Article No. jmbi.1999.3310 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 294, 1351–1362

JMB

Sequence and Structure-based Prediction of Eukaryotic Protein Phosphorylation Sites

Nikolaj Blom[1], Steen Gammeltoft[2] and Søren Brunak[1]*

[1]*Center for Biological Sequence Analysis, Department of Biotechnology, The Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2]*Department of Clinical Biochemistry, Glostrup Hospital, DK-2600 Glostrup Denmark*

**Corresponding author*

Protein phosphorylation at serine, threonine or tyrosine residues affects a multitude of cellular signaling processes. How is specificity in substrate recognition and phosphorylation by protein kinases achieved? Here, we present an artificial neural network method that predicts phosphorylation sites in independent sequences with a sensitivity in the range from 69 % to 96 %. As an example, we predict novel phosphorylation sites in the p300/CBP protein that may regulate interaction with transcription factors and histone acetyltransferase activity. In addition, serine and threonine residues in p300/CBP that can be modified by *O*-linked glycosylation with *N*-acetylglucosamine are identified. Glycosylation may prevent phosphorylation at these sites, a mechanism named yin-yang regulation.

The prediction server is available on the Internet at http://www.cbs.dtu.dk/services/NetPhos/or *via* e-mail to NetPhos@cbs.dtu.dk.

Keywords: phosphorylation; kinase specificity; prediction; protein structure; transcriptional regulation

Introduction

Protein kinases catalyze phosphorylation events that are essential for the regulation of cellular processes like metabolism, proliferation, differentiation, and apoptosis (Koliba & Druker, 1997; Hunter, 1998; Johnson *et al*., 1996, 1998; Pinna & Ruzzene, 1996; Graves *et al*., 1997). This very large family of enzymes share homologous catalytic domains and the mechanism of substrate recognition may be similar despite large variation in sequence. Crystallization studies indicate that a region, between seven and 12 residues in size, surrounding the acceptor residue contacts the kinase active site (Songyang *et al*., 1994).

The specificity of protein kinases is dominated by acidic, basic, or hydrophobic residues adjacent to the phosphorylated residue, but the large variation makes it difficult manually to inspect protein sequences and predict the location of biologically active sites. This prompted us to investigate if the fuzzy sequence patterns can be recognized using artificial neural networks techniques. Neural networks are capable of classifying even highly complex and non-linear biological sequence patterns, where correlations between positions are important. The network recognizes the patterns seen during training, and retains the ability to generalize and recognize similar, but non-identical patterns. Artificial neural networks have been extensively used in biological sequence analysis (Wu, 1997; Baldi & Brunak, 1998). Since determinants of phosphorylation sites probably are no longer than about ten residues, most local sequence alignment tools, such as BLAST and FASTA, will not be useful for detecting phosphorylation sites due to a large number of irrelevant hits in the protein databases, even to non-phosphorylated proteins.

The related proteins p300 and CBP (CREB (cAMP-response-element-binding)-binding protein) integrate molecular signals at the level of gene transcription and chromatin modification. p300 and CBP interact with transcription factors CREB, Jun and Fos, viral oncoproteins E1a and SV40 large T antigen, and kinases $pp90^{RSK}$ and cyclin E-complexed cyclin-dependent kinase (CDK)-2 (Shikama *et al*., 1997; Ait-Si-Ali *et al*., 1998). These interactions may possibly be regulated by reversible phosphorylation of p300/CBP. We demonstrate that regions of p300/CBP, which have been shown to interact with other molecules, contain probable phosphorylation sites. In addition, we describe sites that possibly are regulated by both phosphorylation and glycosylation by *N*-acetylglucosamine (GlcNac), a regulatory mechanism described as a yin-yang dynamic phosphorylation/glycosylation (Hart *et al*., 1995; Hart, 1997).

E-mail address of the corresponding author: brunak@cbs.dtu.dk

0022-2836/99/501351–12 $30.00/0

Results

The general sequence context at experimentally verified phosphorylation sites

Based on the large sets of experimentally verified phosphorylation sites, sequence logos were generated for each of the three acceptor residues, tyrosine, serine, and threonine (Figure 1). The sequence logos emphasize residues that are frequently found in the context of the phosphorylation sites. The logo does not show the specificity determinants for a single kinase, but the overall features of all experimentally verified sites.

Tyrosine sequence logo

For tyrosine phosphorylation sites, we found that tryptophan, a large and rare amino acid, was never found at positions P − 5 to P − 1 relative to the phosphotyrosine (PO), most likely due to steric hindrance (Figure 1(a)). Similarly, cysteine was never found at positions P − 2 and P − 1, indicating that tyrosine phosphorylation is unlikely to occur C-terminally to a disulphide bridge, in agreement with the notion that phosphorylation occurs where the peptide chain is flexible (Tinker *et al.*, 1988). Methionine was never found at position P − 2, whereas it was highly abundant at positions P + 1 and P + 3. Absent residues do not necessarily act as negative determinants for the substrate recognition; this can be shown only by experimental techniques. The presence of acidic residues, aspartic and glutamic acid, in the region from position P − 5 to P − 1 was noted early in the analysis of tyrosine phosphorylation sites (Patschinsky *et al.*, 1982). The relative content of acidic residues varies from 22 % at P − 5 to 34 % at position P − 1.

It is not clear from the sequence logo whether many sites contain consecutive acidic residues. Analysis of 210 tyrosine phosphorylation sites shows that only one sequence contains four consecutive acidic residues at positions P − 4 to P − 1: Tyr_{315} in polyoma virus middle T antigen (EEEEY*MPME)(Zhou & Cantley, 1995). Within positions P − 4 to P − 1, 78 % of the sites have at least one acidic residue, 29 % have at least two, and 8 % have at least three acidic residues. Among the sites with at least two acidic residues, EE and ED are the most frequently occurring dipeptides at positions P − 4 and P − 3. Based on the probability of finding an acidic residue at positions P − 4 to P − 1, the calculated probability of finding two acidic amino acids is 20 %, which is almost equal to the observed frequency of 19.5 %.

The motif [N-P-X-Y*], where X is any amino acid, is recognized by the phosphotyrosine-binding domain (PTB) present in several signaling proteins, e.g. insulin-receptor substrates 1-4 (Pawson, 1995). The observed frequency in the data set of this motif is 4.3 % of 210 sites. However, the calculated frequency, based on the frequency of N at P − 3 and P at P − 2, is 1.3 %, suggesting that significant correlations between these positions occur.

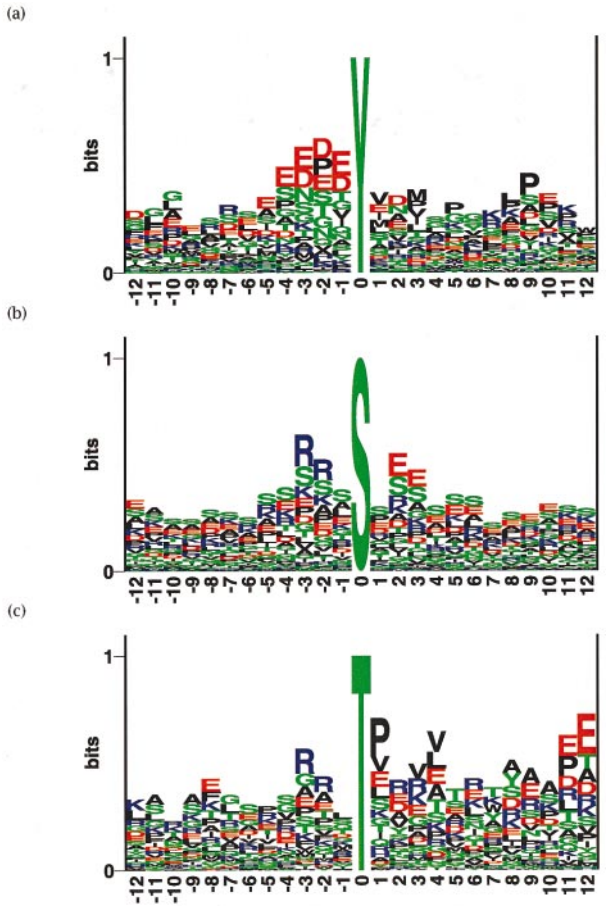

Figure 1. Sequence logos of phosphorylation sites aligned at the phosphoacceptor residue showing the Shannon information (in units of bits): (a) 210 tyrosine phosphorylation sites; (b) 584 serine phosphorylation sites; (c) 108 threonine phosphorylation sites. Note that the central residue has been rescaled to a size of 1 bit (actual size is 4.32 bits).

On the C-terminal side of the phosphotyrosine residue, there is a clear over-representation of hydrophobic residues (M, L, I, or V) and residues with a structural role (G or P). The hydrophobic motif at positions P + 1 to P + 3 has been described in studies of the specificity of the Src-homology 2 (SH2) binding domain (Songyang *et al.*, 1993). The hydrophobic motif [M/L/I/V-X-M/L/I/V] was found in 18.1 % of the tyrosine sites at positions P + 1 to P + 3. The expected frequency of the motif is 16.4 % based on the observed frequencies for each position, indicating that there is no significant difference between the observed and calculated frequencies.

Positions P + 4 to P + 6 are dominated by glycine, indicating that this region may be structurally flexible. Proline is clearly over-represented at positions P + 5 (16 %) and P + 9 (23 %), indicating that the structure of the pep-

tide chain is important. Position P + 7 is dominated by the positively charged basic residues (K/R [28 %]) and position P + 12 by the rare and bulky tryptophan residue (10.8 % compared to 1.2 % on average in natural proteins).

We examined the sites containing either proline at position P + 9 or tryptophan at P + 12 and observed that many sites contained both residues. Most of the sites were known auto-phosphorylation sites in Src-related kinases (e.g. Src, Hck, Tec, Fyn) or in receptor tyrosine kinases (insulin-receptor, IGF-1-receptor, PGDF-receptor, NGF-receptor). In the crystal structure of the insulin receptor kinase domain, the proline and tryptophan residues are located in a buried coil region C-terminally of the activation loop containing the auto-phosphorylated tyrosine residues ($Tyr_{1158,1162,1163}$).

Serine sequence logo

In the sequence logo of serine phosphorylation sites (Figure 1(b)), the basic motif recognized by PKA and PKG at positions P−3 and P−2 was readily observable (31 % and 27 % R + K at positions P − 3 and P − 2, respectively). In the motif of protein kinase C (PKC), basic residues at P + 2 and P + 3 could also be identified. However, these positions were dominated by glutamic acid, which is part of the acidic motif recognized by the casein kinase 2 (CK-2) type kinases. Proline-directed kinases have a preference for proline at position P + 1 (13 % of the residues at P + 1), which could be seen readily from the logo. Rare residues around the serine residue include tryptophan and cysteine. Tryptophan is not observed at positions P − 3 to P − 1 and P + 2, and cysteine never at positions P − 7 and P − 5.

As described earlier for *O*-glycosylation sites, serine residues have a tendency to cluster, and this is observed for the neighboring positions P − 1 and P + 1 also, where serine is the most abundant residue (Wilson *et al*., 1991; Hansen *et al*., 1998).

Threonine sequence logo

In the sequence logo of threonine phosphorylation sites (Figure 1(c)), the basic motif at positions P − 3 and P − 2 was readily observable, as was the proline-directed motif at P + 1. The PKC preference for basic residues at positions P + 2 and P + 3, and the CK-2 preference for acidic residues at the same positions could be observed also.

Tryptophan was never found at positions P − 9 to P − 5 or P + 1 to P + 6, but was over-represented at P + 7 (11 % of the residues). Cysteine was never found at positions P − 5, P − 3 to P + 1 and P + 4 to P + 12, while asparagine was never found at positions P − 3, P − 2 or P + 1. These observations could be due to a smaller amount of data (108 sites), but for cysteine and tryptophan these probably reflect the fact that many phosphorylation sites are located in flexible regions of the target molecule.

The frequent occurrence of valine or leucine at positions P + 3 and P + 4 was investigated further by analyzing sites containing these features. Eight sites contained valine or leucine at both P + 3 and P + 4 (data not shown). We noted that these are involved in cell-cycle-dependent phosphorylation and some of them are known as targets of the CDK-activating kinase. Two of the sites contained tyrosine phosphorylation sites either at position P + 1 or P + 2 relative to the phosphothreonine residue, indicating that they may be targets of dual-specificity kinases. However, not all of the sites contained adjacent tyrosine residues, and therefore the double V/L motif may be important for target recognition of CDK-activating and related kinases. Other kinases that prefer hydrophobic residues at P + 4 include AMP-activated protein kinase and calmodulin-dependent protein kinase I (Dale *et al*., 1995).

Another motif found in protein kinases involved in regulation of cell-cycle was the frequent occurrence of proline at P + 11 and glutamic acid at P + 12. Eleven phosphorylation sites, all in CDKs or related proteins, contained the consensus sequence (T*xxV[V/A]TxxYR[A/S]PE), where the first T indicates the acceptor residue. These sites are clearly related and the conserved residues might reflect the conservation of structural features rather than conservation of specificity determinants.

Phosphorylation sites predicted by neural networks

Sequence motifs composing functional sites in polypeptides can be complex in the sense that positional correlations may play an important role. The amino acids surrounding a phosphorylated residue may not contribute independently as to whether a particular site is activated. This means that simple local alignment methods, or linear weight matrices based on consensus patterns, may be unable with acceptable accuracy to separate true sites from a control set of non-phosphorylated sites. Rules like ''an acidic residue at P − 2 and no cysteine at P + 1 will make this site a probable phosphorylation site'' cannot be taken into account. Two positions can make only independent contributions to the overall score.

Training of the neural networks, using different combinations of input window sizes and training sets, showed that networks containing no hidden units (i.e. linear networks), performed worse than networks containing hidden units (i.e. non-linear networks). This clearly indicated that correlations between the amino acids surrounding a phosphorylated residue are significant in determining whether a particular site is phosphorylated. This was evident also when we compared the experimentally verified data to the kinase patterns in the Prosite database (Bairoch *et al*., 1997). Prosite patterns describe the specificity pattern of a few well-characterized kinases and were never

designed for obtaining a proteome-wide prediction of phosphorylation sites. However, since Prosite patterns are the most widely used approach for predicting phosphorylation sites in novel proteins, we found that it was relevant to compare them with our newly developed method.

As shown in Table 1, the tyrosine-specific Prosite pattern matched only 10 % of the tyrosine phosphorylation sites, indicating that the divergence of tyrosine phosphorylation site motifs is quite high, and hard to describe by a single consensus pattern. The serine and threonine phosphorylation sites were somewhat better identified by the Prosite patterns (sensitivity of 48 % and 38 %, respectively).

With non-linear networks, we were able to obtain much better results. Performance values for the standard approach for each of the three acceptor types are shown in Table 1. Between 65 % and 89 % of the positive sites and 78 % to 86 % of the negative sites were correctly predicted. The correlation coefficient ranged from 0.44 for threonine to 0.47 for tyrosine, and 0.75 for the serine sites. However, a significant number of seemingly false positive predictions had very high scores, indicating that these sites actually had properties similar to those of true phosphorylation sites.

Since the annotation of negative phosphorylation sites is a problem in the databases, and since some negative sites eventually will be shown experimentally to be true phosphorylation sites, we constructed an augmented negative data set. We used the optimal neural network parameters found by the standard approach and selected negative sites from the entire set of acceptor sites that would not conflict with the known experimentally verified phosphorylation sites when initially training the networks.

The augmented data sets contained three to five times more unique negative sites than used in the standard approach. Still, the trained networks had a significantly increased ability to detect the true sites with an improvement of the order of 10-20 %. This indicates clearly that there are false negatives in the publicly available data, and that their effect can be suppressed by this approach. The correlation coefficients now range from 0.82 to 0.97, approaching the optimal value of 1.00. The general test performance on novel data will fall in between these values, and those obtained when the networks trained on the augmented data sets predict the phosphorylation status of residues in the standard unmodified set, see Table 1.

The optimal window sizes were found to be nine residues for tyrosine and threonine, and 11 residues for serine (data not shown). These values are in agreement with the general consensus that the kinase physically contacts a stretch of 7-12 residues surrounding the acceptor residue (Songyang *et al.*, 1994).

Table 1. Predictive performance of the sequence and structure-based methods

Residue	Method	*C*	*Sn* (%)	*Sp* (%)	t_n (%)	Data set
Y						
	Prosite (PDOC0007)	0.28	10	100	100	augm.
	NN-std.	0.47	70	68	78	std.
	NN-augm.	0.59	87	69	74	std.
	NN-augm.	0.92	87	100	100	augm.
	Structural-NN	0.46	87	40	71	augm.
S						
	Prosite (PDOC0004-6)	0.67	48	100	100	augm.
	NN-std.	0.75	89	86	86	std.
	NN-augm.	0.85	96	90	89	std.
	NN-augm.	0.97	96	100	100	augm.
	Structural-NN	0.50	85	65	64	augm.
T						
	Prosite (PDOC0004-6)	0.60	38	100	100	augm.
	NN-std.	0.44	65	52	83	std.
	NN-augm.	0.52	69	58	86	std.
	NN-augm.	0.82	69	100	100	augm.
	Structural-NN	0.38	87	37	59	augm.

For each of the three residue types (Y, S or T), the predictive performance of the sequence and structure-based methods is shown. Prosite patterns, indicated by their relevant codes, and neural networks (NN), trained using either the standard (std.) or augmented (augm.) data sets, are compared. *C* indicates the correlation coefficient, the sensitivity *Sn* indicates the proportion of positive sites correctly predicted (or true positives t_p), the specificity *Sp* indicates the proportion of all positive classifications that are correct and t_n indicates the number of true negative sites. The number of phosphorylation sites was (Y 210; S 584; T 108). The number of negative sites in the data sets was (Y 319/940), (S 584/3266), and (T 380/1283) in the standard and augmented versions, respectively. The Prosite patterns represent PDOC0004 (cAMP and cGMP-dependent kinase) [R/K][R/K]x[S/T*]; PDOC0005 (protein kinase C) [S/T*]x[R/K]; PDOC0006 (CK-2) [S/T*]xx[D/E]; and PDOC0007 (general tyrosine kinase) [R/K]xx[D/E]xxxY* or [R/K]xxx[D/E]xxY*.

Predictions on other phosphoacceptor residues

Serine/threonine kinases are able to phosphorylate either serine or threonine in the same sequence context. An example is protein kinase C, which phosphorylates Thr_{710} of the rat glutamate receptor 1 precursor (PhosphoBase entry B169, [RVRKT*KGKY]), and Ser_{717} of the rat glutamate receptor 2 precursor (PhosphoBase entry B168, [RVRKS*KGKY]). We analyzed whether a neural network trained to predict serine sites would be able to predict correctly on a number of sites, where the motif surrounding the acceptor residue was taken from a known threonine phosphorylation site and *vice versa*. For completeness, the prediction on tyrosine sites was included.

The analysis showed that the tyrosine network is able to predict correctly between 39 % and 41 % of all serine and threonine phosphorylation sites, respectively. The serine network predicts 52 % of the tyrosine sites and 81 % of the threonine sites. The threonine network predicts 19 % of the tyrosine sites and 54 % of the serine sites. Overall, the serine site network is the most general at predicting phosphorylation sites, correctly classifying many of the threonine and tyrosine phosphorylation sites based on their sequence context. The ability of the serine and threonine networks to recognize a larger fraction of each other's sites than of tyrosine sites is consistent with the known specificity overlap of threonine and serine kinases.

Tertiary structure of phosphorylation sites

It is obvious that what the kinase actually recognizes is the three-dimensional structure of the polypeptide at the acceptor residue, and not the primary structure (Johnson *et al*., 1996, 1998; Pinna & Ruzzene, 1996; Songyang *et al*., 1994). From the available protein structure data in PDB, containing phosphorylated sequences, we made a superposition of the local structure of 12 tyrosine phosphorylation sites (Figure 2). Interestingly, nine of the 12 tyrosine side-chains occupied one conformation relative to the C^{α} atom, while three clustered in another specific conformation. This structural conservation appears in otherwise unrelated sequences. Repeating the same procedure for surface-exposed tyrosine residues not predicted to be phosphorylated showed that the tyrosine residue occupied a wide range of conformations.

Since phosphorylation sites are predicted to be located in flexible regions in order to be able to fit the kinase recognition cleft, we examined the known PDB protein structures containing phosphorylation sites to see whether the temperature factor (*B*-factor) was increased. This seems generally to be the case, and for the insulin receptor kinase domain, where phosphorylation is known to occur at Tyr_{1158}, Tyr_{1162}, and Tyr_{1163}, the temperature factor is maximal (maximum occurs at Thr_{1154}) in the loop region just upstream from and close to the three phosphorylation sites.

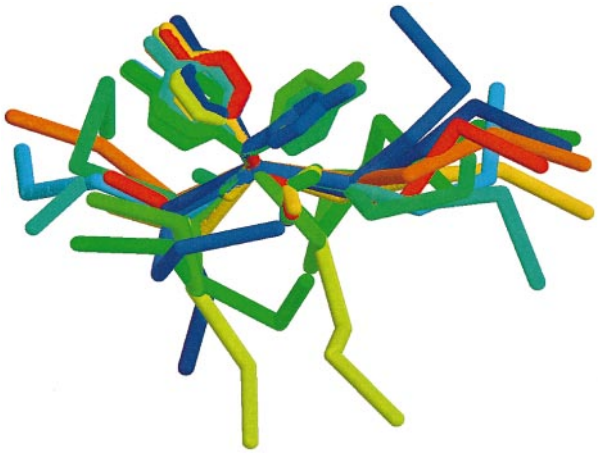

Figure 2. Assembly of 12 tyrosine phosphorylation site structures superimposed at the central tyrosine. Only the central tyrosine side-chain is shown, the rest of the peptides is shown as the C^{α} peptide backbone.

Structure-based prediction

We examined whether a method for prediction of distance matrices or contact maps could indicate whether the local structure of phosphorylated sites is different from that of non-phosphorylated sites. A neural network was trained using a predicted, local contact map of a peptide fragment centered on either positive or negative sites in the data set. Using the same division of the data as described for the sequence-based neural network, we found that the best performance was obtained for a neural network with an input window of 21 residues for the tyrosine and serine data sets, and an input window of 25 residues for the threonine set.

Given that this prediction is based on a prediction, the performance was impressively high, as the sensitivity for positive sites in all three cases was around 85-87 % (see Table 1). The performance on negative sites was less impressive, resulting in a higher level of false positives (between 29 % and 41 % of the negative sites were predicted as positive sites). It must again be emphasized that false positive predictions in some cases may be true phosphorylation sites awaiting experimental demonstration.

In order to compare the performance of the sequence-based and structural-based neural networks on the same data set, we generated a diagram showing the output scores from the sequence-based network and the structure-based network (Figure 3). Ideally, we would find most points on the diagonal, true phosphorylation sites, should obtain a score close to 1.0 from both methods and non-phosphorylation sites should get scores close to 0.0. In reality, the sequence-based network is performing better than the structure-based approach and therefore we would expect points to be better separated in the *X*-dimension than in the *Y*-dimension.

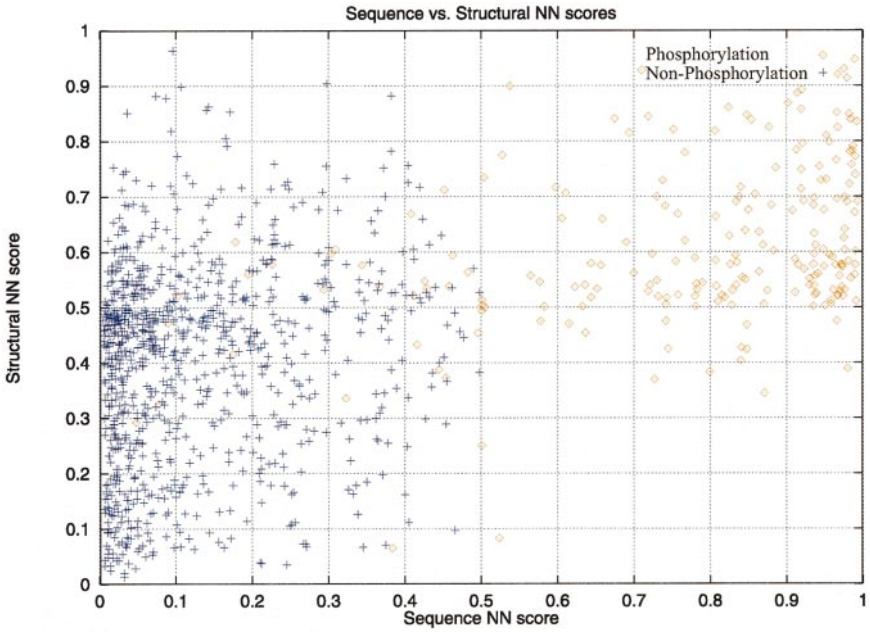

Figure 3. Sequence (*x*-axis) and structure-based (*y*-axis) scores on tyrosine phosphorylation sites. True phosphorylation sites are shown as red diamonds (210 sites), while non-phosphorylation sites are shown as blue crosses (940 sites).

Interestingly, a few of the positive sites incorrectly predicted as negative sites by the sequence-based method were correctly predicted by the structural method. For the tyrosine data set, 17 out of 27 phosphorylation sites incorrectly classified by the sequence-based network were correctly classified by the structure-based network. Some of the sites having a very low sequence score, but a structure score above the threshold of 0.5, include the autophosphorylation site Tyr_{1361} [EHIPY*THMN] of the insulin receptor, Tyr_{762} [QRSLY*DRPA] of the platelet-derived growth factor receptor and Tyr_{826} [VGPGY*LGSG] of the Ret receptor tyrosine kinase. These sites do not match well with the typical consensus features of tyrosine phosphorylation sites, such as an acidic region upstream of the acceptor residue or a strong hydrophobic motif at $P+1$ and $P+3$. Nevertheless, the predictions by the structural network indicate that these tyrosine phosphorylation sites have structural features resembling those of other tyrosine phosphorylation sites.

Putative phosphorylation sites in p300/CBP

p300/CBP is a transcriptional adaptor that interacts with transcription factors and is also a histone acetyltransferase (HAT) (Shikama *et al.*, 1997; Ait-Si-Ali *et al.*, 1998). Especially in the first third and the last third of the 2400+ residues, p300/CBP proteins many protein-protein interactions have been shown to occur, and some of these are believed to be regulated by phosphorylation. Considering the large number of potential phosphoacceptor residues in p300/CBP (CBP contains 394 serine, threonine and tyrosine residues, 16% of the total number of residues) it would be helpful to identify the most probable sites before conducting experiments involving site-directed mutagenesis or in analysis of mass spectra of the modified proteins. Using the prediction networks, we have found putative phosphorylation sites. The amino acid sequences of p300 (accession number Q09472) and CBP (Q92793) are 63% identical, which provided us with an opportunity to compare putative sites in one protein with putative sites at the same aligned position in the other. We concentrated on homologous sites in p300 and CBP with a strong prediction score above 0.9 (Table 2).

Several of the putative phosphorylation sites are completely conserved in both molecules, e.g. Tyr_{620}, Ser_{1497}, and Thr_{1684} (numbering refers to p300). Interestingly, three sites utilize serine as the acceptor residue in one protein and threonine in the other, e.g. Thr_{992}, Ser_{1295} and Thr_{1533} (Table 2).

In the N-terminal third of the molecule, a conserved tyrosine site at position 620 (641 in CBP) may be of importance in the binding of CREB, since interaction in the region from positions 590-669 of CBP has been reported, see Shikama *et al.* (1997) and references therein.

The central part of the molecule (residues 800-1600) includes the catalytic domain of HAT activity of p300/CBP. The putative phosphorylation sites found in this region may be related to regulation of HAT activity.

The C-terminal part includes a region that interacts with viral oncoproteins E1A and SV40 large T, transcription factors c-Fos, c-Jun, JunB, YY1 among others, (Shikama *et al.*, 1997) and references therein.

Table 2. Putative phosphorylation sites in p300/CBP

Protein	Position	NetPhos Score	Acc.res.	Sequence
P300	24	0.984	S	SPALSASAS
CBP	23	0.987	S	SPGFSANDS
p300	620	0.977	Y	EGDMYESAN
CBP	641	0.977	Y	EGDMYESAN
p300	959	0.946	S	PSTSSTEVN
CBP	980	0.953	S	SSVASAETN
p300	992	0.961	T	EPADTQPED
CBP	1013	0.993	S	DPGESKGEP
p300	1135	0.983	T	YNRKTSRVY
CBP	1171	0.983	T	YNRKTSRVY
p300	1289	0.991	S	ENKFSAKRL
CBP	1325	0.975	S	ENKFSAKRL
p300	1295	0.984	S	KRLPSTRLG
CBP	1331	0.972	T	KRLQTTRLG
p300	1346	0.988	S	RFVDSGEMA
CBP	1382	0.993	S	RFVDSGEMA
p300	1396	0.992	S	RVYISYLDS
CBP	1432	0.988	S	RVYISYLDS
p300	1446	0.979	Y	ECDDYIFHC
CBP	1482	0.979	Y	EGDDYIFHC
p300	1497	0.998	S	DRLTSAKEL
CBP	1533	0.998	S	DRLTSAKEL
p300	1516	0.993	S	VLEESIKEL
CBP	1552	0.993	S	VLEESIKEL
p300	1533	0.955	T	REENTSNES
CBP	1568	0.987	S	KKEESTAAS
p300	1684	0.956	T	RWHCTVCED
CBP	1721	0.956	T	RWHCTVCED
p300	1726	0.995	S	AATQSPGDS
CBP	1763	0.995	S	PQSKSPQES
p300	1734	0.993	S	SRRLSIQRC
CBP	1771	0.992	S	SRRVSIQRC
p300	1868	0.935	T	TTPQTPQPT
CBP	1902	0.925	T	STPQTPQPP
p300	2315	0.992	S	QPVPSPRPQ
CBP	2351	0.935	S	APVQSPRPQ
p300	2320	0.959	S	PRPQSQPPH
CBP	2356	0.959	S	PRPQSQPPH
p300	2325	0.984	S	QPPHSSPSP
CBP	2361	0.984	S	QPPH5SPSP
p300	2328	0.996	S	HSSPSPRMQ
CBP	2364	0.993	S	HSSPSPRIQ

NetPhos score is the output score from the ensemble of neural networks trained on that acceptor residue type. The sequence shows the context of the acceptor residue ± four residues.

This region is phosphorylated by cyclin E-Cdk2, and thereby regulating the HAT activity of CBP (Ait-Si-Ali *et al.* , 1998). The region investigated spanned residues 1890-2441 of CBP, thereby including the putative phosphorylation sites at Thr_{1902}, Ser_{2351}, Ser_{2356}, Ser_{2361} and Ser_{2364}. These sites all contain an SP or SXP motif, the first of which is a well-known motif for Cdk-related proline-directed kinases.

Putative yin-yang sites in p300/CBP

The reversible and dynamic modification of a particular serine or threonine residue by either phosphorylation or GlcNac-glycosylation has been named yin-yang regulation (Hart *et al.*, 1995). The addition of the GlcNac sugar moiety prevents the acceptor residue from being phosphorylated and represents one way of inhibiting signals that may otherwise cause abnormal growth or apoptosis. This mechanism is involved in regulation of the tumor suppressor protein p53 and the AP1 transcription factor complex.

Based on known sites of reciprocal phosphorylation/GlcNac-glycosylation, we are currently developing a neural network-based method for the prediction of GlcNac sites on intracellular proteins (R. Gupta, unpublished results), analogous to the methods available for predicting GalNac-*O*-glycosylation sites (Hansen *et al.*, 1998) and GlcNac-*O*-glycosylation sites in *Dictyostelium discoideum* proteins (R. Gupta *et al.*, unpublished results). At present, the latter method, named DictyOGlyc † is probably the method that most closely identifies sites with features similar to known intracellular GlcNac-sites.

† Accessible at http://www.cbs.dtu.dk/services/DictyOGlyc/

To investigate whether some of the putative phosphorylation sites in p300/CBP might also be regulated by GIcNac-glycosylation, we compared the predictions by the phosphorylation site neural networks with the output from the DictyOGlyc server. The sites that have high scores from both methods and that have homologous sites in p300/CBP are listed in Table 3.

Many of the experimentally verified GlcNac sites contain hydrophobic residues, such as proline, alanine or valine, at the flanking positions. This is the case for the five sites reported here, which all contain either serine or threonine followed by proline. This feature indicates that cyclin-dependent kinases or mitogen-activated protein kinases, which have a preference for proline at position P + 1, may be the kinases phosphorylating these residues. Ser_{2315}, Ser_{2336} and Ser_{2346} (p300 numbering) are all preceded by proline and glutamine and followed by proline as well as a basic residue (arginine or histidine). Whether these sites are modified by phosphorylation and glycosylation, as predicted, awaits experimental evidence.

Discussion

The neural network approach presented here for the prediction of phosphorylation sites is top-down, in the sense that an overall, general approach to kinase specificity was taken. This is in contrast to the classical approach, which has been to use a bottom-up philosophy, where the specificity of a single kinase is studied in great detail.

The classical approach is based on determination of the activity of purified protein kinases using *in vitro* assays with either naturally occurring peptides or synthetic peptides. Using a large number of peptides, a consensus sequence of the substrate of a given kinase is obtained (see the review by Pinna & Ruzzene, 1996).

A new approach has been developed that is based on the synthesis either by chemical or cassette mutagenesis methods, of fully or partially degenerate peptide libraries. These libraries are subjected to phosphorylation by a selected kinase, the phosphorylated peptides are separated from the non-phosphorylated ones and sequenced either together or as single entities. The data from libraries are expected to provide optimal sequences within the limits of preselected peptide length and degeneracy (Songyang *et al.*, 1994).

The neural network method described here takes this analysis a step further and includes all phosphorylation sites of a certain acceptor type in one analysis. Using the neural network prediction for the general prediction of putative sites and the knowledge of specificity from the more classical approaches, it might be possible to develop an integrated system for accurately predicting the location of phosphorylation sites and the kinase that is involved.

Mapping phosphorylation sites on proteins is an important step towards understanding the catalytic process itself and the resulting effects on signal transduction events. Prediction methods have several advantages; they are fast, reproducible, publicly available and have been shown to be sufficiently accurate (Nielsen *et al.*, 1997, 1999) for optimizing experiments.

The prediction can assist the experimentalist who wants to design a mutagenesis experiment on a newly found protein of possibly unknown function. The method may be integrated as a part of proteomics identification approach, where the whole protein repertoire from a specific cell type or organism is analyzed for post-translational modifications and functionality.

We do not claim that we have included all known phosphorylation sites in our analysis. Our main concern is that some of the sites that were classified as non-phosphorylated sites in the training data sets may, in fact, be true phosphorylation sites and thereby bias the predictive performance. During the analysis we were made aware of several sites in the PDGF-receptor that are phosphorylated but assigned in the original data set as non-phosphorylated (Dr L. Rönnstrand, personal communication). After reassigning these sites, performance was clearly increased.

To further improve the method, additional knowledge about experimentally verified phos-

Table 3. Putative sites of dynamic phosphorylation/GlcNac-glycosylation of p300/CBP

Protein	Position	NetPhos	DictyOglyc	Acc.Res	Sequence
p300	594	0.719	0.934	T	AIFPTPDPA
CBP	615	0.719	0.934	T	AIFPTPDPA
p300	1849	0.754	0.948	S	QGLPSPTPA
CBP	1884	0.979	0.767	S	QSLPSPTSA
p300	2315	0.992	0.961	S	QPVPSPRPQ
CBP	2351	0.935	0.768	S	APVQSPRPQ
p300	2336	0.876	0.816	S	QPQPSPHHV
CBP	2372	0.914	0.825	S	QPQPSPHHV
p300	2346	0.808	0.928	S	PQTSSPHPG
CBP	2382	0.531	0.952	S	PQTGSPHPG

NetPhos score is the output score from the ensemble of neural networks trained on that acceptor residue (Acc.res) type. DictyOglyc is the score from the GlcNac Dictyostelium prediction server. The sequence shows the context of the acceptor residue ± four residues.

phorylation sites will be needed. Many of the phosphorylation sites described in the literature have not yet been included in PhosphoBase, and thereby not in this analysis. A search in Medline for the keyword phosphorylation yields 69,286 hits (phosphorylation AND site yields 9044 hits; October 1999). A reasonable estimate of the number of papers reporting novel phosphorylation sites may be of the order of 1000-5000. Thus, the continuous increase of phosphorylation site entries in PhosphoBase will make it likely that further improvement of the prediction method can be achieved.

In addition to predicting putative phosphorylation sites, it would be of great value to get a hint as to which kinase is likely to interact at this particular site. We are currently considering different approaches to achieve this goal. The most obvious approach is a simple alignment against extended versions of the known phosphorylation sites in PhosphoBase, from which the phosphorylation site annotation might provide information about the kinase or other interaction partners. A more advanced approach is to generate sequence profiles for all phosphorylation sites being modified by a certain kinase. A sequence profile describes, for each position, the probability of finding each of the 20 amino acids. Predicting the most probable kinase for a novel phosphorylation site would then be a matter of aligning the sequence against the kinase sequence profiles and report the best match. The same type of approach could be used to predict functional domains interacting with phosphoresidues, such as SH2 and PcTB domains.

A third approach would involve training a new neural network to classify substrate sites of known kinases. For example, PhosphoBase contains information about 160 sites for protein kinase A (PKA), 174 sites for PKC and 25 sites for the EGF-receptor tyrosine kinase. This method will then be implemented as a post-processing function to the existing phosphorylation site prediction method. Thus, the initial putative sites predicted in the novel protein will be fed to the kinase-classifying network, which will then output a prediction of the most probable kinase.

Another very important constraint of the highly compartmentalized eukaryotic cell is that many potential protein-protein interactions may never take place because of topological factors, e.g. one protein being membrane-bound, and the other localized in the nucleus. The role of compartmentalized processes for the regulation of the signal transduction kinase cascades has been shown to be very important for the correct target-substrate interactions to occur. During specific phases of the cell cycle, several kinases and phosphatases occur in specific cellular compartments, such as cytoplasmic or nuclear/chromosomal regions (Inagaki *et al*., 1994). Thus, methods available for predicting protein cellular localization might aid in deciding which of the putative sites are most biologically relevant (Andrade *et al*., 1998; Chou & Elrod, 1999; Nakai & Horton, 1999).

Materials & Methods

Data sets extracted from PhosphoBase

Experimentally verified phosphorylation sites were extracted mainly from PhosphoBase (Kreegipuu *et al*., 1999), which is available from http://www.cbs.dtu.dk/databases/PhosphoBase/.

The phosphoproteins were mostly from mammalian sources, with a few examples from viruses or plants. The data set consisted of 584 serine sites (251 protein entries), 108 threonine sites (85 protein entries), and 210 tyrosine sites (98 protein entries). No sites were identical within a 9-mer sequence. Negative examples of phosphorylation sites were assigned by two approaches.

(1) The standard approach. For each of the three acceptor types, a subset of the protein entries were categorized as being well characterized. All acceptor residues in the selected subsets, not reported as being phosphorylated, were assigned as negative sites.

(2) The augmented approach. All acceptor residues in the entire set of protein entries not reported as being phosphorylated, were assigned as negative sites *a priori*. Subsequently, during initial neural network training sessions, all negative sites predicted as positive sites were excluded. The resulting data set thus obtained was used for the final neural network training sessions.

Sequence logos

Sequence logos were used for displaying the position-specific features of complex sequence alignments as described earlier (Schneider & Stephens, 1990; Blom *et al*., 1996). Since phosphorylation sites are quite divergent, sequence logos are better suited at emphasizing the conserved positions than a multiple alignment.

Neural networks and cross-validation

The neural networks were of the standard feed-forward type (Minsky & Papert, 1988; Hertz *et al*., 1991). Details of sequence encoding, error functions, etc., may be found elsewhere (Blom *et al*., 1996). The predictive performance was monitored using the Mathews correlation coefficient (Mathews, 1975) during training and test of the networks.

Phylogenetic trees, indicating the relationship between the proteins in each of the three data sets (serine, threonine or tyrosine), were constructed using multiple alignments and neighbour-joining algorithms of the ClustalW package (Thompson *et al*., 1994) and visualized using the Drawtree program of the Phylip package (Felsenstein, 1989), as illustrated for the tyrosine data set (Figure 4). Based on the trees, each of the three data sets was divided into five parts in order to allow for cross-validated testing. This procedure used four of the five subgroups for training, while the last subgroup (non-sequence similar to the training sets) was used for testing the performance. Five different networks resulted from this approach for each of the three phosphoresidues.

The subgroups included related proteins like e.g. the MAP-kinase family or the Src-related tyrosine kinases (Hck, Blk, Abl, etc.), ensuring that test performance was measured on proteins non-homologous to the training set.

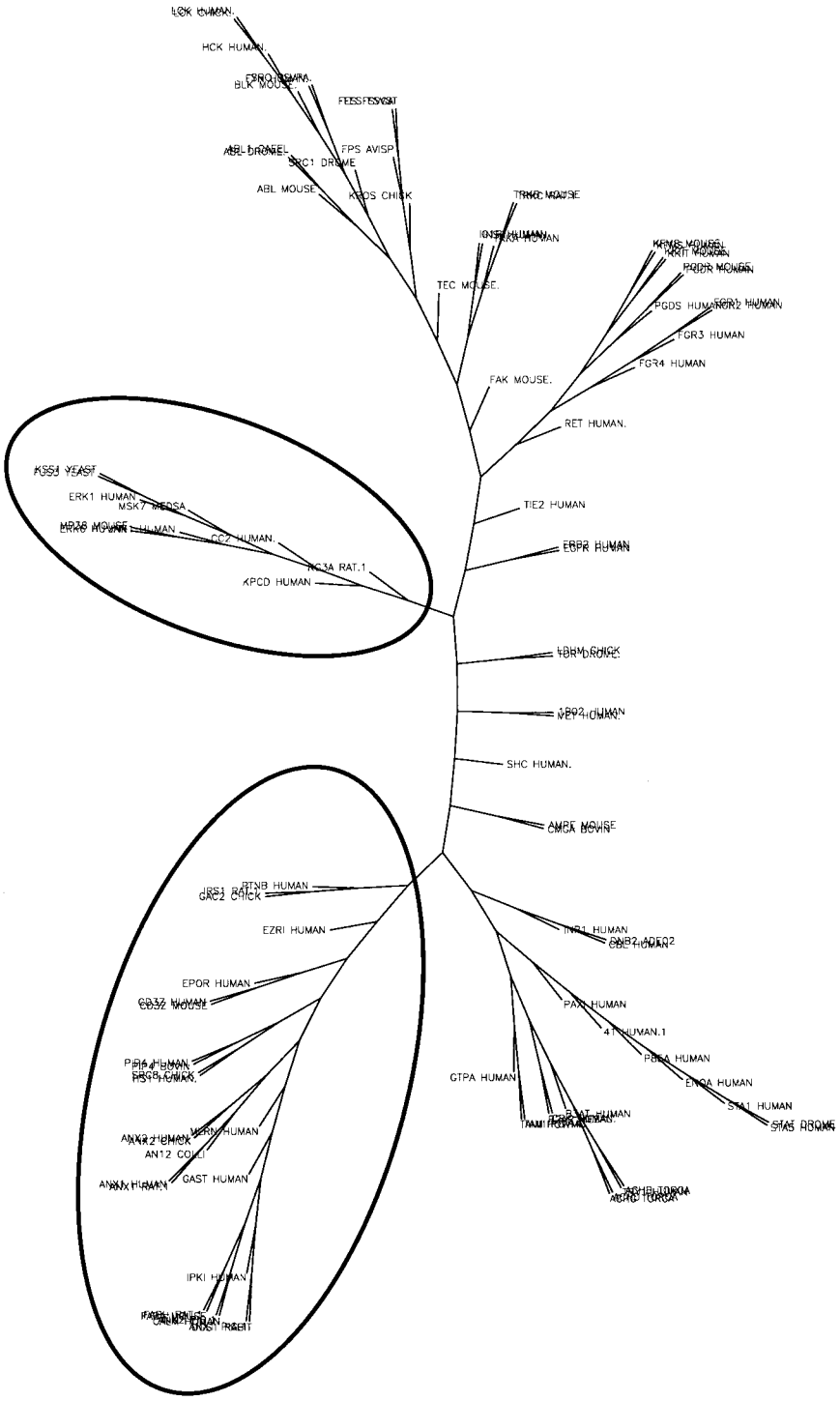

Figure 4. Phylogenetic tree showing the relationship of the phosphoproteins in the tyrosine data set. The circled branches represent subsets of the proteins with similar sequences, exemplifying the division of the data set.

Known tertiary structure of phosphorylation sites

The PDB protein structure database was scanned for proteins with verified phosphorylation sites (Bernstein *et al.*, 1977). Some protein entries contained annotated phosphorylation sites with the phosphate group included in the structure. In other cases, proteins known to be phosphorylated were found without phosphate groups in the crystal structure.

We collected all the phosphorylation sites being represented in protein structures and superimposed those of similar type, e.g. all tyrosine sites, at the acceptor residue using the Insight software package. The coordinates of the CO, C^{α}, and C^{β} atoms were used to superimpose the backbone C^{α} trace of the different peptide fragments. Fragments of length 9 (4 + 1 + 4), centered on the acceptor residue, were used whenever possible.

Neural network based on local tertiary structure

A neural network predicting C^{α} contact maps for peptide sequence inputs was used on peptide fragments from the data sets, up to a length of 33 residues centered at the phosphorylation sites (Lund *et al.*, 1997). For each pair of residues in the input sequence (e.g. residues i and $i+5$), the output from this method gives a probability score indicating the distance between the two residues compared to the average distance for all pairs at this sequence separation. The output was processed in order to extract probabilities pertaining to 9, 13, 17, 21 or 25 residues centered on the phosphoresidue. The same procedure was performed for the non-phosphorylation sites. The resulting data sets consisted of a number of probability parameters for each peptide (indirectly defining its local structure) for different fragment sizes.

Electronic access

The prediction method can be accessed on the Internet at http://www.cbs.dtu.dk/services/NetPhos/ or *via* e-mail by sending the word 'help' to NetPhos@cbs.dtu.dk.

The PhosphoBase database is also available on the Internet at http://www.cbs.dtu.dk/databases/PhosphoBase/.

Acknowledgments

We thank Kristoffer Rapacki for competent computer assistance and Jan Hansen for friendly and helpful discussions. This work was supported by the Danish National Research Foundation.

References

Ait-Si-Ali, S., Ramirez, S., Barre, F., Dkhissi, F., Magnaghi-Jaulin, L., Girault, J., Robin, P., Knibiehler, M., Pritchard, L., Ducommun, B., Trouche, D. & Harel-Bellan, A. (1998). Histone acetyl-transferase activity of CBP is controlled by cycle- dependent kinases and oncoprotein E1A. *Nature,* **396**, 184-186.

Andrade, M., O'Donoghue, S. & Rost, B. (1998). Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.* **276**, 517-525.

Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* **25**, 217-221.

Baldi, P. & Brunak, S. (1998). *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, MA.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Jr, E, F. M., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Blom, N., Hansen, J., Blaas, D. & Brunak, S. (1996). Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Protein Sci.* **5**, 2203-2216.

Chou, K. & Elrod, D. (1999). Protein subcellular location prediction. *Protein Eng.* **12**, 107-118.

Dale, S., Wilson, W. A., Edelman, A. M. & Hardie, D. G. (1995). Similar substrate recognition motifs for mammalian amp-activated protein kinase, higher plant hmg-coa reductase kinase-a, yeast snf1, and mammalian calmodulin-dependent protein kinase i. *FEBS Letters,* **361**, 191-195.

Felsenstein, J. (1989). Phylogeny inference package (version 3.2). *Cladistics,* **5**, 164-166.

Graves, L., Bornfeldt, K. & Krebs, E. (1997). Historical perspectives and new insights involving the MAP kinase cascades. *Advan. Sec. Mess. Phos. Res.* **31**, 49-62.

Hansen, J. E., Lund, O., Tolstrup, N., Gooley, A. A., Williams, K. L. & Brunak, S. (1998). NetOglyc: prediction of mucin type *O*-glycosylation sites based on sequence context and surface accessibility. *Glycoconj. J.* **15**, 115-130.

Hart, G. (1997). Dynamic O-linked glycosylation of nuclear and cytoskeletal proteins. *Annu. Rev. Biochem.* **66**, 315-335.

Hart, G., Greis, K., Dong, L., Blomberg, M., Chou, T., Jiang, M., Roquemore, E., Snow, D., Kreppel, L. & Cole, R. (1995). O-Linked N-acetylglucosamine: the yin-yang of Ser/Thr phosphorylation? Nuclear and cytoplasmic glycosylation. *Advan. Exp. Med. Biol.* **376**, 115-123.

Hertz, J., Krogh, A. & Palmer, R. (1991). *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA.

Hunter, T. (1998). The Croonian Lecture 1997. The phosphorylation of proteins on tyrosine: its role in cell growth and disease. *Phil. Trans. Roy. Soc. ser. B,* **353**, 583-605.

Inagaki, N., Ito, M., Nakano, T. & Inagaki, M. (1994). Spatiotemporal distribution of protein kinase and phosphatase activities. *Trends Biochem. Sci.* **19**, 448-452.

Johnson, L., Noble, M. & Owen, D. (1996). Active and inactive protein kinases: structural basis for regulation. *Cell,* **85**, 149-158.

Johnson, L., Lowe, E., Noble, M. & Owen, D. (1998). The eleventh datta lecture. the structural basis for substrate recognition and control by protein kinases. *FEBS Letters,* **430**, 1-11.

Kolibaba, K. & Druker, B. (1997). Protein tyrosine kinases and cancer. *Biochim. Biophys. Acta,* **1333**, F217-F248.

Kreegipuu, A., Blom, N. & Brunak, S. (1999). PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucl. Acids Res.* **27**, 237-239.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J. & Brunak, S. (1997). Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.* **10**, 1241-1248.

Mathews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta,* **405**, 442-451.

Minsky, M. & Papert, S. (1969, 1988). *Perceptrons*, MIT Press, Cambridge, MA.

Nakai, K. & Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24**, 34-36.

Nielsen, H., Brunak, S. & von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12**, 3-9.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1-6.

Patschinsky, T., Hunter, T., Esch, F. S., Cooper, J. A. & Sefton, B. M. (1992). Analysis of the sequence of amino acids surrounding sites of tyrosine phosphorylation. *Proc. Natl Acad. Sci. USA,* **79**, 973-977.

Pawson, T. (1995). Protein modules and signalling networks. *Nature,* **373**, 573-80.

Pinna, L. A. & Ruzzene, M. (1996). How do protein kinases recognize their substrates? *Biochim. Biophys. Acta,* **1314**, 191-225.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res,* **18**, 6097-6100.

Shikama, N., Lyon, J. & NB, L. T. (1997). The p300/CBP family: integrating signals with transcription factors and chromatin. *Trends Cell Biol.* **7**, 230-236.

Songyang, Z., Shoelson, S. E., Chaudhuri, M., Gish, G., Pawson, T., Haser, W. G., King, F., Roberts, T., Ratnofsky, S., Lechleider, R. J., Neel, B. G., Birge, R. B., Fajardo, J. E., Chou, M. M. & Hanafusa, H., *et al*. (1993). SH2 domains recognize specific phosphopeptide sequences. *Cell,* **72**, 767-778.

Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M. F., Piwnica-Worms, H. & Cantley, L. C. (1994). Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.* **4**, 973-982.

Thompson, J., Higgins, D. & Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.

Tinker, D. A., Krebs, E. A., Feltham, I. C., Attah-Poku, S. K. & Ananthanarayanan, V. S. (1988). Synthetic beta-turn peptides as substrates for a tyrosine protein kinase. *J. Biol. Chem.* **263**, 5024-5026.

Wilson, I. B. H., Gavel, Y. & von Heijne, G. (1991). Amino acid distributions around *O*-linked glycosylation sites. *Biochem. J.* **275**, 529-534.

Wu, C. H. (1997). Artificial neural networks for molecular sequence analysis. *Comput. Chem.* **21**, 237-256.

Zhou, S. & Cantley, L. (1995). Recognition and specificity in protein tyrosine kinase-mediated signalling. *Trends Biochem. Sci.* **20**, 470-475.

Edited by F. E. Cohen

(Received 16 July 1999; received in revised form 12 October 1999; accepted 12 October 1999)