ARTICLE
Journal of Cellular Biochemistry 108:64–74 (2009)

Journal of Cellular Biochemistry

Phosphoproteome Sequence Analysis and Significance: Mining Association Patterns Around Phosphorylation Sites Utilizing MAPRes

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

[1]*Institute of Molecular Sciences and Bioinformatics, Lahore, Pakistan*
[2]*Department of Physics, GC University, Lahore, Pakistan*
[3]*Department of Pathology and Immunology, CMU, University of Genèva, Geneva, Switzerland*
[4]*School of Biological Sciences, University of the Punjab, Quaid-i-Azam Campus, Lahore, Pakistan*
[5]*HEJ Research Institute of Chemistry, University of Karachi, Karachi, Pakistan*

ABSTRACT

Phosphorylation, one of the most common protein post-translational modifications (PTMs) on hydroxyl groups of S/T/Y is catalyzed by kinases and involves the presence or absence of certain amino acid residues in the vicinity of the phosphorylation sites. Using MAPRes, we have analyzed the substrate proteins of Phospho.ELM 7.0 and found that there are both general and specific requirements for the presence or absence of particular amino acids in the vicinity of phosphorylated S/T/Y for both of the phosphorylation data, whether or not kinase information was taken into account. Patterns extracted by MAPRes for kinase-specific data have been utilized to find the consensus sequence motifs for various kinases required to catalyze the process of phosphorylation on S/T/Y. These consensus sequences for different kinase groups, families, and individual members are consistent with those described earlier with some novel consensus reported for the first time. A comparison study for the patterns mined by MAPRes with the results of existing prediction methods was performed by searching for these patterns in the vicinity of phosphorylation sites predicted by different available method. This comparison resulted in 87–98% conformity with the results of the predictions by available methods. Additionally, the patterns mined by MAPRes for substrate sites included 61 kinases, the highest number analyzed so far. J. Cell. Biochem. 108: 64–74, 2009. © 2009 Wiley-Liss, Inc.

KEY WORDS: POST-TRANSLATIONAL MODIFICATIONS OF PROTEINS (PTMS); PHOSPHORYLATION; ASSOCIATION PATTERN MINING; KINASE SUBSTRATES CONSENSUS SEQUENCE; SEQUENCE ANALYSIS OF PHOSPHOPROTEIN

Phosphorylation is a decisive protein modification to carry out and regulate cellular functions. Function modulation by phosphorylation of protein substrates including transcription factors [Ahmad et al., 2006], enzymes [Kaleem et al., 2009], and other proteins [Ahmad et al., 2007] have been investigated in normal and pathological conditions [Hanks, 2003]. The involvement of phosphoproteins in the pathobiology of diseases makes them suitable targets for drug development [Bridges, 2001], such as in the case of type II diabetes [Bridges, 2005] and cancer [Krause and van Etten, 2005]. Drug development targeted against phosphoprotein receptors, enzymes, or signaling proteins requires the complete information regarding the protein's phosphorylation sites and the kinases and phosphatases involved.

Different protein kinases catalyze the transfer of a phosphate group from adenosine triphosphate (ATP) to S, T, or Y residues of substrate protein(s). No single classification for kinases exists that may be accepted universally. One of the most generally accepted classification for kinases is based on the chemical nature of the phosphate acceptor. The primary or secondary OH groups are phosphorylated by S/T kinases (PKA, PKB, and PKC, etc.) and the phenolic group of Y by tyrosine kinases (TKs). There exist other classifications of kinases on the basis of their substrate specificities, amino acid sequence homology and preferred amino acid type in the vicinity of phosphorylation target. A number of studies have shown that protein phosphorylation by a given protein kinase often uses a similar set of amino acids in the vicinity of the phosphorylation site.

Additional Supporting Information may be found in the online version of this article.

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

Received 29 April 2009; Accepted 30 April 2009 • DOI 10.1002/jcb.22220 • © 2009 Wiley-Liss, Inc.
Published online 18 June 2009 in Wiley InterScience (www.interscience.wiley.com).

The concept of consensus sequence actually evolved form such considerations [reviewed in Yeh et al., 2002].

Rapid growth of kinase substrate data for protein phosphorylation in the last decade led to the creation of protein kinase and substrate information databases [Kreegipuu et al., 1999; Diella et al., 2004, 2008; Huang et al., 2005; Lee et al., 2006; Wang et al., 2008]. A statistical analysis of the kinase substrate data, stored in different databases, is useful to develop a biological relationship between sequence, structure, and function of proteins with and without phosphorylation. The preferred substrate of a given kinase is often described as a consensus sequence comprising the S/T or Y phosphorylation sites. However, it is not always possible to define a consensus sequence because of the flexibility in substrate recognition of the kinases. Thus phosphorylation of different proteins can be catalyzed by the same kinase and a given substrate may be phosphorylated by different kinases. To overcome this problem, an alternative approach for describing enzyme–substrate relation is the analysis of the primary sequence of substrate proteins and that of enzymes. One such approach seeks to identify the significantly preferred amino acids in the vicinity of a post-translational modification (PTM) site and then developing a correlation between the significantly preferred amino acids in the vicinity of a PTM site by resorting to the association rule mining technique [Ahmad et al., 2008a]. Besides determining the preferred amino acids in the vicinity of a PTM site, this method establishes whether or not an association between preferred sites and target of phosphorylation actually exists. The patterns mined by MAPRes can be utilized to define a possible consensus among different kinase classification groups. Another iterative statistical approach for extracting, from large-scale data, the protein sequence motifs for phosphorylation has been documented [Schwartz and Gygi, 2005]. The main shortcoming of this approach, however, was that the analysis was restricted to merely extracting the patterns of the substrates for S/T and Y kinases. To overcome this limitation, it is possible to perform an analysis that provides both general and specific patterns for S/T/Y phosphorylation and specifically so for different kinase groups, families, and individual members, by using the MAPRes technique [Ahmad et al., 2008a].

Kinases and or their substrates data has been utilized to develop prediction methods for phosphorylation on S/T/Y without kinase information [Blom et al., 1999, 2004; Iakoucheva et al., 2004] and with kinase information [Obenauer et al., 2003; Kim et al., 2004; Wang et al., 2008] applying one of the machine learning techniques. Hidden Markov Models (HMMs) have also been utilized to develop prediction methods for phosphorylation on S/T/Y by different kinases [Huang et al., 2005]. A comparative study of prediction models based on HMMs with those utilizing different machine learning techniques for predicting the phosphorylation sites showed that the latter provide better prediction accuracy than HMMs [Senawongse et al., 2005]. The prediction accuracy was further improved when machine learning algorithms were based on features extracted by trained HMMs [Senawongse et al., 2005]. Thus, a machine learning algorithm can result in a better prediction accuracy of PTMs (e.g., phosphorylation) when it is based on features extracted with the proper statistical model or when the learning algorithm is trained on the whole PTM data. The approach of MAPRes is also based on the fact that significantly preferred amino acids in the vicinity of PTM sites are utilized for pattern mining/extraction [Ahmad et al., 2008a]. These patterns will be utilized in future for the development and application of an algorithm for training neural networks to achieve maximum prediction accuracy. Previously, an analysis of S/T/Y phosphorylation sites of Phospho.ELM 3.0 [Diella et al., 2004] without kinase information was performed utilizing MAPRes [Ahmad et al., 2008b]. The later version of this database Phospho.ELM 7.0 [Diella et al., 2008] is a phosphorylation sites database that contains 3–4 times more data of experimentally determined protein phosphorylation sites than the previous version (3.0) [Diella et al., 2004]. In this study, phosphorylation data of Phospho.ELM 7.0 [Diella et al., 2008] have been analyzed utilizing MAPRes [Ahmad et al., 2008a] both for all phosphorylated S/T/Y and for different kinase substrate sites. The association patterns mined by MAPRes for all phosphorylated S/T/Y (without kinase information) are consistent with the previous findings with additional patterns compared to the previous analysis [Ahmad et al., 2008b]. Similarly, the patterns mined for kinase-specific phosphorylation sites are also consistent with the previous findings [Obenauer et al., 2003; Blom et al., 2004; Kim et al., 2004; Wang et al., 2008] and also provides novel patterns, to ultimately define the consensus or preferred sequence patterns of substrate required by the different kinases.

MATERIALS AND METHODS

In the present study, two types of analyses were performed. Firstly the sequence patterns were mined generally for all S/T/Y residues, irrespective of the kinase involved, and secondly the sequence patterns were mined for different kinase substrate sites. The analyses by MAPRes included estimation of significantly preferred amino acid residues in the vicinity of the phosphorylated sites, both with and without kinase information. These analyses were performed on peptides of 21 amino acids containing one phosphorylation site (10 amino acids upstream and 10 amino acids downstream the modified S/T/Y) both for all S/T/Y phosphorylation sites and for known substrate sites of specified kinases.

DATASETS PREPARATION

Phospho.ELM is a database of experimentally verified phosphorylation sites in eukaryotic proteins [Diella et al., 2004, 2008]; the whole data of Phospho.ELM 7.0 became available on request from EMBL as a tab-delimited text file. This version contained a data of 4,078 phosphoproteins covering 12,025 S, 2,362 T, and 2,083 Y instances as phosphorylation sites (Table I). These tab-delimited data were cloned in two tables; one for performing general analysis of the sequence environment, irrespective of the kinase involved and the second for analyzing the kinase-specific protein substrate sequences. The table containing the data for general analysis showed a total of 16,470 phosphorylated S/T/Y instances after checking the data for any possible ambiguity or redundancy. This data processing resulted in removal of an entry for histidine phosphorylation. The phospho.ELM 7.0 [Diella et al., 2008] contains a number of instances of phosphorylated S/T/Y without information

TABLE I. Data Summary of Phospho.ELM 7.0 Analyzed

Modified amino acid	S	T	Y	Total
Phosphorylation data without kinase information				
Count	12,025	2,362	2,083	16,470
Kinases information	NI	NI	NI	NI
Phosphorylation data with kinase information				
Count	2,653	800	950	4,402
Kinases information	187	117	77	267

NI, not included in the analysis.

on the kinase. Out of 16,470 phosphorylated S/T/Y, only 3,417 instances of phosphorylated S/T/Y appeared with the kinase information. A total of 269 different kinases were found in the Phospho.ELM 7.0 dataset. However, there were some phosphorylated S/T/Y entries in Phospho.ELM 7.0 which were reported to be catalyzed by more than one kinase. Such S/T/Y entries were sorted out manually for each kinase information, and the isolated entries were added to the respective kinase substrate data. The total number of phosphorylated S/T/Y instances with kinase information therefore increased from 3,417 to 4,403, covering 2,653 S, 800 T, and 950 Y residues (Table I) for sequence analysis of protein substrates for different kinases.

KINASE CLASSIFICATION

The protein phosphorylation data with kinase information was divided into families and groups on the basis of the classification proposed earlier [Hanks and Hunter, 1995; Hanks, 2003]. This methodology is based on homology of kinase catalytic domains that divides and subdivides 269 individual kinases into distinct families and groups (see Supplementary Material Table I). This classification method distributes all kinases into 8 distinct groups and these groups are in turn divided into 83 different families containing 269 individual kinase members. The datasets for analysis of kinase-specific S/T/Y phosphorylation sites/substrates were grouped according to kinase groups, families, and individual members. Analyzing the association patterns by MAPRes, followed by the development of a consensus sequence for individual kinases, families, and groups, will establish general and specific sequence patterns and consensus sequence for each subgroup.

All the datasets for S/T/Y phosphorylation without kinase information and those with kinase information including kinase groups, families, and individual kinases, were applied to peptides of 21 amino acids. These peptides were generated so that 10 amino acids were upstream and 10 amino acids were downstream of the phosphorylated S/T/Y residues. This feature of peptide generation is available in MAPRes.

MAPRes METHODOLOGY AND APPLICATION

The analysis of the general and the kinases-specific phosphorylation datasets by MAPRes included preference estimation and association pattern mining. The preference estimation step by MAPRes included calculating the frequency of each amino acid at every position around phosphorylated S/T/Y in 21-amino acid-long peptides. These frequencies of amino acids were utilized for estimating the deviation parameter of observed and expected frequencies, followed by estimating the significantly preferred amino acid around phosphorylated S/T/Y. These significantly preferred amino acids were then utilized by MAPRes for mining association patterns. Patterns were mined at different support levels and each association pattern was associated with a confidence level. In the present study, the support level reflected the percent data containing the pattern mined by MAPRes, whereas confidence of a mined pattern was measured as a conditional probability of occurrence, as described in the MAPRes methodology [Ahmad et al., 2008a]. MAPRes was applied for preference estimation and association rules mining to phosphorylated S/T/Y Phospho.ELM 7.0 data, with and without kinase information. Association analysis was performed on different possible support values (5–30%).

COMPARISON OF MAPRes PATTERNS FOR GENERAL AND KINASE-SPECIFIC PHOSPHORYLATION DATASETS

The strategy adopted for validating the sequence patterns mined by MAPRes, for general and kinase-specific phosphorylation datasets was identical to that performed previously [Ahmad et al., 2008b] for general phosphorylation data analysis of Phospho.ELM 3.0. For this purpose, sequences of 50 proteins, belonging to diverse categories, were downloaded randomly from the SwissProt database [Boeckmann et al., 2003]. Prediction results were taken from NetPhos 2.0 [Blom et al., 1999], DISPHOS 1.3 [Iakoucheva et al., 2004], and Scansite 2.0 [Obenauer et al., 2003] for all 50 proteins selected from SwissProt. Similarly, the prediction results for kinase substrate sites were performed by NetPhosK 1.0 [Blom et al., 2004], Scansite 2.0 [Obenauer et al., 2003], and KinasePhos 2.0 [Huang et al., 2005]. The surrounding sequence of the positive prediction sites was searched for the association patterns mined by MAPRes both for general and kinase-specific datasets.

RESULTS

PREFERENCE ESTIMATION: GENERAL-DATASET ANALYSIS (PHOSPHORYLATED S/T/Y WITHOUT KINASE INFORMATION)

Preference estimation around the phosphorylated sites (S, T, and Y) included calculating the observed and expected frequency of occurrences of every amino acid residue at each position in the peptides of 21 (-10 to $+10$) amino acids length. The frequency of each amino acid around phosphorylated S/T/Y is shown in Figure 1. Frequency estimation indicates that Pro has the highest frequency of 34.47% at $+1$ position around phosphorylated S (Fig. 1a) and 42.42% around phosphorylated T at $+1$ position (Fig. 1b). While E shows the highest frequency of 14.55% at -3 position around phosphorylated Y (Fig. 1c), the frequency of Pro around phosphorylated Y was the second highest with a value of 13.39% at $+9$ position and then 13.15% at $+3$ position (Fig. 1c). It has also been observed that P at other positions (-10 to -1 and $+2$ to $+10$) varied between 7.31–9.79% around phosphorylated S (Fig. 1a) and 7.15–15.20% around phosphorylated T (Fig. 1b). The deviation parameter values calculated are utilized by MAPRes for estimating the significant preference. The total number of significantly preferred sites/amino acids was 221. Among these significantly preferred sites, 93 were for phosphorylated S, 62 were for phosphorylated T, and 66 for phosphorylated Y (Fig. 2).

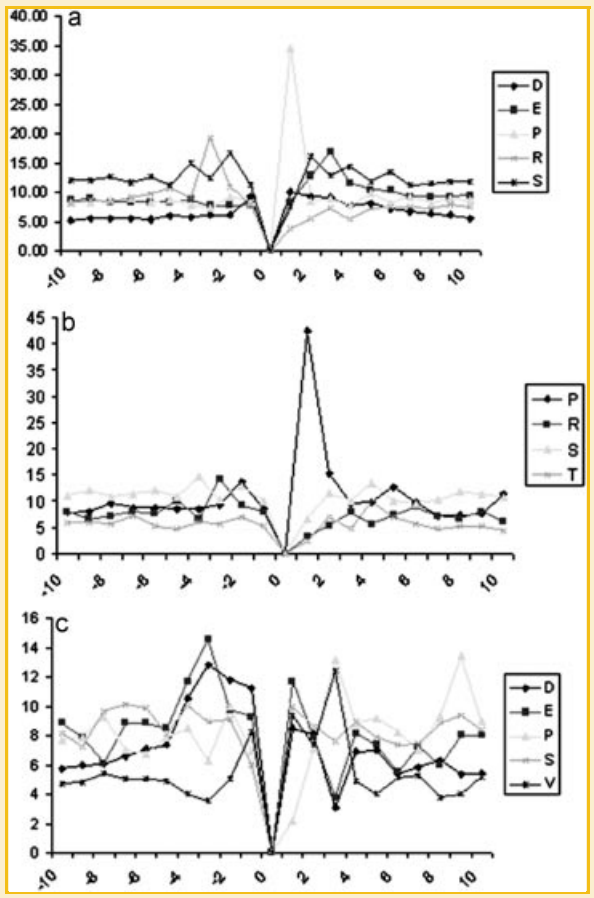

Fig. 1. Graphical presentation of different amino acids with their percent frequency around phosphorylated S (a), T (b), and Y (c).

The results concerning the estimation of significantly preferred amino acids in the vicinity of phosphorylated S showed that D, E, P, R, and S were highly preferred at different positions in the 21 amino acid peptide. Among the significantly preferred amino acids around phosphorylated S, P was the highly preferred one on each of the positions (Fig. 2). Similarly, P, D, R, S, and T were estimated as highly preferred at different positions in −10 upstream and +10 downstream the phosphorylated T residue. Here the P was again found to be significantly preferred on most of the positions (17 out of 20 positions in 21 amino acid peptide length) around phosphorylated T (Fig. 2). In contrast, the significantly preferred amino acids observed around phosphorylated Y were R, S, D, K, E, and G. Despite the significant preference of the acidic amino acids, D and E, around phosphorylated S at various positions, P was found significantly preferred on 12 different positions (Fig. 2).

ASSOCIATION PATTERNS FOR GENERAL-DATASET ANALYSIS

The total association patterns mined by MAPRes from 221 significantly preferred (s-preferred) sites, drawn from 21-amino acid-long peptides of all phosphorylated S/T/Y of Phospho.ELM 7.0 datasets without kinase information amounted to 160. These were determined at varying support levels, ranging from 5% to 30%. Out of these 160 association patterns/rules, 42 were for S, 43 for T, and 75 for Y (Table II) with a confidence range of 74.46–100%, 12.48–33.28%, and 8.93–100%, respectively.

As described earlier, P at position +1 is the most frequent residue, and the most frequent number of association rules with this site were found at 5% support level (Tables III and IV) with a variable confidence levels. Moreover, the other association patterns with Pro at +1 position were also observed on a higher support level (Tables III and IV). A similar trend for preference and association

Preferred Amino Acid Occurence Positon -/+ 10 around Phosphorylated Sites of PhosphoELM_0707 datasets

Amino Acid	S		T		Y	
	Occurance	Position of Preference	Occurance	Position of Preference	Occurance	Position of Preference
A	0		2	9,10,	0	
C	0		1	2,	0	
D	14	-5,-4,-3,-2,-1,1,2,3,4,5,6,7,8,9,	4	-10,-1,2,5,	11	-7,-6,-5,-4,-3,-2,-1,1,2,4,5,
E	12	-10,-9,-4,2,3,4,5,6,7,8,9,10,	1	3,	8	-10,-7,-6,-4,-3,-2,-1,1,
G	5	-10,-4,-3,-1,2,	2	-1,3,	6	-10,-4,-2,5,7,8,
I	0		0		3	-5,-1,3,
K	6	-10,-9,-8,-7,-6,9,	0		2	5,7,
L	0		0		1	3,
M	0		0		2	-9,3,
N	0		0		5	-4,-3,-2,2,4,
P	20	-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,1,2,3,4,5,6,7,8,9,10,	17	-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,1,2,3,4,5,6,10,	12	-9,-8,-6,-4,-2,3,4,5,6,8,9,10,
Q	0		0		1	4,
R	17	-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,3,5,6,7,8,9,10,	13	-10,-8,-7,-6,-5,-3,-2,-1,3,5,6,7,9,	4	-8,-7,4,7,
S	19	-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,2,3,4,5,6,7,8,9,10,	14	-10,-9,-8,-7,-6,-5,-4,-2,2,4,7,8,9,10,	4	-7,-6,-4,1,
T	0		5	-7,-2,2,4,5,	1	-1,
V	0		0		4	-1,1,2,3,
W	0		1	7,	0	
Y	0		2	7,8,	2	5,6,

Fig. 2. Different amino acids statistically preferred in the vicinity of phosphorylated S, T, and Y. It is clear that the amino acids with acidic (D and E), basic (K, R), non-polar neutral (P), and polar neural (S, T) functional groups prevail on most of the positions around phosphorylated S/T/Y, representing the statistically preferred amino acids present in the vicinity of phosphorylated S/T/Y.

TABLE II. General-Dataset Association Pattern Summary

Support level (%)	S	T	Y	Mined patterns
5	4	16	62	82
10	29	21	13	63
15	5	2	–	7
20	1	1	–	2
25	2	2	–	4
30	1	1	–	2
Total	42	43	75	160

patterns mined for phosphorylated Y with P in its vicinity was also observed at various positions (Tables III and IV). However, for phosphorylated Y, other acidic and basic amino acids were found equally important in the different association patterns mined (Tables III and IV). Thus, the most favorable amino acid in the vicinity of phosphorylated S/T is indeed P, with certain specific amino acids required along with the P residue.

KINASE–DATASET ANALYSIS

Analysis of significantly preferred amino acids around phosphorylated substrate sites catalyzed by the kinases of the AGC group showed that R was highly preferred at (−5, −3, and −2) positions for S/T target sites, whereas L at −5 and R at −3 were significantly preferred for the CaM-K group. Creatine kinase (CK) catalyzes the conversion of creatine to phosphocreatine, consuming ATP and generating adenosine diphosphate (ADP) expressed by various tissues. Acidic amino acids, including E at +2, +3, +4, +5 positions and D at +3 position, were found to be significantly preferred in the vicinity of phosphorylated substrate sites catalyzed by CK. MAPRes mined a total of 632 association patterns for phosphorylated S/T/Y substrates of individual kinases (see summary Tables V and VI). Similarly, 442 association patterns were mined at different support levels for substrates of kinase families (see summary Table VII) and 149 patterns for substrates of kinase groups (see summary Tables VII and VIII).

TABLE III. Summary of Association Patterns General-Dataset (Without Kinase Information)

Support level (%)	Modification	Amino acid D	E	G	I	K	L	N	P	Q	R	S	T	V	Total
5	S	–	–	–	–	–	–	–	2	–	–	2	–	–	4
	T	–	–	–	–	–	–	–	10	–	–	6	–	–	16
	Y	11	8	6	3	2	1	5	12	1	4	4	1	4	62
10	S	1	5	–	–	–	–	–	1	–	3	19	–	–	29
	T	–	–	–	–	–	–	–	5	–	2	14	–	–	21
	Y	4	3	–	–	–	1	–	2	–	–	2	–	1	13
15	S	–	1	–	–	–	–	–	1	–	1	2	–	–	5
	T	–	–	–	–	–	–	–	2	–	–	–	–	–	2
20	S	–	–	–	–	–	–	–	1	–	–	–	–	–	1
	T	–	–	–	–	–	–	–	1	–	–	–	–	–	1
25	S	–	–	–	–	–	–	–	2	–	–	–	–	–	2
	T	–	–	–	–	–	–	–	2	–	–	–	–	–	2
30	S	–	–	–	–	–	–	–	1	–	–	–	–	–	1
	T	–	–	–	–	–	–	–	1	–	–	–	–	–	1
Total occurrences of patterns		16	17	6	3	2	2	5	43	1	10	49	1	5	160

TABLE IV. Association Patterns With Support Level (30–5%)

Serial no.	Association patterns	Support level (%)	Confidence level (%)
1	<D,−6>=>Y	5	100
2	<D,−7>=>Y	5	100
3	<E,1>=>Y	5	100
		10	100
4	<E,−1>=>Y	5	100
5	<E,2>=>S	10	100
6	<E,−2>=>Y	5	100
7	<E,−3>=>Y	5	100
		10	100
8	<E,4>=>S	10	100
9	<E,5>=>S	10	100
10	<E,6>=>S	10	100
11	<E,−6>=>Y	5	100
12	<E,−7>=>Y	5	100
13	<G,−2>=>Y	5	100
14	<G,5>=>Y	5	100
15	<G,7>=>Y	5	100
16	<G,8>=>Y	5	100
17	<I,−1>=>Y	5	100
18	<I,3>=>Y	5	100
19	<I,−5>=>Y	5	100
20	<K,5>=>Y	5	100
21	<K,7>=>Y	5	100
22	<L,3>=>Y	5	100
		10	100
23	<N,2>=>Y	5	100
24	<N,−2>=>Y	5	100
25	<N,−3>=>Y	5	100
26	<N,4>=>Y	5	100
27	<N,−4>=>Y	5	100
28	<Q,4>=>Y	5	100
29	<R,4>=>Y	5	100
30	<S,−1>=>S	10	100
31	<S,1>=>Y	5	100
32	<S,3>=>S	10	100
33	<S,−3>=>S	10	100
34	<S,5>=>S	10	100
35	<S,6>=>S	10	100
36	<T,−1>=>Y	5	100
37	<V,1>=>Y	5	100
38	<V,−1>=>Y	5	100
39	<V,2>=>Y	5	100
40	<V,3>=>Y	5	100
		10	100
41	<E,3>=>S	10	89.67972
		15	89.67972
42	<S,2>=>S	10	87.51696
		15	87.51696
43	<R,−3>=>S	10	87.3771744
		15	87.3771744
44	<D,1>=>S	10	87.25702
45	<S,−2>=>S	10	86.7386551
		15	86.7386551
46	<R,−2>=>S	10	85.6579
47	<S,−8>=>S	10	85.44734
48	<S,10>=>S	10	84.85937
49	<S,−10>=>S	10	84.72791
50	<S,7>=>S	10	84.65874
51	<R,−5>=>S	10	84.32147
52	<S,4>=>S	10	84.25971
53	<S,−2><P,1>=>S	5	84.23841
54	<S,9>=>S	10	84.06232
55	<S,−5>=>S	10	83.8993759
56	<S,−9>=>S	10	83.54726
57	<S,8>=>S	10	82.98907
58	<S,−4><P,1>=>S	5	80.896225
59	<P,1>=>S	10	80.53235
		15	80.53235
		20	80.53235
		25	80.53235
		30	80.53235
60	<P,1><S,4>=>S	5	80.2799

TABLE V. Individual Kinase Members' Association Patterns

	Support levels														
	10%			15%			20%		25%			30%			
Kinase members	S	T	Y	S	T	Y	T	Y	S	T	Y	S	T	Y	Total
Abl			4			1		3			2			1	11
AMPK_group	1			1					1			3			6
ATM	1			5	1		1		3	1		1	1		14
Aurora-A	1			3					3			2			9
Aurora-B	1			2					3			1			7
BTK			1			1		2			2			1	7
CaM-KII_group	2			3	1		1		1	1		1	1		11
CaM-KIIalpha	1	1		3	1		1		3	1		2	1		14
CDK_group	5	2		5	2		2		1	2		1	1		21
CDK1	7	4		3	4		2		1	1		1	1		24
CDK2	2	2		2	2		1		1	1		1	1		13
CDK4	1			1					1			1			4
CDK5		1			1		1			1			1		5
CDK7		1			1		1			1			1		5
CK2_group	10	6		3	1		5		3	4		1	2		35
CK2-alpha	15	1		2	2		2		4	2		2	2		32
Csk			1			1		1			1			1	5
DAPK3		1			1		1			1			1		5
DNA-PK		1			1		1			1			1		5
EGFR			2			7		3			1				13
FGFR1			1			1		2			2				6
Fyn			2			2		2							6
GSK-3_group		1			1		1			1			1		5
GSK-3beta		1			1		1			1			1		5
IGF1R			3			2		4							9
IKK_group	1			1					4						6
InsR			2			7		3							12
JAK2			2			2		2							6
Lck			1			8		5							14
LKB1		1			1		1			1			1		5
Lyn			2			1		4							7
MAP2K_group		1			1		1			1			1		5
MAP2K4			1			1		1							3
MAPK_group	1	1		2	1		1		2	1			1		10
MAPK1	1	1		11	1		1		1	1			1		18
MAPK14	1	1		1	1		1		1	1			1		8
MAPK3	1	2		4	2		2		1	1			1		14
MAPK8		1			1		1			1			1		5
MAPK9		1			1		1			1			1		5
MAPKAPK2	1			2					1						4
Met			1			1		1							3
PAK1	2			1					3						6
PDGFR-beta			1			1		2							4
PDK-1		1			1		1			2			1		6
PKA_group	2	3		1	2		1		1	1			1		12
PKA-alpha	1			1					1						3
PKB_group	4	1		2	1		1		1	1			1		12
PKC_group	32	4		12	8		7		3	3			1		70
PKC-alpha	13	1		9	3		3		1	3			3		36
PKC-beta	1			1					1						3
PKC-delta	3			3					3						9
PKD1	1			1					1						3
PKG/cGK_group	1			1					1						3
PKG1/cGK-I	1			1					1						3
PLK1	1			1					1						3
ROCK_group		1			1		1			1			1		5
RSK_group	5			3					4						12
SGK_group	1	1		1	1		1		1	1			1		8
Src			15			5		2							22
Syk			2			4		1							7
ZAP70			1			1		1							3
Total	121	43	42	92	46	46	45	39	58	39	8	17	33	3	632

SUBSTRATE CONSENSUS SEQUENCE FOR DIFFERENT KINASE GROUPS

Consensus sequence refers to those amino acids sequence/pattern(s) that are common in different substrates at specific position(s) catalyzed by the same individual kinase, kinase family, or kinase group. To develop consensus sequence, the association patterns containing peptides of each category were aligned position to position (−8 to +8 and 0 being phosphorylated S/T/Y). Thus, 61 consensus patterns were established for individual kinases, among which novel patterns were uncovered (Fig. 3).

TABLE VI. Kinase Association Patterns Summary

Support level (%)	S	T	Y	Mined patterns
10	121	43	42	206
15	92	46	46	184
20		45	39	84
25	58	39	8	105
30	17	33	3	53
Grand total	288	206	138	632

TABLE VIII. Association Patterns for Kinase Groups

Support levels (%)	Kinase groups								
	AGC	CaM-K	CK	CMGC	Others	STE	TK	TKL	Total
10	2	1	4	19	17	1	24	1	69
15	1	7	17	1	5	20	4	1	56
20	2	1	6	1	1	2	–	–	13
25	1	1	2	2	–	–	–	–	6
30	2	1	1	1	–	–	–	–	5
Total patterns	8	11	30	24	23	23	28	2	149

COMPARISON OF MAPRes PATTERNS FOR GENERAL PHOSPHORYLATION DATASETS

The prediction results of all 50 proteins by NetPhos 2.0 resulted in 1,545 positive predictions, including 991 S, 314 T, and 240 Y. Peptides of 15 amino acids length with seven upstream (−7 to −1) and seven downstream (+1 to +7) amino acids were made for each positive prediction by NetPhos on S/T/Y residues. Searching all 160 association patterns mined by MAPRes in 1,545 peptides with positively predicted sites by NetPhos 2.0 showed that 1,424 peptides contained one or more association patterns in the vicinity of positive predictions. Out of the 1,424 validated sites, 942 were of S, 240 of T, and 240 of Y as positively predicted phosphorylation sites. The comparison results show that more than 92% of the positive predictions of NetPhos 2.0 are consistent with the results of association patterns mined by MAPRes (Table IX). Similarly, the association patterns mined by MAPRes were also searched in the vicinity of the positive prediction of phosphorylation motifs by Scansite 2.0 [Obenauer et al., 2003] for the same 50 proteins. The comparison of the results of MAPRes patterns with those of Scansite 2.0 also revealed a similarity of over 92% (Table IX). Comparison of MAPRes results with those of DISPHOS 1.3 showed that over 94% of the positive prediction sites contain one or more association patterns in their vicinity, which represents a very high degree of validation percentage (Table IX).

COMPARISON OF MAPRes PATTERNS FOR KINASE-SPECIFIC PHOSPHORYLATION DATASETS

The prediction results by NetPhosK 1.0 of the same 50 proteins for kinase substrate sites were compared with kinase-specific association patterns mined utilizing MAPRes, by searching these patterns in the vicinity of substrate sites of different kinases (S/T/Y). NetPhosK predicted a total of 2,841 positive sites (with 1,861 S, 815 T, and 165 Y) catalyzed by 17 kinases. The vicinity of the substrates of 17 kinases positively predicted by NetPhosK 1.0 was compared with the association patterns mined by MAPRes for the same 17 kinases. Out of 2,841 peptides, with a length of 15 amino acids, 2,785 (with 1,853 S, 772 T, and 160 Y−) contained one or more association patterns in the vicinity of substrates of all 17 kinases predicted by NetPhosK 1.0. This validation comparison showed over 94% consistency with NetPhos K 1.0 (Table X). Motif Scan by Scansite for the same 50 proteins, utilizing medium stringency, resulted in 738 (including 354 S, 154 T, and 230 Y) positive sites catalyzed by 42 kinases. The comparison of these predictions with the association patterns mined by MAPRes resulted in 98% consistency, as the vicinity of 723 (including 354 S, 148 T, and 221 Y) out of 738 (including 354 S, 154 T, and 230 Y) predicted sites catalyzed by 42 kinases contained one or more association pattern(s) mined by MAPRes (Table X). Prediction results by KinasePhos 2.0 on the same 50 proteins resulted in 34,721 (with 20,285 S, 4,053 T, and 10,383 Y) positive sites catalyzed by 52 kinases. Comparison of these prediction results with the association patterns mined by MAPRes showed that the vicinity of 30,546 (including 19,487 S, 3,083 T, and 7,976 Y) sites out of 34,721 (with 20,285 S, 4,053 T, and 10,383 Y) positively predicted sites catalyzed by 52 kinases contained one or more association pattern mined by MAPRes (Table X). Thus, 88% KinasePhos 2.0 predictions are in conformity with the association patterns mined by MAPRes.

TABLE VII. Summary of the Patterns Mined for Kinase-Specific Data

Phospho.ELM 7.0	Pattern mined support level					Total patterns
	10%	15%	20%	25%	30%	
Member-kinase (268)	206	184	84	105	53	632
Family-kinase (61)	210	173	35	13	11	442
Group-kinase (8)	69	56	13	6	5	149

DISCUSSION

Protein phosphorylation is a key event for the transmission of information along signaling pathways. Dynamic phosphorylation often causes activation and deactivation of proteins through temporary conformational changes. New developments in the field of phosphoproteomics have helped a great deal to demonstrate the complex cellular signaling networks regulated by phosphorylation. Understanding the mechanism of each step in the signaling pathways regulated by phosphorylation and dephosphorylation requires complete kinase and phosphatase information. But there remain important challenges faced by wet lab approaches to map the phosphorylation sites with exact kinase information and study the biological/physiological implications by the dynamic nature of phosphorylation, particularly in vivo. In Phospho.ELM 7.0 [Diella et al., 2008], the largest protein phosphorylation database, there are 16,470 instances of experimentally determined phosphorylation sites, but those with kinase information are just 3,417, which represents only 21% of the total phosphorylation data known. Statistical analyses, computational methods, and developing the prediction models are necessary to develop consensus or preferred sequence patterns for the involvement of a kinases to phosphorylate

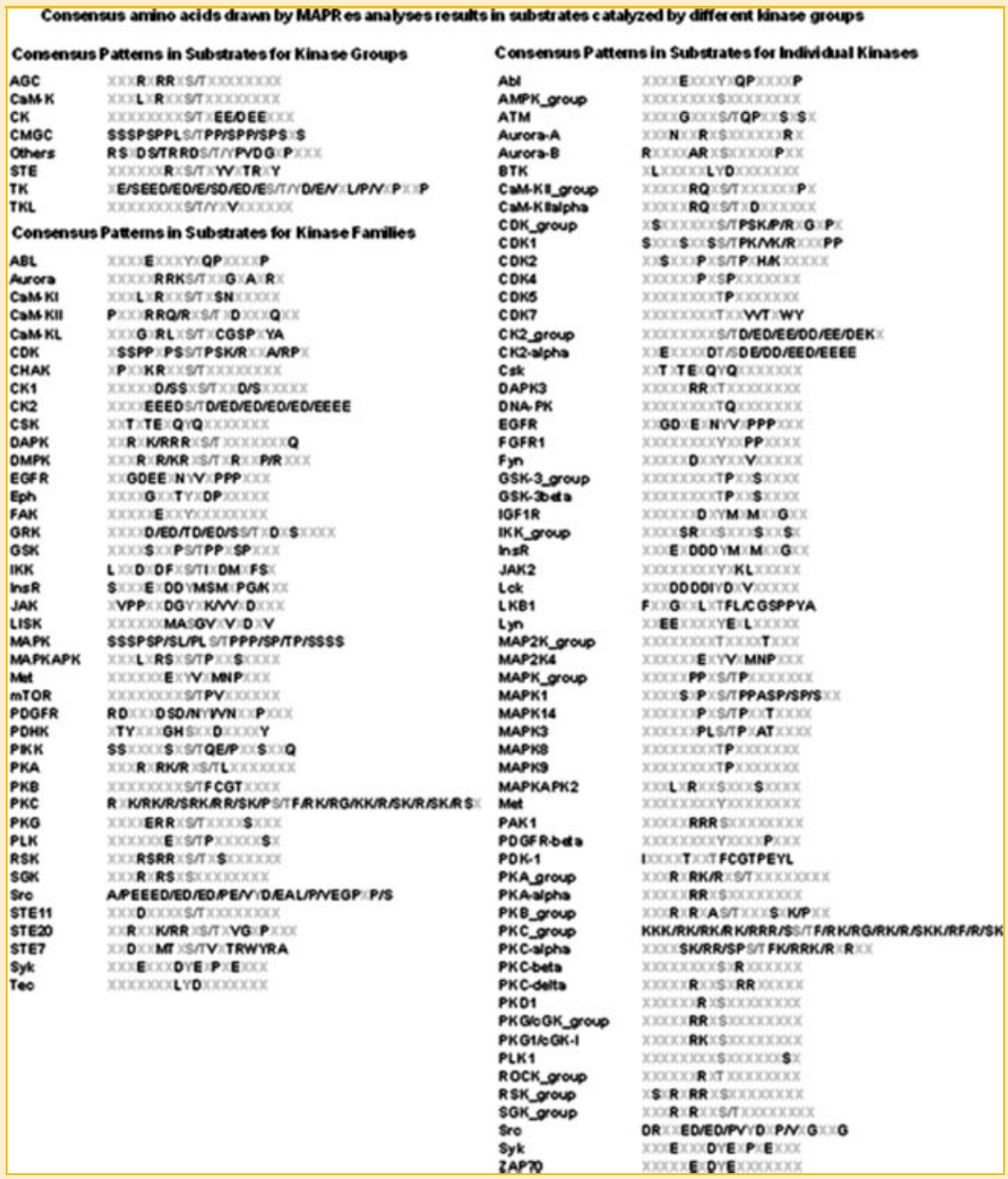

Fig. 3. Consensus sequence patterns for kinase groups, families, and individual members. The consensus sequence patterns have been developed from the data of the association patterns mined by the MAPRes developing an association among the preferred sites with the phosphorylated site (S/T/Y). Specific types of amino acid prevail in the vicinity of the substrate sites of the different kinase groups and the absence of others. For example, only a basic amino acid R can be seen in the vicinity of AGC group, whereas, acidic D and E are the only amino acids present in the consensus for CK group, representing the general requirement for kinase groups. However, the substrates of kinase families and individual members of the same group represent different types of amino acids, which show the specific requirement of the families and individual kinases.

TABLE IX. Validation of the Patterns Mined by MAPRes for General Datasets of Phosphorylation Sites With the Results of Other Methods

		Predictions				Association patterns				
General-dataset validation	Proteins	S	T	Y	Total sites	S	T	Y	Total patterns	Validate (%)
NetPhos 2.0	50	991	314	240	1,545	942	242	240	1,424	92.17
Scansite 2.0	50	991	312	240	1,543	942	240	240	1,422	92.16
DISPHOS 1.3	46	856	319	186	1,361	823	279	182	1,284	94.34

TABLE X. Validation of Kinase-Specific Patterns by MAPRes With the Results of Other Methods

		Predictions				Association patterns				
Kinase-dataset validation	Proteins	S	T	Y	Total sites	S	T	Y	Total patterns	Validate (%)
NetPhosK (17-Kinase)	50	1,861	815	165	2,841	1,853	772	160	2,785	98.03
Scansite (42-Kinase)	50	354	154	230	738	354	148	221	723	97.97
KinasePhose 2.0 (52-Kinase)	50	20,285	4,053	10,383	34,721	19,487	3,083	7,976	30,546	87.98

S/T/Y in presence or absence of certain amino acids in their immediate vicinity. These studies will eventually help understanding the cellular signaling networks, and the data resulting from such studies will be valuable to experimentalists for the design of directed studies.

Previously, Phospho.ELM 3.0 [Diella et al., 2004] was analyzed with MAPRes [Ahmad et al., 2008a]. In a previous version of Phospho.ELM 3.0, the kinase data was too scanty to run an analysis for mining association patterns [Ahmad et al., 2008b]. The analysis was therefore restricted to mining association patterns for all phosphorylated S/T/Y without kinase information. In the present version of Phospho.ELM 7.0, the phosphorylation data with kinase information is still well below the total data but contains sufficient elements to run an analysis for 61 individual kinases, which is the highest number of substrates and kinases ever included by different theoretical/computational studies.

The association pattern mining results showed that a maximum number of patterns were mined at 10% support level (Table II), which is consistent with established procedures in data mining. An increase in support level appreciably decreases the number of patterns mined, whereas a decrease in support level to 5% results in a slight decrease in the number of association pattern in the case of S and T (Table II). Whereas, at 5% support and the number of patterns mined were maximum for Y (Table II) compared to 10% support and no pattern was mined above 10% support level for Y (Table II). This shows that the optimum support level for S/T phosphorylation data resulting in maximum numbers of pattern mining was 10%, whereas for Y phosphorylation data, it was 5%. This trend shows the existence of diverse sequence requirements for phosphorylation on S/T/Y catalyzed by different kinases. The reduced number of patterns mined at 5% support compared to those supported by 10% of the data may be due to reduction of the phosphorylated S/T sites data catalyzed by the same kinase or kinase group. On the other hand, an appreciable reduction or absence of patterns mined by MAPRes at higher support level may be due to mixing of the substrate data catalyzed by various kinds of kinases and their isoforms.

Association analysis results showed that the length of association patterns is reduced with an increase in the support level and at 10% or more, the length of association patterns becomes equal to a single amino acid for all phosphorylated S, T, and Y. Such association pattern of a single amino acid length will be more useful than a significantly preferred site in the vicinity of phosphorylated S/T/Y, as each association pattern has a confidence level (conditioned probability of occurrence) (Table IV), and significantly preferred sites do not (Fig. 2). Another usefulness of such short patterns consisting of a single amino acid in the vicinity of phosphorylated S/T/Y (at higher support level) rests on that it shows a general preference of a specific amino acid in the vicinity of phosphorylation sites. The patterns consisting of more than one amino acids mined at lower support level include a combination of patterns (of single amino acids) that were mined at higher support level (Table V), confirming that for phosphorylation of S/T/Y, there is a general requirement for vicinal amino acids, while for other amino acids, there is a specific requirement. For instance, a pattern mined at 30% support level for phosphorylated S was P1=>S with a confidence level of 80.53% and another pattern at 10% support level was S4=>S with 84.26% confidence, whereas a combination of these two patterns was found at 5% support level as <P1><S4>=>S with a confidence of 80.28% (Table V). Here, the two single amino acid patterns (P1=>S and S4=>S) at higher support level represent their general preference for S phosphorylation catalyzed by a larger group of kinases. On the other hand, a pattern which is a combination of these two patterns (<P1><S4>=>S) supported by 5% of the phosphorylation data shows specific preference for a smaller kinase group or individual kinase. Another interesting and unique trend for combination of the patterns, consisting of single amino acids (mined at higher support levels) with larger patterns mined at lower support level, was found in case of patterns mined for phosphorylated T. For instance, each of the association patterns of single amino acid with S at −2, −4, −6, −7, −8, and −9 positions in the vicinity of phosphorylated T were mined at 10% support and another pattern of single amino acid with P at +1 position was mined at all support levels from 10% to 30%, but the patterns of two amino acids with combination of P+1 with all S−2, −4, −6, −7, −8, and −9 (e.g., <P+1, S−2>=>T, <P+1, S−4>=>T) were mined at 5% support level in the vicinity of phosphorylated T (Table V). Consequently, these general and specific patterns mined at different support levels show the involvement of more than one kinase or its isoforms for catalyzing the phosphorylation of S/T/Y with similar sequence in the vicinity. A comparison of the results of association patterns, mined by MAPRes for the present version (7.0) of Phospho.ELM with those of the previous (i.e., 3.0) reported earlier [Ahmad et al., 2008b] shows that most of the patterns are identical to the previous ones. In case of the patterns mined in the vicinity of phosphorylated S, all previous (28 patterns mined for version 3.0) patterns were included in the present analysis with an addition of five new patterns. But in case of the patterns mined, for T and Y, by MAPRes in the current analysis, they were much more numerous than the previous ones with an addition of many new patterns (Tables II and V).

Comparison of the results from the two types of association analyses mined by MAPRes, both for substrates with and without kinase information, showed that the minimum support level for which association patterns were mined was 5% (patterns with

maximum length) for general datasets, whereas it was 10% (patterns with maximum length) for kinase-specific analyses. Maximum support level for which association patterns were mined was 30% (patterns with minimum length) for general datasets, whereas 50% (patterns with minimum length) support level was maximum for association patterns mined for kinase-specific datasets. When comparing the results of the two analyses (general and kinase-specific datasets), a similarity of patterns mined for groups, families, and individual kinases is noted with those of the general datasets. A general trend has also been observed in kinase-specific analyses, namely that the patterns mined by MAPRes show association of specific types of amino acids with kinase groups, families, and individual members, whereas others are absent in the vicinity of phosphorylated sites (Fig. 3). Patterns mined for different substrates catalyzed by different kinase groups exhibit a general requirement for catalyzing phosphorylation reaction on –OH group of S/T/Y. Whereas, the patterns mined for substrates of kinase families and individual members reflect a specific requirement in addition to the general one for their presence in the vicinity of the phosphorylated S/T/Y (Fig. 3). For instance, at group level the patterns mined show the presence of only basic (R) amino acids and the absence of all others for substrates of AGC group, whereas the patterns mined for substrates of kinase families and individual members contain diverse amino acids including neutral non-polar (A, G), neutral polar (S, Q), basic (L), and acidic (E) amino acids (Fig. 3). Similarly, the patterns mined for substrates of CaMK group contain a majority of basic amino acids (R) along with some neutral non-polar amino acids (A, L) and neutral polar amino acids (S) (Fig. 3), but the patterns mined for substrates of kinase families and individual members show involvement of other amino acids (Fig. 3) specifically required for phosphorylation of S/T. The patterns for substrates of the CK group exhibit the presence of only acidic amino acids (D, E) (Fig. 3), while the patterns for substrates of the TK group exhibit the presence of acidic (D/E) amino acids in addition to neutral non-polar (L, P, V), neutral polar (S) amino acids at different positions (Fig. 3) generally; however, the patterns for substrates of kinase families and individual members show more specific amino acids required for their phosphorylation on Y residues (Fig. 3). The amino acid with an aliphatic side chain, P at different positions, was extensively mined in the patterns for substrates of CMGC group along with a frequent occurrence of the neutral polar amino acid S at various positions, which showed a general (Fig. 3) as well as a specific (Tables III and IV) requirement in substrates of some kinase families and individual members (Fig. 3). The patterns for substrates of other kinase families and members indicate involvement of other amino acids specifically for the phosphorylation of S/T. The patterns for substrates of the STE group show the presence of diverse amino acids within the kinase group, families, and individual members (Fig. 3). All such patterns can be found with few exceptions in the results of general datasets (Table IV), probably because of the large difference in the amount of data of the two analyses types. However, the general trend of the two analyses remains the same. These results allow drawing the conclusions that there are some general requirements for the presence or absence of specific amino acids in the vicinity of phosphorylated S/T/Y and some specific requirements for the presence or absence of amino acids for phosphorylation catalyzed by different kinase groups, families, and members.

The association patterns mined for datasets with and without kinase information were compared with the results of previous studies, including statistical and computational approaches. Several studies earlier reported or suggested an important role of amino acid residues present in the vicinity of phosphorylated S/T/Y residues [Yaffe et al., 2001; Iakoucheva et al., 2004; Qazi et al., 2006; Ahmad et al., 2008b]. The occurrence of acidic, basic, hydrophobic, and charged amino acids in the vicinity of phosphorylated S/T/Y was described earlier [Iakoucheva et al., 2004], which was also showed by the patterns of the present analyses (Table IV; Fig. 3). It has also been documented that the presence of the positively charged basic amino acids R and L on positions −2, −3, and −5 and of negatively charged acidic amino acids including E and D straddling at −4 to +4 positions in the vicinity of phosphorylated sites of proteins catalyzed by S/T kinases are important determinants of protein phosphorylation on S/T residues [Yaffe et al., 2001]. This trend is also evident in the present analyses, both in general and kinase-specific phosphorylation data analyses (Table IV; Fig. 3). Other amino acids, such as M, V, I, F, and L, were reported to be favored at +3 position in mammals [Yaffe et al., 2001], and P residues have been documented to be important at various positions (−2, +5, +9) for Y phosphorylation [Iakoucheva et al., 2004], which was also found in the present analyses by MAPRes (Table IV; Fig. 3). Moreover, S phosphorylation was described to be associated with S residues at −4, +2, +4 positions [Iakoucheva et al., 2004]. These S residues, along with other amino acids on other positions around phosphorylated S and T, were also detected in the present analyses (Table IV; Fig. 3). In addition to the amino acids described earlier [Yaffe et al., 2001; Iakoucheva et al., 2004; Qazi et al., 2006; Ahmad et al., 2008b], we also found novel patterns and data trends, as described above. Thus, the MAPRes results are in conformance with previous findings with some new insights for preferred amino acid patterns present in the vicinity of phosphorylated S/T/Y.

Other comparisons of the patterns mined by MAPRes included searching the patterns in the vicinity of positively predicted phosphorylation sites by different methods available, both for general predictions without information of kinase and kinase-specific predictions. The comparison of MAPRes at protein sequence level with the results of different phosphorylation prediction methods for general datasets show more than 90% conformity (Table IX). Similarly the comparison of MAPRes analysis results for kinase-specific substrate sites with those of available prediction methods show a very high conformity level, ranging from 87% to 98% (Table X). This high rate of conformity points to the accuracy of the algorithm and the technique of data mining applied in MAPRes. Additionally, all the kinase-specific prediction methods available at present cover a maximum of 52 kinases for which predictions or analysis can be performed. MAPRes has analyzed the phosphorylation of the substrate data for 61 kinases, implying that some new and novel associations of amino acids in the vicinity of phosphorylated S/T/Y have been uncovered. Additionally, the association of these patterns to 61 out of 267 kinases is because only one-fourth of the Phopho.ELM 7.0 data is given with kinase information, and if this data covered the kinase information for all of the entries, the number

of kinases associated with specific amino acid patterns in the vicinity of phosphorylation sites would be much higher than 61.

Protein sequence patterns mined by MAPRes both for general and kinase-specific datasets are similar and comparable to the previous findings and predictions. The patterns mined by MAPRes cannot be considered as classification or predictions, but they exhibit a correlation approach of the modification sites with the amino acids in their vicinity. Each association pattern mined at a specific support level is coupled with a confidence value, a conditioned probability for the occurrence, is indicative of the validity of the patterns mined by MAPRes. These extracted patterns have been utilized to establish the general consensus sequence patterns for kinase groups and specific consensus patterns for families and individual members. Therefore, the results of this study should be useful to molecular biologists, biochemists, and medical scientists, as they provide information on the general and specific requirement for amino acids in the vicinity of phosphorylated S/T/Y catalyzed by different kinase groups, families, and members.

ACKNOWLEDGMENTS

NUD and ARS acknowledge the partial financial support of Pakistan Academy of Sciences and WHO-EMRO for this study.

REFERENCES

Ahmad I, Hoessli DC, Walker-Nasir E, Rafik SM, Shakoori AR, Nasir-ud-Din. 2006. Oct-2 DNA binding transcription factor: Functional consequences of phosphorylation and glycosylation. Nucleic Acids Res 34:175–184.

Ahmad I, Hoessli DC, Gupta R, Walker-Nasir E, Rafik SM, Choudhary MI, Shakoori AR, Nasir-ud-Din. 2007. In silico determination of intracellular glycosylation and phosphorylation sites in human selectins: Implications for biological function. J Cell Biochem 100:1558–1572.

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

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

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

Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S. 2004. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 4:1633–1649.

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. Nucleic Acids Res 31:365–370.

Bridges AJ. 2001. Chemical inhibitors of protein kinases. Chem Rev 101: 2541–2571.

Bridges AJ. 2005. Therapeutic challenges of kinase and phosphatase inhibition and use in anti-diabetic strategy. Biochem Soc Trans 33:343–345.

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 Bioinform 5:79.

Diella F, Gould CM, Chica C, Via A, Gibson TJ. 2008. Phospho.ELM: A database of phosphorylation sites—Update 2008. Nucleic Acids Res 36:D240–D244.

Hanks SK. 2003. Genomic analysis of the eukaryotic protein kinase superfamily: A perspective. Genome Biol 4:111.

Hanks SK, Hunter T. 1995. The eukaryotic protein kinase superfamily: Kinase (catalytic) domain structure and classification. FASEB J 9:576–596.

Huang HD, Lee TY, Tzeng SW, Horng JT. 2005. KinasePhos: A web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res 33:W226–W229. (Web Server issue).

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 Acids Res 32:1037–1049.

Kaleem A, Ahmad I, Hoessli DC, Walker-Nasir E, Saleem M, Shakoori AR, Nasir-ud-Din. 2009. Epidermal growth factor receptors: Function modulation by phosphorylation and glycosylation interplay. Mol Biol Rep 36:631–639.

Kim JH, Lee J, Oh B, Kimm K, Koh I. 2004. Prediction of phosphorylation sites using SVMs. Bioinformatics 20:3179–3184.

Krause DS, van Etten RA. 2005. Tyrosine kinases as targets for cancer therapy. N Engl J Med 353:172–187.

Kreegipuu A, Blom N, Brunak S. 1999. PhosphoBase, a database of phosphorylation sites: Release 2.0. Nucleic Acids Res 27:237–239.

Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, Wang TH. 2006. dbPTM: An information repository of protein post-translational modification. Nucleic Acids Res 34:D622–D627. (Database issue).

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

Qazi WM, Ahmed M, Hoessli DC, Ahmad I, Khawaja I, Wajahat T, Kaleem A, Nasir E-W, 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.

Schwartz D, Gygi SP. 2005. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat Biotechnol 23:1391–1398.

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.

Wang M, Li C, Chen W, Wang C. 2008. Prediction of PK-specific phosphorylation site based on information entropy. Sci China C Life Sci 51:12–20.

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.

Yeh RH, Lee TR, Lawrence DS. 2002. From consensus sequence to high-affinity ligands: Acquisition of signaling protein modulators. Pharmacol Ther 93:179–191.