Minireview

Statistical analysis of protein kinase specificity determinants

Andres Kreegipuu[a], Nikolaj Blom[b], Søren Brunak[b], Jaak Järv[a,*]

[a] *Institute of Chemical Physics, University of Tartu, 2 Jakobi Str., EE-2400 Tartu, Estonia*
[b] *Center for Biological Sequence Analysis, Building 208, The Technical University of Denmark, DK-2800 Lyngby, Denmark*

Received 16 April 1998

Abstract The site and sequence specificity of protein kinases, as well as the role of the secondary structure and surface accessibility of the phosphorylation sites on substrate proteins, was statistically analyzed. The experimental data were collected from the literature and are available on the World Wide Web at http://www.cbs.dtu.dk/databases/PhosphoBase/. The set of data involved 1008 phosphorylatable sites in 406 proteins, which were phosphorylated by 58 protein kinases. It was found that there exists almost absolute Ser/Thr or Tyr specificity, with rare exceptions. The sequence specificity determinants were less strict and were located between positions −4 and +4 relative to the phosphorylation site. Secondary structure and surface accessibility predictions revealed that most of the phosphorylation sites were located on the surface of the target proteins.

Key words: Protein phosphorylation; Protein kinase; Substrate specificity; Sequence logo

1. Introduction

Protein kinases are responsible for the regulation of a variety of physiological processes via phosphorylation of different structurally and/or functionally distinct proteins [1,2]. It has been estimated that about 2% of the vertebrate genome encodes protein kinases, which probably makes these proteins the most abundant family of enzymes in living cells [3]. Hence the selectivity of regulation of biochemical processes via protein phosphorylation should greatly depend upon the precision of the molecular recognition of the appropriate target proteins by protein kinases. Therefore the understanding of the substrate recognition mechanisms of protein kinases and comparison of substrate specificity of these enzymes has remained one of the key questions in the study of regulatory phosphorylation.

Following the concept that substrate specificity of protein kinases is essentially determined by the chemical nature of the phosphorylation site and the protein structure around the phosphorylatable residue, the substrate specificity of these enzymes can arbitrarily be discussed in different parts. Firstly, the site (residue) specificity of the enzymes determines the type of phosphoacceptor residue. Secondly, the importance of the protein primary structure around the phosphorylatable site is generally recognized and attempts have been made to describe the sequence specificity of protein kinases in terms of consensus sequence motifs. By definition the consensus sequence refers to the primary structure elements present in all substrates of a particular kinase [4]. Finally, molecular recognition of secondary and tertiary structure elements of the target proteins by protein kinases can be assumed, although until recently this aspect was mentioned rather seldom.

Several attempts have been made to formulate the consensus sequences for the most frequently studied protein kinases (for review see [4–7]). However, the steadily growing number of identified phosphorylation sites has made the concept vague, as for several protein kinases a significant number of rather different phosphorylation sites has been described. Recently the published experimental data on phosphorylation of protein substrates were collected and systematically presented in a database, described by Blom et al. [8]; they are available on the World Wide Web at http://www.cbs.dtu.dk/databases/PhosphoBase/.

In the present review this data source provided a unique possibility to statistically analyze the distribution of the naturally encoded amino acids in different positions around the phosphorylatable sites with the purpose of characterizing the substrate specificity patterns of different kinases. Using the same data we also made an attempt to analyze the influence of the secondary and tertiary structure of proteins on their phosphorylation. As the three-dimensional structure for most of the substrate proteins is still unknown, several prediction methods for the secondary and tertiary structure assignments, based on protein primary structure, were used for this purpose.

2. Procedures

2.1. Data

The experimental data on phosphorylation of serine, threonine, and tyrosine residues in proteins were systematically collected from literature sources and complemented with the primary structure data of these proteins from the SwissProt [9] and PIR [10] databases. All the collected data were converted into a uniform format and were listed in Phosphobase, a database of phosphorylation sites in proteins [8]. The currently available set of data involved 1008 phosphorylatable sites located in 406 substrate proteins. Most of these phosphorylation sites were from proteins originating from eukaryotic cells, but some prokaryotic and viral proteins were also involved. Phosphorylation sites for 58 distinct protein kinases were involved (Table 1) but for many phosphoresidues the enzyme responsible for phosphorylation has yet to be identified. It can be seen that for 10 enzymes more than 20 phosphorylation sites were described and for another 16 enzymes 5–19 phosphorylatable sites were identified. For all other protein kinases fewer than five phosphorylation sites have been analyzed which hindered a proper statistical analysis. Thus for many

*Corresponding author. Fax: (372) (7) 465 247.
E-mail: jj@chem.ut.ee

PII: S0014-5793(98)00503-1

Table 1
Protein kinases having more than one substrate protein in the analysis data set

Protein kinase (abbreviation)	Number of substrate proteins	Number of substrate sites[a]	Site specificity
Protein kinase C (PKC)	86	179 (148/31)	S/T
cAMP-dependent protein kinase (PKA)	95	161 (148/13)	S/T
Casein kinase II (CKII)	40	76 (64/12)	S/T
Calmodulin-dependent protein kinase II (CaM-II)	25	35 (30/5)	S/T
cGMP-dependent protein kinase (PKG)	17	32 (25/7)	S/T
Casein kinase I (CKI)	10	27 (24/3)	S/T
Cell division cycle protein kinase p34cdc2	12	44 (32/12)	S/T
Glycogen synthase kinase 3 (GSK3)	8	18 (14/4)	S/T
H1 histone kinase	4	11 (4/7)	S/T
Histone kinase	3	10 (10/0)	S/T
Mitogen-activated protein kinase (MAPK)	7	9 (8/1)	S/T
Phosphorylase kinase (PhK)	7	9 (8/1)	S/T
Rhodopsin kinase (RK)	2	9 (5/4)	S/T
AMP-dependent protein kinase (AMP-PK)	6	9 (9/0)	S/T
G protein-coupled receptor kinase 5 (GRK5)	2	8 (5/3)	S/T
S6 kinase (S6K)	4	7 (7/0)	S/T
Double-stranded DNA kinase	2	6 (0/6)	S/T
MAPKAP kinase-2	2	5 (5/0)	S/T
Growth factor-regulated kinase (ERT PK)	4	4 (2/2)	S/T
Myosin light chain kinase (MLCK)	3	4 (2/2)	S/T
Raf kinase	2	4 (4/0)	S/T
Multifactor protein kinase (MFPK)	2	3 (1/2)	S/T
Calmodulin-dependent protein kinase I (CaM-I)	2	3 (3/0)	S/T
MAP kinase p42 (p42mapk)	2	2 (1/1)	S/T
Glycogen synthase kinase 4 (GSK4)	2	2 (2/0)	S/T
Proline-directed PK	2	2 (2/0)	S/T
Epidermal growth factor receptor (EGFR)	11	25	Y
Tyrosine kinase Src	18	25	Y
Insulin receptor (INSR)	5	22	Y
Tyrosine kinase Abl	4	6	Y
Platelet-derived growth factor kinase (PDGFR)	3	6	Y
Tyrosine kinase Lck/Fyn	3	4	Y
Tyrosine kinase Fes	2	2	Y
Tyrosine kinase gag-fps	2	2	Y

[a]The number of experimentally identified phosphorylation sites. In the case of Ser/Thr specific protein kinases the ratio of serine and threonine phosphorylation sites is shown in parentheses (number of serine sites/number of threonine sites).

protein kinases only the preliminary estimate of the substrate specificity pattern can be obtained.

2.2. Statistical analysis of primary structure

The statistical analysis of the phosphorylatable sites' primary structure was made for protein kinases for which at least 20 phosphorylation sites were determined (Table 2). The sequences of these sites were aligned with respect to the phosphoresidue and the frequencies of appearance of amino acids at particular positions were calculated for each enzyme.

2.3. Sequence logos

Aligned sequences were also displayed as sequence logos (Fig. 1), a concept introduced by Schneider in 1990 [11]. Each position in the sequence alignment corresponds to a column in the sequence logo, the height of which characterizes the degree of conservation at this position. The height of the bar is equal to the Shannon information content [12], calculated for the 20 genetically encoded amino acids. When only one amino acid is found at a fixed position in all alignment sequences, i.e. the position is entirely conserved, the maximal height of the column is $\log_2 20 = 4.32$ bits. In contrast, when all 20 amino acids are equally probable at a position the height of the bar is 0 bits. The entire set of information content values forms a curve that represents the importance of various positions in the sequence. The height of each letter in a bar is proportional to the relative frequency of the corresponding amino acid residue. Thus the sequence logo would be a useful tool for presenting common primary structure features of peptides.

2.4. Secondary structure prediction

Since the spatial structures of many phosphorylated proteins were not available, three secondary structure prediction algorithms (PHDSec [13], nnpredict [14], and predator [15] algorithms) were used in the analysis. All these methods take the primary sequence as input and predict for each residue either a helical (H), an extended (E) or an unordered (coiled, C) conformation. Secondary structure prediction for 406 distinct proteins was carried out by all three prediction algorithms and the results are summarized in Table 3.

2.5. Surface exposure prediction

Surface exposure of the phosphorylation sites was estimated by the PHDAcc solvent accessibility prediction algorithm [16]. The algorithm predicts each residue in the sequence to be in a buried (B) or exposed (E) state in a water solution. Surface exposure of all amino acids in the 406 proteins were calculated and the results obtained for the phosphorylatable serine, threonine and tyrosine residues are summarized in Table 4.

3. Specificity for the phosphorylatable amino acid

The present survey confirms the existence of predominantly

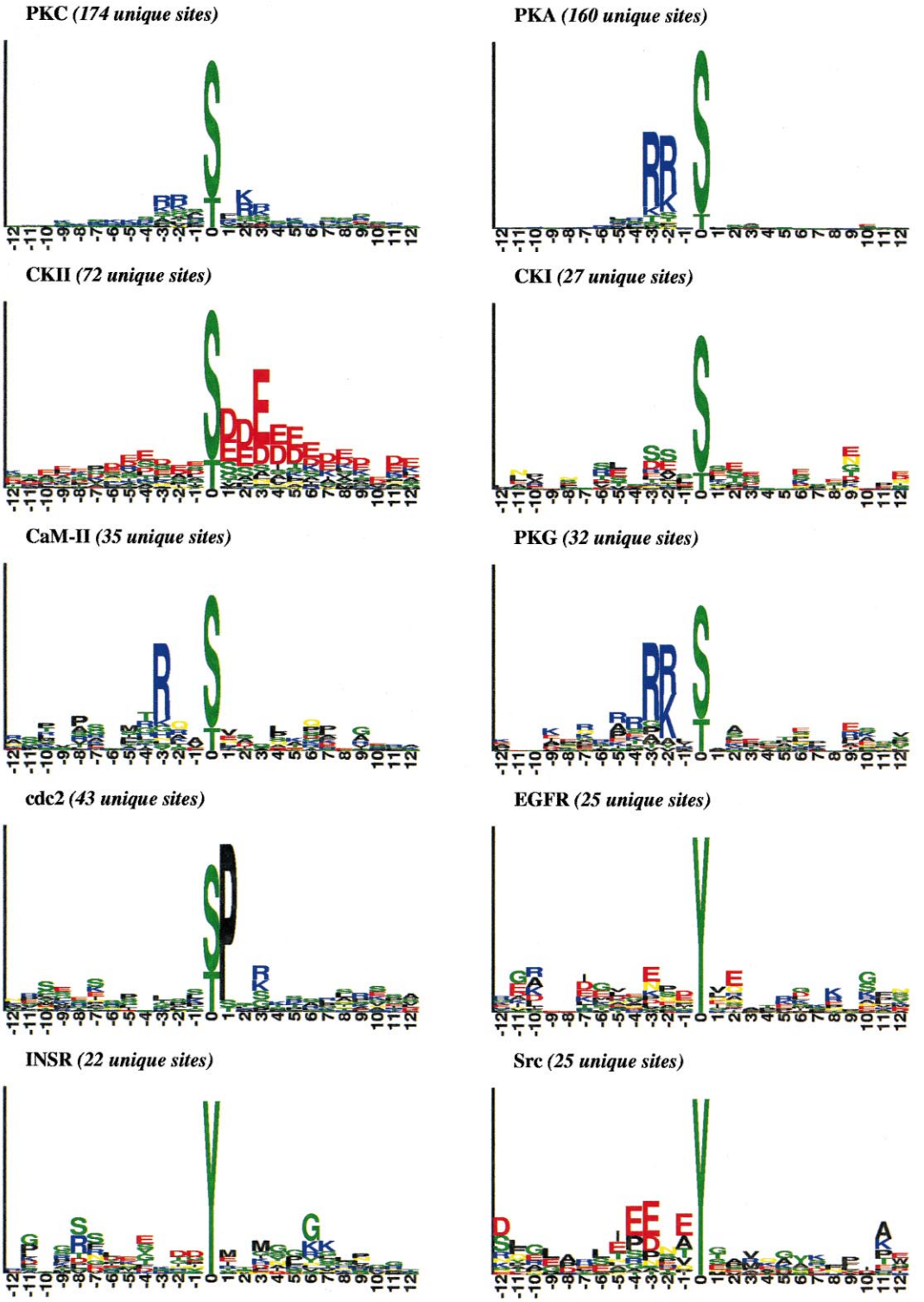

Fig. 1. Sequence logos for protein kinases having more than 20 unique (i.e. no two identical 25-residue fragments) phosphorylation sequences within the analyzed data. For explanation of sequence logos see Section 2. Hydrophobic residues are colored black, acidic red, basic blue, Gln and Asn yellow, and other hydrophilic residues green.

two types of protein kinases, phosphorylating either serine/threonine residues or tyrosine residues (Table 1). The dual specificity protein kinases seem to be rather exceptional and only some kinases have been found to phosphorylate both aliphatic and aromatic hydroxyl groups [17]. However, in most of these cases the phosphorylation efficiency for Ser/Thr and Tyr residues is very different and the dual specificity appears only under functionally irrelevant conditions. A remarkable exception is MAP kinase kinase which phosphorylates both Thr and Tyr residues separated only by one residue in the sequence of MAP kinase [7].

The collected data show that the majority of the Ser/Thr

Table 2
Overview of sequence specificities of protein kinases having more than 20 substrate sites in the used data set

Protein kinase	Position[a]							
	−4	−3	−2	−1	+1	+2	+3	+4
PKC	–	**R 30%**	R 26%	–	F 15%	K 28% R 27%	R 21% K 15%	S 20%
PKA	R 16%	**R 69%**	**R 59%** K 18%	–	L 15%	S 16%	–	–
CKII	E 22% S 15%	D 15% S 15%	E 15%	–	**D 36%** E 26%	**D 36%** E 26%	**E 63%** D 16%	E 29% D 28%
cdc2	–	L 21%	S 16%	S 19%	**P 93%**	–	R 27% K 20% S 16%	K 16%
CaM–II	R 17% T 23%	**R 71%**	Q 26%	A 26% V 17%	V 23% E 17%	A 17% S 17%	G 17%	L 23% P 17%
PKG	**R 31%**	**R 72%**	**R 47%** **K 41%**	L 16%	A 16%	A 25% S 16%	E 22% P 16%	G 16%
CKI	S 19%[b]	**S 35%**[b] D 19%	S 30%[b] E 22%	S 22%[b] L 19%	E 19% S 19%[b] L 15%	E 22% S 22%[b] T 15%[b]	E 19% G 15% S 15%[b]	S 19%[b]
EGFR	E 16% G 16%	**E 32%** N 20%	P 20% Q 16% S 16%	D 20% E 20%	V 20% L 16%	**E 36%** Q 16%	A 16%	P 16% T 16%
Src	**E 44%** P 20%	**E 48%** D 20%	N 24% S 16%	**E 36%** A 20% T 16%	G 20% E 16% S 16% T 16%	A 20% P 16%	V 24% M 16% R 16%	Q 20% E 16% K 16%
INSR	E 23%	–	D 23%	D 23%	M 23% E 18%	–	**M 32%**	G 23%

[a]Statistics about relative frequency of amino acids in the neighborhood of the phosphoacceptor residue. All residues with a frequency of 15% or higher are shown, residues with a frequency pf 30% or higher are shown in boldface.
[b]In many cases previous phosphorylation of these serine and threonine residues is required for CKI substrate sites as this enzyme displays acidophilic substrate specificity.

specific protein kinases have a preference for serine residues. The number of discovered threonine phosphorylation sites is in most cases 3–10 times less than the number of serine sites (Table 1). This fact cannot be explained by different natural occurrence of these amino acids since the ratio of serine and threonine content in proteins is 1.3:1 as calculated from sequence data for more than 60 000 different proteins, annotated in the SwissProt database [9].

Secondly, it can be seen in Table 1 that there seem to be rather big differences in Ser and Thr specificity between individual kinases. For example, these differences can be seen already in the case of protein kinase (PK) C and PKA, which are the most extensively studied enzymes of this class. In natural substrates of these enzymes the serine/threonine ratios were 4.8 and 11.4, respectively, showing that PKC is much more tolerant towards threonine than PKA.

4. Specificity for phosphorylation site sequence

For 54 distinct protein kinases the sequence of at least two phosphorylation sites can be found in the Phosphobase. Analysis of these data revealed that each of these enzymes can phosphorylate sites of rather different structure and the presence of some uniform peptide sequence motif in these sites was not obligatory for the catalytic reaction. This means that the substrate specificity of these protein kinases cannot be adequately presented in terms of a consensus sequence, assuming conservation of some elements of the primary structure of the substrate proteins. This conclusion can be illustrated by sequence logos in Fig. 1, calculated for the 10 most thoroughly investigated enzymes.

On the other hand, the same logos show that in all cases some areas of higher conservation of the primary structure can be identified. It is remarkable that these more conservative areas were located rather close to the phosphorylatable site and in most cases remained between positions −4 and +4. Thus the protein kinases seem to recognize only some limited part of the peptide around the phosphorylatable residue. This simplifies the analysis and agrees with the general finding that these enzymes can effectively and specifically phosphorylate also short synthetic peptides. On the other hand, this finding did not exclude the presence of some remote recognition areas, determined by spatial structure of the substrate proteins.

The frequency of appearance of amino acids between positions −4 and +4 around the phosphorylatable amino acid was analyzed for the most often investigated enzymes (Table 2). It can be seen that only in a few cases the same amino acid appeared in more that half of the phosphorylated sequences,

Table 3
Results of secondary structure predictions of phosphorylation sequences

Residue	Method	Coil (C)	Helical (H)	Extended (E)
Serine	nnpredict	81%	12%	7%
	predator	68%	24%	8%
	PHDSec	76%	17%	6%
Threonine	nnpredict	78%	13%	9%
	predator	61%	27%	12%
	PHDSec	70%	18%	12%
Tyrosine	nnpredict	66%	21%	13%
	predator	58%	24%	18%
	PHDSec	75%	14%	11%

Average fractions (between positions −4 and +4) of different structural assignments.

Table 4
Results of solvent accessibility predictions of phosphorylation sequences

Residue	Method	Exposed (E)	Buried (B)	Not predicted (N)
Serine	PHDAcc	70%	25%	5%
Threonine	PHDAcc	68%	25%	6%
Tyrosine	PHDAcc	66%	29%	6%

Average fractions (between positions −4 and +4) of different states.

and only in one case 93% of 44 substrate sites contained a conserved proline in position +1 (cdc2 kinase). However, even in this case the probability of finding a fixed sequence of two (or more) amino acids becomes very low. Therefore it can be assumed that substrate specificity of protein kinases is not based on recognition of some particular peptide sequence, but most likely is based on complex interaction between the substrate and enzyme molecules, governed by the sum of different specificity determining factors. In this case the same effectiveness of the phosphorylation reaction can be achieved by a variety of combinations of amino acids around the phosphorylatable residue and for an adequate modeling of this phenomenon these specificity determining factors should be quantified for each amino acid and each position of the recognized sequence. Neither of the methods applied here provided such a possibility.

The statistical analysis and the sequence logos of aligned phosphorylation site sequences confirmed that most of the Ser/Thr protein kinases can be divided into three classes, including basophilic (e.g. PKA, PKC, CaM-II kinase), acidophilic (e.g. casein kinases I and II) and proline-directed protein kinases (e.g. cyclin-dependent kinases, MAP kinases). Most of the analyzed tyrosine kinases were acidophilic. This probably means that electrostatic forces are important for substrate specificity of many protein kinases. These forces, however, should be supported by complementarity of the substrate and enzyme molecules, depending to a great extent on the spatial structure of the peptide chain around the phosphorylatable group. This factor probably explains the crucial role of the structurally important proline residues in the specificity of some protein kinases.

5. Secondary structure and surface exposure of phosphorylation site motifs

Secondary structure of peptide fragments, containing the phosphorylatable serine, threonine and tyrosine residues, was predicted for all protein kinase substrates and the results obtained are summarized in Table 3. It can be seen that all the used prediction procedures yielded quite similar results, although they are based on different principles. Secondly, it is interesting that most of the phosphorylatable sites were located in coil structures, while the ordered helical and extended structures were much less presented. Nevertheless, one cannot conclude from this study that all the phosphorylation sites must be in coiled conformation since 20–30% of the sequences were predicted to be in a helical or extended structure. The present analysis also revealed that the phosphorylatable serine residues seem to be more frequently predicted in coiled conformation than threonine residues. On the other hand, both serine and threonine were located in more coiled sequences than tyrosine. This may be related to different bulkiness or hydrophobicity of these amino acid side chains, probably influencing the spatial protein structure.

A similar secondary structure prediction was also made for serine, threonine and tyrosine residues in studied proteins that have not been reported as being phosphorylation sites. This analysis revealed that the presence of these amino acids in coiled structures seems to be a rather general trend. However, the frequency of coil formation by the appropriate peptide segments was in all cases 10–30% lower than in the case of phosphorylatable sites. This means that the secondary structure of proteins may also be important for recognition of the phosphorylation sites, although its influence seems to be much weaker that the effects of the peptide primary structure.

The same data set was used for prediction of the surface accessibility of the phosphorylatable serine, threonine and tyrosine residues in the protein substrates for protein kinases. The location of these residues in protein tertiary structure was described using the solvent accessibility predictions as described above. This approach classifies residues as either buried in the protein core or exposed on the surface.

It can be seen in Table 4 that about 65–70% of all the phosphorylatable residues were predicted to be on the surface of the protein. On the other hand, if this analysis was made in the case of all serine, threonine and tyrosine residues in the studied protein substrates, almost equal probability of internal and external location of these residues was obtained: 45–50% for Ser and Thr and around 55% for Tyr. This difference in predicted location of these amino acids supports the understanding that the phosphorylatable sites should be located on the surface of the substrate proteins to be accessible for protein kinases. This means that the polypeptide chain in the proximity of these reaction sites should possess a relatively unordered conformation and be flexible enough to fit the enzyme binding site.

6. Conclusions

In the framework of this survey the statistical analysis of the primary, secondary and tertiary structure aspects of substrate specificity of protein kinases was made. As might be expected, these enzymes reveal clear preferences within each class of these structural elements. However, it is also clear that no absolute specificity determinants related to structure of the substrate proteins can be defined. This means that protein kinases belong to a group of enzymes of rather broad specificity. On the other hand, the specificity patterns of these enzymes should be rigid enough to provide the necessary precision of targeting and selectivity of the regulatory phosphorylation phenomena.

This selectivity can be illustrated by the fact that even within the analyzed set of substrates, involving 406 distinct proteins with 14 849 Ser, 11 095 Thr and 6535 Tyr residues, only 647 serines, 150 threonines, and 196 tyrosines were reported to be phosphorylated.

The analysis points to the fact that the specificity determining factors can be most clearly identified at the level of the

primary structure of the phosphorylatable sites, while the higher structures seem only to support the recognition process. For example, the acidophilic or basophilic surrounding of the phosphorylatable residues in many substrate proteins clearly supports the external location of the appropriate peptide segments, making them highly hydrophilic. Combination of the proline residues with some ionogenic residues is also a clear reason for formation of irregular and externally located peptide coils.

The results of the present survey indicate that the qualitative description of protein kinase specificity can hardly be used for prediction of the phosphorylatable sites in substrate, even if the spatial structure of these proteins is considered.

Acknowledgements: This work was supported by a Grant from the Estonian Science Foundation ESF 3348, by the INCO Copernicus program (Contract IC-CT96-0900) and by the Danish National Research Foundation.

References

[1] Krebs, E.G. (1985) Biochem. Soc. Trans. 13, 813–820.
[2] Pawson, T. (1994) FASEB J. 8, 1112–1113.
[3] Hunter, T. (1994) Semin. Cell Biol. 5, 367–376.
[4] Kennelly, P.J. and Krebs, E.G. (1991) J. Biol. Chem. 266, 15555–15558.
[5] Kemp, B.E. and Pearson, R.B. (1990) Trends Biochem. Sci. 15, 342–346.
[6] Pearson, R.B. and Kemp, B.E. (1991) Methods Enzymol. 200, 62–81.
[7] Pinna, L.A. and Ruzzene, M. (1996) Biochim. Biophys. Acta 1314, 191–225.
[8] Blom, N., Kreegipuu, A. and Brunak, S. (1998) Nucleic Acids Res. 26, 382–386.
[9] Bairoch, A. and Apweiler, R. (1998) Nucleic Acids Res. 26, 38–42.
[10] Barker, W.C., Garavelli, J.S., Haft, D.H., Hunt, L.T., Marzec, C.R., Orcutt, B.C., Srinivasarao, G.Y., Yeh, L.S.L., Ledley, R.S., Mewes, H.W., Pfeiffer, F. and Tsugita, A. (1998) Nucleic Acids Res. 26, 27–32.
[11] Schneider, T.D. and Stephens, R.M. (1990) Nucleic Acids Res. 18, 6097–6100.
[12] Shannon, C.E. (1948) Bell Syst. Tech. J. 27, 379–423 and 623–656.
[13] Rost, B. and Sander, C. (1994) Proteins 19, 55–72.
[14] Kneller, D.G., Cohen, F.E. and Langridge, R. (1990) J. Mol. Biol. 214, 171–182.
[15] Frishman, D. and Argos, P. (1996) Protein Eng. 9, 133–142.
[16] Rost, B. and Sander, C. (1994) Proteins 20, 216–226.
[17] Lindberg, R.A., Quinn, A.M. and Hunter, T. (1992) Trends Biochem. Sci. 17, 114–119.