PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation

Peter V. Hornbeck[1], Indy Chabra[2], Jon M. Kornhauser[1], Elzbieta Skrzypek[1] and Bin Zhang[1]

[1]Cell Signaling Technology, Beverly, MA, USA
[2]ForScience Inc., Stony Brook, NY, USA

PhosphoSite™ is a curated, web-based bioinformatics resource dedicated to physiologic sites of protein phosphorylation in human and mouse. PhosphoSite is populated with information derived from published literature as well as high-throughput discovery programs. PhosphoSite provides information about the phosphorylated residue and its surrounding sequence, orthologous sites in other species, location of the site within known domains and motifs, and relevant literature references. Links are also provided to a number of external resources for protein sequences, structure, post-translational modifications and signaling pathways, as well as sources of phospho-specific antibodies and probes. As the amount of information in the underlying knowledgebase expands, users will be able to systematically search for the kinases, phosphatases, ligands, treatments, and receptors that have been shown to regulate the phosphorylation status of the sites, and pathways in which the phosphorylation sites function. As it develops into a comprehensive resource of known *in vivo* phosphorylation sites, we expect that PhosphoSite will be a valuable tool for researchers seeking to understand the role of intracellular signaling pathways in a wide variety of biological processes.

Received	5/12/03
Revised	30/1/04
Accepted	2/2/04

Keywords: Database / Kinase / Phosphorylation / Phosphorylation site / Signal transduction

1 Introduction

Protein phosphorylation is a critical and dynamic protein modification that regulates a broad spectrum of regulatory processes including development [1], the cell cycle [2], metabolic pathways [3], oncogenic transformation [4], immunological responsiveness [5, 6], and memory [7, 8]. Over 510 distinct protein kinases [9] and 100 phosphoprotein phosphatases [10] are encoded in the human genome, and upwards of 40% of cellular proteins appear to be phosphorylated during some stage of growth and differentiation (Hornbeck, unpublished observations). Many proteins are phosphorylated on scores of sites, making it likely that there are minimally 100 000 distinct phosphorylation sites in the mammalian proteome. This diversity of protein phosphorylation sites, combined with the biomedical importance of protein phosphorylation [4], the advent of high-throughout technologies for phosphorylation site discovery (Rush *et al.*, unpublished; [11]), and the increasingly large number of research reports on protein phosphorylation, argue for the need for a comprehensive database dedicated to protein phosphorylation.

PhosphoSite is a curated database dedicated to aggregating information on physiological protein phosphorylation sites. Its goal is to identify and organize information about all *in vivo* phosphorylation sites in human and mouse proteomes and to provide information and resources that will facilitate phosphorylation research. Version 1.0 contains the peptide sequences of phosphorylation sites, their location within known domains and motifs, links to useful resources, selected literature references, and sources of reagents for studying these sites. New information will be added dynamically into PhosphoSite, and incremental additions to the application will be introduced as updates or new releases.

Two of the large-scale public protein sequence databases, SWISS-PROT [12] and PIR (Protein Information Resource) [13], include curated information on post-trans-

Correspondence: Dr. Peter V. Hornbeck, Cell Signaling Technology, 166B Cummings Center, Beverly, MA 01915, USA
E-mail: phornbeck@cellsignal.com
Fax: +1-978-867-2400

Abbreviations: FAK, focal adhesion kinase; **NCBI**, National Center for Biotechnology Information; **OMIM**, Online Mendelian Inheritance in Man; **PDB**, Protein Data Bank

lational modifications including protein phosphorylation. Their mandate to curate all features of all proteomes is so broad, however, that it is difficult to keep their coverage of a specialized area like mammalian phosphorylation up-to-date. For example, SWISS-PROT includes minimal information about hundreds of known sites and thousands of predicted phosphorylation sites, but their coverage of sites is not comprehensive.

A number of other large curated databases focus on whole proteomes. Incyte's Proteome BioKnowledge Library provides encyclopedic information for all proteins from many organisms, while HPRD (Human Protein Reference Database) [14] seeks to become a comprehensive resource of information about human proteins. We have intentionally decided against duplicating efforts that are already well advanced in other database projects. For instance, since there are already a number of excellent protein interaction databases [18, 19], we only capture protein-protein interactions when they are directly influenced by the phosphorylation status of the protein. This strategy allows us to focus our resources on phosphorylation events.

A number of databases are dedicated to protein phosphorylation, but all have mandates that differ significantly from that of PhosphoSite. PhosNet [15] and Scansite [16] are used to predict possible sites of phosphorylation in protein sequences but contain little, if any, curated information. While predicted phosphorylation sites can be useful to researchers, both programs have limitations in their ability to accurately predict *in vivo* phosphorylation sites. Another resource, PhosphoBase [17] enumerates 1400 *in vivo* and *in vitro* eukaryotic phosphorylation sites, while PhosBase is a database of phosphorylation sites in bacteria and archaea.

Information is curated from literature reports as well as from high-throughput phosphorylation site discovery programs. Information extracted from the public domain will be freely available to educational users and will require subscription by commercial entities. Some proprietary information may also be made available on a subscription basis. The revenue generated by subscriptions will help to pay for the long-term survival and growth of PhosphoSite.

2 Methods and strategic considerations

2.1 Mouse and man

PhosphoSite is a human- and mouse-centric database. We have chosen these two species because of their preeminence in biomedical research, because their basic genomes have been almost fully deciphered, and to limit the scope of PhosphoSite to a manageable size. Although sites from other mammalian species will be included in the underlying knowledge base, they are included primarily to help users predict potential phosphorylation sites in man and mouse.

2.2 *In vivo vs. in vitro*

By *in vivo* we mean "in living cells". This includes cell lines, primary cells, and tissue. If evidence proves that the site is phosphorylated *in vivo*, then the site is curated into the database. There is an exception to this rule, however. Information about *in vitro* phosphorylation sites is included in PhosphoSite if the site, or its homolog in orthologs or isoforms, is known to be phosphorylated *in vivo*. In this case, knowledge about the kinase(s) that can phosphorylate the site *in vitro* may help the investigator infer what kinases might control the phosphorylation *in vivo*.

2.3 Splice-variant isoform information

We anticipate that in the coming years, understanding the biology of differential expression of splice-variant isoforms derived from the same gene will be an increasing focus of biomedical research. For this reason, we curate information about phosphorylated splice-variant isoforms. When information about the phosphorylation of an isoform is present in PhosphoSite, the user must view an "isoform" page before being directed to a phosphoprotein page. The "isoform" page functions as a "lookup table" for researchers interested in a particular isoform, and also provides links to information about isoformous phosphorylation sites. When possible, we use the same convention used by SWISS-PROT in describing isoforms. The longest isoform is usually considered the parent protein in PhosphoSite, while shorter isoforms are indicated by the suffixes "iso2 . . . isoN".

2.4 Quality control

The quality of information entered into PhosphoSite is controlled by a number of mechanisms including: automatic extraction and parsing of information associated with SWISS-PROT or NCBI (National Center for Biotechnology Information) RefSeq accession numbers; internal PhosphoSite phosphorylation site residue-protein sequence mismatch control; consistent naming conventions including controlled vocabulary descriptors for critical fields and the use of Find buttons; editorial and final user feedback processes; and regular quality control editing of master lists.

2.5 Collaborations

The developers of Scansite [16] and PhosphoSite have arrangements to reciprocally share resources. The Scansite program predicts sites within proteins that are likely to be phosphorylated by specific protein kinases using quantitative binding affinity data derived from the oriented peptide library technique [20]. The sharing of information between PhosphoSite and Scansite represents an ideal synergy: PhosphoSite provides information about observed *in vivo* sites, while Scansite provides additional predictive information about these sites. We are also planning to collaborate with SWISS-PROT in populating mammalian phosphorylation information into SWISS-PROT.

2.6 Updates

Initially, many phosphorylation sites will have only one journal article associated with them. This may not be the original article that proved that the site is phosphorylated *in vivo*. As PhosphoSite matures, however, we intend to include the reports that originally demonstrated each phosphorylation event *in vivo* as well as those that illuminate the biological function of each site. This process will be considerably enhanced if knowledgeable users submit the relevant records to us. We will make every effort to quickly add these references into PhosphoSite.

2.7 PhosphoSite software development

2.7.1 System architecture

PhosphoSite was developed using J2EE (Java 2 Platform, Enterprise Edition) following the client/server paradigm. PhosphoSite is a three-tiered application in which interactions are enabled between the client, the server, and the underlying database. The client on the web tier communicates with the PhosphoSite Application on the middle tier, which is powered by BEA's WebLogic 8.1 Application Server using http. The PhosphoSite Application communicates with the information tier, which is powered by the Oracle 9i relational database management system using Java Database Connectivity. The J2EE architecture provides platform-independence and enables any user with a web browser and an Internet connection to access PhosphoSite. Distinct user interfaces are served to different types of users, including final users, internal users, curators, editors, and administrators. Access to different interfaces and stored data is controlled by individual user's privileges. Java Server Pages and Servlets are used to generate dynamic content on the HTML interfaces. To encapsulate business logic away from the presentation layer, and increase transactional efficiency, Enterprise JavaBean objects have been used for managing data flow to and from the database. This architectural design enables PhosphoSite to deliver a wide range of functionalities while providing brisk response times and allowing for flexibility in future development.

2.7.2 Database

The processing of information during curation as well as the database table structure have been designed in a modular fashion, which allows related types of information to be processed together, establishing a flexible and extensible database structure. Information is stored in a collection of tables that are integrated into a relational database using Oracle 9i. The tabular schema has been clustered based on the types of biological information on each phosphorylation site that are being stored, queried and displayed. The hierarchical organization of these tables has been logically developed with the condition that one protein may have multiple phosphorylation sites, and each phosphorylation site may have multiple pieces of relevant information. The schema includes a large number of tables containing different master lists, fostering the standardization of curation entries and ameliorating the task of maintaining large ontologies. The columnar structure within each table has been designed with emphasis on biologically relevant information down to its lowest common denominator, wherever possible. Such an organizational scheme enables powerful queries, facilitates interconnectivity with other bioinformatics resources and also makes the data more amenable to future algorithmic analyses and novel user functionalities.

2.7.3 Supporting applications

The annotation and maintenance of PhosphoSite has been facilitated by automating various tasks. Much of the basic data about proteins including sequence, alternative names, and associated links, is directly imported from other databases. Agents have been developed to retrieve, parse and store this information in an automated fashion. For example, the SWISS-PROT parser requires only that the curator enter the proper accession number. The parser then automatically retrieves information including the protein sequence, protein name and synonyms, gene name, function statement, and links to EMBL and PDB (Protein Data Bank). Additionally, agents have been built to retrieve and integrate information from

PubMed, GenPept [21], and Pfam [22]. Agents that can handle bulk information from large reference submissions or data from high-throughput experiments have also been developed to support PhosphoSite's population.

2.8 Curation

The specialists that curate PhosphoSite are all research scientists highly trained in cell biology and related disciplines. All curators have earned either doctorates or masters degrees in cell biology or related disciplines with extensive experience in the art and science of scientific curation. All editors have doctoral degrees (PhD or MD/PhD) with extensive post-doctoral training and a significant publication record in the area of signal transduction and protein phosphorylation. Curation involves three steps prior to final submission: identifying appropriate records, curating information from records, and editing curated information prior to final submission. This process is rigorously controlled at all steps to assure the accuracy of the information that is submitted into PhosphoSite. Each record must be separately curated and edited, providing multiple opportunities to verify the accuracy of the aggregated information. Most of the data to be entered by the curator is in the form of expandable lists of defined vocabularies. This process facilitates building our internal ontologies, and will permit powerful queries and data mining.

2.8.1 Identification and submission of information to PhosphoSite

PubMed abstracts are searched semi-automatically with multiple intelligent search algorithms to identify reports that potentially characterize specific phosphorylation sites in human and mouse or related species. More than 60 000 abstracts were identified and downloaded for further analysis. The retrieved abstracts were further analyzed, parsed, and hierarchically organized using Perl scripts. Abstracts that describe mammalian phosphorylation on particular amino acids (serine, threonine, and tyrosine) are given highest priority, while those that describe phosphorylation of a protein but do not mention a specific phosphorylation site are given lower priority. The selected abstracts are then read by our curators to eliminate false positives and to select those that appear to characterize physiological phosphorylation sites.

PhosphoSite users are strongly urged to participate in this process by submitting missing sites to the editors. The "How to Contribute" tab in the top right of PhosphoSite pages provides easy-to-use forms for submitting missing or incorrect information to the editors. All submitted information will be screened and, if appropriate, quickly curated. Users who submit more than five records that each contains one or more sites that are new to PhosphoSite will be publicly acknowledged as contributors if they wish. Other sources of phosphorylation site information include the unpublished output from high-throughput phosphorylation site discovery programs. Initially this information will be shown to end users only when the site has been previously described in the public domain. Access to this proprietary information will, however, be made available on a subscription basis.

2.8.2 Curating information

The first step in the process of curating information from a record about new phosphorylation sites is to verify the sequence surrounding the phosphorylated residue. The original article or a related article that explicitly provides its amino acid sequence is used. The species of the phosphoprotein is then confirmed, and its corresponding SWISS-PROT or RefSeq [21] accession number is entered into the database. The program then automatically retrieves, parses, and curates information from SWISS-PROT or NCBI, enhancing the speed and accuracy of the curation process. The retrieved information includes sequence, species, and related links (OMIM (Online Mendelian Inheritance in Man), GeneCards, PDB, *etc.*). The curator then identifies the relevant phosphorylation sites in the main protein and the corresponding sites in its orthologs and, when necessary, in its isoforms.

After retrieving the basic protein information, the full article is read to extract classes of information including: (i) Characterization of the phosphorylated site. This includes the methods used to identify *in vivo* phosphorylation sites (mass spectroscopy, mutagenesis, *etc.*), the tissues and cell types in which the experiments were conducted, and the kinases and phosphatases that control the phosphorylation state *in vitro*. (ii) Upstream regulation. This includes the *in vivo* kinases, phosphatases, ligands, receptors, and pathways implicated in regulating the phosphorylation of the site; GO (gene ontology) descriptors of biological process; and the types of experiments used to support the experimental conclusions. (iii) Downstream consequences. This includes enzymatic activation or inhibition, protein degradation, molecular interactions, altered cellular localization, *etc.* Information about the influence of phosphorylation upon downstream molecular interactions and functional consequences of these inter-

actions will be captured. Curated information includes names of interacting proteins, the consequence(s) of the interaction, and tissues or cell lines in which experiments were conducted.

Since most types of experimental evidence about putative *in vivo* kinases/phosphatases and downstream effects are inferential, our inclusion of experimental methodologies should help the end user evaluate the strength of the conclusions. For instance, if an author infers that a particular phosphorylation reaction is regulated by protein kinase C and bases this inference solely on the treatment of cells with pharmacological activators such as phorbol esters, then the inference may be considered weak at best. The initial public version of PhosphoSite will contain little of this inferential information; rather it focuses on identifying as many known *in vivo* sites as possible.

2.8.3 Editing

All curated records are reviewed for accuracy and consistency by an editor before final submission to PhosphoSite. The editor verifies or corrects the curated information from each record and, after final submission, confirms that the phosphorylation sites and their orthologs and isoforms are displayed properly to the user.

3 Results

The initial release of PhosphoSite provides users a web-based interface to a large amount of information on phosphorylation sites. This interface is designed to be simple and intuitive to use while providing the flexibility to perform complex searches through different categories. At the time of writing this manuscript, PhosphoSite displays 2389 phosphorylation sites on 881 phosphoproteins; this information has been curated from 1807 research publications. The information content of PhosphoSite continues to increase rapidly. In order to achieve a comprehensive representation of known phosphorylation sites within a reasonable time frame, we have initially followed an expedited curation strategy. We have initially focused on extracting the fundamental information about phosphorylation sites from the majority of curated publications; we will subsequently curate more detailed information (including upstream signaling and downstream consequences of phosphorylation events) from each reference. Therefore, as the representation of phosphorylation sites in PhosphoSite nears completion in the near future, the depth of information about these phosphorylation events, and their roles in signal transduction pathways and biological responses, will continue to expand.

3.1 Home Page

To locate proteins, their phosphorylation sites, and associated information, a user can do a simple search for the name of the protein (Fig. 1, left). Complex searches (Fig. 1, right) can combine terms such as protein name, author, PubMed ID, SWISS-PROT ID, phosphorylated residue number, domain, motif, or sequence. A search for a domain returns only those domains in PhosphoSite that contain a phosphorylation site. While most search terms can be used in multiple combinations, a sequence search cannot be done with any other category. Future versions of Phosphosite will allow additional searches, such as finding all *in vivo* substrates of a particular kinase.

3.2 Search Results page

A simple search for the protein Fak opens a Search Results page showing human and mouse proteins that match the query (Fig. 2). The user can select a protein from the Search Results list to go to the relevant Phosphoprotein Page. A protein name search is conducted through both main protein names and synonyms. This is done to anticipate varying naming conventions used by different users.

3.3 Phosphoprotein Page

Selecting a protein name, *for example*, FAK (focal adhesion kinase, human), from the Search Results page (Fig. 2) takes the user to the FAK Phosphoprotein Page (Fig. 3). This page is the focal point for a wealth of information about proteins and their phosphorylation sites. Basic information about the protein is displayed in an Overview section which contains a brief description of the protein, the Reference Accession number for the sequence used in PhosphoSite, Alternative Names, predicted Molecular Weight, and a Protein Type description. If the parent protein is a kinase, the taxonomic classification of the kinase is listed. The group name is hyperlinked to its family tree to allow the user to easily determine the taxonomic relationships of the specified kinase (see Fig. 4). An interactive Isoelectric calculator is provided that allows the user to view the calculated Isoelectric points for the parent protein with multiple phosphorylated serines, threonines, and tyrosines. The basic Isoelectric calculator was kindly provided by Dr. Elizabeth Gasteiger of the Swiss Institute of Bioinformatics. The pK_a values used for phosphate were: pK_{a_1} = 1.0 and pK_{a_2} = 6.1.

Below this are useful links to a number of external sources for information about the protein, including: the sequence reference databases LocusLink [23] and SWISS-PROT;

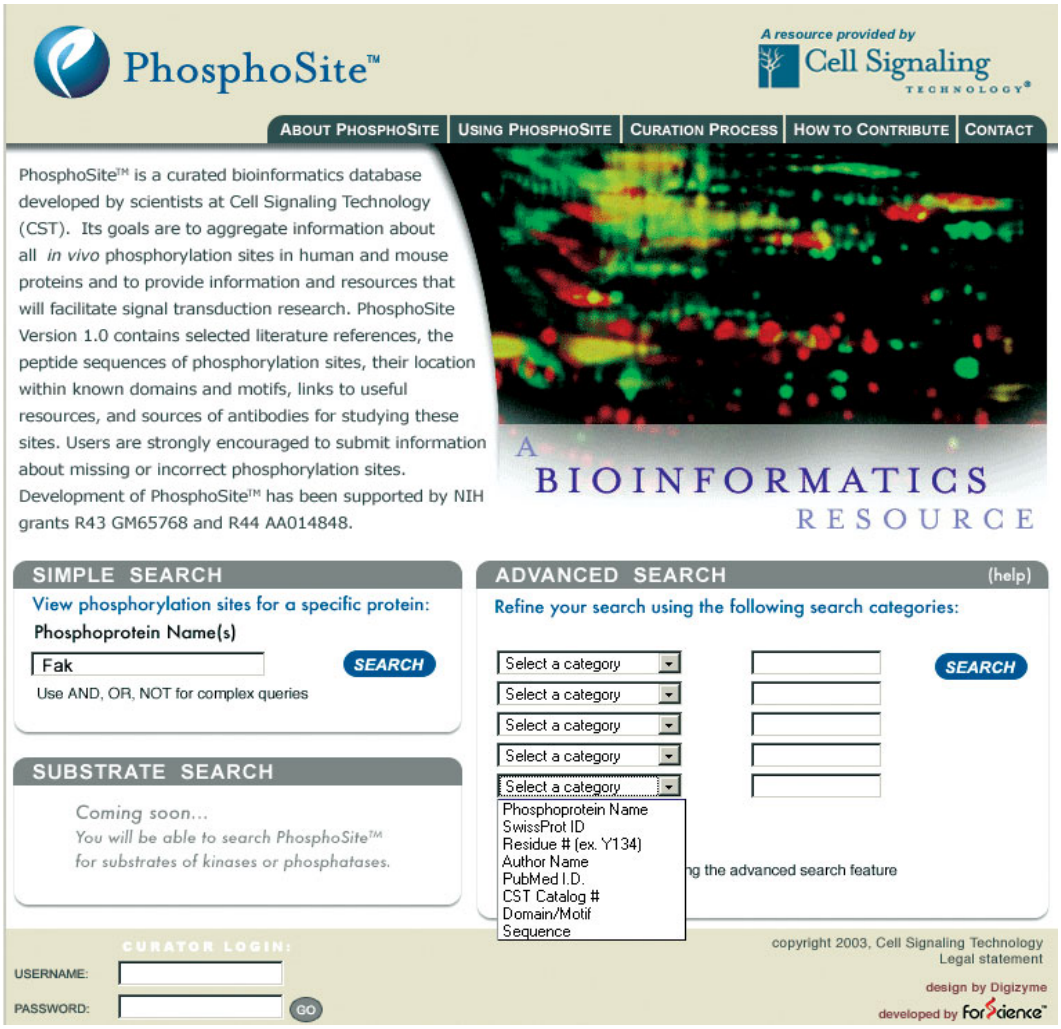

Figure 1. PhosphoSite Home Page. In this example, the interface is poised to submit a simple search for "Fak". Advanced searches can include Phosphoprotein Name, SWISS-PROT ID, Residue Number, Author name, PubMed ID, Cell Signaling Technology catalog number, Domain/Motif, and a peptide sequence of between 6 and 30 residues.

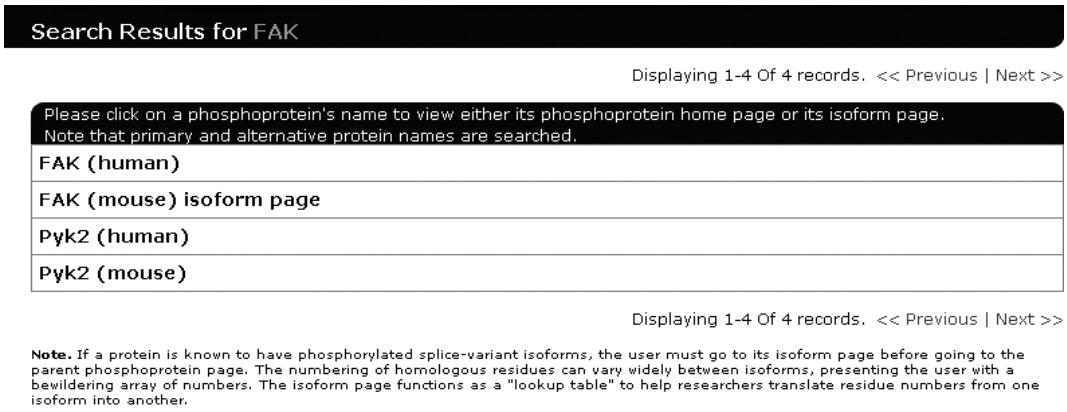

Figure 2. Search Results Page. The search string "Fak" was matched against all primary and alternative names in the database. See text for discussion.

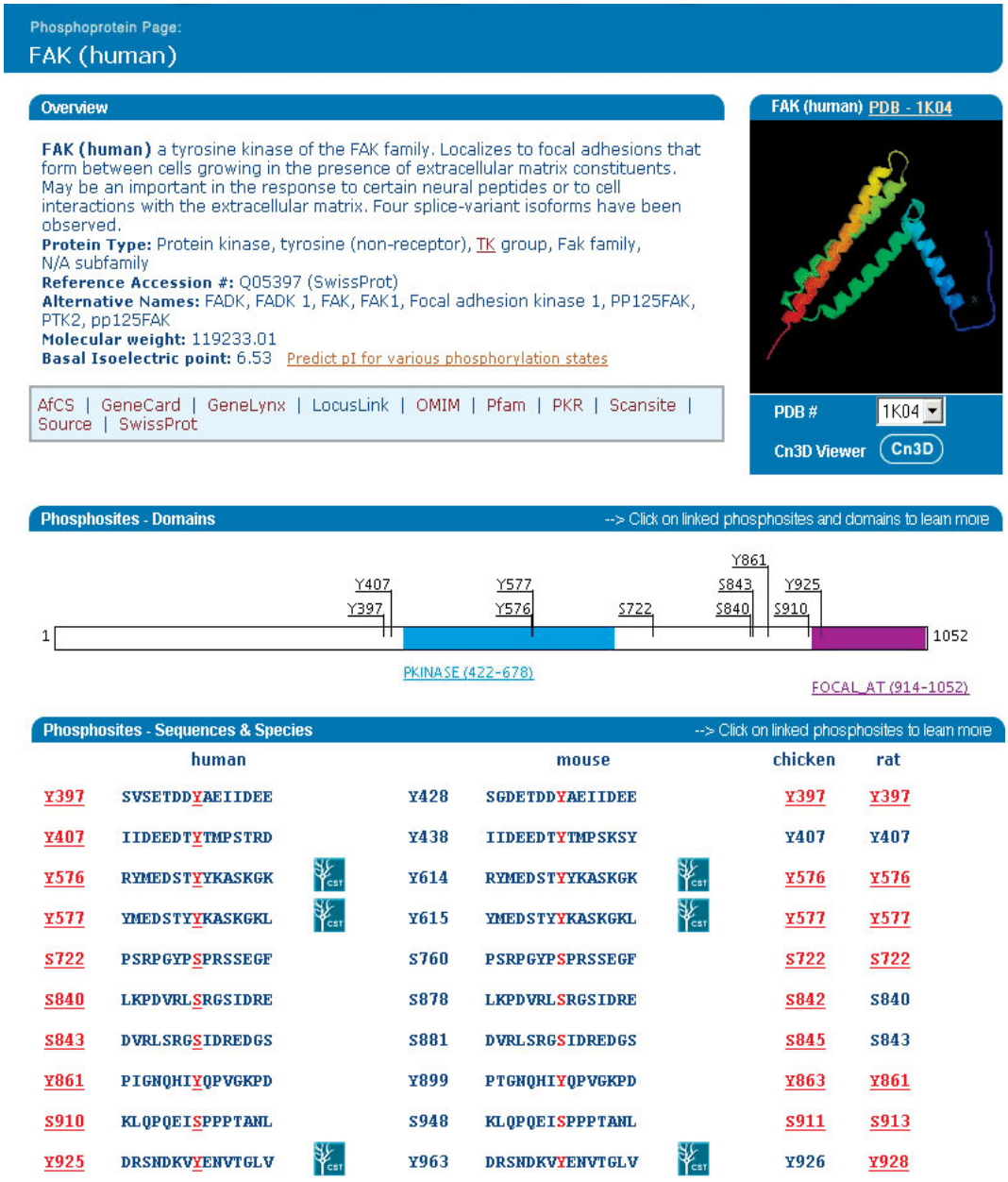

Figure 3. Phosphoprotein Page. The phosphoprotein Page for FAK (human) is shown here. The red hyperlinked residues in the lower section are linked to their corresponding Phosphorylation Site Pages. Clicking on the kinase group link "TK" in the "Protein Type" field will open up the corresponding kinase family tree (see Fig. 4).

the signal transduction databases the Protein Kinase Resource (PKR) (http://pkr.sdsc.edu/) and the Alliance for Cell Signaling (AfCS) [24]; disease information from OMIM (http://www.ncbi.nlm.nih.gov/omim/); human expression data from SOURCE [25]; secondary structure and domain information from Pfam [22]; and predictive information about phosphorylation sites from Scansite. A viewer for protein structures is integrated into this page, with links to protein structure databases PDB and MMDB. The structural displays will be significantly enhanced in subsequent versions of the database.

In the middle section of the Phosphoprotein Page, Phosphosites-Domains, a graphical representation of the protein shows domain and motif information imported from Pfam, and the locations of phosphorylation sites on

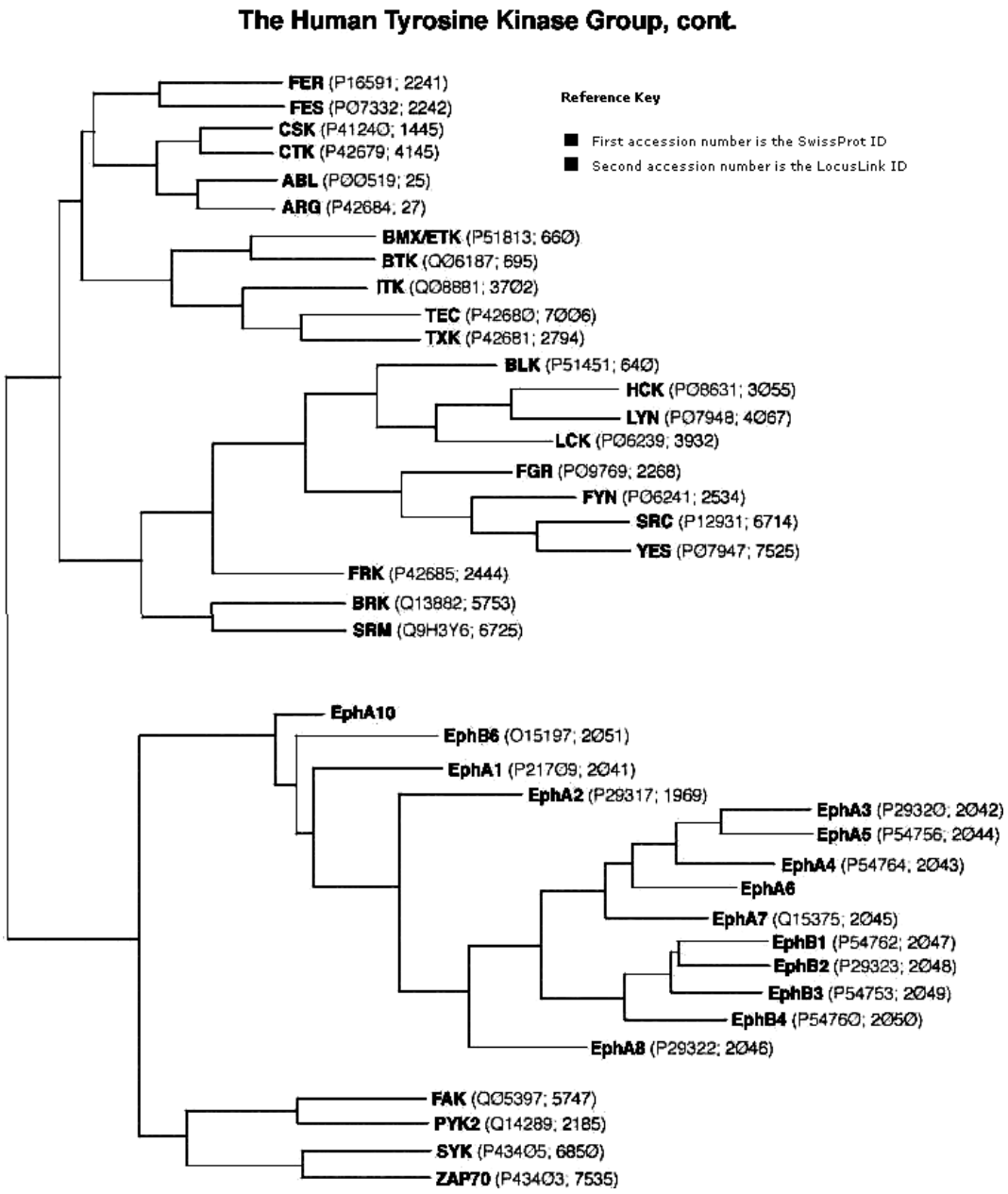

Figure 4. Human Kinase Group Tree. The family tree of the group of human tyrosine kinases that contains FAK. Clicking on the kinase group link "TK" in the "Protein Type" field (see Fig. 3) will open up this page. See text for further discussion.

the proteins. Domains shown are directly linked to Pfam for complete information, and phosphorylation sites are linked to their respective Phosphorylation Site pages (Fig. 5). The lower section of the Phosphoprotein Page displays the sequences flanking phosphorylation sites in the protein and their orthologous sites in human or mouse in a tabular format, with links to the relevant Phosphorylation Site page. Those residues that are red and hyperlinked are known to be phosphorylated *in vivo*. In addition to human and mouse, links are provided to publications describing phosphorylation of orthologous sites in other mammalian species. Clicking on a link takes the user to the Phosphorylation Site page (Figure 5). A small icon next to a site indicates that phosphosite-specific antibodies or other probes are available for this site; clicking on the icon will link to the source of the specific probe.

Phosphorylation Site Page:
THR231 - tau iso9 (human)

Phosphorylation Site Information

KKVAVVRTPPKSPSS SwissProt

View predicted information about this site: Scansite

Blast this site: against NCBI SwissProt PDB

Phosphorylation controlled by

Putative upstream kinases:	**Cdk5** (3)
Kinase, *in vitro*:	**GSK3-beta** (5), **Cdk5** (5)
Ligands:	**MSH** (1), **nicotine** (1), **beta-amyloid 42** (1)

References --> Click on PubMed ID to view Curated Information or Abstract

1	Wang HY, Li W, Benedetti NJ, Lee DH **Alpha 7 nicotinic acetylcholine receptors mediate beta-amyloid peptide-induced tau protein phosphorylation.** J Biol Chem 2003 Aug 22; 278(34): 31547-53 12801934
2	Mitchell A, Brindle N **CSF phosphorylated tau--does it constitute an accurate biological test for Alzheimer's disease?** Int J Geriatr Psychiatry 2003 May; 18(5): 407-11 12766916
3	Noble W, et al. **Cdk5 is a key factor in tau aggregation and tangle formation in vivo.** Neuron 2003 May 22; 38(4): 555-65 12765608
4	Giasson BI, et al. **The environmental toxin arsenite induces tau hyperphosphorylation.** Biochemistry 2002 Dec 24; 41(51): 15376-87 12484777
5	Liu F, Iqbal K, Grundke-Iqbal I, Gong CX **Involvement of aberrant glycosylation in phosphorylation of tau by cdk5 and GSK-3beta.** FEBS Lett 2002 Oct; 530(1-3): 209 12387894
6	Taniguchi T, et al. **Phosphorylation of tau is regulated by PKN.** J Biol Chem 2001 Mar 30; 276(13): 10025-31 11104762
7	Sengupta A, et al. **Phosphorylation of tau at both Thr 231 and Ser 262 is required for maximal inhibition of its binding to microtubules.** Arch Biochem Biophys 1998 Sep; 357(2): 299-309 9735171
8	Illenberger S, et al. **The endogenous and cell cycle-dependent phosphorylation of tau protein in living cells: implications for Alzheimer's disease.** Mol Biol Cell 1998 Jun; 9(6): 1495-512 9614189
9	Singh TJ, et al. **Protein kinase C and calcium/calmodulin-dependent protein kinase II phosphorylate three-repeat and four-repeat tau isoforms at different rates.** Mol Cell Biochem 1997 Mar; 168(1-2): 141-8 9062903
10	Mawal-Dewan M, et al. **Identification of phosphorylation sites in PHF-TAU from patients with Guam amyotrophic lateral sclerosis/parkinsonism-dementia complex.** J Neuropathol Exp Neurol 1996 Oct; 55(10): 1051-9 8858002

Figure 5. Phosphorylation Site Page. Contains information specific to a particular phosphorylation site. This example shows information about Threonine 231 in human tau isoform 9. See text for discussion.

3.4 Phosphorylation Site Page

Clicking on a red hyperlinked residue in the lower section of the Phosphoprotein Page (Fig. 3) or the Isoform Page (see Fig. 6) will open the Phosphorylation Site Page that provides information specifically about that phosphorylation site (Fig. 5). In the Phosphorylation Site Information section, a Scansite link will submit the phosphorylation site and its surrounding sequence (+/−7 residues) to Scansite to predict kinases that are likely to phosphorylate the site, and likely interactions with other signaling proteins. The initial query is done at high stringency, but the query may be resubmitted at lower stringencies within the Scansite window. Links are also provided to the Phosphorylation Site Page for orthologous residues, and to a BLAST search of the sequence surrounding the phosphorylation site against NCBI, SWISS-PROT, or PDB. A second section of this page, which is not yet widely implemented, summarizes key curated information about the phosphorylation site, including upstream kinases, ligands and treatments that control its phosphorylation, and downstream effects of the phosphorylation event. Refer-

Isoform Page: this protein has phosphorylated isoforms shown below

tau (human) <-- Click on protein name to view its protein page

Protein Type: Cytoskeletal protein
Reference Accession #: P10636 SwissProt
Alternative Names: MAPT, MTAPT, MTBT1, Microtubule-associated protein tau, Neurofibrillary tangle protein, PHF-tau, PNS-Tau, Paired helical filament-tau
Molecular weight: 78746.36
Basal Isoelectric point: 6.61 Predict pI for various phosphorylation states

tau iso2 - Expressed in neurons in fetal brain.

Reference Accession #: P10636-2 SwissProt
Alternative Names: fetal-tau
Molecular weight: 32813.04
Basal Isoelectric point: 10.48 Predict pI for various phosphorylation states

tau iso3 - Expressed in neurons in fetal and adult brain; three tau/MAP repeats, no N-terminal inserts.

Reference Accession #: P10636-3 SwissProt
Alternative Names: Tau-A, tau 23, tau 3, tau 352, tau-3
Molecular weight: 36628.88
Basal Isoelectric point: 9.76 Predict pI for various phosphorylation states

tau iso5 - Expressed in neurons in fetal and adult brain; three tau/MAP repeats, two N-terminal inserts.

Reference Accession #: P10636-5 SwissProt
Alternative Names: Tau-C, tau 39, tau 410, tau-3, tau-3L
Molecular weight: 42471.98
Basal Isoelectric point: 7.26 Predict pI for various phosphorylation states

tau iso6 - Expressed in neurons in adult brain specifically; four tau/MAP repeats, no N-terminal inserts.

Reference Accession #: P10636-6 SwissProt
Alternative Names: Tau-D, tau 24, tau 383, tau-4
Molecular weight: 39875.61
Basal Isoelectric point: 9.84 Predict pI for various phosphorylation states

tau iso9 - Expressed in neurons in adult brain specifically; four tau/MAP repeats, two N-terminal inserts; longest CNS isoform.

Reference Accession #: NP_005901 GenPept
Alternative Names: Tau-F, tau 40, tau 441, tau-4, tau-4L
Molecular weight: 45849.9
Basal Isoelectric point: 8.38 Predict pI for various phosphorylation states

Phosphosites - Isoforms --> Click on linked phosphosites to learn more

tau		tau iso2	tau iso3	tau iso5	tau iso6	tau iso9
S45	TDAGLKESPLQTPTE			S45		S46
T49	LKESPLQTPTEDGSE			T49		T50
T469	DGKTKIATPRGAAPP	T58	T94	T152	T94	T153
T491	ATRIPAKTPPAPKTP	T80	T116	T174	T116	T175
T497	KTPPAPKTPPSSGEP	T86	T122	T180	T122	T181
S514	SGDRSGYSSPGSPGT	S103	S139	S197	S139	S198
S515	GDRSGYSSPGSPGTP	S104	S140	S198	S140	S199
S518	SGYSSPGSPGTPGSR	S107	S143	S201	S143	S202
T521	SSPGSPGTPGSRSRT	T110	T146	T204	T146	T205
T528	TPGSRSRTPSLPTPP	T117	T153	T211	T153	T212
S530	GSRSRTPSLPTPPTR	S119	S155	S213	S155	S214
T533	SRTPSLPTPPTREPK	T122	T158	T216	T158	T217
T547	KKVAVVRTPPKSPSS	T136	T172	T230	T172	T231
S551	VVRTPPKSPSSAKSR	S140	S176	S234	S176	S235
S578	NVKSKIGSTENLKHQ	S167	S203	S261	S203	S262
S636	VDLSKVTSKCGSLGN	S194	S230	S288	S261	S320
S672	RVQSKIGSLDNITHV	S230	S266	S324	S297	S356
S712	GAEIVYKSPVVSGDT	S270	S306	S364	S337	S396
S716	VYKSPVVSGDTSPRH	S274	S310	S368	S341	S400
S720	PVVSGDTSPRHLSNV	S278	S314	S372	S345	S404
S725	DTSPRHLSNVSSTGS	S283	S319	S377	S350	S409
S729	RHLSNVSSTGSIDMV	S287	S323	S381	S354	S413
S738	GSIDMVDSPQLATLA	S296	S332	S390	S363	S422

Figure 6. Isoform page. If information about phosphorylated isoforms has been curated into PhosphoSite, then the user will be directed first to an Isoform Page before being able to go to the Phosphoprotein Page of the parent molecule. This example shows the Isoform Page for tau (human).

ences that document this phosphorylation site are listed in a bibliography at the bottom of the page, with links to their abstracts in PubMed. In future versions of PhosphoSite, each literature reference will be linked to an additional page providing detailed information curated from individual papers.

3.5 Isoform Page

In cases where phosphorylated isoforms of a protein have been described, such as tau (human), users cannot go directly to the tau (human) Phosphoprotein Page; rather they must first go to the tau Isoform Page (Fig. 6). Each

phosphorylated isoform is briefly described and the phosphorylated residues on each are displayed. The Isoform Page serves two purposes. First, in cases where isoforms are known to be differentially or specifically phosphorylated in particular biological contexts, its associated information is linked with the phosphorylation site in the relevant isoform(s). Second, in cases where the numbering of residues cited in publications refers to a particular isoform, the Isoform Page serves the pragmatic function of providing a look-up table to translate the residue numbering used in different publications, or between a particular paper and the parent sequence shown in PhosphoSite.

4 Concluding remarks

PhosphoSite is the first bioinformatics resource to provide a comprehensive collection of known *in vivo* protein phosphorylation sites. While the database in its current version should prove to be highly useful to researchers in many fields, future developments are planned to exploit the full potential of the data aggregated in the database. First, the sheer numbers of *in vivo* phosphorylation sites contained in PhosphoSite are likely to increase dramatically in the near future, as the rapidly expanding number of new sites discovered using high-throughput methods are added into the database. In addition, as more complete experimental data from literature references are curated into PhosphoSite, users will be able to systematically search for the kinases, phosphatases, ligands, and receptors that have been shown to regulate the phosphorylation status of the sites, and the pathways in which the phosphorylation sites function. Beyond this, new capabilities to be developed in PhosphoSite will allow higher levels of analysis of the cellular signaling networks involving phosphorylation events. Other developments will include protein structural viewers that will allow users to interactively examine the topology of phosphorylation sites. Perhaps most excitingly, as increasing numbers of phosphorylation events that have been correlated with human disease are identified by researchers and collected into the database, PhosphoSite may be invaluable for identifying, analyzing, and possibly predicting new relationships and commonalities between signaling pathways whose deregulation can lead to disease.

We gratefully acknowledge the NIGMS (R43 GM65768) and NIAAA (R44 AA014848) for supporting this work, John Obenauer for reading the manuscript, Mike Comb and Roby Polakiewicz for their support, and Mike Yaffe, Lew Cantley and Tony Hunter for their helpful suggestions.

5 References

[1] Shi, Y., Massague, *J. Cell.* 2003, *113*, 685–700.

[2] Swedlow, J. R., Hirano, T., *Mol. Cell.* 2003, *11*, 557–569.

[3] Leff, T., *Biochem. Soc. Trans.* 2003, *31*, 224–227.

[4] Blume-Jensen, P., Hunter, T., *Nature* 2001, *411*, 355–365.

[5] Mustelin, T., Tasken, K., *Biochem. J.* 2003, *371*, 15–27.

[6] Gauld, S. B., Dal Porto, J. M., Cambier, J. C., *Science* 2002, *296*, 1641–1642.

[7] Lisman, J., Schulman, H., Cline, H., *Nat. Rev. Neurosci.* 2002, *3*, 175–190.

[8] Selcher, J. C., Weeber, E. J., Varga, A. W., Sweatt, J. D. *et al.*, *Neuroscientist* 2002, *8*, 122–131.

[9] Manning, G., Whyte, D. B., Martinez, R., Hunter, T. *et al.*, *Science* 2002, *298*, 1912–1934.

10] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W. *et al.*, *Science* 2001, *291*, 1304–1351.

11] Aebersold, R., Mann, M., *Nature* 2003, *422*, 198–207.

12] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C. *et al.*, *Nucleic Acids Res.* 2003, *31*, 365–370.

13] Wu, C. H., Yeh, L. S., Huang, H., Arminski, L. *et al.*, *Nucleic Acids Res.* 2003, *31*, 345–347.

14] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z. *et al.*, *Genome Res.* 2003, *13*, 2363–2371.

15] Blom, N., Gammeltoft, S., Brunak, S. J., *Mol. Biol.* 1999, *294*, 1351–1362.

16] Obenauer, J. C., Cantley, L. C., Yaffe, M. B., *Nucleic Acids Res.* 2003, *31*, 3635–3641.

17] Kreegipuu, A., Blom, N., Brunak, S., *Nucleic Acids Res.* 1999, *27*, 237–239.

18] Bader, G. D., Betel, D., Hogue, C. W., *Nucleic Acids Res.* 2003, *31*, 248–250.

19] Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J. *et al.*, *Nucleic Acids Res.* 2001, *29*, 239–241.

20] Songyang, Z., Cantley, L. C., *Methods Mol. Biol.* 1998, *87*, 87–98.

21] Pruitt, K. D., Tatusova, T., Maglott, D. R., *Nucleic Acids Res.* 2003, *31*, 34–37.

22] Bateman, A., Birney, E., Cerruti, L., Durbin, R. *et al.*, *Nucleic Acids Res.* 2002, *30*, 276–280.

23] Pruitt, K. D., Maglott, D. R., *Nucleic Acids Res.* 2001, *29*, 137–140.

24] Li, J., Ning, Y., Hedley, W., Saunders, B. *et al.*, *Nature* 2002, *420*, 716–717.

25] Diehn, M., Sherlock, G., Binkley, G., Jin, H. *et al.*, *Nucleic Acids Res.* 2003, *31*, 219–223.

