Nucleic Acids Research, 1999, Vol. 27, No. 1 237–239

PhosphoBase, a database of phosphorylation sites: release 2.0

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

Institute of Chemical Physics, University of Tartu, 2 Jakobi St., EE2400 Tartu, Estonia and [1]Center of Biological Sequence Analysis, Institute of Biotechnology, Building 208, The Technical University of Denmark, DK-2800 Lyngby, Denmark

Received October 5, 1998; Accepted October 7, 1998

ABSTRACT

PhosphoBase contains information about phosphorylated residues in proteins and data about peptide phosphorylation by a variety of protein kinases. The data are collected from literature and compiled into a common format. The current release of PhosphoBase (October 1998, version 2.0) comprises 414 phosphoprotein entries covering 1052 phosphorylatable serine, threonine and tyrosine residues. The kinetic data from peptide phosphorylation assays for ~330 oligopeptides is also included. The database entries are cross-referenced to the corresponding records in the Swiss-Prot protein database and literature references are linked to MedLine records. PhosphoBase is available via the WWW at http://www.cbs.dtu.dk/databases/PhosphoBase/

INTRODUCTION

Protein phosphorylation is a key event in many signal transduction pathways of biological systems. Protein kinases recognize and phosphorylate specific amino acid residues (mainly serine, threonine or tyrosine) in the substrate proteins. The research of protein phosphorylation has been essential in understanding intracellular signaling pathways. The number of identified phosphorylation sites is steadily growing—several thousand are now known. However, this seems to be only a small fraction of all potential phosphorylation possibilities since the estimated fraction of phosphoproteins may be as high as 30–50% of the total protein repertoire (1).

We have collected information about known protein phosphorylation sites from the literature and compiled these data into a dedicated database—PhosphoBase. The aim of the database is to be an extensive data source and repository for the researchers working in the fields related to protein phosphorylation and signal transduction.

The content and format of PhosphoBase were described in the 1998 Database Issue of *Nucleic Acids Research* (2). In release 2.0 of PhosphoBase the number of entries has more than doubled—it now contains 414 phosphoprotein entries including detailed data about 1052 phosphorylatable serine, threonine and tyrosine residues. The oligopeptide part of the database is also remarkably larger than in the previous release and there is now information about phosphorylation assays for more than 330 different peptides.

The general format of PhosphoBase has remained essentially unchanged (2), only a few minor additions and changes have been implemented to improve data presentation, parsing and handling. We have also enhanced the WWW user interface to the database such that in addition to accessing general information and retrieval of the whole database, the users can now browse and query PhosphoBase online via a search engine. Direct links from PhosphoBase entries to the corresponding Swiss-Prot (3) protein entries and MedLine abstracts have been established.

CHANGES IN THE FORMAT OF PhosphoBase

PhosphoBase is presented as a structured text file consisting of separate entries for each phosphoprotein. Each entry is further organised into several subsections and divided into individual fields that are identified by specific keywords. The detailed format specification of the database has been described previously (2).

As a modification to the previous format, the accession codes of subsections for native protein phosphorylation sites and synthetic peptides now have different formats. In this way subsections for big protein phosphorylation sites and peptides can easily be distinguished. Also, as synthetic peptides are often derived from phosphorylation site sequences in phosphoproteins, their relation to the parent protein phosphorylation site is now indicated. Thus, phosphoresidues in native proteins are referred to by an accession code consisting of the entry code plus A–Z (e.g. A001-A, A001-B, ...) whereas accession numbers for oligopeptide data are now based on the related phosphorylation site motif—peptides, whose sequences are derived from phosphoresidue designated as A001-A have identifiers A001-A01, A001-A02, A001-A03, ...; peptides derived from the motif A001-B have identifiers A001-B01, A001-B02, ..., respectively.

The PhosphoBase data are mainly collected from original research reports and all the information is tagged to the respective papers. In release 2.0 we have added MedLine record identification numbers (if applicable) in the REFERENCE field for each bibliographic reference for direct access to the abstract in the MedLine database.

*To whom correspondence should be addressed. Tel: +45 45 252 477; Fax: +45 45 931 585; Email: nikob@cbs.dtu.dk

DATABASE CONTENTS

The amount of data in PhosphoBase has more than doubled over the past year. The current release of PhosphoBase (October 1998, release 2.0) consists of 414 phosphoprotein entries. There are proteins from 67 different organisms—mainly from vertebrate species but several plant, insect, bacterial and viral phosphoproteins are also present (Table 1). These proteins contain altogether 1052 experimentally determined phosphorylation sites, including 688 serine, 168 threonine and 196 tyrosine residues which are phosphorylatable by about 100 different protein kinases. The number of phosphorylatable peptides included in PhosphoBase has also remarkably increased in comparison with the previous release and is now more than 300. The information in the database is compiled from more than 500 bibliographic sources.

PHOSPHORYLATION MOTIFS FOR PROTEIN KINASES

The mechanism of substrate recognition by protein kinases has been one of the major challenges in phosphorylation research over the years. It has been found that the primary sequence flanking the phosphoacceptor plays an important role in defining the potency of the substrate (4). However, numerous studies have revealed that the local amino acid sequence is certainly not the sole determinant of substrate specificity but several other factors, including local spatial structure and surface accessibility, also play a role (1).

Table 1. Overview of PhosphoBase content (release 2.0): number of phosphoprotein entries from different species and synthetic peptides

Protein type	Sequences
Vertebrate	**321**
incl. proteins from:	
Homo sapiens (human)	127
Rattus norvegicus (rat)	55
Bos taurus (bovine)	48
Mus musculus (mouse)	32
Oryctolagus cuniculus (rabbit)	26
Gallus gallus (chicken)	23
Insect	**7**
Plant	**9**
Other eukaryotic	**16**
Bacterial	**22**
Viral	**22**
All proteins	**397**
All peptides	**329**
Total number of sequences in PhosphoBase	**726**

The remarkable collection of phosphorylation sequence data in PhosphoBase provides a good basis for the analysis of substrate sequence preferences of protein kinases. We performed an

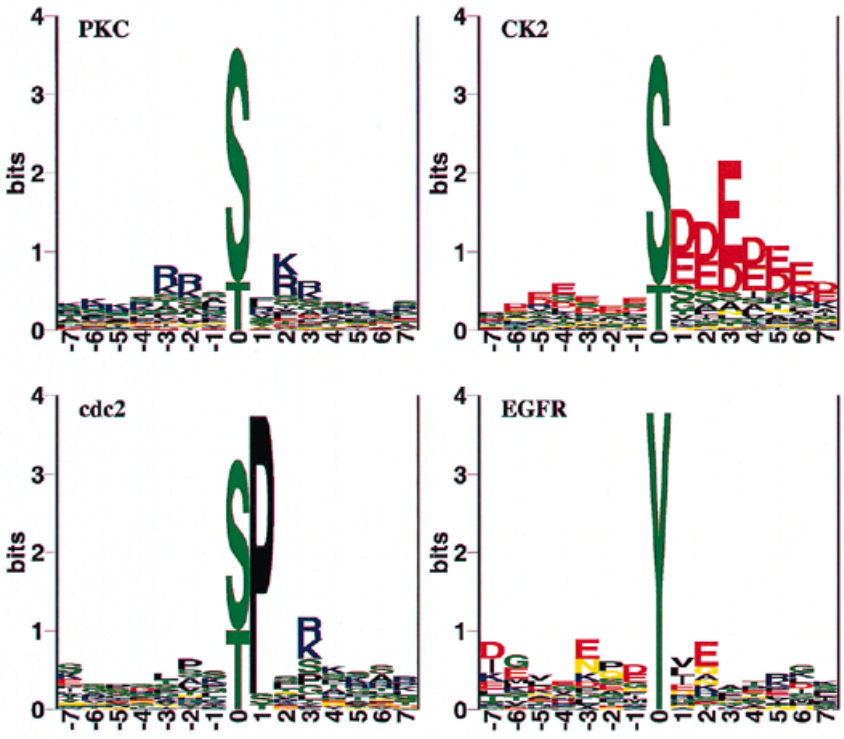

Figure 1. Conservation of primary structure features around the phosphoacceptor residue for a selection of protein kinases. The sequence logos of substrate sequence motifs are shown for a basophilic protein kinase, protein kinase C (PKC, 181 sites), an acidophilic protein kinase, protein kinase CK2 (79 sites), a proline-directed protein kinase, protein kinase cdc2 (57 sites), and a protein tyrosine kinase, epidermal growth factor receptor (EGFR, 25 sites). For further explanation of information content and sequence logos, see refs 7 and 8.

analysis of substrate sequence motifs for a few extensively studied protein kinases. The results, in the form of sequence logos, are shown in Figure 1. In general, the results agree with previous knowledge of substrate specificity of these protein kinases and are compatible with current consensus sequence patterns (5). Primary structure preferences can be strongly pronounced (e.g. Pro residue at +1 position for cdc2, basic residues for PKA, PKC) or weakly pronounced (e.g. most tyrosine kinases). However, it seems that kinases rarely have absolute requirements for the substrate protein sequence. A more thorough analysis of substrate specificity based on PhosphoBase data has been published (6).

DATABASE ACCESS

The PhosphoBase public WWW server is available at http://www.cbs.dtu.dk/databases/PhosphoBase/ . The database can be retrieved from the server as a single text file consisting of ~25 000 lines (~1 Mb in release 2.0). The release notes of the database and supplementary material such as database format specification, list of used abbreviations, statistics about database content, etc., are also available. By using the enhanced PhosphoBase WWW server all the contents of the database can be queried and browsed via a new search engine. Database users have an option to directly access Swiss-Prot (3) protein entries and NCBI Pubmed MedLine records cross-referenced in PhosphoBase via hypertext links.

All relevant updates, corrections or new information concerning PhosphoBase from researchers working in the field of phosphorylation should be submitted to the authors. Users of PhosphoBase are encouraged to cite this paper.

ACKNOWLEDGEMENTS

We thank Kristoffer Rapacki for competent computer assistance. This work was supported by the Danish National Research Foundation and by Estonian Science Foundation Grant 3348.

REFERENCES

1 Pinna,L.A. and Ruzzene,M. (1996) *Biochim. Biophys. Acta*, **1314**, 191–225.
2 Blom,N., Kreegipuu,A. and Brunak,S. (1998) *Nucleic Acids Res.*, **26**, 382–386.
3 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
4 Kemp,B.E. and Pearson,R.B. (1991) *Methods Enzymol.*, **200**, 121–135.
5 Kennelly,P.J. and Krebs,E.G. (1991) *J. Biol. Chem.*, **266**, 15555–15558.
6 Kreegipuu,A., Blom,N., Brunak,S. and Järv,J. (1998) *FEBS Lett.*, **430**, 45–50.
7 Shannon,C.E. (1948) *Bell Syst. Tech. J.*, **27**, 379–423 and 623–656.
8 Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.