes for Ser/Thr/Tyr Phosphorylation Data

For Ser phosporylation data			For Thr phosporylation data			For Tyr phosporylation data			For Tyr phosporylation data		
Association patterns	Confidence	Support	Association patterns	Confidence	Support	Association patterns	Confidence	Support	Association patterns	Confidence	Support
<P,1>	78.19	15	<P,1>	21.81	15	<P,9>	37.32	15	<D,-2>	100.00	5
<R,−3>	85.67	15	<P,2 >	30.37	15	<D,−3>	100.00	10	<S,−8>	22.45	5
<S,−4>	69.85	15	<R,−3>	14.33	15	<D,−2>	100.00	10	<G,−10>	100.00	5
<S,−2>	84.57	15	<P,1><P,2>	38.24	10	<D,−1>	100.00	10	<Q,4>	100.00	5
<E,2>	100.00	10	<P,1><S,4>	22.41	5	<E,−4>	100.00	10	<R,−8>	25.36	5
<E,3>	100.00	10	<P,1><S,8>	26.59	5	<E,−3>	100.00	10	<R,−7>	30.11	5
<E,5>	100.00	10	<S,−4><P,1>	22.60	5	<E,−2>	100.00	10	<R,4>	100.00	5
<P,1>	78.19	10	<S,−9><P,1>	20.47	5	<E,−1>	100.00	10	<P,9>	37.32	5
<R,−3>	85.67	10	<P,1><P,5>	23.87	5	<E,1>	100.00	10	<R,7>	27.82	5
<R,−2>	86.01	10	<P,1><P,4>	100.00	5	<L,3>	100.00	10	<P,8>	27.51	5
<S,−10>	100.00	10	<P,−2><P,1>	24.56	5	<P,3>	33.42	10	<S,−7>	22.46	5
<S,−9>	81.61	10	<P,−8><P,1>	27.78	5	<P,8>	27.51	10	<S,−6>	23.02	5
<S,−8>	77.55	10	<G,3><P,10>	54.17	5	<P,9>	37.32	10	<S,−4>	15.56	5
<S,−7>	77.54	10	<A,9><P,10>	100.00	5	<P,10>	21.63	10	<V,−1>	100.00	5
<S,−6>	76.98	10	<P,5><A,9>	100.00	5	<R,−7>	30.11	10	<V,1>	100.00	5
<S,−5>	100.00	10	<G,3><A,9>	100.00	5	<S,−8>	22.45	10	<V,3>	100.00	5
<S,−4>	69.85	10	<P,1><P,2>	38.24	5	<S,−7>	22.46	10	<R,6>	19.49	5
<S,−3>	100.00	10	<P,5><P,10>	43.84	5	<S,−6>	23.02	10	<N,−3>	100.00	5
<S,−2>	84.57	10				<S,−4>	15.56	10	<Y,6>	100.00	5
<S,−1>	100.00	10				<V,1>	100.00	10	<G,5>	100.00	5
<S,2>	100.00	10				<V,3>	100.00	10	<I,3>	100.00	5
<S,3>	100.00	10				<D,−4>	100.00	5	<K,7>	100.00	5
<S,4>	83.87	10				<E,1>	100.00	5	<L,3>	100.00	5
<S,5>	100.00	10				<E,−1>	100.00	5	<P,10>	21.63	5
<S,6>	100.00	10				<E,−2>	100.00	5	<N,−4>	100.00	5
<S,7>	100.00	10				<E,−3>	100.00	5	<G,−4>	100.00	5
<S,8>	79.95	10				<E,−4>	100.00	5	<N,−2>	100.00	5
<S,9>	100.00	10				<E,−5>	100.00	5	<N,2>	100.00	5
<S,10>	100.00	10				<D,2>	23.85	5	<P,−10>	27.05	5
<S,−4><P,1>	77.40	5				<D,−1>	100.00	5	<P,−2>	18.43	5
<S,−2><P,1>	84.69	5				<D,−3>	100.00	5	<P,3>	33.42	5
<P,−2><P,1>	75.44	5				<D,−6>	56.36	5	<P,5>	18.80	5
						<D,−7>	100.00	5	<M,3>	100.00	5
							100.00	5			

TABLE III. Association Patterns Mined by MAPRes for GalNAc Modification on Ser/Thr

Association patterns	Modifying residue	Confidence	Support %
<P,3>=>S	S	19.67	25
<P,−1>=>S	S	28.23	20
<P,3>=>S	S	19.67	20
<P,4>=>S	S	100.00	20
<P,−6>=>S	S	28.07	15
<P,−3>=>S	S	25.83	15
<P,−1>=>S	S	28.23	15
<P,3>=>S	S	19.67	15
<P,4>=>S	S	100.00	15
<P,5>=>S	S	100.00	15
<P,7>=>S	S	20.16	15
<S,−10>=>S	S	37.50	15
<S,−9>=>S	S	28.42	15
<S,−7>=>S	S	100.00	15
<S,−6>=>S	S	100.00	15
<S,−4>=>S	S	100.00	15
<S,−2>=>S	S	100.00	15
<S,2>=>S	S	100.00	15
<S,6>=>S	S	100.00	15
<S,10>=>S	S	100.00	15
<T,−10>=>S	S	100.00	15
<T,−7>=>S	S	22.95	15
<T,1>=>S	S	100.00	15
<T,3>=>S	S	100.00	15
<T,4>=>S	S	100.00	15
<P,−3><P,−1>=>S	S	48.65	10
<A,−4>=>T	T	85.21	25
<P,−5>=>T	T	85.26	25
<P,3>=>T	T	80.33	25
<A,−4><G,−2>=>T	T	100.00	20
<P,−5><A,−4>=>T	T	97.80	20
<A,−4><P,3>=>T	T	94.85	20
<G,−2><P,3>=>T	T	100.00	20
<P,−5><P,3>=>T	T	88.68	20
<P,−5><A,−4><G,−2><P,3>=>T	T	100.00	15
<A,−10><A,−4><A,1><A,4><A,10>=>T	T	100.00	10
<G,−10><P,−6><A,−4><P,3><D,4>=>T	T	100.00	10
<T,−8><A,−4><H,−3><V,−1><R,6>=>T	T	100.00	10

subset of those mined at lower support value and a union of all patterns mined at different support levels, dropping the identical patterns, resulted in 91 unique patterns. MAPRes analyses results for Ser, Thr and Tyr are given below separately.

Serine. In the vicinity of phosphorylated Ser, 70 significantly preferred sites were found. For the observed frequency, five significantly preferred sites of major importance were found. Distribution of s-preferred and non-preferred sites in different observed frequency range has been summarized in Table IV.

A total of 32 association patterns were mined by MAPRes on 5%, 10%, and 15% support level (Table II) and after removing the identical patterns mined at different support levels total unique patterns mined were 28. The confidence range of association patterns for Ser was from 69.8% to 100% (Table II). Minimum confidence (69.8%) was found for Ser at −4 position on 15% support level (Table II). Pro at position +1 was the most frequently occurring residue among the association patterns mined for Ser phosphorylation sites (Table II).

Threonine. Around phosphorylated Thr, 35 significantly preferred sites were found; five sites of major importance were noted for observed frequency. The distribution, of s-preferred and non-preferred sites, in different observed frequency range is given in Table IV.

Association patterns mined, for Thr phosphorylation sites, were 18 in total on 5%, 10%, and 15% support level (Table II). Removing the identical patterns at different support levels resulted in 17 unique association patterns for Thr phosphorylation sites. The confidence range for association patterns mined for phosphorylated Thr was from 14.32% to 100% (Table II). Out of these 18 association rules, four had 100% confidence value with 5% support

TABLE IV. Distribution of S-Preferred and Non-Preferred Sites for Ser, Thr and Tyr Phosphorylation

	Observed frequency	S-preferred site	Non-preferred sites
Phosphorylated Ser	≥9.14 and ≤33.43	30	0
	≥4.069 and ≤9.14	39	150
	≥0 and <4.069	0	200
Phosphorylated Thr	≥11.51 and ≤ 40.22	14	0
	≥3.62 and <11.51	20	212
	≥0 and <3.62	0	173
Phosphorylated Tyr	≥9.90 and ≤ 15.63	21	0
	≥3.57 and < 9.90	30	207
	≥0 and <3.57	0	161

level (Table II). Similar to phosphorylated Ser, the most often occurring residue for phosphorylated Thr was Pro at +1 position. Eleven association rules were found containing Pro at +1 position (Table II). The confidence level of association patterns containing Pro at +1 was from 20.46% to 100%.

Tyrosine. For phosphorylated Tyr, 52 significantly preferred sites were found. With respect to the observed frequency, five sites of major importance were found. Sites with observed frequency greater or equal to 9.90% were all significantly preferred, but there were some other sites which were found to be significantly preferred with an observed frequency lower than 9.90% for instance Met at position −5 was significantly preferred with an observed frequency of 3.579%. Association patterns for phosphorylated Tyr were 67 mined by MAPRes on 5%, 10%, and 15% support level. There were 21 identical patterns that were mined at different support levels. Thus 46 unique patterns were derived from all three support levels dropping the identical patterns. The range of confidence level of association patterns for Tyr phosphorylation sites was from 15.55% to 100%. Out of the 67 association patterns/rules, 40 had 100% confidence with support level 5%, 10%, and 15%. Minimum confidence (15.55%) was for Ser at position −4 (S,-4 => Y) on 5% support level. The observed frequency of Ser at position −4 was 11.45%, the highest value amongst all significantly preferred sites for Ser modification by phosphate.

In summary, Pro at position +1 and Arg at position -3 are the most important residues in the vicinity of phosphorylayted Ser and/or Thr (Table II, Fig. 1a). Additionally, Ser at all positions (−10 to +10) is moderately important for phosphorylated Ser (Table II, Fig. 1a). Whereas Ser, for phosphorylated Thr, is also moderately important at many positions on right and left (Table II, Fig. 1a). Acidic residue like Glu also seems to be important moderately at various positions (most importantly on positions +2, +3 and +5) around phosphorylated Ser (Table II, Fig. 1a). But, for phosphorylated Thr, another important residue at various positions is Pro that is mined in many association patterns for Thr (Table II) and is also depicted in Fig. 1a. Importance of Pro around phosphorylated Tyr is apparent only on +3 and +9 positions. Carboxyl group containing amino acids like Asp and Glu at various positions in close vicinity of phosphorylated Tyr have been the most important amino acids (Table II, Fig. 1a). Other moderately important amino acid residue for Tyr phosphorylation emerged is Val at +1, +3 and −1 positions (Table II, Fig. 1a). Similarly, Leu at +3 is also moderately important for Tyr phosphorylation. Ser at various positions (mostly at distant positions) is also apparent to play an important role in Tyr phosphorylation.

OGlycBase Analysis Results

MAPRes was used to analyze *O*-glycosylated sites (Ser and Thr). Analysis resulted in 38 association patterns at 10%, 15%, 20%, and 25% support levels. These patterns contained identical elements, when mined at different support levels. Removal of these identical elements resulted in 34 unique patterns for GalNAc modification of Ser and Thr.

Serine. Preference estimation of the amino acids in the vicinity of *O*-GalNAc modified Ser resulted in 35 significantly preferred sites. Association patterns mining yielded a total of 26 patterns for *O*-GalNAc modified Ser, which, upon removal of identical elements became 22 unique patterns for *O*-GalNAc modified Ser

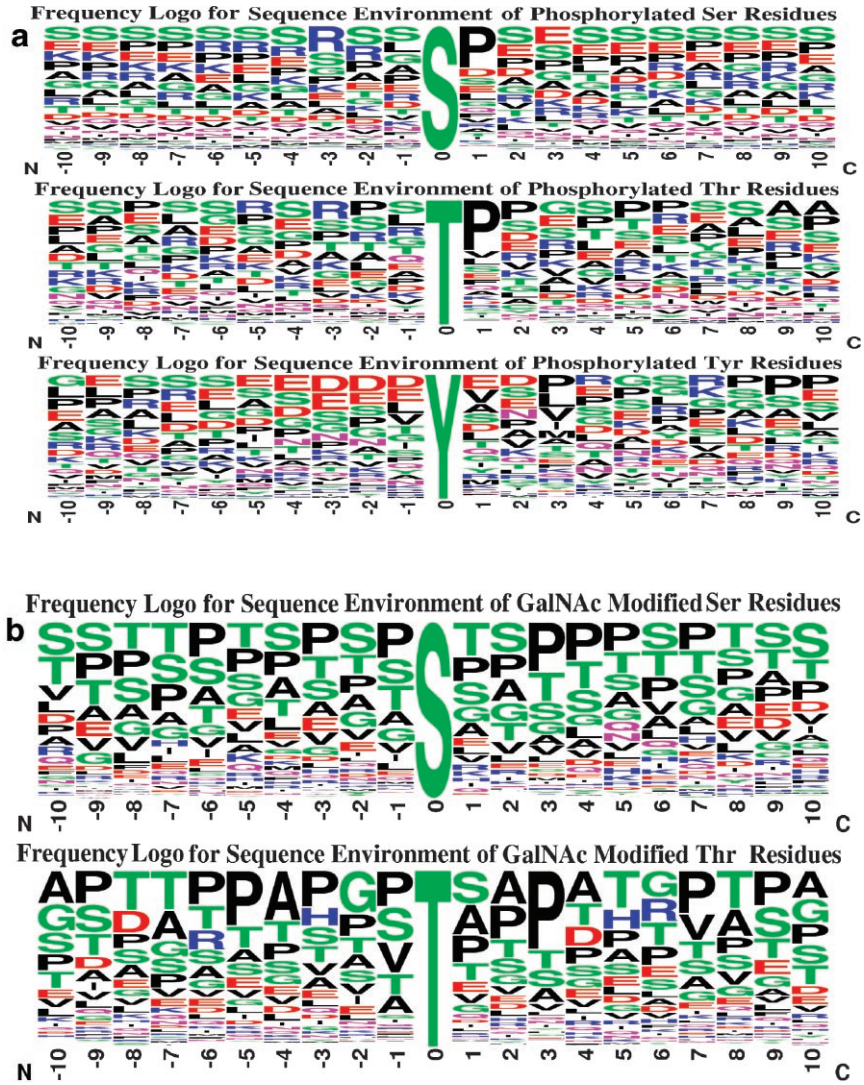

Fig. 1. Sequence logo for every 20 amino acids at each position around phosphorylated/glycosylated amino acids showing frequency of each amino acid around (−10 to +10) modified Ser/Thr/Tyr. The frequency of amino acids has been shown in bits. The size of amino acid at any position shows the exact frequency of an amino acid at a specific position. **a**: Shows the sequence logo of all the peptides containing phosphorylated Ser, Thr and Tyr that were extracted from Phospho.ELM (version 3) before analyzing by MAPRes. It is apparent that non-polar and neutral amino acids are the most frequent residues in close vicinity of phosphorylated Ser/Thr whereas polar, basic and acidic amino acids are less frequent in close vicinity, which is same as the results of MAPRes for phosphorylated Ser/Thr described in Table II. In case of Tyr modification the situation is reverse. Here the acidic amino acids are more frequent in close vicinity of modified Tyr and non-polar, neutral polar and basic amino acids are less frequent in the close vicinity but are important at distant positions probably due to the effect of resonating ring structure. Same trend can also be observed for MAPRes results in Table II. **b**: The sequence logo of all peptides containing GalNAc modification sites (Ser/Thr) drawn from OGlycbase 6.0. Here the frequency of all amino acids around modified Ser/Thr by GalNAc show that Pro at +3 position is the most frequent amino acid both for Ser and Thr modification by GalNAc. Similarly, Pro, Ser and Thr are also frequent on various positions around the modification site, but there is difference in Ser and Thr sequence context for acidic and basic amino acids. These acidic and basic amino acids are important for Thr only and not for Ser. Same is also evident from the Table III for patterns of GalNAc mined by MAPRes.

residues. These association patterns are summarized in Table III. It was observed that the confidence level range of the mined association patterns was 5.79–100%.

Threonine. Number of significantly preferred sites found in the neighboring of *O*-GalNAc modified Thr was 67. Association rules mined for *O*-GalNAc modified Thr were 12. These patterns had a confidence range from 80.33% to 100% (Table III).

It is evident that presence of Pro at various positions is very important for Ser and Thr

modification by *O*-GalNAc (Table III, Fig. 1b), and becomes the most important on +3 position both for Ser and Thr and moderately important on −1 and −3 positions for Ser which are consistent with the previous findings [Christlet and Veluraja, 2001]. Similarly Gly at −2 and Val at −1 seem to be moderately important for Thr modification by GalNAc (Table III, Fig. 1b) and are consistent with other earlier findings [Elhammer et al., 1993; Hansen et al., 1995]. Ser and Thr at different positions appear to be important both for Ser and Thr modification by *O*-GalNAc.

DISCUSSION

Protein phosphorylation is a dynamic PTM that may last for minutes and is reversed rapidly. Phosphorylation of protein is a basic source of functional switches of proteins by changes in activity of proteins due to its dynamic nature. Phosphorylation regulates cellular signal transductional events and other cellular processes like metabolism, proliferation, differentiation and apoptosis [Kolibaba and Druker, 1997]. Substitution of –OH group of Ser/Thr/Tyr by phosphate, catalyzed by different kinases that are the most diverse class of enzymes in human and other organisms, results in temporary conformational changes in proteins that are important for their activation and deactivation.

Prior to the general analysis on phosphorylated Ser, Thr and Tyr reported in this study, phosphorylation sites were analyzed for those kinases which commonly phosphorylate Ser, Thr and Tyr. The Phospho.ELM (version 3.0) was scanned for kinases which commonly phosphorylate Ser, Thr and Tyr. The scan resulted in three kinases, AMPK, MAP2K4 and TGF-β R. Their statistics has been summarized in Table V. Number of Ser/Thr/Tyr phosphorylation sites for each kinase was too low to run the analysis.

TABLE V. Kinases Common to Ser, Thr and Tyr Phosphorylation

	Number of phosphorylation sites		
Kinase	Ser	Thr	Tyr
AMPK	26	3	1
MAP2K4	1	1	1
TGF-beta R, typeII	7	1	3

Similarly, Phospho.ELM was once again scanned to define kinases which were only involved in the phosphorylation of Tyr acceptor. Scan resulted in 70 kinases. An association pattern mining was performed by MAPRes on these 70 kinase datasets. First step of the analysis was to estimate preferred sites of each kinase. Results of preference analysis showed that out of 70 kinases preferred sites for only 11 kinases were established (Table VI). But among these 11 kinases, three including ZAP70, FAK, BTK had only one preferred site (Table VI). These preferred sites and number of peptides were insufficient to perform efficient association analysis. As a consequence of inadequate data for kinase based association mining, the analysis using MAPRes with different support levels was performed for phosphorylation data generally without dividing each of the acceptor S/T/Y for different kinases.

It was observed that with the increase of minimum support level, the length of association pattern is decreased. When the minimum support limit was set to 15%, the length of association patterns was reduced to one amino acid only in the surroundings of phosphorylation sites (Ser, Thr and Tyr). This observation has no affect on the utility and practicality of association patterns relevant to modification. Possible reason for large number of association rules mined for phosphorylation is due to large number of diverse class of kinases involved in phosphorylation of acceptor S/T/Y.

Comparison of association patterns, mined by MAPRes, was performed with different available methods. Firstly the comparison of the mined patterns was made with the available literature. A number of studies had pointed out important residues that take part in

TABLE VI. Statistics of Modification Sites Catalyzed by Tyrosine Kinases

Kinase	Peptides	S-preferred residues
1. SRC	98	10
2. EGFR	50	4
3. IR	37	4
4. SYK	34	8
5. ABL	27	4
6. LYN	22	2
7. LCK	22	2
8. JAK2	21	2
9. ZAP70	12	1
10. FAK	12	1
11. BTK	12	1

the phosphorylation of Ser/Thr/Tyr residues [Yaffe et al., 2001; Iakoucaheva et al., 2004]. These studies reported that phosphorylation sites frequently cluster within functionally important protein domains, and that phosphorylation of protein depends, that is on the distribution of acidic neighboring, hydrophobic and charged amino acids [Iakoucaheva et al., 2004]. It has been documented that phosphorylation of proteins (in mammals and bacteria) by Ser/Thr kinases depends on the presence of positively charged Arg/Lys residues on positions −2, −3 and −5 and negatively charged Glu and Asp spanning −4 to +4 positions [Yaffe et al., 2001]. Met, Val, Ile, Phe and Leu have been reported to be favored at +3 position in mammals [Yaffe et al., 2001], and Pro residues have been documented to be important at various positions (−2, +5, +9) for Tyr phosphorylation [Iakoucaheva et al., 2004]. Additionally, Ser phosphorylation tends to occur when Ser residues are present in neighboring environment that is at −4, +2, +4 [Iakoucaheva et al., 2004] which are in conformity with the association patterns mined by MAPRes (Table II). Similar conformance with previous findings [Iakoucaheva et al., 2004] was also observed for non-preferred sites which were not processed during association patterns mining. Thus this stage of comparison through literature is in agreement with the previous findings.

Then the association patterns searched in the neighboring of positively predicted phosphorylation sites by NetPhos2.0 resulted in a cumulative percentage of conformity equal to 80.42% for all positively predicted S/T/Y acceptor sites. Among S/T/Y acceptor sites the percentage of conformance was the highest (97.92%) for Y and the lowest (58.33%) for T but for S the conformance with the predictions of NetPhos2.0 was in the middle that is 83.15%. During this comparison more than 80% patterns mined by MAPRes (Table VII) were found in the vicinity of positively predicted sites by NetPhos 2.0 in all 30 proteins. The percentage of conformance might be much higher if the two methods, that is Netphos2.0 (data used from phosphobase 2.0 with nine amino acid length of peptides) and MAPRes (data used from Phospho.ELM (version 3.0) with 21 amino acid length of peptides), had used the same data of phosphorylation sites and peptide size.

TABLE VII. Percent Conformity of the Patterns Mined by MAPRes With Other Methods

	Percent conformity of patterns for phosphorylation sites					Percent conformity of patterns for *O*-GalNAc sites					
	Ser	Thr	Tyr	Total	Percentage of conformity		Ser	Thr	Total	Percentage of conformity
Total patterns by MAPRes at different support	32	18	67	117		Total patterns by MAPRes at different support	26	12	38	
Unique (non-redundant) patterns	28	17	46	91		Unique (non-redundant) patterns	22	12	34	
Patterns found in NetPhos2.0 predicted peptides	17	16	40	73	80.21	Patterns found in NetOGlyc 3.1 predicted peptides	20	11	31	91.18
Patterns found in DISPHOS 1.3 predicted peptides	28	16	37	81	89.01	Patterns found in Oglyc predicted peptides	21	9	30	88.23
Patterns found in ScanSite 2.0 predicted peptides	20	6	36	62	68.13	Average percentage of conformed patterns				89.7
Patterns found in Prosite motif scan searched peptides	28	17	40	85	93.4					
Average percentage of conformed patterns					82.68					

Similarly the association patterns were also searched in the neighboring of positively predicted S/T/Y acceptor sites by DISPHOS 1.3 [Iakoucaheva et al., 2004] that resulted in 83.87% conformance. The data source of DISPHOS 1.3 and that for MAPRes analysis was the same. But the results of MAPRes analysis for the other two methods, which used phosphorylation data sources other than Phospho.ELM including *ScanSite* [Obenauer et al., 2003] and *Prosite MotifScan* [Falquet et al., 2002] were also much encouraging. The association patterns, when searched in the vicinity of Scansite phosphosites predictions of all 30 protein sequences, resulted in more than 76% conformity. For 30 protein sequences total 55 sites were predicted as possible phosphosites at high sensitivity and out of the 55 phosphopeptides containing predicted phosphosite, 42 were with at least one association pattern in the vicinity of phosphosites. The phosphopeptides predicted on medium and low sensitivity were much more in number. Inclusion of these phosphopeptides resulted in an increased percentage of conformity (more than 80%) with the patterns mined by MAPRes. Similarly, association patterns' search in the vicinity of all the positive sites for Ser, Thr, and Tyr phosphorylation searched by *Prosite MotiScan* at *MyHits* resulted in total conformity of 79.28% (total 502 sites in number and amongst them 398 sites contained one or more pattern in their vicinity mined by MAPRes).

Cumulative comparison data from all methods for all 30 proteins showed that the most important amino acid in the vicinity of phosphorylated Ser/Thr is Pro at +1 position that is evident as 31.12% for Ser and 33.33% for Thr of the phosphorylation prediction data compared with all other methods contained the association pattern with Pro at +1 position (Fig. 2a,

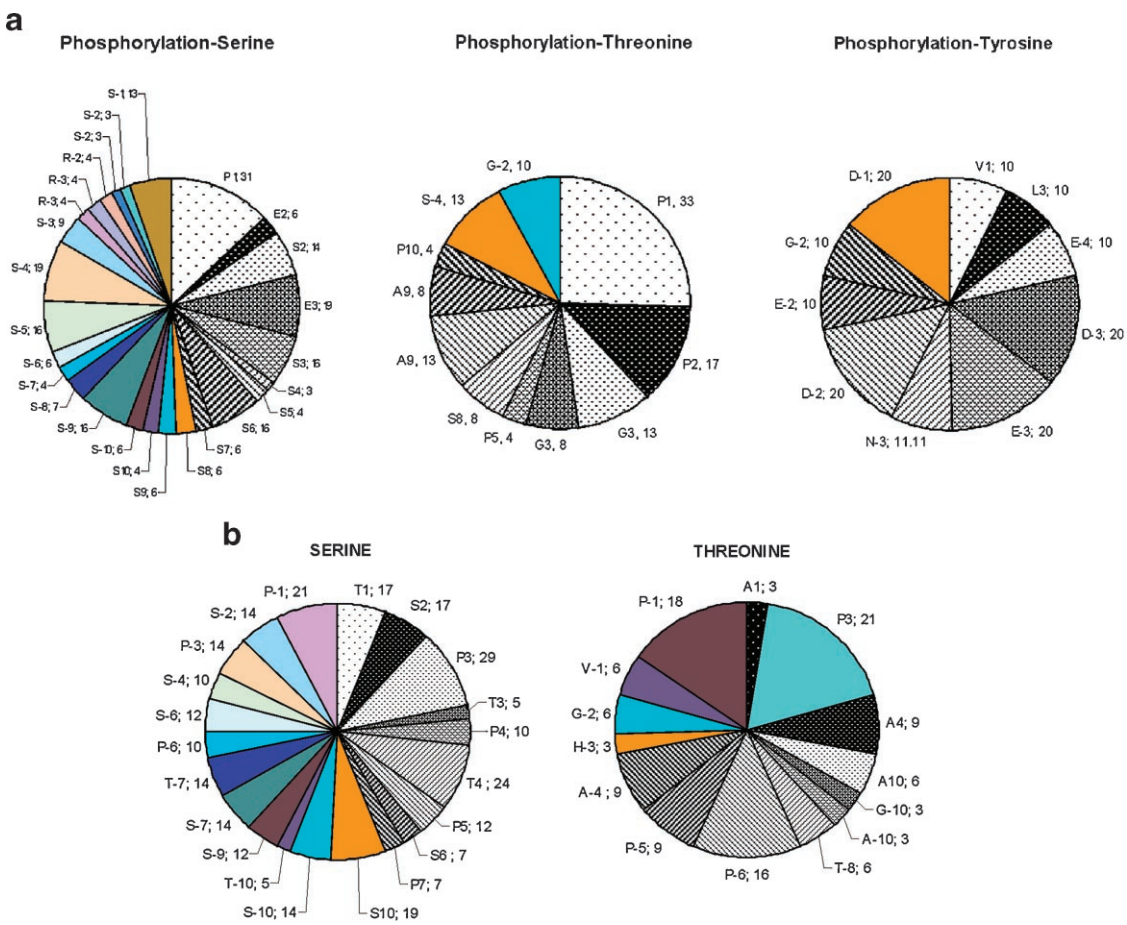

Fig. 2. Percentage values of the validated data of all 30 proteins by different methods for MAPRes patterns are also in agreement with the Fig. 1, and with data in Tables II and III of MAPRes results. **a**: Validation data information for MAPRes patterns for Phosphorylated Ser, Thr and Tyr. **b**: Validation data information for MAPRes patterns for GalNAc modified Ser and Thr.

Table VII). Besides Pro some acidic amino acids such as Glu (with 18.84% validation data containing Glu at position +3 for Ser phosphorylation that is lesser for Thr phosphorylation) also seem to be important. Whereas, Asp seems to be less important as percent validation for Asp data at any position around phosphorylated Ser/Thr was much low. Serine is the amino acid that is important at every position (−10 to +10) around phosphorylated Ser and Thr (Table VII, Fig. 1a and Fig.2b). In case of Tyr phosphorylation, acidic amino acids Glu and Asp in the close vicinity are very important (Table VII, Fig. 1a). These observations suggest that most of the kinases prefer Pro in close vicinity for phosphorylating Ser/Thr in combination either with –OH group of other Ser, –COOH group of Glu, or $-NH_2$ group of Arg in its vicinity. In contrast, most of the Tyr phosphorylating kinases require –COOH group of Glu and Asp, in combination with non-polar aliphatic group containing amino acids such as Leu, Val, Ala, Gly, in the vicinity of Tyr for its phosphorylation.

Glycosylation is another important post-translational modification event. MAPRes algorithm mined 38 association patterns at support levels 10%, 15%, 20%, and 25% for the *O*-GalNAc data. This is somewhat unexpected as there is only one enzyme, *O*-GalNAc-T (*O*-GalNAc transferase) catalyzing the transfer of *O*-GalNAc to –OH group of Ser/Thr. However, this diversity in association patterns mined by MAPRes can be justified by the presence of 24 genes of *O*-GalNAc-T and at least 15 mammalian isoform members of GalNAc-T family [reviewed in Ten Hagen et al., 2003]. Thus different isoforms of *O*-GalNac-T may have different catalyzing preferences to same or different acceptor substrate. It has also been observed that out of 38 association patterns only 14 of them had confidence value of 100% at support 25%. Comparison of the patterns mined for *O*-GalNAc modification sites with the neighboring environment of positive prediction sites by NetOGlyc 3.1 showed 72% conformance while the conformance with the prediction results of the same 30 proteins by *Oglyc* was approximately equal to 80% (Table VII). Conformance with the literature search was also encouraging.

The amino acids that have been proposed to be highly preferred for the glycosylation process are Pro, Thr, Ser, and Ala [Christlet and Veluraja, 2001; Julenius et al., 2005] and the same pattern can be observed among the patterns mined by MAPRes for *O*-GalNAc data (Table III). Additionally, limited influence of Glu, Val, Gly, Met, Ile, Gln, Trp, Asp, Arg, Phe, Tyr, Lys, Cys, Asn, and Leu has also been observed [Julenius et al., 2005]. Preference and association analysis of *O*-GalNAc modification sites by MAPRes suggests that Glu, Asp, Val, Gly, Arg, Leu, His in the vicinity of *O*-GalNAc have significant bearing on *O*-glycosylation (Table III, Fig. 2b). In case of Ser modification by GalNAc, the sequence environment is different form that of Thr, as described earlier [Hansen et al., 1995]. It is clear from Table VII and Fig. 1a and Fig.2b that Pro at position +3 is the most important residue for GalNAc modification for both Ser and Thr (Table VII, Fig. 1b) modification by *O*-GalNAc. The difference in GalNAc modification for Ser and Thr lies in the fact that some amino acid residues are favored for Thr and less favored for Ser. These include basic amino acids Arg, and His, acidic amino acids like Asp and non-polar aliphatic group-containing amino acids Val, Ala and Gly (Table VII, Fig. 1b). This implies that the isoforms of GalNAc-T that glycosylate Ser by *O*-GalNAc require Pro in combination with –OH group of Ser or Thr in the vicinity, but the isoforms of GalNAc-T that glycosylate Thr by *O*-GalNAc require Pro in combination with $-NH_2$ group of Arg and His and or –COOH of Asp along with one or more non-polar aliphatic side chain-containing amino acids.

Attempt was also made to perform analysis on *O*-GlcNAc modification data, as *O*-GlcNAc modification alternates with phosphorylation at sites that control the functional behavior of many proteins. However, preference estimation analysis of *O*-GlcNAc was not adequate to mine association rules, because of insufficiently available data. Data for the analysis of *O*-GlcNAc modification sites was obtained from OGlycBase and Swiss Prot. In this analysis, 63 *O*-GlcNAc modification sites were used (Ser = 29 sites and Thr = 34 sites). Preference estimation analysis resulted in four significant sites for *O*-GlcNAc modified Ser and 16 significant sites for *O*-GlcNAc modified Thr. Thus the association analysis for this insufficient data was not possible.

These analyses propose that association patterns with Pro, Ser, Glu and Arg play important role in the vicinity of phosphorylated Ser/Thr,

whereas acidic amino acids like Glu and Asp appear to prevail in close vicinity of phosphorylated Tyr (Table II, Fig. 1a and Fig. 2b). Association patterns with Pro, Ser, Thr Val, Gly, and Ala contribute significantly in the vicinity of *O*-GalNAc glycosylation for both Ser and Thr generally but other amino acids such as Arg, His and Asp are also important specifically for Thr. These analysis results by MAPRes provide important insights for developing efficient prediction methods, by combining the approach of learning the sequence window of an experimentally known PTM site with the association patterns mined by MAPRes by presenting to neural networks. Additionally the results of MAPRes analysis for the phospho- and glyco-proteome show a positive and acceptable conformity with the existing methods, indicating, that the algorithm utilized by MAPRes for mining association patterns is an efficient and first method of its kind for analyzing PTM sequence data.

ACKNOWLEDGMENTS

Nasir-ud-Din acknowledges support from Pakistan Academy of Sciences, HEC, Pakistan and EMRO-WHO for this research work.

REFERENCES

Attwood T. 2000. The quest to deduce protein function from sequence: The role of pattern databases. Int J Biochem Cell Biol 32:139–155.

Baisse B, Galisson F, Giraud S, Schapira M, Spertini O. 2007. Evolutionary conservation of P-selectin glycoprotein ligand-1 primary structure and function. BMC Evol Biol 7:166.

Blom N, Gammeltsoft S, Brunak S. 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294:1351–1356.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL. Nucl Acids Res 31:365–370.

Bork P, Dansekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. 1998. Predicting function: From genes to genome and back. J Mol Biol 283:707–725.

Christlet THT, Veluraja K. 2001. Database analysis of *O*-glycosylation sites in proteins. Biophys J 80:952–960.

Creighton C, Hanash S. 2003. Mining gene expression databases for association rules. Bioinformatics 19: 79–86.

Diella F, Cameron F, Gemund C, Linding R, Via A, Kuster B, Sicheritz-Ponten T, Blom N, Gibson T. 2004. Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics 5:79.

Elhammer AP, Poorman RA, Brown E, Maggiora LL, Hoogerheide JG, Kezdy FJ. 1993. The specificity of UDP-GalNAc: Polypeptide N-acetylgalactosaminyltransferase as inferred from a database of *in vivo* substrates and from the *in vitro* glycosylation of proteins and peptides. J Biol Chem 268:10029–10038.

Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A. 2002. The PROSITE database, its status in 2002. Nucl Acids Res 30:235–238.

Georgii E, Richter L, Rückert U, Kramer S. 2005. Analyzing micro array data using quantitative association rules. Bioinformatics 21:ii123–ii129.

Hansen JE, Lund O, Engelbrecht J, Bohr H, Nielsen JO, Hansen JE. 1995. Prediction of *O*-glycosylation of mammalian proteins: Specificity patterns of UDP-GalNAc: Polypeptide *N*-acetylgalactosaminyltransferase. Biochem J 308:801–813.

Huang HD, Lee TY, Tseng SW, Horng JT. 2005. KinasePhos: A web tool for identifying protein kinase-specific phosphorylation sites. Nucl Acids Res 33:W226–W229.

Hynes RO. 2007. Cell-matrix adhesion in vascular development. J Thromb Haemost 1:32–40.

Iakoucaheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. Nucl Acids Res 32:1037–1049.

Ji L, Tan KL. 2004. Mining gene expression data for positive and negative co-regulated gene clusters. Bioinformatics 20:2711–2718.

Julenius K, Mølgaard A, Gupta R, Brunak S. 2005. Prediction, conservation analysis and structural characterization of mammalian mucin-type *O*-glycosylation sites. Glycobiology 15:153–164.

Kolibaba KS, Druker BJ. 1997. Protein tyrosinekinases and cancer. Biochim Biophys Act 1333:F217–F248.

Konstantinopoulos PA, Karamouzis MV, Papavassiliou AG. 2007. Post-translational modifications and regulation of the RAS superfamily of GTPases as anticancer targets. Nat Rev Drug Discov 6:541–555.

Li S, Liu B, Zeng R, Cai Y, Li Y. 2006. Predicting *O*-glycosylation sites in mammalian proteins by using SVMs. Comput Biol Chem 30:203–208.

Macher BA, Yen TY. 2007. Proteins at membrane surfaces-a review of approaches. Mol Biosyst 3:705–713.

Obenauer JC, Cantley LC, Yaffe MB. 2003. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucl Acids Res 31:3635–3641.

Oyama T, Kitano K, Satou K, Ito T. 2002. Extraction of knowledge on protein-protein interaction by association rule discovery. Bioinformatics 18:705–714.

Senawongse P, Dalby AR, Yang ZR. 2005. Predicting the phosphorylation sites using hidden markov models and machine learning methods. J Chem Inf Model 45:1147–1152.

Ten Hagen KG, Fritz TA, Tabak LA. 2003. All in the family: The UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases. Glycobiology 13:1R–16R.

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.