382–386 *Nucleic Acids Research, 1998, Vol. 26, No. 1*

PhosphoBase: a database of phosphorylation sites

Nikolaj Blom*, Andres Kreegipuu[1] and Søren Brunak

Center for Biological Sequence Analysis, Department of Chemistry, Building 207, The Technical University of Denmark, DK-2800 Lyngby, Denmark and [1]Institute of Chemical Physics, University of Tartu, 2 Jakobi Str., EE2400 Tartu, Estonia

Received September 2, 1997; Revised and Accepted October 3, 1997

ABSTRACT

PhosphoBase is a database of experimentally verified phosphorylation sites. Version 1.0 contains 156 entries and 398 experimentally determined phosphorylation sites. Entries are compiled and revised from the literature and from major protein sequence databases such as SwissProt and PIR. The entries provide information about the phosphoprotein and the exact position of its phosphorylation sites. Furthermore, part of the entries contain information about kinetic data obtained from enzyme assays on specific peptides. To illustrate the use of data extracted from PhosphoBase we present a sequence logo displaying the overall conservation of positions around serines phosphorylated by protein kinase A (PKA). PhosphoBase is available on the WWW at http://www.cbs.dtu.dk/databases/PhosphoBase/

INTRODUCTION

Phosphorylation of enzymes, receptors and other proteins is one of the most important signaling mechanisms in the regulation of cellular processes at the molecular level. Protein kinases catalyze the transfer of the γ-phosphate of a nucleoside triphosphate (usually ATP) to an acceptor residue, usually serine, threonine or tyrosine in the substrate protein (1,2).

It has been estimated that the mammalian genome encodes >2000 different protein kinases (3). Experiments by 2D-gel electrophoresis indicate that as much as 30–50% of the proteins in a eukaryotic cell may be phosphorylatable (4). Since there may be ~10 000–30 000 different protein species in a eukaryotic cell, the average kinase probably has more than one target protein. Considering that only ~5% of the proteins in the SwissProt database (5) have phosphorylated residues in their annotation, it is obvious that only a small fraction of the active phosphorylation sites in proteins has been identified so far.

Researchers working on phosphorylation site determination will use different approaches and levels of detail when reporting the results. Various experimental methods, e.g., direct sequencing or mass spectrometry, are employed. Data will usually state that residue *k* in protein X is phosphorylated, but provide no kinetic data or information about the kinase in question.

As substrate specificity of protein kinases is determined mainly by the amino acid sequence in proximity of the phosphorylatable residue, many protein kinases are known to phosphorylate synthetic oligopeptides with kinetics comparable to those of the intact protein substrates (6). Peptide substrates are usually related to phosphoacceptor sites in the natural phosphoproteins. Kinetic assays with peptides of different length and mutations are valuable tools for studying substrate specificity of protein kinases. Unfortunately much of the kinetic data derived from *in vitro* peptide studies is not available from a single source.

We have created a framework for submission of data on specific phosphorylation sites in proteins. These may be simple positional annotations or detailed kinetic parameters. We also take into consideration that several studies examine mutated motifs of a given phosphorylation site and thereby point to the key residues in the kinase-substrate interaction. The ability of PhosphoBase to incorporate detailed information about a given phosphorylation site in terms of interactions and kinetic parameters in an easily readable format makes it a unique tool. We anticipate that PhosphoBase will be a valuable tool for molecular biologists in search of phosphorylation sites in proteins, for biochemists in planning new studies in the field of substrate specificity of protein kinases, and also for theoretical studies in making conclusions and computational predictions on the specificity of phosphorylation reactions.

PROTEIN KINASE A SEQUENCE MOTIFS

The main types of phosphorylation in eukaryotic cells are of tyrosine and serine/threonine residues. Since not all serine, threonine and tyrosine residues in a phosphoprotein are phosphorylated, kinases must display some degree of specificity. Many studies indicate that this specificity is determined by the primary sequence surrounding the phosphorylatable residue (7–9). However, most kinases studied are able to accept variations in the surrounding sequence to a smaller or larger degree and related kinases may display overlapping, yet different specificities.

To study the specificity of protein kinase A/cAMP-dependent protein kinase (PKA), we extracted all sequences from PhosphoBase annotated as being phosphorylated by PKA. Forty different sequences were aligned at the phosphorylated serine, see sequence logo in Figure 1. The basophilic nature of PKA is readily seen, as arginine and lysine dominate positions –3 and –2. This is in agreement with the consensus pattern described in the Prosite database (10), pattern PDOC0004, which has the form '[R,K][R,K].S(p)', where [R,K] means arginine or lysine, '.' means any residue and S(p) is the phosphoserine. However, 9 of the 40 sequences, corresponding to 22.5%, did not match this pattern and would thus remain undetected in a search using only the Prosite pattern.

*To whom correspondence should be addressed. Tel: +45 45 252 477; Fax: +45 45 934 808; Email: nikob@cbs.dtu.dk

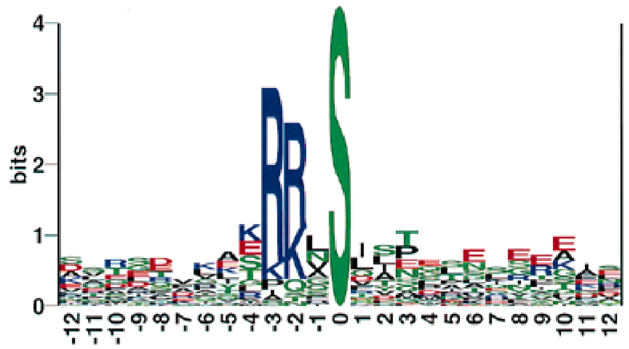


Figure 1. Protein kinase A (PKA) substrate site sequence logo (13,14) showing the Shannon information content (15). A set of 40 unique phosphorylation sites of window size 25 were aligned around the phosphorylated serine at position 0. The height of each column reflects the bias in the distribution of the amino acid residues found at that position. A totally conserved position has a height of 4.32 bits [$\log_2(20)$]. The letters correspond to the one-letter code for amino acid residues and the height of a single letter corresponds to its relative abundance at that position. Neutral and polar amino acids are shown in green, basic amino acids in blue, acidic amino acids in red and hydrophobic in black.

Furthermore, the PKA logo shows a tendency for basic residues at position –4 as well as hydrophobic residues at position +1 (Ile, Leu, Val). C-terminally to the phosphoserine, at positions +6 to +10, an abundance of acidic glutamate (E) residues are found. This region is most likely not recognized by the catalytic domain of PKA, but may still play a role in defining the phosphorylatable site, perhaps by making it more surface accessible.

The general tendency of many kinases to display a broad range of specificity calls for more sophisticated methods for predicting the location of phosphorylation sites (N.Blom, S.Gammeltoft, J.Hansen and S.Brunak, manuscript in preparation). Analysis of different sites phosphorylated by a particular kinase is now easily possible using PhosphoBase. This will hopefully lead to prediction methods that will be able to deal with the complex consensus patterns of phosphorylation sites.

DATA SOURCES

Data was collected from SwissProt (5) and PIR (11) protein databases, literature studies and personal experiments (University of Tartu). Phosphorylation sites annotated in protein databases as ‘potential’, ‘probable’ or ‘by similarity’ were not included.

DATABASE FORMAT

Overall description

Version 1.0 of PhosphoBase contains 156 entries and 398 experimentally determined phosphorylation sites.

PhosphoBase was designed to incorporate data on several levels of detail. The main part describes the latest revision dates, the name of the protein and species and cross-references to other databases. The positions of serine, threonine or tyrosine residues that have been described as phosphorylation sites are listed, followed by the actual sequence with a visual indication of the positions.

The second part of the entry presents detailed phosphorylation information. Literature references to the original phosphorylation site identification reports, information about the kinase catalyzing the phosphorylation reaction and kinetic data about peptides related to a particular phosphoprotein are listed.

All fields in the main part are required for each entry although they may be empty. In the second part of the entry most fields are optional as they may not apply to the phosphorylation site in question. The notation ‘//’ marks the end of each entry.

Two entries are shown in Figure 2. Note that in the entry for pyruvate kinase (A005), several detailed studies on natural and mutated peptides are reported, whereas the entry for src kinase (A006) describes mainly the position of several phosphorylation sites.

Description of fields

Main section

ACCESSION PhosphoBase accession code (single letter + 3 digits)

DATE Dates for creation and updates

PROT_ID Protein name

SPECIES Species name (latin and common)

DB_XREF Database cross-reference to SwissProt, PIR or GenBank

SERINE Position of phosphorylated serines. Parentheses indicate the peptide accession code (see PEPTIDE section) described below and is constructed from the ACCESSION code plus [A–Z]

THREONINE Same as above, but for threonine

TYROSINE Same as above, but for tyrosine

Sequence section

SEQUENCE Number of residues followed by the actual sequence in 80 residues per line. Then follows the assignment field, where S, T and Y denote phosphorylated serine, threonine and tyrosine, respectively. A ‘.’ means that no information about this position is provided. It does not indicate that this position is never phosphorylated.

Phosphorylation section

PEPTIDE The first field contains the peptide accession number which is constructed from the ACCESSION number plus [A–Z]. The second field contains the beginning and ending positions in the natural protein and the third field contains the actual peptide sequence. The third field contains the word ‘natural’ if the protein was examined as a whole or if details about the peptide used are unknown.

MUTATION Mutations from native protein sequence (optional). In case of studies on mutated peptides, this field indicates which residues were changed, e.g., 43(S→T), meaning that serine 43 was changed to threonine.

EXPERIMENT The first field contains the number of the reference pertaining to this experiment, while the second field contains the type and position of the residue being described (e.g., S–43).

KM Value of kinetic constant K_m (optional)

VMAX Value of kinetic constant V_{max} (optional)

KCAT Value of kinetic constant K_{cat} (optional)

KINASE Protein kinase which phosphorylates the residue described in EXPERIMENT (optional). Common abbreviations are used. In case of ambiguity, refer to the Help section on the PhosphoBase WWW-pages (see below).

ASSAY Conditions of kinetic experiments, e.g., pH, temperature, enzyme activity (optional)

INTERACTION Possible interaction partners of the phosphorylated residue (e.g., SH2 or PTB domains, other kinases) (optional).

EXP_COMMENT Experimental comment. Comments to indicate any other important details pertaining to EXPERIMENT (optional)

Reference section

REFERENCE [N] relates to the first field described for each EXPERIMENT section.

COMMENT Overall comments to indicate any other important details (optional)

FUTURE VERSIONS

At present, PhosphoBase includes data on phosphorylated serine, threonine or tyrosine residues in eukaryotic proteins. However, phosphorylation may also occur on histidine, lysine or arginine residues (12) or may occur in prokaryotic proteins. In any case, the format proposed here will easily handle these cases if needed.

Other relevant subjects which might be incorporated in future versions include information about pseudosubstrate peptides and data on phosphatase interactions at given sites.

ACCESS AND CITATION

PhosphoBase is made publicly available on the WWW at http://www.cbs.dtu.dk/databases/PhosphoBase/ . PhosphoBase depends on the quality and use of the data provided. Therefore, we encourage people in the field of phosphorylation or related areas to submit any relevant updates, corrections or new information to PhosphoBase, which will be accordingly updated.

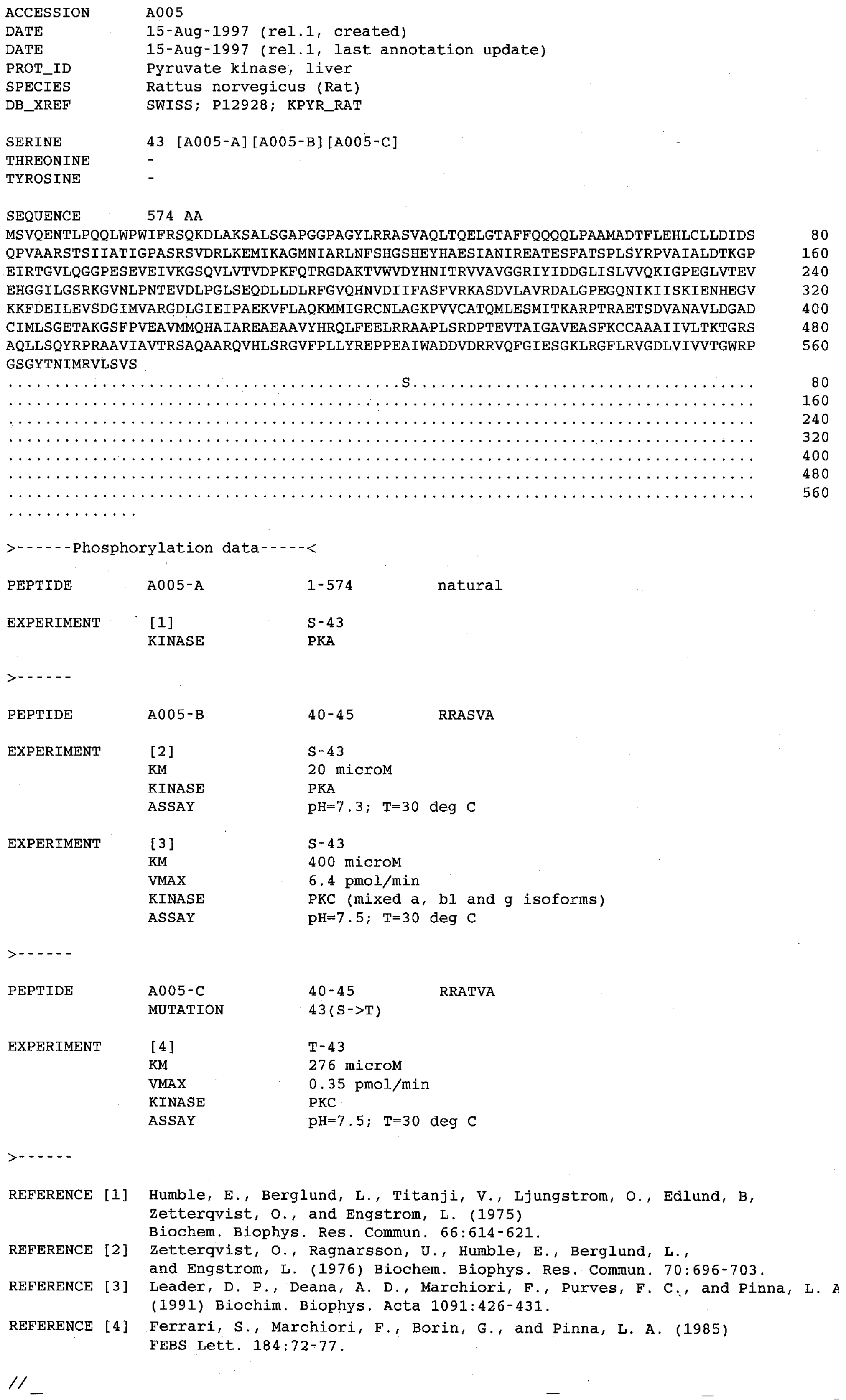
We encourage users of PhosphoBase to cite this paper.

ACKNOWLEDGEMENTS

We thank Kristoffer Rapacki, Hans Henrik Stærfeldt and Kristian de Lichtenberg for competent computer assistance, Jaak Järv, Mart Loog and Katrin Sak for useful suggestions about the database format. This work was supported by the Danish National Research Foundation.

REFERENCES

1 Krebs,E.G. and Beavo,J.A. (1979) *Annu. Rev. Biochem.*, **49**, 923–959.
2 Pawson,T. (1994) *FASEB J.*, **8**, 1112–1113.
3 Hunter,T. (1994) *Semin. Cell Biol.*, **5**, 367–376.
4 Pinna,L.A. and Ruzzene,M. (1996) *Biochim. Biophys. Acta*, **1314**, 191–225.
5 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36. [See also this issue *Nucleic Acids Res.* (1998) **26**, 38–42.]
6 Kemp,B.E. and Pearson,R.B. (1991) *Methods Enzymol.*, **200**, 121–135.
7 Cantley,L.C. and Songyang,Z. (1994) *J. Cell Sci.*, **18** (Supplement), 121–126.
8 Tegge,W., Frank,R., Hofmann,F. and Dostmann,W.R. (1995) *Biochemistry*, **34**, 10569–10577.
9 Loog,M., Eller,M., Ekman,P., Engstrom,L., Eriksson,S., Jarv,J., Ragnarsson,U. and Toomik,R. (1995) *Bioorganic Chem*., **22**, 328–336.
10 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **25**, 217–221.
11 George,D.G., Dodson,R.J., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Sidman,K.E., Srinivasarao,G.Y., Yeh,L.S.L., Arminski,L.M., Ledley,R.S., Tsugita,A. and Barker,W.C. (1997) *Nucleic Acids Res.*, **25**, 24–27. [See also this issue *Nucleic Acids Res.* (1998) **26**, 27–32.]
12 Matthews,H.R. (1995) *Pharmacol. Ther.*, **67**, 323–350.
13 Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
14 Blom,N., Hansen,J., Blaas,D. and Brunak,S. (1996) *Protein Sci.*, **5**, 2203–2216.
15 Shannon,C.E. (1948) *Bell System Tech. J.*, **27**, 379–423 and 623–656.



```
ACCESSION       A005
DATE            15-Aug-1997 (rel.1, created)
DATE            15-Aug-1997 (rel.1, last annotation update)
PROT_ID         Pyruvate kinase, liver
SPECIES         Rattus norvegicus (Rat)
DB_XREF         SWISS; P12928; KPYR_RAT

SERINE          43 [A005-A][A005-B][A005-C]
THREONINE       -
TYROSINE        -

SEQUENCE        574 AA
MSVQENTLPQQLWPWIFRSQKDLAKSALSGAPGGPAGYLRRASVAQLTQELGTAFFQQQQLPAAMADTFLEHLCLLDIDS    80
QPVAARSTSIIATIGPASRSVDRLKEMIKAGMNIARLNFSHGSHEYHAESIANIREATESFATSPLSYRPVAIALDTKGP   160
EIRTGVLQGGPESEVEIVKGSQVLVTVDPKFQTRGDAKTVWVDYHNITRVVAVGGRIYIDDGLISLVVQKIGPEGLVTEV   240
EHGGILGSRKGVNLPNTEVDLPGLSEQDLLDLRFGVQHNVDIIFASFVRKASDVLAVRDALGPEGQNIKIISKIENHEGV   320
KKFDEILEVSDGIMVARGDLGIEIPAEKVFLAQKMMIGRCNLAGKPVVCATQMLESMITKARPTRAETSDVANAVLDGAD   400
CIMLSGETAKGSFPVEAVMMQHAIAREAEAAVYHRQLFEELRRAAPLSRDPTEVTAIGAVEASFKCCAAAIIVLTKTGRS   480
AQLLSQYRPRAAVIAVTRSAQAARQVHLSRGVFPLLYREPPEAIWADDVDRRVQFGIESGKLRGFLRVGDLVIVVTGWRP   560
GSGYTNIMRVLSVS
..........................................S.....................................    80
................................................................................   160
................................................................................   240
................................................................................   320
................................................................................   400
................................................................................   480
................................................................................   560
..............

>------Phosphorylation data-----<

PEPTIDE         A005-A          1-574           natural

EXPERIMENT      [1]             S-43
                KINASE          PKA

>------

PEPTIDE         A005-B          40-45           RRASVA

EXPERIMENT      [2]             S-43
                KM              20 microM
                KINASE          PKA
                ASSAY           pH=7.3; T=30 deg C

EXPERIMENT      [3]             S-43
                KM              400 microM
                VMAX            6.4 pmol/min
                KINASE          PKC (mixed a, b1 and g isoforms)
                ASSAY           pH=7.5; T=30 deg C

>------

PEPTIDE         A005-C          40-45           RRATVA
                MUTATION        43(S->T)

EXPERIMENT      [4]             T-43
                KM              276 microM
                VMAX            0.35 pmol/min
                KINASE          PKC
                ASSAY           pH=7.5; T=30 deg C

>------

REFERENCE [1]   Humble, E., Berglund, L., Titanji, V., Ljungstrom, O., Edlund, B,
                Zetterqvist, O., and Engstrom, L. (1975)
                Biochem. Biophys. Res. Commun. 66:614-621.
REFERENCE [2]   Zetterqvist, O., Ragnarsson, U., Humble, E., Berglund, L.,
                and Engstrom, L. (1976) Biochem. Biophys. Res. Commun. 70:696-703.
REFERENCE [3]   Leader, D. P., Deana, A. D., Marchiori, F., Purves, F. C., and Pinna, L. A
                (1991) Biochim. Biophys. Acta 1091:426-431.
REFERENCE [4]   Ferrari, S., Marchiori, F., Borin, G., and Pinna, L. A. (1985)
                FEBS Lett. 184:72-77.

//
```

Figure 2. (Above and opposite) Two examples of entries in PhosphoBase.

```
PB_ACCESSION    A006
DATE            21-Aug-1997 (rel.1, created)
DATE            21-Aug-1997 (rel.1, last annotation update)
PROT_ID         Proto-oncogene tyrosine-protein kinase src
SPECIES         Gallus Gallus (chicken)
DB_XREF         SWISS; P00523; SRC_CHICK

SERINE          11 [A006-A], 16 [A006-B], 47 [A006-C]
THREONINE       -
TYROSINE        415 [A006-D], 526 [A006-E]

SEQUENCE        532 AA
GSSKSKPKDPSQRRRSLEPPDSTHHGGFPASQTPNKTAAPDTHRTPSRSFGTVATEPKLFGGFNTSDTVTSPQRAGALAG   80
GVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLTTGQTGYIPSNYVAPSDSIQAEEWYFGKITRRESERL  160
LLNPENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFSSLQQLVAYYSKHADGLCHRL  240
TNVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLKPGNMSPEAFLQEAQVMKKLRHEK  320
LVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGEMGKYLRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCK  400
VADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVERGYRM  480
PCPPECPESLHDLMCQCWRRDPEERPTFEYLQAFLEDYFTSTEPQYQPGENL
..........S....S..............................S.................................   80
................................................................................  160
................................................................................  240
................................................................................  320
................................................................................  400
..............Y.................................................................  480
.............................................Y......


>-----Phosphorylation data-----<

PEPTIDE         A006-A          1-532       natural

EXPERIMENT      [1]             S-11

>------

PEPTIDE         A006-B          1-532       natural

EXPERIMENT      [2]             S-16

>------

PEPTIDE         A006-C          1-532       natural

EXPERIMENT      [2]             S-47

>------

PEPTIDE         A006-D          1-532       natural

EXPERIMENT      [3]             Y-415
                KINASE          autophosphorylation

>------

PEPTIDE         A006-E          1-532       natural

EXPERIMENT      [4]             Y-426

>------

REFERENCE [1]   Gould K.L., Woodgett J.R., Cooper J.A., Buss J.E., Shalloway D.,
                Hunter T.(1985), Cell 42:849-857
REFERENCE [2]   Hardie G.  and Hanks S. (1995)
                "The Protein Kinase Factsbook", Academic Press, London
REFERENCE [3]   Smart J.E., Oppermann H., Czernilofsky A.P., Purchio A.F.,
                Erikson R.L., Bishop J.M. (1981)
                Proc. Natl. Acad. Sci. U.S.A. 78:6013-6017
REFERENCE [4]   Cooper J.A., Gould K.L., Cartwright C.A., Hunter T.(1986)
                Science 231:1431-1434
//
```