KINEXUS

IN SILICO PREDICTION SERVICES IKSP

Version 10SE1

IN SILICO SERVICES

PROTEIN KINASE SPECIFICITY PREDICTION CUSTOMER INFORMATION PACKAGE

Toll free: 1-866-KINEXUS or 604-323-2547
Facsimile: 604-323-2548
E-mail: info@kinexus.ca
www.kinexus.ca

KINEXUS IN SILICO PROTEIN KINASE SPECIFICITY PREDICTION SERVICES

Table of Contents

PDF Page No.

1. Introduction 2

2. Background on the Kinase Substrate Prediction Algorithms 4
   1. Background.. 4
   2. Kinase Phospho-Site Specificity 5
   3. Prediction of Position-specific Scoring Matrices (PSSM) for Kinases without Substrate Data.. 8
   4. Examples of PSSM's for Human Protein Kinases 12
   5. Comparison of Optimal Recognition Sequences for Protein Kinases from Prediction and Consensus from Empirically-derived Substrate Phospho-site Alignments 15
   6. Derivation of PSSM's for Yeast Protein Kinases 17

3. Follow Up Services 19

4. Forms to be Completed 20

Appendices

5. Service Order Form (IKSP-SOF) 21
6. Service Identification Form (IKSP-SIF).. 22

1. INTRODUCTION

The identification of protein kinase substrates is important for understanding the architecture and operations of cell signalling networks. The vast majority of the proteins phosphorylated by specific protein kinases in humans and other species remain unknown despite more than 4 decades of intense effort. At Kinexus, we have identified at least 10 thousand kinase-substrate phospho-site pairs from our literature and database searches, but we think that the actual number exceeds 10 million. With the emergence of protein kinases as key components of normal and pathological cellular regulation, it is important to define how these enzymes recognize their target proteins. Over 400 diseases in humans have been linked to defects in kinase signalling systems. In cancer, it is evident that perhaps 20% of the 516 known human kinases, if mutated so that they are inappropriately inactivated, activated or acquire altered specificity, can facilitate the development of cancer. We expect that many of the oncogenic mutations that facilitate cancer arise from alterations in the amino acids that define the specificity of protein kinases and in the amino acids in the substrate phospho-sites that provide for kinase recognition. With the sequencing of tens of thousands of human genomes that are predicted in the coming decade, establishing linkages between kinases, phospho-sites and gene mutations will provide a major advance in the application of personalized medicine. Moreover, improved understanding of kinase signalling systems in other experimental model systems will provide valuable insights into cellular regulation in these organisms.

To foster the discovery of new and potentially important kinase-substrate interactions, Kinexus has created a unique and cost-effective platform of integrated bioinformatics and proteomics services that permit the discovery and validation of kinase-substrate connections. The “in silico” prediction services offered by Kinexus have been created using proprietary software that has been developed by Kinexus and our academic partners at the University of British Columbia and Simon Fraser University. These algorithms for kinase substrate prediction utilized our in-house databases of the carefully aligned sequences of 488 human protein kinase catalytic domains and 10,000 known kinase-substrate phospho-site pairs. From this data and the application of our algorithms, we have been able to generate predicted phospho-site amino acid frequency matrices for any “classical” protein kinase for which the catalytic domain sequence is known. With our recent “cracking of the kinase code” Kinexus can now predict the sequences of optimal peptide substrates for most protein kinases. Conversely, we can take any known or putative phospho-site sequence and generate an individual score for each of over 492 human protein kinases for their ability to target the sequence for phosphorylation. While our prediction algorithm is not perfect and has limitations, it is the most accurate and versatile bioinformatics method available for this purpose at this time.

Our In Silico Kinase Specificity Prediction (IKSP) Service was originally developed to predict the importance of each of the amino acids surrounding the phosphorylation sites of substrates of typical human protein kinases. However, our research has revealed that our algorithms may have wider application to other species, including those as diverse as budding yeast as demonstrated later in this information package. It is also useful for prediction of the possible effects of mutation of human kinases within their catalytic domains on their substrate specificities. With this service, clients simply provide the name and Uniprot ID or NCBI accession numbers for the desired protein kinase, and Kinexus provides back a table with the expected probability frequencies of each of the 20 amino acids, 7 amino acids before and after the phospho-acceptor amino acid. The identification of positive and negative determinants in the surrounding amino acids of phosphorylation sites permits the creation of position-specific scoring matrices (PSSM) for each protein kinase. This has allowed Kinexus to produce PSSM’s for 488 human protein kinase domains to test any known or putative phosphorylation site as a substrate for all of these protein kinases in silico.

Towards the ambitious goal of creation of a comprehensive map of most of the connections between protein kinases and their substrates in the human proteome, it is necessary to identify the vast majority of the human phospho-sites. We have been able to identify over 21,000 different human proteins from careful analysis of the Uniprot database. From this, we have determined that there is a maximum upper limit of approximately 1.8 million candidate phospho-sites with serine, threonine and tyrosine as the phospho-acceptor. We have performed hydrophobicity scoring analyses to narrow down the more promising phospho-site sequences. From the inspection of over 93,000 known human phospho-sites, it is quite evident that the vast majority possess hydrophobicity scores of less than 0. As a second filter, we have separately aligned all of these known serine, threonine and tyrosine phospho-site sequences to create three P-site matrices. We have scored all 1.8 million candidate phospho-sites against these P-site matrices to identify those phospho-sites that best match the predicted amino acid distributions around known phospho-sites with either serine, threonine or tyrosine as the phospho-acceptor residues. Finally, we have also used the deduced PSSM’s for 492 human protein kinases using our algorithms and individually scored each of these against the 1.8 million candidate phospho-sites to identify those that are the best matches. Through analyses of the 1.8 million candidate phospho-sites with the application of all of these filters in combination, it appears the human phosphoproteome contains at least 700,000 distinct phospho-sites. We believe the phosphoproteomes of other simple model organisms such as yeast, Drosophila and C. elegans also feature over 100,000 phospho-sites.

2. BACKGROUND ON THE KINASE SUBSTRATE PREDICTION ALGORITHMS

1. Background

In the near future, Kinexus will publish more information about our proprietary protein kinase substrate prediction algorithm in our upcoming Kinetica Online Resource (www.kineticaonline.ca) when it is launched in late 2010. The following provides a fairly detailed overview on how it was developed and its effectiveness.

Proteins like kinases feature functional domains, which are substrings of protein sequences that can evolve, and even exist independently of the rest of the protein chain. The most common domain in protein kinases, known as its catalytic domain, carries out the actual phosphorylation of protein substrates. Of 516 known protein kinases in humans, 478 feature one or more highly conserved catalytic domains of about 250 amino acids in length.

There are specificity-determining residues (SDRs) distributed throughout the catalytic domain of the kinases that interact with the side chains of amino acids neighbouring and including the phospho-acceptor residue on the protein substrate. A phospho-site sequence on a substrate can involve amino acid residues even 6 or more positions N- and C-terminal to the phospho-acceptor residue. Kinase-substrate binding commonly involves a semi-linear phospho-site sequence fitting into a kinase active site in the vicinity of the SDRs.

Atypical kinases have completely different structures when compared to the typical protein kinases. Atypical protein kinases do not possess a catalytic domain similar to those found in the typical kinases and appear to have evolved separately. No equivalent catalytic domain has been computed for them using alignment techniques. We have determined the amino acid substrate specificities of several of the PI3KR kinases, including mTOR/FRAP, ATM, ATR and DNAPK. We have predicted the locations of SDRs in 488 human kinase catalytic domains and generated position-specific scoring matrices (PSSM) for each kinase.

There are many previous attempts to predict kinase specificities for protein substrate recognition and identify potential phosphorylation sites in the human or yeast proteomes. These methods are usually based on computing consensus kinase recognition sequences, PSSM matrices or machine learning methods. Scansite [www.scansite.mit.edu], artificial neural networks (ANN) and support significant efforts for modeling cell phosphorylation networks. NetworKIN [www.networkin.info/search.php] uses artificial neural networks and PSSM to predict kinase domain specificities and uses protein-protein interaction databases such as STRING [www.string.embl.de] to increase the accuracy of the prediction. For a kinase and a substrate that are connected or if there is a short path linking them in the STRING database, these are better candidates to be selected in a phosphorylation network than those kinase-substrates which are absent in STRING. However, NetworKIN covers only 108 kinases of the 516 known human kinases NetworKIN does not compute the kinase phosphorylation specificity of those kinases where there is no consensus phospho-site specificity data. NetPhorest [www. netphorest.info/index.php] has wider coverage compared to NetworKIN with 179 kinases. NetPhorest is similar to Networkin and uses a combination of ANN and PSSM matrices for prediction, but it puts related kinases in the same groups and assumes that all the kinases in the same group have identical kinase phosphorylation specificities.

All the mentioned methods have two major problems. Firstly, they can only compute specificity of those kinases that are available in the kinase-phosphorylation site pair databases. Secondly, they are highly dependent on the number of confirmed phosphorylation sites available for each kinase. The training dataset for all these methods is usually from PhosphoSitePlus [www.phosphosite.org] and PhosphoELM [http://phospho.elm.eu.org/] which store information on kinase-phospho site pairs. At this juncture, PhosphoSitePlus has gathered 80,967 phosphorylation sites in 11,134 distinct proteins, while Phospho.ELM features about 42,000 sites in 8,718 proteins. For the vast majority of these phospho-sites, the kinases that phosphorylate them is not known. Under 10,000 kinase-substrate phospho-sites have been described.

2. Kinase Phospho-Site Specificity

Generally, there is a recognition pattern in the phospho-peptide amino acid sequence that a specific kinase will phosphorylate. We shall refer to this pattern as its kinase phospho-site specificity. This pattern is usually represented by amino acid profile (frequency) matrices that show the frequency of each of the 20 common amino acid at each position surrounding the phospho-acceptor residue of the phospho-peptide. Optimal consensus sequences are also another way of representation of kinase specificity in which the most important amino acids at each position are presented. Other methods such as PSSM matrices and machine learning methods (eg. ANN) generate a score for a given kinase and phospho-peptide. Higher scores show that a kinase is more likely to phosphorylate that phospho-site it is a better match than a lower scoring phospho-site sequence

As we mentioned above, to determine a recognition pattern for a kinase within a phospho-site, all of the aforementioned methods rely on kinase-phospho-site pairs that are found in databases such as PhosphoSitePlus and Phospho.ELM. It is known that phospho-peptides should have at least nine amino acids (centering at phosphorylation site with four amino acids in right and left of the site) to represent the pattern properly, but we decided to explore phosphopeptides of 15 amino acid lengths, because by inclusion of more amino acids we may obtain further information about the specificities for some kinases. After computing the profile matrices of several hundred kinases we determined that there can be some additional specificity information from the added positions -7, -6, 6, and 7 (where 0 is the phosphorylation site, - means left and + means right of the phospho-site). However, increasing the length of the phospho-peptide more than 15 may lead to the higher noise in the training data and makes the prediction task harder.

We created a new PSSM matrix to predict kinase phospho-site specificities, which is computed in the 3 steps described below.

Profile matrix of each kinase. We first compute the probability matrix, called the *profile matrix* for each kinase. Assume that kinase *k* phosphorylates *n* different phospho-site regions $\{p_1, p_2, ...p_n\}$ of length 15, which are found experimentally. The profile matrix P_k of kinase *k* is 21 x 15 matrix, where rows represent amino acids (including unknown amino acid or space ‘x’) and columns represent amino acid residue positions in phospho-site region.

Background frequency of amino acids. Next, we compute the probability of each type of amino acid on the surface of proteins. We refer to it as the *background frequency* of each amino acid type and denote them by B(*i*), where $1 \le i \le 21$. To compute background frequency we use all of 93,000 confirmed phospho-site sequences in the human proteome. The reason is that all of these confirmed phospho-sites are on the surface of the protein, and hence, they can be a good sampling of the protein surface. By looking at the profile matrix of the kinases, we determined that positions -3, -2, 0, and 1 are particularly important and therefore biased for kinase recognition, since all of them had a very low entropy. Therefore, we excluded these positions for the computation of the background frequency of each type of amino acid.

PSSM matrix of each kinase. With a profile matrix for each kinase and the background frequency of each amino acid type, the PSSM matrix for each kinase is typically computed using log odds ratio measure:

$$M(i,j) = \log \frac{P(i,j)}{B(i)} \qquad (1)$$

where $1 \leq i \leq 21$ and $1 \leq j \leq 15$. The problem with this method is that there are many zeros in the profile matrices P which are computed using experimental data, therefore the resulted PSSM matrix will have many $-\infty$ and consequently will not be smooth enough for the prediction. Usually different smoothing techniques [11] are applied here to avoid zeros and $-\infty$'s, but we used a different approach which results in improved matrices for prediction. Equation (2) shows the way we compute PSSM matrix of each kinase.

$$M(i, j) = \mathrm{sgn}(P(i, j) - \mathrm{B}(i)) \cdot |P(i, j)- B(i)|^{1.2} \qquad (2)$$

The exponent 1.2 was determined experimentally to produce the best results.

The logic behind this method is similar to log odds ratio. If the probability of one amino acid i is bigger in the profile matrix than the background frequency then that amino acid is a positive determinant, while if it is less than background frequency it is a negative determinant to be recognized by that specific kinase. For a given candidate phospho-site sequence for a specific kinase, we expect to see more positive and less negative determinant amino acids to predict it as a substrate.

Score of each peptide. With a PSSM matrix Mk for kinase k, we can compute how likely a given candidate phospho-site region $p = a_1a_2 \ldots a_{15}$ is going to be phosphorylated by kinase k. This value is called *kinase specificity score S* and is computed as follows.

$$S(k, p) = \sum_{j=1}^{15} Mk\,(a_j\,, j) \qquad (3)$$

Figure 1. Some of the best characterized protein kinases with respect to their substrate amino acid specificities are shown for the 10 most critical amino acids in their catalytic domains that appear to interact with or influence binding to the –3 amino acid residue position relative the phospho-acceptor amino acid on the protein substrate. The positions of the kinase amino acids are based on the precise alignment in the catalytic domain. Positively charged amino acids are represented as blue, Histidine (H) as light blue (because it much less positively charged that Arginine (R) and Lysine (K)), negatively charged amino acids as red, hydrophobic amino acids as green, and Proline (P) as purple.

Protein Kinase Name	Uniprot ID	Kinase Domain Amino Acid Positions										Substrate AA at -3
		2	27	89	120	152	156	172	195	196	209	
AKT1	P31749	E	Y	E	E	C	E	A	Q	D	E	R
AKT3	Q9Y243	D	Y	E	E	C	E	A	Q	D	D	R
AurA (STK15)	O14965	E	I	E	E	C	D	K	N	T	E	R
AurB (STK12)	Q96GD4	E	I	T	E	C	D	K	A	S	D	R
CAMK2α	Q9UQM7	Q	E	E	E	A	G	P	E	D	P	R
CHK2 (CHEK2)	O96017	I	K	E	E	C	T	A	H	R	I	R
MAPKAPK2	P49137	V	K	E	E	C	Y	S	N	H	P	R
p70S6K (RPS6KB1)	P23443	E	I	E	E	C	E	A	E	N	K	R
PAK1	Q13153	T	E	E	D	V	Y	K	E	N	E	R
PKACα (PRKACA)	P17612	E	H	E	E	C	E	A	D	Q	K	R
PKCα (PRKCA)	P17252	N	L	E	D	C	D	S	E	D	N	R
PKCβ (PRKCB1)	P05771	N	L	E	D	C	D	S	E	D	N	R
PKCδ (PRKCD)	Q05655	I	Y	L	D	C	D	S	D	D	T	R
PKCε (PRKCE)	Q02156	N	V	E	D	C	D	S	D	N	D	R
PKCζ (PRKCZ)	Q05513	D	I	E	D	C	N	S	I	T	P	R
PKD1 (PRKCM)	Q15139	F	D	E	E	V	A	S	D	E	P	R
PKG1 (PRKG1)	Q13976	N	T	D	E	C	E	S	P	D	M	R
ROCK1	Q13464	E	V	E	D	V	D	E	D	S	L	R
RSK1 (RPS6KA2) Dm2	Q15418	E	E	E	S	C	N	A	G	P	S	R
RSK2 (RPS6KA3) Dm2	P51812	E	E	E	S	C	N	A	P	D	S	R
SGK	O00141	H	F	E	E	C	E	T	Q	D	P	R
ERK1	P27361	T	R	N	S	V	W	S	K	H	W	P
ERK2 (MAPK1)	P28482	T	R	N	S	V	W	S	K	H	W	P
GSK3β	P49841	T	L	V	Q	I	Y	S	D	S	T	P
CDK1 (CDC2)	P06493	T	V	S	Q	V	W	P	D	S	L	L
Plk1	P53350	V	V	E	G	C	N	E	S	C	E	L
JAK2 Dm2	O60674	I	E	I	K	Q	P	A	A	L	R	I
CK2α1 (CSNK2A1)	P68400	Q	K	D	H	V	Y	S	G	H	H	E
Lyn	P07948	K	K	L	A	K	K	K	G	R	R	E
Src	P12931	R	R	L	A	K	K	K	G	M	R	E
Abl1	P00519-2	T	T	A	R	K	K	K	G	I	R	D
BARK1 (ADRBK1)	P25098	S	M	E	A	V	G	S	H	K	A	D
CK1α1 (CSNK1A1)	P48729	K	E	M	D	T	D	R	L	K	E	D
EGFR	P00533	K	P	S	R	K	K	Q	G	I	R	D
Lck	P06239	K	K	I	A	K	K	K	G	M	R	D
Syk	P43405	L	T	D	R	K	K	K	G	M	R	D
PDK1 (PDPK1)	O15530	K	E	E	E	V	Q	S	G	N	E	T
CK1δ1 (CSNK1D)	P48730	R	E	L	D	T	R	R	L	K	E	S
GSK3α	P49840	T	L	I	Q	I	Y	S	D	S	T	G
JNK1 (MAPK8)	P45983	Q	N	H	S	V	Y	N	R	D	D	A
	Relative Importance	43	24	39	90	38	62	28	47	27	47	

3. Prediction of Position-specific Scoring Matrices (PSSM) for Kinases without Substrate Data

In this section, we introduce our algorithm for prediction of PSSM matrices based on their catalytic domains. The idea is that those catalytic domains in different kinases that have similar SDRs tend to have similar patterns in the phospho-site sequences of their substrate proteins. To quantify the similarity of catalytic domains of kinases we performed multiple sequence alignment (MSA) of catalytic domains using the ClustalW algorithm The result of the MSA is not quite accurate as it has many gaps and inserts, so the alignments were manually modified. We perform this alignment on all 488 catalytic domains of the typical protein kinases. The length of each kinase catalytic domain after MSA is 247 positions, which includes any gaps or inserts as a single position. For 229 of the domains in the alignment we also compute consensus sequences using 9,125 confirmed kinase–phospho-site pairs. Figure 1 represents key amino acid positions of the catalytic domain after MSA of some of the best characterized kinases for which the most phospho-sites in substrates are known. These amino acid positions appear to be the most critical for recognition of the –3 amino acid residue position before the phospho-acceptor site in the substrate peptide sequence.

In the following parts we use the examples in Figure 1 and will explain how mutual information and charge information are used to find SDRs on the catalytic domains of the kinases.

Mutual Information. Each position in catalytic domains or consensus sequences can be considered as a random variable which can take 21 different values. Both random variables can take any of the 20 amino acids. In addition, the random variables in domains can also take gap value ~, while the random variables in consensus sequence can take 'x' value. In information theory the mutual information of two random variables is a quantity that measures the mutual dependence of the two variables. We can use this measure here to find out which two positions in consensus and catalytic domain are highly correlated. Formally, the mutual information of two discrete random variables X and Y is defined as:

$$I(X, Y) = \sum_{X \in A} \sum_{y \in B} p(x.y) \log \frac{p(x,y)}{p_1(x) p_2(y)} \qquad (3)$$

where $p(x, y) = \mathrm{P}(X = x, Y = y)$, $p1(x) = \mathrm{P}(X = x)$, and $\mathrm{p2(y)} = \mathrm{P}(Y = y)$. The higher mutual information, the more the random variables are correlated. In our context, X is a position in the kinase catalytic domain, Y is a position in the consensus sequence, A is a set of amino acids plus ~ and B is a set of amino acids plus 'x'.

Amino Acids in Protein Kinase Catalytic Domain

Amino Acids in Phospho-Site Region of Protein Substrate

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	X
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	-2	-2	-1	0	1	-1	2	-1	0	0	0	0	2	0.5	0.5	-1	0	0.5	0
E	0	0	-2	-2	-1	0	1	-1	2	-1	0	0	0	0	2	0.5	0.5	-1	0	0.5	0
F	0	0	-1	-1	2	0	-1	2	-1	2	0	0	0	0	-1	0	0	2	0	0	0
G	0	0	0	0	0.5	-1	0.5	0	0.5	0	0	0.5	0	0.5	1	0	0	0	1	0.5	0
H	0	0	1	1	0	0	-1	0	-1	0	0	0	0	0	-1	0	0	0	0	0	0
I	0	0	-1	-1	2	0	-1	2	-1	2	0	0	0	0	-1	0	0	2	0	0	0
K	0	0	2	2	-1	0	-1	-1	-2	-1	0	0	0	0	-2	0	0	-1	0	0	0
L	0	0	-1	-1	2	0	-1	2	-1	2	0	0	0	0	-1	0	0	2	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	2	2	-1	0	-1	-1	-2	-1	0	0	0	0	-2	0	0	-1	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0.5	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0.5	0
V	0	0	-1	-1	2	0	-1	2	-1	2	0	0	0	0	-1	0	0	2	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0.5	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2. Residue interaction matrix R. Rows shows the amino acids in the phospho-site regions of substrates and columns are amino acids in the catalytic domain of kinases. Negatively charged amino acids are orange, positively charged amino acids are blue, hydrophobic amino acids yellow, proline as purple, and phosphorylatable amino acids are represented as gray. 'x' also corresponds to the absence of an amino acid residue, which occurs for phospho-sites located at the N- and C-termini of proteins. This table was derived from knowledge of the structure and charge of the individual amino acid side chains and the nature of their predicted interactions.

Charge Information. Negatively charged amino acid side-chains interact favorably with the side chains of positively charged amino acids, and hydrophobic amino acids with hydrophobic ones. Therefore, if a position in the catalytic domains (see Figure 1) tends to have many negatively charged amino acids and a position in the consensus sequences tends to have more positively charged amino acids, it is likely that these two positions are interacting with each other. Therefore, we define *charge dependency* $C(X,Y)$ of two positions (random variables), one in kinase catalytic domains (X) and the other in consensus sequences (Y), as follows.

$$C(X, Y) = \sum_{i=1}^{n} R(x_i, y_i) \qquad (4)$$

where n is the number of kinases with consensus pairs (in our case 229 kinases were used). R is also residue interaction score of two different amino acids (Figure 2), x_i is the amino acid of the i^{th} kinase at position X of the catalytic domain and y_i is the amino acid of the corresponding consensus sequence at position Y.

Residue interaction matrix shown in Figure 2 estimates the strength of a bond created between amino acids in the average case independent of their distance. Negatively (positively) charged amino acids repel themselves (score -2 in the interaction matrix R) and they attract positively (negatively) charged amino acids (score +2). Histidine (H) has a smaller positive charge than Lysine (K) and Arginine (R). Therefore, scores for it are +1 for interacting with negatively charged amino acids and -1 for interacting with positively charged amino acids. Hydrophobic amino acids attract each other (score +2) while they repel both positively and negatively charged amino acids. S, T and Y residues have a weak tendency to bind to each other (score 0.5), while they are completely neutral with the other amino acids (score 0). For all the amino acids discussed so far, it is not relevant whether they are in the kinase catalytic domain or phosphosite region. In both situations the score is the same, which makes the interaction matrix symmetric. However, Glycine (G) is favored to be in the phospho-site region, because it is a small amino acid residue that creates a pocket on the surface of the region and permits the catalytic domain of the kinase come closer to the phospho-site region. The reason that we did not consider effect of G in the catalytic domain is that we are less clear about the 3D structure of the most kinase catalytic domains, while phospho-site regions are likely to be linear or semi-linear.

In general, the amino acid positions 69, 135, and 161 in the catalytic domains of protein kinases are quite conserved with negatively charged amino acids. These amino acids are found in the highly conserved kinase subdomains V, VII and VIII, respectively. In the example shown in Figure 1, since at (-3) position of the consensus sequences of the peptide substrates of these kinases are mostly positively charged amino acids (e.g. arginine (R) commonly appears), these three positions have a high charge dependency score C and would be considered as strong candidate positions for interaction with (-3) position of the phospho-site regions of the protein substrate. However, in addition to being very well conserved, the amino acids at these positions in the catalytic domain do not correlate well with the (-3) position of the phospho-site regions of a large number of known substrates for kinases where their specificities are well characterized (i.e., when the (-3) position features a negatively charged or neutral amino acid, positions 69, 135, and 161 are still often negatively charged). Therefore, we needed a criterion to combine correlation and charge dependency measures. The following equation combines these two measures.

$$Ce(X,Y) = \sum_{X \in A} \sum_{y \in B} R(x.y)\ .\ p(x.y) \log \frac{p(x,y)}{p_1(x)p_2(y)} \qquad (5)$$

where $Ce(X,Y)$ is called *correlation–charge dependency* of two positions X in catalytic domains and Y in consensus sequences.

Using this hybrid criterion $Ce(X,Y)$ in our example, amino acid position 120 gets the maximum correlation charge dependency in Figure 1. It might be expected that for a particular amino acid position in a substrate surrounding a phospho-site, SDRs in the kinase catalytic domain might reside near each other. For example, positions 120 and 121 might be preferred to positions 120 and 220. However, in the 3D structure of the protein kinase domain, amino acids that are well separated in the primary structure could be situated next to each other, whereas the side chains of neighboring amino acid residues could be orientated in opposite directions. In view of such exceptions, we did not model it mathematically in our equations and algorithms. The following algorithm computes the best SDRs (positions X in the catalytic domain) for each kinase consensus sequence position Y and their interaction probabilities $\mathbf{P}(Y|X) = \mathrm{P}(X, Y)\ /\ \mathrm{P}(X)$ using correlation–charge dependencies. By examination of the x-ray crystallographic 3D structures of 11 protein kinases co-crystallized with peptide substrates, we determined that usually at most seven SDRs may interact with a single amino acid position within the substrate phospho-site region. Therefore, we set the value m in Algorithm 1 to 7. From 516 known human protein kinases, 478 kinases are typical kinases with 488 known catalytic domains and the remaining 38 kinases are all atypical kinases and we have appreciable phospho-site specificity data only for four of them.

Algorithm 1 Computing SDRs

```
    Input: 229 human kinase catalytic domains and their consensus sequences. Parameter $m \leq 247$.
    Output: SDRs and their interaction probabilities for each position in the phospho-site region.
for $j \leftarrow 1, 15$ do
    Let $Y_j$ be the $j$th position in consensus sequences
    for $i \leftarrow 1, 247$ do
        Let $X_i$ be the $i$th position in catalytic domains
        Compute $C_c(X_i, Y_j)$
    end for
    Order positions $X_i$ based on $C_c(X_i, Y_j)$ (decreasingly)
    Let $Z_{j,k}$ be the $k$th position in this order.
    Output $Z_{j,1}, \ldots, Z_{j,m}$ as SDRs for
    position $Y_j$ and interaction probabilities
    $\mathbf{P}(Y_j | Z_{j,1}), \ldots, \mathbf{P}(Y_j | Z_{j,m})$.
end for
```

Next, we present Algorithm 2 which computes the profile and PSSM matrices for 488 catalytic domains in 478 different human kinases and uses the SDRs determined by Algorithm 1. The formula in Line 5 of the Algorithm 2 is based on the observation that those interactions which have higher correlation–charge dependency are more important in estimation of profile matrices.

Algorithm 2 Prediction of PSSM matrices of all kinases.

```
    Input: SDRs and interaction probabilities from Algorithm 1 and 488 catalytic domains.
    Output: Profile and PSSM matrices of all kinase catalytic domains.
1:      > Estimation of the profile matrix of each kinase.
2:  for $k \leftarrow 1, 488$ do
3:      for $j \leftarrow 1, 15$ do
4:          > Estimation of interaction probabilities
5:      Compute $\mathbf{P}(Y_j | Z_{j,1}, Z_{j,2}, \ldots, Z_{j,m})$ as
```

$$\frac{\sum^{m}{}_{l=1}\ C_c(Z_{j,,l,}, Y_j)\ \mathbf{P}(Y_j\ |Z_{j,1})}{\sum^{m}{}_{l=1}\ C_c(Z_{j,,l,}, Y_j)}$$

```
        end for
        Store 21 x15 computed values in profile matrix $P_k$
end for
                > Computing PSSM matrices.
Compute the background frequencies $B$ using the idea described earlier for kinase phospho-site specificity.
Compute the PSSM matrix of each kinase using Equation (2).
```

4. Examples of PSSM's for Human Protein Kinases

We have applied our algorithms to deduce PSSM for 488 human protein kinase domains. To illustrate the power and accuracy of these analyses, six examples of well characterized and distinct protein kinases are provided on the next two pages in Figure 3. For each kinase, we first show the PSSM that is predicted with our algorithm purely from the primary amino acid sequence of the kinase. In the second PSSM, the calculated values for the scoring of each amino acid surrounding the phospho-site is based on the alignment of the amino acid sequences of phospho-sites in known in vitro substrates. Careful comparison of these predicted and empirically derived PSSM's reveals them to be remarkably similar despite the fact that the experimentally generated PSSM uses phospho-site sequences from many in vitro substrates that were not necessarily optimal for the kinases shown.

For 309 kinases we gathered 9,125 confirmed phospho-sites from websites such as PhosphoSite and Phospho.ELM and from the scientific literature. PSSM profiles these kinases were computed using the method of alignment of the amino acids in the phospho-sites targeted by each kinase (empirical matrices) and also with Algorithm 2 (predicted matrices). We performed both methods to evaluate the over all reliability of the predicted matrices. To measure the difference of predicted matrices with empirical matrices we use sum of squared differences. We obtained 309 error values, each one representing how close the predicted matrix is to the empirical one. The vast majority of the predicted matrices were extremely similar to those generated by known substrate alignments. In addition to the numerical results, we confirmed that Algorithm 2 was successful for predicting all the assignments of the serine–, threonine–, and tyrosine–phospho-acceptor specificities correctly.

Figure 3. The PSSM's that are generated using our kinase substrate prediction algorithm (left tables) and from the alignments of phospho-sites in known in vitro substrates of the same 6 protein kinases (right tables) are compared. The short name and Uniprot identification numbers are provided for each of these protein kinases. The "0" position corresponds to the phospho-acceptor amino acid.

Akt1/PKBα
P31749

Predicted Data

	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
A	-3	-4	-5	-4	-7	-3	-2	0	-5	-2	-3	-4	-3	-2	-1
C	4	4	2	2	0	1	3	0	2	7	2	4	4	4	4
D	-2	-3	-2	2	-2	-4	2	0	-5	-1	-3	-3	-3	-2	-2
E	-5	-6	-8	-3	-10	-8	-5	0	-7	-5	-7	-6	-6	-4	-5
F	2	16	0	1	0	0	1	0	46	1	4	5	2	2	2
G	5	5	-1	-2	-6	-7	-1	0	-6	-4	2	-4	-2	-2	-2
H	3	2	1	1	-1	1	1	0	0	1	1	2	2	5	3
I	1	0	0	0	-2	-1	1	0	-1	0	0	2	1	1	1
K	0	0	-4	0	0	-6	2	0	-1	0	-2	-2	0	-2	-2
L	-4	-5	2	-5	-4	0	9	0	9	-5	-5	-4	0	-3	-2
M	3	2	2	1	0	0	3	0	1	2	2	2	6	3	4
N	1	0	0	1	-2	-2	1	0	-1	1	4	0	1	1	1
P	-4	-5	-5	-7	-9	1	-5	0	0	-4	-4	-6	-3	-1	-5
Q	0	0	-1	0	-3	-1	-1	0	-1	-1	-1	0	0	0	0
R	-2	-1	39	0	105	30	-3	0	-6	12	15	0	-3	-3	0
S	-11	-11	-12	11	-9	-1	-11	58	-14	-6	-8	5	-10	-7	-10
T	-1	0	-2	-3	-4	11	-2	6	-4	-2	-3	0	2	0	-1
V	-1	0	-2	-1	-5	-4	-1	0	1	0	0	1	-1	-1	0
W	5	4	3	3	1	1	3	0	2	3	3	4	4	5	5
Y	2	1	1	1	-1	0	1	1	0	1	1	1	2	4	4

Empirical Data

	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
A	0	4	-6	-3	-6	2	2	0	3	-1	3	2	-1	0	-3
C	0	-1	2	0	-1	0	2	0	0	0	0	0	1	0	0
D	-6	-3	-8	-1	-7	-5	-5	0	-3	4	1	-5	-1	-6	1
E	-4	-6	-10	-1	-12	-9	-8	0	-1	0	2	-4	-1	-2	-1
F	-1	8	-3	-2	-2	-1	0	0	12	-2	4	0	2	0	2
G	7	3	-8	-3	-10	-7	-1	0	-2	3	-1	2	2	-2	0
H	-2	-2	-1	0	-2	0	6	0	-2	-2	3	0	3	0	0
I	1	-1	-3	-1	-3	-1	-1	0	6	0	0	0	0	1	0
K	-3	1	-2	1	-6	0	-1	0	-6	-4	-3	-2	-3	4	-2
L	1	0	-5	-2	-9	0	-2	0	2	-8	-1	0	0	-1	-1
M	-1	2	-1	2	-2	0	2	0	9	0	0	-1	-2	1	2
N	-2	-2	-4	0	-4	-2	6	0	0	2	2	0	0	0	0
P	3	0	-3	8	-12	-3	-6	0	-8	3	-6	-3	0	1	-3
Q	4	3	-2	0	-5	-1	0	0	-3	-3	-2	0	-3	0	2
R	3	3	147	16	201	15	1	0	-9	3	-2	1	5	2	1
S	0	-4	-16	-3	-18	9	2	54	0	1	6	8	-3	6	2
T	4	0	-5	-3	-7	12	0	12	0	5	-2	0	-3	-1	-5
V	-2	-1	-7	1	-7	0	3	0	2	0	0	1	1	-1	2
W	0	1	-1	-1	-1	-1	1	0	2	0	-1	0	1	0	0
Y	-1	0	-2	-1	-1	-1	0	0	0	0	-3	1	0	-1	1

EGFR
P00533

Predicted Data

	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
A	-3	-3	-1	-5	-5	-5	-3	0	-3	-2	-2	-2	-3	-2	-1
C	4	4	3	2	2	2	1	0	2	4	3	5	4	4	4
D	0	-2	0	0	35	18	19	0	0	0	-3	-2	-2	-2	-2
E	0	-4	-5	19	11	1	0	0	-5	-3	-6	-4	-1	-5	-5
F	4	2	1	0	1	1	0	0	5	1	31	2	2	2	2
G	-3	2	-1	-4	-5	-3	-4	0	-4	-4	-3	-3	0	-3	-3
H	2	2	2	1	1	3	0	0	0	2	1	2	3	3	3
I	1	1	0	0	0	0	1	0	-1	0	0	1	1	3	1
K	0	-1	-3	-4	-4	-4	0	0	-5	-1	-3	-2	-2	-2	-2
L	-3	-3	-1	-6	-3	-3	7	0	0	-5	-5	-4	-4	-3	-3
M	3	3	2	1	2	1	2	0	2	2	3	3	3	3	3
N	1	1	0	10	-1	1	1	0	-1	11	0	3	1	1	1
P	-5	-3	-4	-6	-8	0	-7	0	4	-2	-2	-1	1	2	-5
Q	0	1	0	-1	-1	-1	-1	0	0	0	0	0	0	0	0
R	-2	-1	8	-4	-2	-3	-4	0	-6	-1	-3	-2	-3	-3	1
S	-11	-11	-9	-3	-13	-10	-12	13	-13	-9	-9	-7	-10	-9	-10
T	1	0	-1	-1	-2	-1	-3	5	-3	0	-2	1	-1	-1	-1
V	-1	0	-1	-1	-3	-2	0	0	25	0	-1	0	-1	-1	0
W	5	4	4	3	2	3	2	0	2	4	3	4	5	5	5
Y	2	2	3	1	0	2	0	50	1	2	1	2	2	3	4

Empirical Data

	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
A	0	0	3	-4	-1	5	-4	0	-1	-4	5	-1	-1	-9	3
C	-1	-1	-1	-1	-1	-1	1	0	3	1	-1	-1	1	-1	-1
D	1	8	15	11	18	8	6	0	-1	3	1	-3	-3	-3	-1
E	11	2	5	24	14	2	2	0	0	0	-10	2	5	-7	0
F	3	1	-1	6	-1	-1	1	0	-3	-1	15	6	-1	1	3
G	-2	10	5	5	-4	0	-4	0	-2	-4	-7	-7	-4	14	0
H	0	2	0	0	2	2	-2	0	0	0	0	2	0	-2	-2
I	5	-4	-2	-4	-2	-2	14	0	5	5	7	2	-4	0	0
K	-3	-6	-1	-3	1	-3	-6	0	-6	-1	-6	-3	5	11	5
L	-5	0	2	-5	0	-8	-5	0	7	-11	0	-5	2	4	-5
M	2	0	0	-2	-2	-2	-2	0	-2	0	7	0	-2	5	7
N	-2	0	0	5	7	5	5	0	2	17	-2	2	5	5	-4
P	-6	-1	-1	-6	-6	5	-4	0	-6	-4	8	5	11	-1	0
Q	-3	1	-3	1	4	9	-3	0	4	6	-1	1	4	1	-3
R	0	-7	-7	-7	-7	-10	0	0	-7	0	-7	-2	-5	-7	-2
S	-9	-9	-6	-6	-12	-9	-4	0	-6	0	-12	-6	-12	0	3
T	16	2	-2	-2	-7	-2	4	0	-5	-2	-2	4	4	-5	4
V	-4	0	0	-2	2	2	-2	0	17	-2	5	0	-2	-4	-7
W	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	1	-1	-1	-1
Y	-3	0	-3	-3	0	-3	4	61	4	-3	0	0	-3	0	-3

ERK2
P28482

Predicted Data

	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
A	-3	-2	-4	-4	-4	-6	-4	0	-7	-1	2	-3	-3	-3	0
C	4	5	3	3	1	1	2	0	0	4	3	3	4	4	4
D	-3	-2	-2	0	-1	-3	-3	0	-4	0	-3	-4	-3	-1	-2
E	-5	-4	-6	-6	-8	-2	-7	0	-10	-7	-6	-7	-3	-5	-5
F	2	2	1	3	2	-1	2	0	2	1	2	1	2	2	2
G	1	0	-1	-4	-4	-7	-4	0	-7	-5	-1	-4	-3	-2	-3
H	5	2	1	1	0	0	0	0	0	1	1	1	2	2	2
I	1	2	0	0	-2	-1	2	0	0	0	0	0	1	3	1
K	-1	-2	-3	-4	-4	-4	-2	0	-6	-1	-1	-1	-2	-3	-2
L	-4	-4	1	-5	-3	-1	26	0	0	-5	-5	-4	-3	0	-3
M	3	3	2	2	4	1	5	0	0	2	3	2	3	3	3
N	1	1	0	0	-2	-1	0	0	-2	2	1	1	1	1	1
P	-2	-3	8	0	45	39	1	0	99	-2	10	6	0	-1	0
Q	0	0	-1	0	-3	1	-1	0	-3	-1	0	-1	0	0	0
R	-2	-1	4	-4	14	3	-2	0	-7	1	-1	-2	-3	-4	-1
S	-9	-6	-11	18	-15	-12	-14	54	-14	-9	-9	-4	-9	-10	-8
T	3	-1	-1	-2	-3	6	-3	7	-5	18	-2	8	10	0	-1
V	-1	0	-2	-2	-4	-4	0	0	-3	0	-1	0	-1	-1	0
W	4	4	3	3	1	1	2	0	1	3	3	3	4	4	5
Y	2	2	2	1	0	0	0	6	0	1	1	1	2	9	3

Empirical Data

	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
A	1	2	0	-1	1	1	3	0	-8	-1	4	0	2	4	4
C	0	0	0	1	0	0	0	0	-1	1	0	0	1	0	0
D	-1	-1	-3	-2	-3	-4	-6	0	-8	-5	-3	-1	-2	-3	0
E	-2	-3	-4	-3	-4	-8	-8	0	-12	-6	-4	-7	-4	-6	-4
F	0	1	0	-1	0	-1	0	0	-2	0	1	0	1	2	0
G	0	2	-1	1	3	-6	3	0	-10	2	-2	1	1	1	-1
H	-1	0	0	-2	-1	-1	-1	0	-2	0	1	0	0	-1	-1
I	1	1	2	0	0	1	0	0	-4	0	-2	0	-1	0	0
K	1	-3	-3	-1	-4	-5	-2	0	-8	-3	-2	-2	-5	-2	-4
L	-3	-3	3	-1	0	5	7	0	-10	2	1	1	-2	0	0
M	0	1	1	0	-1	-1	4	0	-2	0	0	1	0	0	2
N	3	1	1	2	-1	-3	-1	0	-4	0	-2	-1	-2	1	-1
P	5	3	9	6	15	62	5	0	203	4	7	9	6	4	4
Q	1	0	-1	0	1	-2	0	0	-4	1	2	0	-1	0	-1
R	-5	-6	-2	-4	-3	-5	-1	0	-9	0	-2	-4	0	0	-3
S	-1	2	2	3	0	-6	-4	44	-18	0	1	1	7	3	0
T	5	2	2	3	3	-1	5	22	-6	7	1	5	1	0	2
V	-3	0	-1	1	0	1	-2	0	-6	1	0	0	-1	-1	0
W	0	-1	-1	0	0	0	0	0	0	0	0	0	0	-1	0
Y	0	0	0	0	-1	-1	0	1	-2	-1	1	0	0	0	2

			-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7			-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
PKAα P17612	Predicted Data	A	-2	-3	-4	-4	-7	-2	-3	0	-5	-1	-1	-3	-3	-2	-2	Empirical Data	A	1	1	0	1	-4	-2	1	0	0	2	3	2	-1	-2	-1
		C	4	4	2	2	0	1	2	0	2	4	3	4	4	4	4		C	0	0	0	0	-1	0	0	0	0	1	1	1	0	0	1
		D	-2	-2	-2	2	-3	-4	4	0	-4	-2	-1	-2	-3	-2	-2		D	-3	-2	-4	-5	-8	-8	-4	0	-2	0	-2	-2	-2	-1	-2
		E	-5	-6	-7	-4	-7	-8	-7	0	-8	1	-6	-4	-6	-2	-5		E	-1	-2	-5	0	-10	-11	-7	0	-5	-3	1	0	0	-2	-2
		F	2	5	1	1	0	0	1	0	19	1	3	2	2	2	2		F	0	1	2	0	-2	0	0	0	7	1	1	2	0	1	0
		G	0	1	0	-3	-6	-6	-2	0	-5	-4	1	-3	-2	-2	-2		G	1	1	-2	0	-7	-6	4	0	-2	-2	3	-2	1	0	-2
		H	3	2	1	1	-1	2	1	0	1	1	1	2	2	5	3		H	-1	1	0	0	-1	-1	0	0	-1	0	0	0	0	1	1
		I	1	1	0	0	-2	-1	0	0	1	0	1	1	1	1	1		I	0	0	1	0	-2	-3	0	0	4	0	-1	3	1	0	0
		K	0	0	-4	0	0	-5	1	0	-1	-1	-1	-1	1	-2	-2		K	1	3	2	3	8	14	0	0	-4	-4	-2	-2	0	-1	1
		L	-3	-4	1	-5	-3	-3	10	0	11	-4	-5	-2	-3	-3	-1		L	-1	-1	1	0	-6	-5	6	0	8	0	0	2	0	2	3
		M	3	3	2	2	0	1	7	0	1	2	4	3	3	3	3		M	0	0	0	0	-1	-1	1	0	1	-2	0	0	0	0	0
		N	1	1	0	0	-2	-1	1	0	0	2	0	2	1	1	1		N	0	0	-1	-1	-3	-1	1	0	-1	0	0	-1	-1	2	0
		P	-3	-2	-4	-6	-9	0	-4	0	-3	-6	-2	-5	-1	-2	-4		P	0	-2	-3	-4	-5	-10	0	0	-3	1	-2	-3	-1	-1	-1
		Q	0	0	-1	1	-3	1	-1	0	1	-1	0	0	0	0	0		Q	1	1	1	3	-1	-2	-1	0	-1	2	-2	2	0	-1	0
		R	-2	-1	30	0	106	35	-2	0	-5	4	1	-1	-2	-3	-2		R	1	0	10	3	100	86	1	0	-2	-1	-2	0	-1	1	0
		S	-11	-11	-12	7	-12	-1	-12	58	-9	-7	-5	-3	-8	-6	-10		S	-3	-2	-3	0	-11	-8	-2	54	-4	1	0	-5	0	-3	-3
		T	-1	-1	-1	-2	-5	3	-3	6	-3	0	-2	0	5	-1	-1		T	0	-1	-1	1	-5	-1	-1	11	0	-1	0	0	0	-1	0
		V	0	0	-2	-1	-5	-4	0	0	3	1	2	0	-1	-1	0		V	-1	1	-1	-3	-2	-5	1	0	5	4	-1	3	1	1	1
		W	5	5	3	3	1	2	3	0	3	4	3	4	4	5	5		W	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0
		Y	3	2	1	1	-1	0	1	1	1	1	2	2	2	4	6		Y	0	-1	0	-1	-1	-2	0	0	1	-1	0	0	1	1	1
PKCα P17252	Predicted Data	A	-2	-2	-4	-4	-8	-4	-1	0	-6	-2	-4	-4	-3	-2	-2	Empirical Data	A	0	0	-2	-1	-2	0	4	0	-2	-3	-1	-1	0	1	0
		C	4	4	3	3	0	0	2	0	1	3	2	2	4	4	4		C	1	0	0	-1	0	0	-1	0	0	0	0	0	0	0	1
		D	-2	-2	-1	0	0	-2	1	0	-5	1	-3	-4	-3	-2	-2		D	-4	-3	-5	-4	-6	-5	-3	0	-5	-4	-2	-3	-2	-2	-3
		E	-5	-5	-7	-4	-11	-9	-7	0	-7	-7	-7	-5	-6	-4	-4		E	-3	-4	-6	-5	-9	-8	-6	0	-7	-10	-2	-6	-6	-1	-2
		F	2	2	1	14	-1	-1	1	0	70	0	1	1	2	2	2		F	2	1	1	3	1	-2	0	0	10	0	-1	1	2	0	2
		G	0	1	-3	-3	-8	-7	-4	0	-7	-5	3	-5	-3	-2	-2		G	-1	0	-1	0	-4	-1	1	0	-3	-6	3	-1	-2	-1	0
		H	3	3	1	1	-1	0	1	0	0	1	1	1	2	10	3		H	0	1	0	0	0	-2	0	0	-1	-1	0	0	0	1	0
		I	1	1	0	0	-3	-1	0	0	-1	-1	0	1	1	1	1		I	0	1	0	1	-1	-1	-1	0	2	0	0	-1	0	0	0
		K	0	-2	-4	0	2	-6	24	0	2	2	-2	18	2	-3	-2		K	0	4	5	3	7	3	3	0	5	33	6	8	6	4	0
		L	-3	-4	4	-5	-5	-7	2	0	1	-5	-5	-4	-1	-3	-1		L	-2	0	5	0	-4	-4	0	0	4	-4	-3	-1	3	0	0
		M	3	3	2	2	0	0	4	0	0	1	2	1	5	3	3		M	2	0	1	0	-1	0	1	0	1	-1	0	0	1	0	2
		N	1	1	1	0	-3	-2	0	0	-2	1	0	0	1	1	1		N	2	-1	0	0	-2	0	2	0	-1	-1	0	0	0	0	0
		P	-1	-4	-6	-7	-11	-3	-5	0	0	-4	0	-5	-3	-3	-4		P	-5	-1	-1	-5	-5	-4	-1	0	-9	-6	-3	-3	-3	-2	0
		Q	0	0	-1	-1	-4	-1	-1	0	-3	-1	-1	-1	0	0	0		Q	2	0	-1	0	0	2	-1	0	3	-1	-2	0	0	1	0
		R	-3	-1	24	-3	125	97	-4	0	-7	32	25	8	-2	-4	-1		R	1	4	4	5	42	30	4	0	0	34	17	3	1	2	-1
		S	-10	-11	-10	6	-11	-13	-12	59	-14	-9	-10	-3	-11	-7	-11		S	-1	-3	-3	0	-5	-3	-2	53	-5	-9	-7	2	-1	-6	-3
		T	-1	0	-2	-2	-6	0	-2	6	-4	-3	-3	-2	5	0	-1		T	1	0	0	1	0	2	-1	11	0	-1	-1	0	0	0	0
		V	0	3	-2	-1	-5	-4	-2	0	2	-1	-1	-1	-1	-1	0		V	-1	0	-1	-1	-1	0	-2	0	4	-2	-1	-1	0	0	0
		W	5	5	3	3	0	1	3	0	1	2	3	3	4	4	5		W	1	0	1	0	0	0	0	0	0	0	0	0	-1	0	0
		Y	3	2	1	1	-1	0	1	1	0	0	1	1	2	3	6		Y	1	0	0	0	-1	0	0	0	2	-1	1	1	1	1	0
Src P12931	Predicted Data	A	-3	-2	-2	-4	-6	-4	-4	0	-4	-1	-5	0	-2	-2	-1	Empirical Data	A	-1	-3	1	-4	-1	0	-3	0	0	1	-1	-3	-2	-2	-3
		C	4	6	4	3	1	2	2	0	2	5	2	5	4	4	5		C	0	-1	0	0	-1	0	0	0	-1	0	0	0	0	0	0
		D	1	-2	0	0	35	15	3	0	-3	-2	-3	-3	-2	-2	-2		D	2	2	2	8	9	8	4	0	6	0	-2	1	0	-2	-1
		E	0	-2	-5	20	15	0	-6	0	-4	-5	-8	-5	-3	-5	-5		E	1	0	2	9	9	3	0	0	5	1	-6	1	-2	0	4
		F	6	2	2	3	2	0	0	0	2	2	1	2	2	2	3		F	1	3	1	-1	-2	-1	0	0	-1	-1	4	0	0	1	3
		G	-3	0	-2	-4	-4	0	-6	0	3	-3	-4	-3	-1	-3	-1		G	-2	3	3	5	0	4	-2	0	8	0	-1	1	6	1	4
		H	4	2	2	1	0	2	0	0	1	2	1	2	3	4	3		H	0	0	0	0	0	2	3	0	0	0	0	1	0	0	0
		I	1	1	1	0	-1	0	13	0	0	0	0	1	1	1	1		I	0	2	3	0	-2	-1	6	0	1	-1	3	0	-1	1	1
		K	-2	-2	-3	-4	-3	-4	-3	0	-4	-1	-4	-1	-2	-2	-2		K	0	-1	-2	-2	-4	-6	-4	0	-1	-2	-6	-3	1	2	0
		L	-4	-3	1	-5	-7	-5	2	0	-3	-4	0	-3	-2	-3	-3		L	-1	0	-1	-2	-6	-3	0	0	-3	-3	7	-4	-4	1	-2
		M	3	3	3	2	0	1	5	0	3	3	2	3	3	3	4		M	0	0	0	-1	0	-1	0	0	-1	2	1	0	0	1	2
		N	1	1	1	5	-1	4	0	0	0	2	0	2	1	1	1		N	-1	-1	0	2	2	3	0	0	-1	4	0	1	0	0	-1
		P	-5	-4	-3	-4	-9	-5	-8	0	7	0	21	-2	0	-1	-4		P	0	-1	-1	-4	-2	5	-4	0	-9	-3	6	1	0	0	-1
		Q	0	1	0	-1	-2	-1	1	0	2	0	0	0	0	0	0		Q	0	1	1	0	3	0	-1	0	4	2	-1	1	2	1	0
		R	-3	-2	1	-5	0	2	-4	0	-5	4	-5	0	-3	-3	-1		R	2	0	-2	0	-1	-3	-6	0	-3	-3	-5	0	-1	-2	-1
		S	-11	-9	-10	-5	-7	-11	-13	6	-13	-9	-14	-7	-9	-10	-10		S	-5	-9	-5	-6	0	-4	-8	1	-5	-3	-7	-7	-5	-4	-2
		T	0	2	0	-2	-3	0	-1	1	-2	-1	-3	0	-1	4	-1		T	0	-1	0	-1	0	-3	2	1	-1	2	-2	-1	2	0	-2
		V	-1	0	-1	-2	-3	-2	15	0	13	-1	21	0	-1	-1	0		V	1	2	-2	-2	1	-1	11	0	0	2	9	5	0	1	-1
		W	5	4	4	3	2	3	2	0	3	4	2	4	5	5	5		W	1	0	-1	0	0	0	0	0	1	1	0	1	0	0	0
		Y	2	2	3	1	0	2	1	58	1	2	1	2	2	3	4		Y	6	3	2	1	0	0	5	62	5	2	3	2	4	0	0

5. Comparison of Optimal Recognition Sequences for Protein Kinases from Prediction and Consensus from Empirically-derived Substrate Phospho-site Alignments

To further evaluate the reliability of our kinase substrate prediction matrices, we compared the optimal consensus sequences for kinase recognition based on the predicted PSSM and those generated from the alignment of known substrate phospho-sites for each kinase. Figure 4 below provides the deduced optimal phosphorylation site sequences for 45 of the best characterized human protein kinases for their substrate selectivities. The predicted optimal phospho-site sequences feature more information about amino acid residues that could contribute to binding to the target kinase. Careful inspection of this figure reveals an amazing convergence in the predicted optimal phosphorylation site sequences generated by these two methods for the large majority of the protein kinases. However, for some kinases there are some discrepancies. It should be appreciated that the consensus alignment method is based on the use of sequences of phospho-sites in physiological proteins. These phospho-sites, while phosphorylated in vitro by a kinase, may not necessarily feature the best amino acid sequences for recognition by that kinase. Physiological phospho-sites have probably evolved to permit the recognition by multiple protein kinases in vivo.

Figure 4. The predicted most favorable amino acids including and surrounding the phosphorylation sites (P-sites) targeted by 45 well characterized protein kinases are provided. The predicted optimal P-sites using our algorithms and the catalytic domain sequences of the kinases are shown on the left. The consensus P-sites based on alignment of known in vitro substrates of these kinases is given on the right. The number of P-site peptides that were used to generate the consensus phosphorylation site sequences is indicated in a middle column. Amino acids letters provided in upper case represent the most critical amino acids in positions for kinase recognition; lower case letters correspond to lesser important amino acids. The “0” position corresponds to the phospho-acceptor amino acid. The short name and Uniprot identification numbers are provided for each of these protein kinases.

		Predicted Optimal Recognition Sequences																**Empirical Consensus Recognition Sequences**														
Kinase Name	Uniprot ID	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	# P-site peptides	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
Abl	P00519-2	e	x	x	n	D	d	e	Y	f	r	P	x	e	x	x	73	x	x	x	n	d	x	x	Y	x	x	P	x	x	x	x
Akt1/PKBα	P31749	g	f	R	s	R	R	l	S	F	r	r	s	x	x	x	158	g	f	R	r	R	r	x	S	f	x	x	s	x	x	x
Akt3/PKBγ	Q9Y243	g	f	R	s	R	R	l	S	F	r	r	s	x	x	x	45	x	f	R	x	R	t	x	S	f	x	x	x	x	x	x
AurA	O14965	x	x	r	s	R	R	r	S	l	V	k	k	x	x	x	40	x	x	n	x	r	R	x	S	l	x	p	x	k	x	x
AurB	Q96GD4	x	s	r	s	R	R	r	S	l	V	k	k	t	x	x	48	k	g	s	x	r	R	r	S	x	v	x	k	x	x	x
BARK1 (GRK2)	P25098	x	v	r	d	R	T	D	S	p	v	p	S	t	x	x	49	x	x	x	d	x	x	x	S	d	x	x	g	x	x	x
CaMK2α	Q9UQM7	x	p	L	s	R	q	d	S	f	d	r	s	x	x	x	178	x	x	l	s	R	q	x	S	l	d	x	x	x	x	x
CDK1	P06493	x	x	r	s	r	P	l	S	P	x	K	k	x	l	x	393	x	x	x	x	l	p	x	S	P	x	k	k	x	x	x
CDK2	P24941	x	x	r	s	R	P	l	S	P	x	K	k	p	l	x	201	l	l	x	x	x	p	x	S	P	g	K	k	x	l	x
CDK5	Q00535	x	x	r	s	r	P	d	S	P	x	K	k	x	x	x	147	x	x	x	s	x	p	l	S	P	x	k	x	x	x	x
Chk2	O96017	p	x	L	s	P	p	l	S	p	r	p	k	x	x	s	42	p	x	l	x	R	x	x	S	q	x	x	k	x	x	x
CK1α1	P48729	g	g	l	s	R	R	a	S	p	x	g	x	x	s	x	102	g	x	x	x	d	d	d	S	x	x	g	x	x	x	x
CK1δ	P48730	g	x	l	s	S	r	a	S	p	x	s	x	x	s	x	66	x	x	x	x	ps	x	a	S	x	x	x	x	x	s	x
CK2α1	P68400	x	a	r	e	e	p	d	S	D	d	E	E	e	e	e	483	x	x	x	e	e	e	d	S	D	d	E	e	e	e	e
EGFR	P00533	x	x	r	e	D	d	d	Y	v	n	f	x	p	p	x	58	x	x	x	e	d	x	x	Y	v	n	f	x	x	x	x
ERK1	P27361	x	x	p	S	P	P	L	S	P	t	p	p	t	x	x	292	x	x	p	p	p	P	l	S	P	t	p	t	x	x	x
ERK2	P28482	x	x	p	s	P	P	L	S	P	t	p	p	t	x	x	410	x	x	p	x	p	P	l	S	P	t	p	p	x	x	x
Fyn	P06241	x	x	x	e	D	d	v	Y	p	r	p	x	x	x	x	120	x	x	x	x	x	x	v	Y	x	x	v	x	x	x	x
GSK3α	P49840	x	x	l	s	r	P	l	S	P	p	p	s	p	p	x	40	x	x	x	s	g	p	p	S	P	p	p	S	p	x	x
GSK3β	P49841	x	x	r	s	r	P	p	S	P	p	p	s	p	x	x	175	x	x	x	S	p	p	p	S	P	p	p	S	p	x	x
InsR	P06213	x	x	x	e	r	d	d	Y	v	d	v	x	x	x	x	73	x	x	x	e	x	d	d	Y	m	x	M	x	x	x	x
JAK2	O60674	x	x	r	s	r	p	d	Y	E	x	v	x	r	x	x	49	x	x	x	x	l	d	x	Y	I	k	I	x	x	x	x
JNK1	P45983	x	x	r	S	R	L	L	S	P	x	p	x	x	l	x	122	x	x	x	s	a	l	x	S	P	x	a	x	x	x	x
JNK2	P45984	s	x	r	S	R	L	L	S	P	t	p	s	t	x	x	57	s	x	x	s	x	L	l	S	P	x	x	x	x	x	x
Lck	P06239	e	x	l	e	D	d	l	Y	v	r	p	x	p	p	x	98	x	x	x	x	d	d	l	Y	x	x	v	x	x	x	x
Lyn	P07948	e	g	r	e	D	d	l	Y	f	x	p	t	x	p	x	85	x	e	x	x	e	n	I	Y	e	x	p	x	x	p	x
MAPKAPK2	P49137	t	x	L	s	R	r	p	S	l	r	r	s	x	x	x	59	x	x	L	x	R	s	p	S	l	x	x	x	x	x	x
p38α	Q16539	p	x	l	S	R	P	L	S	P	x	p	p	t	l	e	178	x	x	x	x	x	P	l	S	P	x	x	p	x	x	x
p70S6K	P23443	x	x	R	s	R	R	l	S	f	s	s	s	x	x	x	40	x	x	R	s	R	t	x	S	x	s	s	x	x	x	x
PAK1	Q13153	x	t	r	s	R	R	k	S	v	r	g	k	l	x	x	73	x	t	x	x	R	R	r	S	v	x	g	x	l	x	x
PDK1	O15530	r	s	g	s	R	s	l	S	F	c	g	t	t	e	y	43	x	x	g	x	t	t	x	T	F	C	G	T	p	e	Y
PKACα	P17612	x	x	R	s	R	R	l	S	l	r	s	s	t	x	x	734	x	x	r	x	R	R	l	S	l	x	x	x	x	x	x
PKCα	P17252	x	x	r	s	R	R	k	S	F	R	r	k	x	h	x	523	x	x	x	x	R	R	x	S	f	K	r	k	k	x	x
PKCβ1	P05771	x	x	r	s	R	R	k	S	F	R	r	r	k	x	x	86	x	x	x	f	r	r	k	S	f	R	x	x	x	x	x
PKCδ	Q05655	x	x	r	s	R	R	k	S	F	r	r	r	x	x	x	109	x	x	x	x	R	R	x	S	f	R	r	x	x	x	x
PKCε	Q02156	x	g	R	s	R	R	K	S	F	R	r	r	k	x	x	72	x	x	x	k	R	R	k	S	f	R	r	r	x	x	x
PKCζ	Q05513	x	g	R	s	R	R	K	S	F	R	r	r	x	s	x	71	x	x	x	x	r	r	k	S	F	r	r	x	x	x	x
PKD1	Q15139	r	s	L	s	R	R	l	S	l	v	f	f	x	x	x	37	x	p	L	x	R	r	x	S	x	v	x	x	x	x	x
PKG1	Q13976	g	r	R	s	R	R	l	S	f	d	r	s	x	s	x	85	x	x	r	x	R	R	x	S	x	x	x	v	x	x	x
Plk1	P53350	x	l	r	s	R	p	d	S	f	t	p	x	x	p	y	97	x	x	l	x	l	d	d	S	x	v	x	x	x	x	x
ROCK1	Q13464	x	g	R	s	R	R	l	S	v	k	v	s	x	x	x	66	x	x	r	x	r	R	x	S	x	v	x	x	x	x	x
RSK1	Q15418	g	r	R	s	R	r	l	S	f	k	x	s	x	s	w	66	x	x	R	x	R	x	x	S	x	x	x	x	x	x	x
SGK	O00141	x	x	R	s	R	s	l	S	f	r	v	r	t	x	x	46	x	x	R	s	R	s	x	S	x	x	x	x	x	s	x
Src	P12931	x	x	x	e	D	d	v	Y	p	r	p	x	x	x	x	385	x	x	x	e	e	d	v	Y	g	x	v	x	x	x	x
Syk	P43405	e	e	r	e	D	d	D	Y	E	r	P	x	e	p	x	68	x	e	d	d	d	d	D	Y	E	x	p	x	e	x	x

6. Derivation of PSSM's for Yeast Protein Kinases

In view of the high degree of success that we obtained with the application of our algorithms to diverse human protein kinases for prediction of their specificities, we tested the ability of our algorithms to predict PSSM's for budding yeast based on their primary amino acid structures. In a recent study involving a consortium of yeast research laboratories, the substrate specificities of 60 recombinant yeast protein kinases were investigated with the use of peptide arrays [Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. Mok J, Kim PM, Lam HY, Piccirillo S, Zhou X, Jeschke GR, Sheridan DL, Parker SA, Desai V, Jwa M, Cameroni E, Niu H, Good M, Remenyi A, Ma JL, Sheu YJ, Sassi HE, Sopko R, Chan CS, De Virgilio C, Hollingsworth NM, Lim WA, Stern DF, Stillman B, Andrews BJ, Gerstein MB, Snyder M, Turk BE. (2010) Sci Signal. 3(109):ra12.] In Figure 5, we show the experimentally obtained optimal consensus sequences of 50 of the yeast protein kinases as deduced from the PSSM's provided in the Mok et al. (2010) study. We omitted those 10 yeast protein kinases for which no obvious specificities were evident or phosphorylated amino acids appeared to be a major part of the PSSM's. Also in Figure 5, we provide the optimal amino acid sequences for these same yeast protein kinases as predicted from the PSSM's generated from the primary amino acid sequences of their catalytic domains with our algorithms. Comparison of the optimum kinase substrate specificities of the 50 kinases from Saccharomyces cerevisiae determined by the two methods reveals a very high degree of concordance. Interestingly, while basic amino acids on the N-terminal side of the phospho-acceptor amino acid residue for most of the yeast kinases represented strong positive recognition determinants, the unusual amino acid tryptophan was often identified as a positive recognition determinant for yeast kinases. In the human protein kinases, such a specificity for tryptophan was not as evident.

In view of these findings, there is a strong prospect that our algorithms can be used to successfully predict the substrate specificities of typical kinases from very diverse animal, plant and fungi species. At the very least, our method provides the starting structures of promising parent substrate peptides that could be used to generate analogues for further improvement in kinase recognition and binding. Such peptide analogues can be synthesized on peptide arrays and tested with our In Vitro Kinase and Phosphopeptide Testing (IKPT) Services (http://www.kinexus.ca/ourServices/substrate_profiling/phosphopeptide_kinase.html).

Kinase	Uniprot	Predicted Optimal Recognition Sequences										Empirical Consensus Recognition Sequences									
		-5	-4	-3	-2	-1	0	1	2	3	4	-5	-4	-3	-2	-1	0	1	2	3	4
Akl1	P38080	L	R>h	R>P	RQ	Y>RK	T>S	G>ip	G>py	Y	x	L>mi	k	pr	Qh	rkh	T>S	G	g	y	w
Ark1	P53974	L	R>k	R	R>Q	Y>rk	T>S	G>p	Y>gr	fry	I	li	rk	R	rq	yrh	TS	G	x	x	(pT)
Atg1	P53104	R	R>kw	L>R	R>qsi	R	S>T	V	Y>ir	Y	Y	x	w	lm	s	x	S	cfiv	yfimvw	yflw	d
CAK1	P43568	rld	R>W	R>I	S>r	R>K	S>T	DP	YI	Y	ylkt	rw	x	r	ft	wyf	T>S	nah	yivrc	rvc	r
CDC15	P27636	R	R>Q	R>t	D>rq	kry	S>T	Fv>ig	Y	Y	y	x	rpn	rk	yspng	gknpqr	T>s	rmi	Rk	kr	kr
Cdc28	P00546	R>lk	w/h	R>lw	S>P	YR	S>T	P	P	P	K>t	x	w	c	stpml	r	S>T	P	prk	prks	krh
Cdc5	P32562	r/l	R>w	R	D>P	H>R	S>T	F>I	I>Y	fn	ly	w	r	rwy	D>en	hikrtv	ST	F>my	yidlv	hlvwy	av
Cdc7	P06243	DL	R	R	I>R	Ky	S>T	D>P	R	psey	klt	yn	x	rhd	lir	krw	S>T	d(pt)	rc	(pT)	x
CLA4	P48562	R	qrhw	R>Tw	R	R>K	S>T	W>dv	Y>r	F>R	ky	rk	rw	Rk	R	rhlwy	S>T	wvmlif	ylif	fw	x
Cmk1	P27466	L>R	R>k	R	R	K>I	S>T	F	Y>I	F>sye	lkt	L>fiv	rlkig	R>k	rmtka	km	T>S	hm(py	ylic	f(py)w	v(py)
Cmk2	P22517	L>R	R	R	R>I	K>rl	S>T	F	yri	YF>se	y	L	ri(py)	R>k	R>k	fpry	T>S	fy	yc	yfw	y(py)f
Fus3	P16892	R	Wr	RW	P>s	r	ST	P	E	e	ekt	w	(pt)	x	p>stl	x	TS	P	r(pt)	e(pt)	pw
Gcn2	P15442	r	wr	R	E	R>kl	T>S	F>dgi	YP	S	Y	w	w	x	ed	hil	TS	f	yri	s	x
GIN4	Q12263	L>r	R	R	S>R	K>rl	S>t	fdh	I>ypr	N	L	lkimr	rkh	Rk	S>tr	kmpr	S>T	m(pt)	wy	ndfh	li
Hog1	P32485	RI	W>r	R>w	P>N	ryk	S>T	P	Pyi	E	Ety	x	x	x	plt	q	TS	P	g	ep	e
HSL1	P34244	L	R	R	S>R	K>r	S>t	fd	I>Rpy	Nyf	L	y	pk	R>k	S>t	knpr	ST	qr(pt)	y	nd	li
Ipl1	P38991	R>I	R>k	R	R	LK>r	S	I>G	P>Y	Y	t	rk	rk	Rk	Rks	lmy	S>T	film	x	p	x
Kcc4	P25389	L	R	R	S>R	K>rl	S	fdh	I>Y	N	L	ry	rk	R	srtv	x	S>T	(pt)	ic	n	l(pt)
Kin1	P13185	R>I	RH	R	R>N	R>YK	S>T	Fgh	Yi	DY	L	k	wh	fhwy	nst	r	S	qr(pt)	(pT)l	dnc	Liv
Kin3	P22209	rl	W>R	R	dpri	R>h	S>T	PF	R>Y	R>fy	yte	(py)	w	flw	rsa	rk	S>T	rvi	ry	rkg	rk
Kin4	Q01919	RI	Rh	R	R>S	Yklr	S>T	fih	Ypr	N	L	lim	Rk	Rkh	Rks	krp	S	rmf	yr	nf	lfmr
KSP1	P38691	RI	Rq	R>T	R>P	R>kl	S>T	P>wi	E>YR	F	E>y	(pt)(py	(pt)(py)	R>k	n	K>rg	ST	rkn	rnp	(pT)kr	prg
Kss1	P14681	lr	R	RW	Prsn	kry	S>T	P	E	ey	ey	x	hw	w	Ps	prda	ST	P	x	(pt)(py)	kc
MEK1	P24719	RL	R>Q	R>S	R>s	RK>L	S>T	lgfpd	Y>e	Y>r	yl	R>k	rh	R	nrpc	rkhqpa	T>S	kqr	dae	pq(pt)	d(pt)
Mps1	P54199	I	rw	R	D	RK	ST	pvwg	R	ysd	Y	pt)(py	pt)(py)	rd(pt)	dert(pt)	d	TS	imv	hky	l	r
PHO85	P17157	R>kl	wrh	R>wl	SPR	Y>r	S>T	P>d	P>ly	P>s	K	x	y	x	SPTC	GM	ST	P	rkpm	ilm	g
Pkh2	Q12236	R>kl	R>kh	R	R	K>R	S>T	F	Y>R	Y>De	Y	rh(pt)	(pt)	R>k	rcts	kyr	S>T	fm	yc	Y	Y
Prk1	P40494	L	R	R	QR	Y>k	T>R	G>p	G>y	yf	y	L>im	rk	rk	qr	yhf	T>S	G	x	x	x
PRR1	P28708	R>Ld	R	R	R>q	K>RY	S>T	FD	Y>G	e	E	iv	r	rk	S>rt	S>rk	S>t	fr	st	y	li
Psk2	Q08217	R>L	R	R	R	K>R	S>T	H>F	Y	ys	Y>k	hrk	rk	R	R	kr	ST	h	ywl	yp	k(pt)
PTK2	P47116	R>L	R>k	R>I	R>pqr	KR>I	S>T	FW>g	Y>rei	YF	yk	li	r	R	rktva	kr	T>S	l	yh	yp	y
RAD53	P22216	R>I	RW>q	R>t	NR>di	KHR>y	S>T	F>P	P>ry	Y>P	Y>L	ap	w	rka	tsc	rpg	S>T	F>m	ry	r	kpq
SCH9	P11792	KR>I	R	R	R	LK>R	S>T	fi	pyg	Y>f	Y>e	kr	rk	R	R	L>m	S>T	fiv	(pt)	h	(pt)
SKM1	Q12469	R>I	q>hw	R>T	R>I	R>kly	S>T	W>vd	Y>pr	F	lety	rk	r	rkh	R>h	rhlpwy	S>T	wrmli	yvmih	wki	q
SKY1	Q03656	R>Ik	R>Q	R	R>S	KR	S>T	P>fgdi	Y>pi	N>yef	L>y	r	rk	R>(pt)	rsn	kr	S>t	pl(pt)	yl(pt)	rl	l
SLT2	Q00772	R	R>W	R>w	P>D	R>kh	S>T	P	Ey>ip	e>y	EY	r	r	R	srtpa	rkmqa	S>T	P	vc	(py)	x
SNF1	P06782	R>I	R	R	R>D	K>R	S>T	I>fh	Y>p	N>fdry	L	li	rkh	R>k	rtsvc	rkhy	S>t	fmcy	i	nf	L>im
STE20	Q03497	rk	qr	R/T	R	R>K	S>T	W>pv	Y	F	let	rk	r	rkw	R	rkwy	S>t	wmry	ywfh	fwy	x
TOS3	P43637	R>I	W	R>w	N>P	Y>Rk	S>t	pf	Y>I	Y	Y	rky	whk	ryh	nra	yrkh	S>t	c(pt)	y(pt)	ydn	yl
TPK1	P06244	R	R	R	R>s	KR>I	S>T	IF>pv	RY	yfser	lt	kr	rk	R	R>s	rk	S>T	iv	x	x	x
TPK2	P06245	R>I	R	R	R>q	RK	S>T	IF>pv	Y>r	yfse	tl	rk	r	R	R>s	pgnqr	S	q	cdgp	degnp	dgp
TPK3	P05986	R	R	R	R	RK	S>T	I>F	Y>r	y>fs	tl	rk(pt)	rkg	R	R>s	rpnkg	S>T	iv	d	dq	d(pt)
VHS1	Q03785	lr>k	R>q	R	R>P	YKR	S>T	GW	YE	y>fse	ylet	M	rkv	R	strq	kh	ST	x	st	x	f
YAK1	P14680	R>kld	R	R	R	RK	S>T	P	Y	Y	Y	r	r	R	rs	rkh	ST	P>r	yv	yvp	yw
YDL025C	Q12100	R>I	R>Q	R>t	R>p	K>R	S>T	F>I	Y>Pg	y	Y	L	r	R>k	rkt	krp	ST	rhn	dnq	p	x

Figure 5. The predicted most favorable amino acids including and surrounding the phosphorylation sites (P-sites) targeted by 50 well characterized yeast protein kinases are provided. The predicted optimal P-sites using our algorithms and purely based on the catalytic domain sequences of the kinases are shown on the left. The experimentally deduced consensus P-sites based on analysis of phosphorylation of peptide arrays with recombinant active forms of these same yeast kinases is given on the right. Amino acids letters provided in upper case represent the most critical amino acids in the positions for kinase recognition; lower case letters correspond to lesser important amino acids. (pS), (pT) and (pY) correspond to phospho-serine, phospho-threonine and phospho-tyrosine, respectively. The "0" position corresponds to the phospho-acceptor amino acid. The short name and Uniprot identification numbers are provided for each of these protein kinases.

3. FOLLOW UP SERVICES

In the identification of potential substrates for human kinases as part of our In Silico Protein Kinase Match Prediction (IKMP) Services, we perform our screening against 700,000 putative phospho-sites, which includes 93,000 experimentally observed phospho-sites. With one of our In Silico Kinase Match Prediction (IKMP-PK) Services, we can take either a known or putative phospho-site and find the most promising protein kinase candidates that target this phospho-site. Such information for the top 50 best matching kinases is freely available on-line for over 90,000 known human phospho-sites in our PhosphoNET KnowledgeBase (www.phosphonet.ca). However, this custom service is warranted if clients wish to predict kinases for uncharacterized or mutated phospho-sites in humans and other species. We also provide a ranking of the best 100 rather than 50 kinases if their prediction scores are greater than 0. Mutations of the amino acids residues in phosphorylation sites may alter result in recognition by additional protein kinases. Most protein kinases appear to have broad and overlapping substrate specificities. Many phospho-sites are likely to be targeted by multiple kinases.

With our In Silico Kinase Match Prediction (IKMP-PS) Service, we start with a human protein kinase of special interest to our clients and identify the top 1000 or 5000 known and predicted human phospho-sites that are likely to be targeted by that kinase. This is achieved by scoring over 700,000 known and putative phospho-site sequences with the specific PSSM matrix that we have produced from about 492 for various human protein kinases. One effective strategy to further identify the best kinase-substrate matches from this information is to observe which putative substrates have multiple predicted phospho-sites for a given kinase. The more phospho-sites with individual higher scores that are located close to each other on the same protein, the greater the probability that the protein is a bona fide substrate for the kinase in vivo. To take advantage of this bioinformatics service for our Introductory Price of only US$89 for 1000 phospho-sites or US$179 for 5000 phospho-sites, clients just have to provide the name and Uniprot ID number of the human protein kinase that they are interested in. The results of these analyses are also provided back to clients by e-mail in an MS-Excel spreadsheet. More detailed information on most known phospho-sites, including their evolutionary conservation, is freely available in PhosphoNET (www.phosphonet.ca).

Our In Silico Kinase Match Prediction (IKMP) Services permit the prediction of promising kinase-substrate pairs. With our Custom Peptide Synthesis Services (http://www.kinexus.ca/ourServices/proteinAndPeptide/peptidesynthesis/index.html), Kinexus can produce synthetic peptides both in soluble form and on arrays with the amino acid sequences that correspond to phospho-sites of high interest. Furthermore, with our In Vitro Kinase and Phosphopeptide Testing (IKPT) Services (http://www.kinexus.ca/ourServices/substrate_profiling/phosphopeptide_kinase.html), Kinexus can test the ability of over 350 different protein kinases to phosphorylate these peptides or recombinant proteins that feature these phospho-sites. This can permit the narrowing down of the best candidate kinases for important phospho-sites and experimental support for important kinase-substrate connections in cell signalling pathways. With our custom cell culture, immunoblotting and mass spectrometry services, we can further aid our clients in the experimental validation of these kinase-substrate interactions.

To find out more about how we can further assist our clients in these types of customized studies, please contact our Customer Services Representatives to learn more about our custom proteomics services.

4. FORMS TO BE COMPLETED

All of the forms necessary to use the In Silico Protein Kinase Specificity Prediction Services are provided in the Appendices section of this Customer Information Package. Fillable MS-Word versions of these forms are directly downloadable from the Kinexus website at http://www.kinexus.ca/ourServices/substrate_profiling/kinase_match.html and by request by e-mail or by phone. Please contact our Technical Service Representatives by e-mail at info@kinexus.ca or by phone at 604-323-2547 Ext. 1 for all enquiries related to technical/research issues, work orders, service fees or request of fillable order forms.

All customers are required to complete the following forms for each order placed:

A. Service Order Form (IKSP-SOF)

Please ensure:

- Shipping address and contact name and numbers are specified
- Pricing information is completed as outlined in Section C on the Service Identification Form (IKSP-SIF)
- Any promotional vouchers or quotations are listed in the billing sections
- Include a Purchase Order, Visa or MasterCard number for payment
- This form is certified correct and signed and dated

B. Service Identification Form (IKSP-SIF)

For each sample submitted, please ensure the following:

- In Section A, the customer should assign a unique Client Screen Identification Name so that they can track their order internally
- In Section B, the customer should provide the names of the target kinases, their Uniprot or NCBI accession numbers, and the species from which the kinases originate

Upon completion and transmission of these forms to Kinexus, we will endeavor to return the results of our analyses back to you within 3 working days.

KINEXUS

Form: IKSP-SOF

IN SILICO KINASE SPECIFICITY PREDICTION SERVICES SERVICE ORDER FORM

Kinexus Order Number
For Kinexus internal use only.
____

Customer Information □ Repeat Customer **OR** □ New Customer

□ Dr. □ Mr. □ Ms.

Name of Authorized Representative or Principal Investigator *Title/Position*

Company Name or Institute *Department*

Street Address

City *State or Province* *Country* *Zip or Postal Code*

Email Address *(Area Code) Telephone Number* *(Area Code) Facsimile Number*

Contact Person (if different from Authorized Representative) *Email Address* *(Area Code) Telephone Number*

IKMP Service Reports

RESULTS SENT BY EMAIL TO: □ Authorized Representative/Investigator **AND/OR** □ Contact Person

Ordering Information

In Silico Protein Kinase Specificity Prediction Services are offered for the prediction of a position-specific scoring matrix (PSSM) for the amino acid sequence specificity of a target kinase for its substrates

NOTE 1: THIS FORM MUST BE ACCOMPANIED BY AN IKSP-SIF FORM
NOTE 2: EACH IKSP ANALYSIS IS SEPARATELY COSTED

All prices in U.S. Funds

No. of IKSP-PK analyses – 1 PSSM for each typical protein kinase ____ @ US $179 per analysis = $______

Total number of IKSP analyses ordered:______

Quotation or Reference Number:__________ - $______

TOTAL COST FOR THIS ORDER = $______

FOR CANADIAN CUSTOMERS ONLY:
Add an additional 12% to the above total for HST (No. 893907329 RT0001): + $________________ = $____________

TOTAL AMOUNT PAYABLE IN U.S FUNDS

Payment Method

□ Purchase Order Accepted from companies and institutes with approved credit. P.O. Number: __________

□ Visa OR □ MasterCard

Print Cardholder Name *Visa Number* *Expires (M/Y)* *Cardholder Signature*

Billing Information □ Send invoice to customer at above address **OR** □ Send invoice to accounts payable contact:

□ Dr □ Mr □ Ms

Accounts Payable Contact Name *Company Name or Institute*

Street Address *City*

State or Province *Country* *Zip or Postal Code* *(Area Code) Telephone Number*

Authorization - Customer agrees to payment within 30 days of receipt of an invoice for these services that has been issued following their completion:

Print Name of Authorized Representative or Principal Investigator *Authorized Signature* *Date (D/M/Y)*

How did you originally hear about the IKSP Services? □ ***Direct Mail*** □ ***Email*** □ ***Web Site*** □ ***Advertisement*** □ ***Referral*** □ ***Conference or Trade Show*** □ ***Other***

KINEXUS

Form: IKSP-SIF

IN SILICO KINASE SPECIFICITY PREDICTION SERVICE IDENTIFICATION FORM

KINEXUS ORDER NUMBER

____

NAME: ______________________________________
(Authorized Representative or Principal Investigator)

COMPANY/INSTITUTE: ______________________________________

STANDARD IN SILICO PROTEIN KINASE SPECIFICITY PREDICTION SERVICES REQUESTED:

Use this form to order our Standard In Silico Protein Kinase Specificity Prediction (IKSP) Services currently offered by Kinexus. Please check the appropriate tick boxes. If you need assistance, please contact a technical service representative by calling toll free in North America 1-866-KINEXUS (866-546-3987) or by email at info@kinexus.ca.

□ STANDARD SERVICES REQUESTED: IKSP

Prediction Services with Kinexus' proprietary algorithms and databases to permit generation of position-specific scoring matrix for a kinase of choice

A protein kinase name and its corresponding Uniprot ID Number or NCBI Accession Number must be provided by the client

KINEXUS ID NUMBER

(Bar Code Identification Number)

For Kinexus Internal Use Only.

A. CLIENT SCREEN ID NAME:

Customer ID:

Provide ID name of your choice for this order for your own reference

B. IKSP SERVICE INFORMATION: *Use this form to provide descriptions of up to 10 kinase targets. Complete a second IKSP-SIF sheet if additional analyses with more protein kinases are required. Consult the IKSP Customer Information Package for more details.*

□ **IKSP Service** – Generation of a PSSM for one distinct protein kinases.

	Short Name for Kinase	Uniprot ID No. or NCBI Accession No.	Species
□ **Protein Kinase 1**			
□ **Protein Kinase 2**			
□ **Protein Kinase 3**			
□ **Protein Kinase 4**			
□ **Protein Kinase 5**			
□ **Protein Kinase 6**			
□ **Protein Kinase 7**			
□ **Protein Kinase 8**			
□ **Protein Kinase 9**			
□ **Protein Kinase 10**			

C. PRICING:

□ IKSP service for 1 analysis (1 kinase) = $179

Use this pricing information for completion and submission of Service Order Form IKSP-SOF.

______________________________ ______________________________ ______________________________

Name of person completing this form *Signature* *Date (Y/M/D)*