D240–D244 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 25 October 2007*
doi:10.1093/nar/gkm772

Phospho.ELM: a database of phosphorylation sites—update 2008

Francesca Diella[1], Cathryn M. Gould[1], Claudia Chica[1], Allegra Via[2] and Toby J. Gibson[1,*]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany and [2]Center for Molecular Bioinformatics, Dept. of Biology, Tor Vergata University, Rome, Italy

Received July 18, 2007; Revised September 4, 2007; Accepted September 13, 2007

ABSTRACT

Phospho.ELM is a manually curated database of eukaryotic phosphorylation sites. The resource includes data collected from published literature as well as high-throughput data sets.

The current release of Phospho.ELM (version 7.0, July 2007) contains 4078 phospho-protein sequences covering 12 025 phospho-serine, 2362 phospho-threonine and 2083 phospho-tyrosine sites. The entries provide information about the phosphorylated proteins and the exact position of known phosphorylated instances, the kinases responsible for the modification (where known) and links to bibliographic references. The database entries have hyperlinks to easily access further information from UniProt, PubMed, SMART, ELM, MSD as well as links to the protein interaction databases MINT and STRING.

A new BLAST search tool, complementary to retrieval by keyword and UniProt accession number, allows users to submit a protein query (by sequence or UniProt accession) to search against the curated data set of phosphorylated peptides.

Phospho.ELM is available on line at: http://phospho.elm.eu.org

INTRODUCTION

Protein phosphorylation is one of the most-studied post-translational modifications: it has been estimated that up to one-third of the proteins may be modified by protein kinases (1). This ubiquitous regulatory mechanism controls many biological processes, including cellular growth, differentiation and DNA repair (2).

Knowing the phosphorylated residues in proteins is central to understanding the various signaling events in which they partake; therefore much effort has been invested in trying to identify and characterize phosphorylation sites. Traditional methods for measuring protein phosphorylation, such as mutational analysis and Edman degradation chemistry on phosphopeptides, have the disadvantage of being relatively time consuming and laborious, requiring large amounts of purified protein. On the other hand, mass spectrometry-(MS)based methods have emerged as powerful tools for the analysis of post-translational modifications due to higher sensitivity, selectively and speed. Over the past few years MS, combined with enrichment strategies for phosphorylated proteins e.g. isotope-coded affinity tags (ICAT) (3), stable isotopic amino acids in cell culture (SILAC) (4) and isobaric reagent iTRAQ (5), has been increasingly employed to identify novel phosphorylation sites. One consequence of this change in phosphorylation research is that bioinformatics resources need to be adapted and expanded to accommodate the new data.

For the thousands of phosphorylation sites identified by phosphoproteomic MS the information on which kinase phosphorylates them, and consequently the pathway in which they act, is still missing. To improve the link between experimentally identified phosphorylation sites and protein kinases, Linding and collaborators (6) have recently used the Phospho.ELM data set to develop and train a method, NetworKIN (http://networkin.info/) that combines computational methods for predicting which group of kinases are likely to phosphorylate a given site with information about signaling pathways and protein interaction data.

The analysis of protein phosphorylation by MS will clearly prove to be an invaluable source of information for understanding cellular signaling. For this reason, we consider it increasingly important to create and maintain publicly available phospho-protein databases, where the exponentially increasing number of known phosphorylation sites (7–12) can be easily accessed by the research community.

MATERIALS AND METHODS

The Phospho.ELM database

The content and the format of Phospho.ELM have been previously described in Diella *et al.* (13). While the general

*To whom correspondence should be addressed. Tel: +49 6221 3878398; Fax: +49 6221 3878517; Email: toby.gibson@embl.de

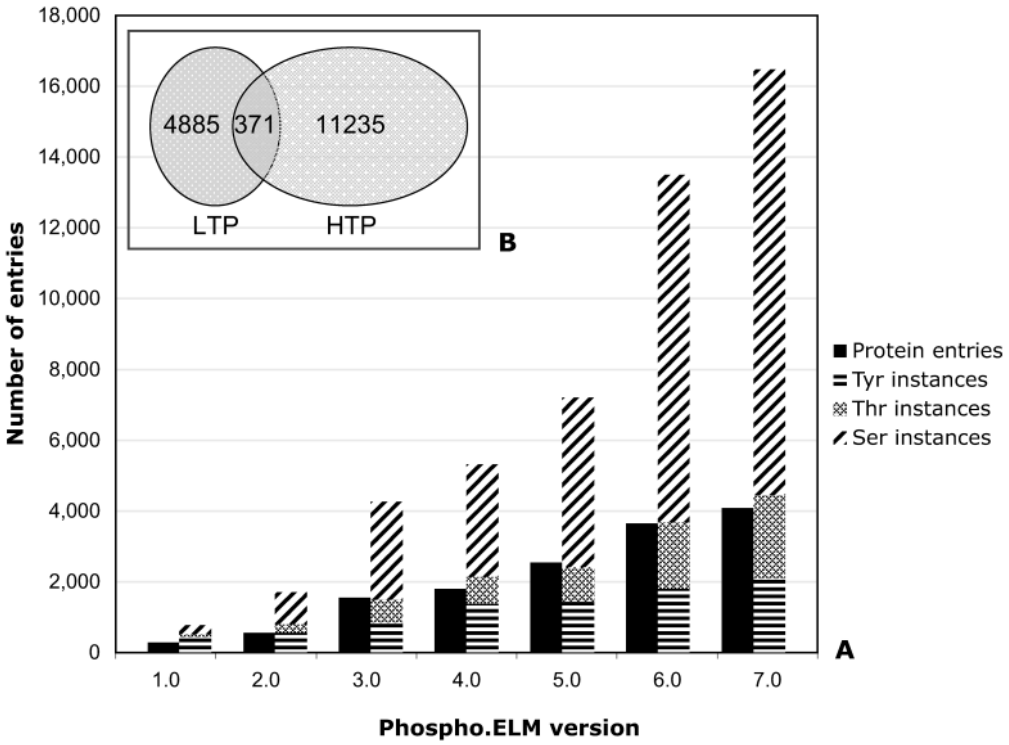

Figure 1. The plot shows the growth of the Phospho.ELM data set beginning with version 1.0 in December 2003 (panel A). The exponential growth of the phosphorylation instances from Version 5.0 is mainly due to incorporation of the high-throughput data sets. The overlapping of the instances derived from low-throughput (LTP) and high-throughput (HTP) experiments is also shown (panel B).

format of the database has remained essentially unchanged, some additions have been implemented to improve the data retrieval and presentation. The updated version also contains a much larger number of phosphorylation sites (see Figure 1), a new search tool based on sequence comparison and a Web Services interface.

The user can query the database by protein name, UniProt accession number/identifier, kinase name or binding motif to get a list of all known phosphorylation sites (instances) in a specific protein. The main results page summarizes information about the substrate protein (e.g. a brief description of the protein, protein type, the UniProt protein identification number), the phosphorylation sites contained within it and its surrounding amino acids (+/−10). The annotations to each instance include (where available) the PubMed reference, the kinase(s) phosphorylating the given site, the phospho-peptide binding domain(s) and a link to the ELM server (14) to retrieve further information about the kinase. Also where available, hyperlinks are provided to protein structures containing phosphorylated residues (15). Recently, Zanzoni and collaborators (16) have developed Phospho3D, a database of three-dimensional structures of phosphorylation sites, which stores data derived from the Phospho.ELM database and is focused on the annotation of structural information at the residue level.

Additional information for each protein kinase substrate includes the subcellular compartment [annotated with the Gene Ontology terms (17)], the tissue distribution and a list of interaction partners derived from the MINT (18) and STRING databases (19). The STRING interactors are shown in a summary graphic (network) that opens in a pop-up window. The network views provide links to the STRING database, where the information relative to the interactors is described in detail.

Data set

The current release of the Phospho.ELM data set (version 7.0, July 2007) contains 4078 phospho-protein sequences covering 12 025 phospho-serine, 2362 phospho-threonine and 2083 phospho-tyrosine sites with a total of 16 470 sites. The dataset is currently limited to metazoan species. This is partly due to our annotation capacity and partly because the kinases and nomenclature are so different in other lineages that they should be placed in separate databases. Although no animal species is purposely excluded from the data, currently human (11 197 phospho-sites) and mouse (2073 phospho-sites) are the most representative species due to the prevalence of their use as model organisms in biological research e.g. phosphoproteomic MS analyses have been mainly performed on human/mouse cell lines/tissues.

For each phospho-site we report if the phosphorylation evidence has been identified by small-scale analysis (low throughput; LTP) that typically focus on one or a few proteins at a time or by large-scale experiments (high throughput; HTP), which mainly apply MS techniques. It is noteworthy that in our data set there is a small overlap between instances identified by LTP and HTP experiments (Figure 1). This implies that most of the

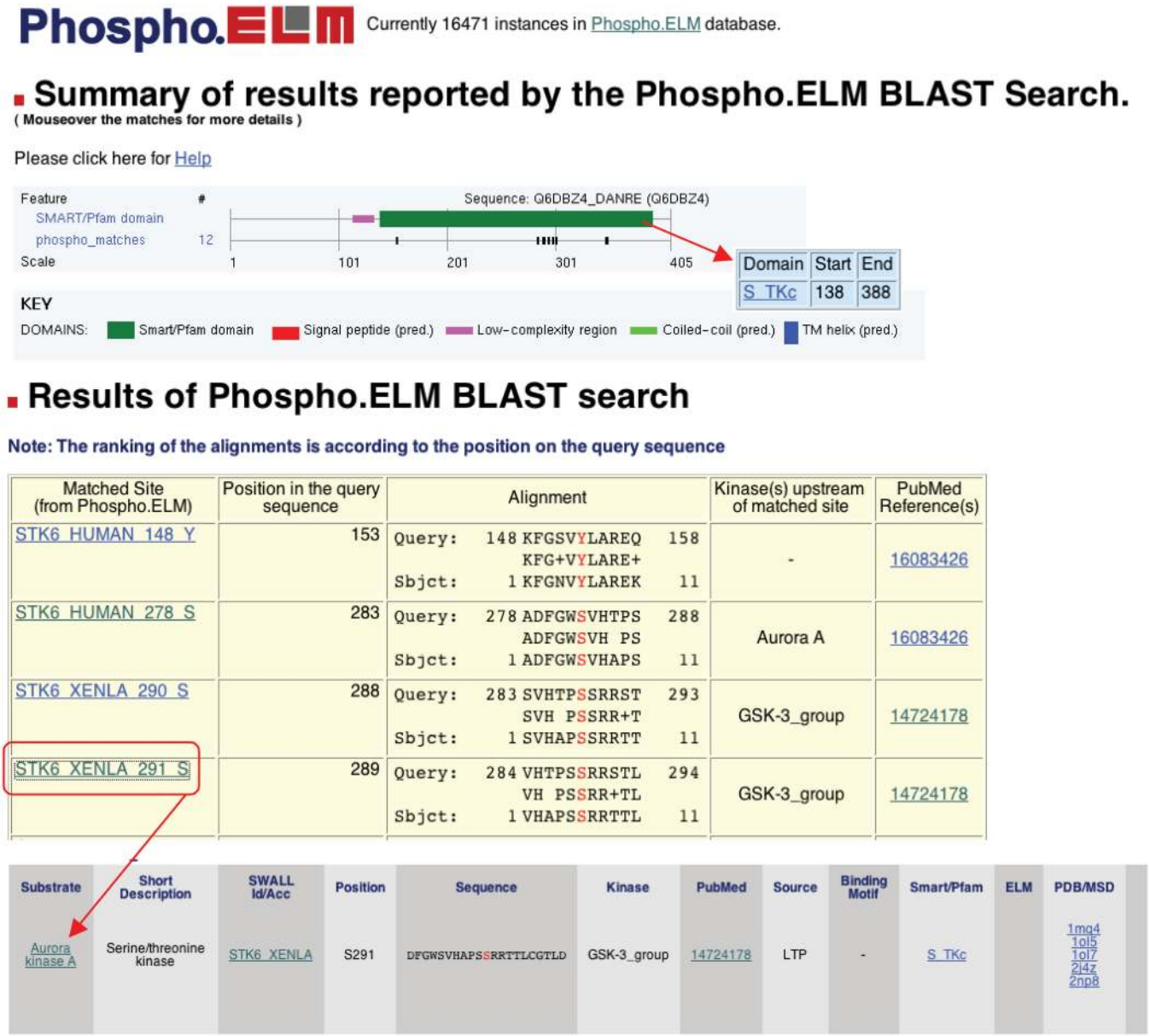

Figure 2. Output example of a PhosphoBLAST Search using as query the *Danio rerio* Aurora A kinase sequence. The summary graphic shows the phospho-hits on the query sequence and features from SMART. Details about the matches are shown below in the results table. Clicking on the 'subject name' the users can retrieve additional information about the matched Phospho.ELM phosphorylated sites, including the flanking sequence, the PubMed reference, the kinase responsible for the phosphorylation (where known) and links to additional information for the substrate and other relevant databases.

human phosphoproteome remains to be discovered. Figure 1 also shows that the rate of identification of additional phosphorylation sites on proteins has been increasing at a much faster rate than identification of novel phosphoproteins (e.g. see the srmm2 protein, UniProt accession Q9UQ35). While revealing that many more proteins are heavily phosphorylated than was previously known, it may be worth investigating whether the data also imply a strong bias in the proteins retrieved in the MS experiments.

The kinase responsible for the phosphorylation is known for ~21% of the Phospho.ELM instances. Currently, more than 250 kinases are annotated in the database (for a detailed list of the kinases see the related information at the Phospho.ELM home page).

The PhosphoBLAST search tool

A BLAST search has been implemented which is complementary to the retrieval by keyword or UniProt accession/identifier. This tool identifies phospho-peptides contained in the query sequence that match those stored in Phospho.ELM (Figure 2). It consists of a two-step process: a BLAST (20) search and a parsing of the

BLAST output. The BLAST program performs a sequence-similarity search against the Phospho.ELM data set of peptides (16 471), which have been experimentally proven to contain phospho-residues. It returns a set of local gapped alignments between the query sequence peptides and the phospho-peptides. In the parsing stage, those matches that present more than 70% sequence similarity and that conserve the phospho-residue in the same position as the corresponding phospho-peptide are selected. The final output shows the list of chosen matches, with their alignments and links to database records.

The PhosphoBLAST tool does not aim at predicting phosphorylation motifs in the query protein and is primarily useful for retrieving phosphorylation sites that are conserved in related proteins (whether orthologs or paralogs). Nevertheless, unrelated query proteins occasionally yield matching phosphorylation sites in Phospho.ELM that can be equally interesting: it will be up to the user to consider carefully the possible biological meaning (e.g. shared kinase and/or phospho-peptide-binding domain specificities) associated with these match(es).

Web service

In order to facilitate remote tool integration, a Web Service to access the phospho.ELM database programmatically has been implemented and is available at: http://phospho.elm.eu.org/webservice/phosphoELMdb.wsdl.

The WSDL (Web Service Description Language) (21) file is WS-I compatible. The WS-Interoperability Basic Profile (22) proposes a set of rules to achieve interoperability of web services between different platforms. The WSDL file implements an XML wrapped document/literal style (23). The backend code is implemented in Java and runs on Axis2 (24) inside a Tomcat servlet container (25).

The functionality provided by the Web Service encompasses the current interface functionality with some additional filters. The extra options implemented in the Web Service are to search by PubMed ID and to retrieve all instances with a PDB entry assigned to them.

Database access

Phospho.ELM is developed and deployed with open source software (26). Software is developed in Python including some modules from the BioPython project (27) to retrieve information from UniProt and PubMed. The web interface software uses the CGImodel framework (28).

The data set is publicly available for academic users. Phospho.ELM can be accessed on the public Apache2 powered website at: http://phospho.elm.eu.org.

SUMMARY

Since its inception in 2004, the Phospho.ELM data set has been adopted for numerous bioinformatics tools and pipelines e.g. the protein kinase-specific prediction server GPS (group-based phosphorylation scoring method) (29), the RLIMS-P, a rule-based text-mining program designed to extract information on phosphorylation sites from abstracts (30), PhosphoregDB, a database of tissue and sub-cellular distribution of mammalian protein kinases and phosphatases (31), and NetworKin, a computational approach which combines consensus sequence motifs and contextual data to predict which kinases phosphorylate experimentally identified phosphorylation sites (6).

While anticipating that the size of the Phospho.ELM data set will constantly grow, we consider that the resource should be kept relatively lean in terms of the categories of data to be incorporated. On the other hand, links to external resources are under regular review and likely to be augmented from time to time. For example, resources such as KEGG (32) and Reactome (33) that annotate cell signaling networks are increasing their pathway coverage and it will clearly become essential to provide links to such resources. In the near future we intend to equip Phospho.ELM with links to the predicted kinase-substrate relations from the NetworKIN database (R.Linding, *et al.*, submitted for publication).

ACKNOWLEDGEMENTS

We would like to acknowledge all the Phospho.ELM users who, by reporting missing sites or sending us their data sets, have contributed to improve the database. We wish to thank the EU EMBRACE grant (LHSG-CT-2004-512092) for funding. Many thanks to Ivica Letunic and Arnaud Ceol for technical support. We are grateful to Lars Juhl-Jensen and Rune Linding for their insightful comments and suggestions. We are thankful to Niall Haslam for critical reading of the manuscript. Funding to pay the Open Access publication charges for this article was provided by EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Johnson,S.A. and Hunter,T. (2005) Kinomics: methods for deciphering the kinome. *Nat. Methods*, **2**, 17–25.
2. Hunter,T. (2000) Signaling–2000 and beyond. *Cell*, **100**, 113–127.
3. Gygi,S.P., Rist,B., Gerber,S.A., Turecek,F., Gelb,M.H. and Aebersold,R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.*, **17**, 994–999.
4. Ong,S.E., Blagoev,B., Kratchmarova,I., Kristensen,D.B., Steen,H., Pandey,A. and Mann,M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics*, **1**, 376–386.
5. Ross,P.L., Huang,Y.N., Marchese,J.N., Williamson,B., Parker,K., Hattan,S., Khainovski,N., Pillai,S., Dey,S. *et al.* (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics*, **3**, 1154–1169.
6. Linding,R., Jensen,L.J., Ostheimer,G.J., van Vugt,M.A., Jorgensen,C., Miron,I.M., Diella,F., Colwill,K., Taylor,L. *et al.* (2007) Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**, 1415–1426.
7. Beausoleil,S.A., Jedrychowski,M., Schwartz,D., Elias,J.E., Villen,J., Li,J., Cohn,M.A., Cantley,L.C. and Gygi,S.P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl Acad. Sci. USA*, **101**, 12130–12135.

8. Brill,L.M., Salomon,A.R., Ficarro,S.B., Mukherji,M., Stettler-Gill,M. and Peters,E.C. (2004) Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry. *Anal. Chem.*, **76**, 2763–2772.
9. Nousiainen,M., Sillje,H.H., Sauer,G., Nigg,E.A. and Korner,R. (2006) Phosphoproteome analysis of the human mitotic spindle. *Proc. Natl Acad. Sci. USA*, **103**, 5391–5396.
10. Olsen,J.V., Blagoev,B., Gnad,F., Macek,B., Kumar,C., Mortensen,P. and Mann,M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, **127**, 635–648.
11. Rush,J., Moritz,A., Lee,K.A., Guo,A., Goss,V.L., Spek,E.J., Zhang,H., Zha,X.M., Polakiewicz,R.D. *et al.* (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.*, **23**, 94–101.
12. Villen,J., Beausoleil,S.A., Gerber,S.A. and Gygi,S.P. (2007) Large-scale phosphorylation analysis of mouse liver. *Proc. Natl Acad. Sci. USA*, **104**, 1488–1493.
13. Diella,F., Cameron,S., Gemund,C., Linding,R., Via,A., Kuster,B., Sicheritz-Ponten,T., Blom,N. and Gibson,T.J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.
14. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M., Ausiello,G., Brannetti,B. *et al.* (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.
15. Boutselakis,H., Dimitropoulos,D., Fillon,J., Golovin,A., Henrick,K., Hussain,A., Ionides,J., John,M., Keller,P.A., Krissinel,E. *et al.* (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res.*, **31**, 458–462.
16. Zanzoni,A., Ausiello,G., Via,A., Gherardini,P.F. and Helmer-Citterich,M. (2007) Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res.*, **35**, D229–D231.
17. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
18. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
19. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
20. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
21. Web Services Description Language http://www.w3.org/TR/wsdl (30 August 2007, date last accessed).
22. Web Services Interoperability Basic Profile Version 1.0 http://www.ws-i.org/Profiles/BasicProfile-1.0-2004-04-16.html (30 August 2007, date last accessed).
23. Document/Literal Wrapped style http://www-128.ibm.com/developerworks/webservices/library/ws-whichwsdl/ (30 August 2007, date last accessed).
24. Axis 2 http://ws.apache.org/axis2 (30 August 2007, date last accessed).
25. Tomcat http://tomcat.apache.org/ (30 August 2007, date last accessed).
26. PostgreSQL http://www.postgresql.org (30 August 2007, date last accessed).
27. BioPython http://BioPython.org (30 August 2007, date last accessed).
28. Ramu,C., Gemund,C. and Gibson,T.J. (2000) Object-oriented parsing of biological databases with Python. *Bioinformatics*, **16**, 628–638.
29. Zhou,F.F., Xue,Y., Chen,G.L. and Yao,X. (2004) GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem. Biophys. Res. Commun.*, **325**, 1443–1448.
30. Hu,Z.Z., Narayanaswamy,M., Ravikumar,K.E., Vijay-Shanker,K. and Wu,C.H. (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **21**, 2759–2765.
31. Forrest,A.R., Taylor,D.F., Fink,J.L., Gongora,M.M., Flegg,C., Teasdale,R.D., Suzuki,H., Kanamori,M., Kai,C. *et al.* (2006) PhosphoregDB: the tissue and sub-cellular distribution of mammalian protein kinases and phosphatases. *BMC Bioinformatics*, **7**, 82.
32. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
33. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.