PhosphaBase: An Ontology-Driven Database Resource for Protein Phosphatases

K. J. Wolstencroft,[1]* R. Stevens,[2] L. Tabernero,[1] and A. Brass[1,2]
[1]*School of Biological Sciences, University of Manchester, Manchester, United Kingdom*
[2]*School of Biological Sciences, Department of Computer Science, University of Manchester, Manchester, United Kingdom*

ABSTRACT **PhosphaBase is an ontology-driven database resource containing information on the protein phosphatase family. It is the first public resource dedicated to protein phosphatases, which are enzymes that perform dephosphorylation reactions. In conjunction with the phosphorylation action of protein kinases, phosphatases are involved in important control and communication mechanisms in the cell. They have also been implicated in many human diseases, including diabetes and obesity, cancers, and neurodegenerative conditions. PhosphaBase aims to centralize the growing base of knowledge in the phosphatase research domain. The resource is built around a formal, domain-specific DAML+OIL ontology, and the data are collected from heterogeneous biological sources using Gene Ontology terms as a means of data extraction. The overall ontology-driven architecture provides a robust structure with distinct advantages for sustainability and provides the potential for the development of diagnostic tools, as well as a data repository. Proteins 2005;58:290–294.** © 2004 Wiley-Liss, Inc.

Key words: ontology; protein family; database sustainability; automated classification

INTRODUCTION

In the public domain, there are a growing number of resources dedicated to specific protein families, for example, the Protein Kinase Resource (PKR),[1] TRANSFAC,[2] the transcription factor database, and GPCRDB,[3] the G protein–coupled receptor database. These resources provide central data repositories specifically tailored to the requirements of the research communities they support. They also act as a focus for that community for the sharing and exchange of information.

Here, we present PhosphaBase, an ontology-driven protein family resource containing information on protein phosphatases. This is the first public resource specifically dedicated to protein phosphatases. The architecture of the system, using ontology to manage the data, provides a new approach to data capture and data management that has distinct advantages over traditional database systems in terms of sustainability and automation. In addition, PhosphaBase provides other services and links to related work. For example, a local BLAST[4] program allows users to compare their protein sequence with all the phosphatases in the database, and information about forthcoming phosphatase conferences is available.

PhosphaBase is available online (http://www.bioinf.man.ac.uk/phosphabase).

The Protein Phosphatase Family

Phosphorylation and dephosphorylation reactions are important mechanisms of control and communication in most cellular processes, including metabolism, homeostasis, cell signaling, transport, muscle contraction, and cell growth. These reactions allow the cell to respond to external stimuli, such as hormones and growth factors, as well as to cellular stress and cytokines.[5–10]

Regulation of the phosphorylation state of cellular targets is controlled by the interaction of protein kinases and protein phosphatases. Since phosphorylation events play a part in almost all biological processes, protein kinases and protein phosphatases are important targets for medical, pharmaceutical, and molecular biology research. Both protein kinases and protein phosphatases have been implicated in many human diseases, including cancers, diabetes and obesity, and neurodegenerative conditions.[11–20] Therefore, specific resources dedicated to these protein families are essential to capture the growing knowledge in these areas. There are already several resources dedicated to protein kinases, for example, the PKR[1] or KinBase.[21] However, up until now, there has not been a public resource specifically dedicated to protein phosphatases.

Database Content

The content and data structure of PhosphaBase was constructed in close collaboration with phosphatase biologists. Since the database will primarily be serving the phosphatase research community, the data contained within it had to reflect the needs of that community. As a result, the data in PhosphaBase range from the genomic level to three-dimensional (3D) protein structures. The current release has over 2800 phosphatase entries from 345 different species.

All data in PhosphaBase have been populated with data extracted from peer-reviewed literature and from publicly

available databases [e.g., UniProt,[22] Protein Data Bank (PDB),[23] LocusLink,[24] etc.]. PhosphaBase reflects the collective knowledge of the phosphatase research community and does not attempt to alter or modify those views.

MATERIALS AND METHODS

PhosphaBase Architecture Overview

PhosphaBase is an ontology-driven resource. It is constructed around two separate ontologies, a phosphatase domain–specific ontology (the PhosphaBase ontology) to manage the structure and content of the database, and the Gene Ontology[25] to facilitate extraction of relevant data from external biological sources. The PhosphaBase ontology was built using the ontology editor OILED[26] in the description logic language DAML+OIL.[27]

The structure of the PhosphaBase ontology was used to form the PhosphaBase relational database schema. Each of the properties necessary for ontology descriptions formed database tables, or fields within tables, mapping the ontology to the database.

PhosphaBase data are stored in a MySQL relational database with a java servlet, web-accessible user interface. Since java is platform-independent and both MySQL and java are freely available, this will facilitate the installation of local copies of the resource by users should they wish in the future.

Figure 1 shows the overall architecture of the PhosphaBase system.

Data Extraction

The data in PhosphaBase have been extracted from a number of different publicly available biological databases. Automatically extracting all relevant data without introducing any irrelevant data, and without omitting results, presented a number of problems. The same molecule can have multiple names, which can have multiple derivative abbreviations; for instance, protein-tyrosine phosphatase 1C has also been called protein-tyrosine phosphatase nonreceptor type 6, PTP-1C, hematopoietic cell protein-tyrosine phosphatase, 70Z-SHP, SH-PTP1, PTP-1C, and PTP1C. Key word searches that do not include all synonyms may result in the omission of data. Another problem common to most biological databases is that some of the information, often the most pertinent or important, is stored in free text, which is computationally unreadable.

To overcome these problems in PhosphaBase, we do not rely on free-text and key word searches to extract data from biological resources. Instead, data are extracted using terms from the Gene Ontology.

Gene Ontology

The Gene Ontology project (GO) began after whole-genome sequencing initiatives revealed that almost all gene products were common to Eukaryotes. The aim of GO is to provide a controlled vocabulary to describe these common gene products, thereby ensuring a shared understanding of terminology in the scientific community. There are now over 17,500 terms in GO, and many biological
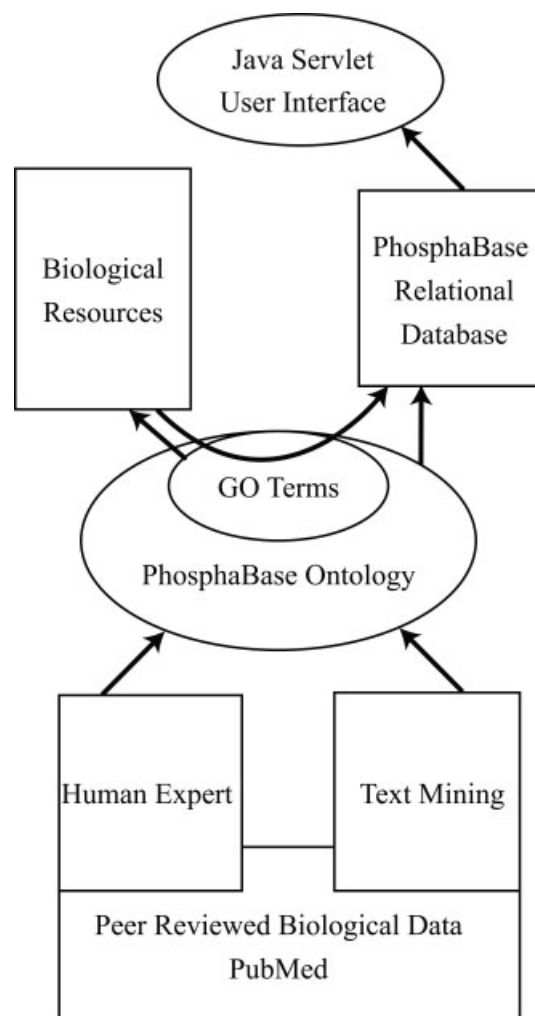

Fig. 1. A diagram representing the PhosphaBase architecture. The domain-specific ontology is the central organizational point of PhosphaBase. It was constructed using peer-reviewed literature and contains GO terms relating to protein phosphatases. These GO terms facilitate the extraction of data from biological resources into the database. The java servlet user interface directly accesses the database.

databases have been working to annotate their entries to GO terms. Consequently, GO is becoming an important factor in the unification of heterogeneous biological resources. Problems associated with free-text searching are greatly reduced by using the GO terms to extract relevant data. A single GO term has associated with it all know synonyms for a particular biological molecule, which means that the user can extract all relevant information without having to know all of those synonyms. For example, searching proteins annotated to the GO term "protein tyrosine phosphatase activity (GO: 0004725)" returned 201 results. Out of the 201 proteins, 67 had names that did not include the key words "tyrosine phosphatase" or "PTP." Therefore, performing a similar search using key words would have missed all of these proteins, unless the names of each were incorporated into the key word list.

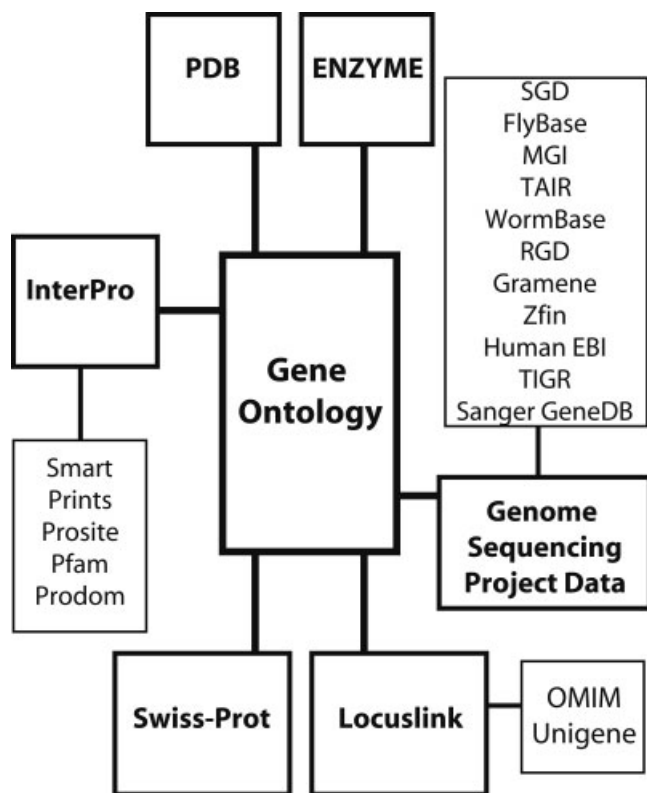Figure 2 shows the resources currently accessible using GO terms.

Fig. 2.   A diagram showing biological resources that have GO annotations and are therefore accessible using GO terms. Bold lines denote direct accessibility. Plain lines denote resources that are integrated to those that are directly accessible.

The PhosphaBase Ontology

The GO is made up of three separate ontologies: molecular function, biological process, and cellular component. While terms from these ontologies can provide a lot of information on individual gene products, many areas are not covered from a protein family perspective. For example, information on structural elements and motifs is an important consideration when comparing closely related protein families. Similarly, GO does not address genotype–phenotype relationships and the involvement of a gene product in a disease state, but instead focuses only on the "normal" state. For these reasons, we concluded that in order to properly represent the phosphatase research domain, we had to build a domain-specific ontology that would encompass those areas identified by the community as important.

We modeled the phosphatase research domain in a DAML+OIL ontology. DAML+OIL is a formal, description logic language. In such a description logic system, terms (or classes) in the ontology can be explicitly defined by their relationships with other terms and by the attributes assigned to them. For example, classes can be represented in taxonomic hierarchies; an enzyme "is a" protein and a phosphatase "is an" enzyme. They can also be described by their properties; a tyrosine phosphatase "dephosphorylates" tyrosine residues, protein phosphatase type 2A "contains" *exactly 3* subunits and requires "no

metal ions" for activity. In this way, phosphatases can be described by their functional and structural properties.

In the PhosphaBase ontology, each phosphatase subfamily is described in terms of its enzyme activity, its domain and subunit structure, known substrates, inhibitors and activators, and functional motifs. The descriptions allow us to distinguish any phosphatase subfamily from any other. The result is that the computer captures the understanding of what a phosphatase protein is in a comparable manner to that of a domain expert. For example, if human experts were asked to define what a protein phosphatase type 2B is, they would use the functional and structural properties of the molecule. They might say that it is a serine/threonine phosphatase, that it has two subunits, calcineurin A and calcineurin B, and that for enzyme activity it requires the presence of calmodulin and calcium. The definition in the DAML+OIL ontology gives the same information. The structure of the DAML+OIL means that there are no computationally unreadable free-text definitions. The definitions are produced from the properties and attributes of each ontology class.

Automated Classification

Using a formal DAML+OIL ontology in PhosphaBase not only supports a robust data repository for phosphatase-related data but also enables the development of a knowledge base and diagnostic tool for phosphatase classification. To give an example of this functionality, the classical protein tyrosine phosphatase family (the PTP family) is divided into 8 receptor-type transmembrane subtypes and 9 nontransmembrane subtypes.[28] The GO terms for PTP proteins, however, do not account for these subtypes. GO divides the PTPs into receptor-type and nonreceptor-type PTPs, as that is their functional distinction. Therefore, extra analysis would be required to classify each PTP annotated to the GO terms into the different subtypes. The PhosphaBase ontology can be used to automatically classify tyrosine phosphatases into these groupings by comparing the domain architecture.

Membership in a particular tyrosine phosphatase subtype is dependent upon the organization of functional domains; for example, receptor-type subtypes have a transmembrane domain and either 1 or 2 tyrosine phosphatase catalytic domains, whereas nonreceptor subtypes only have 1 tyrosine phosphatase catalytic domain and no transmembrane domain. Several nonreceptor-type subtypes contain an F ezrin-radixin-moesin (FERM) homology domain, but each contains a different number of PSD-95, Dlg, and ZO-1 (PDZ) domains. By performing a domain/motif scan of each sequence using a protein pattern database, such as InterPro,[29] the domain architecture of each sequence can be determined. A comparison of the domain architecture of each sequence and the domain architecture "rules" for membership in each PTP subtype in the ontology can be achieved using a description logic "reasoner," such as FaCT,[30] over the ontology. Expressing these relationships in the formal DAML+OIL structure allows the reasoner to assign a place for each protein in the family hierarchy.

One of the most useful applications of an automated classification system such as this would be the annotation of phosphatases in new genomes.

RESULTS
Database Searching and Results

PhosphaBase is publicly available over the Internet. Searches can be performed using a series of preformed or "canned" queries. A *quick search* facility allows users to search for specific phosphatases by name or accession number, or for a specific phosphatase subfamily. An *advanced search* facility allows for more varied and complex data queries, including disease associations, genomic positions, and GO term associations for cellular location, biological process, and molecular function.

Search results can be returned as a simple list of database hits, or as a series of short summaries, including gene names, protein names, and organism and subfamily information. Results are color-coded depending on the classification of each phosphatase match. For example, tyrosine phosphatases are red, serine/threonine phosphatases are green, and dual-specificity matches are yellow. This allows the user to see "at a glance" the type of molecules that match their query.

Multiple Views

All information in the database about any particular phosphatase can be retrieved from any canned query result using the *Alternative Views* button bar on each results page. There are 6 alternative views:

Protein View contains all information specifically related to the protein itself. This includes name(s) and identifiers, family classifications, subunit/domain structure, amino acid sequence(s), catalytic activity, inhibitors, activators, substrates, cellular locations, and cellular processes.

Gene View contains all information specifically related to the gene. This includes name(s) and identifiers, genomic location, and LocusLink identifiers.

References currently contains all references relating to the submission of a phosphatase to a biological sequence database (UniProt, Trembl).

Database Links provide a list of other biological databases containing the molecule and provide links to these external sites.

Disease gives descriptions of disease states relating to phosphatases from the Online Mendelian Inheritance in Man (OMIM)[31] database and links to OMIM.

Structure provides descriptions of existing 3D structures and provides links to the PDB.

Example Queries

Since PhosphaBase is currently the only resource dedicated to phosphatase family data, queries can be performed over the data that are difficult and/or not possible elsewhere. The tutorial page in the PhosphaBase resource details query construction and the types of queries possible. Queries can be performed for specific phosphatase molecules, for example, "All dual-specificity phosphatase 2 proteins." Alternatively, they can be for more generic subfamily characteristics, for example, "What are the known inhibitors for members of the protein phosphatase type 2B subfamily?" Queries can also focus on functional or biological process information (e.g., "All human phosphatases involved in adhesion"). The origins of query results can be traced back to the peer-reviewed literature through links to PubMed.

Database Sustainability

One of the greatest challenges in biological data management is the maintenance of a resource over time. Often, one or more full-time database curators have to be appointed to keep databases updated and annotated, and in some disciplines, the amount of data generated per day or per week is vast. While this method is effective, it is prone to funding problems, especially in the academic environment. If funding is discontinued, the resource becomes outdated quickly. The use of ontology to manage PhosphaBase should prevent such problems. Data extraction from external resources using GO terms reduces protein synonyms and alternative nomenclature problems, enabling automated update strategies. When external resources issue new releases, the data are parsed for GO molecular function terms relating to protein phosphatases, and any new data are extracted to the database. This considerably reduces the number of human hours required to keep PhosphaBase updated and should sustain the life of the resource.

DISCUSSION

PhosphaBase is a resource designed primarily for those working on protein phosphatases but is expected to serve a broader scientific community. The protein phosphatase field is a growing area of research due to the diverse and numerous roles that phosphatases play in cellular regulation and disease. The amount of data being generated is increasing at a high rate and needs to be centralized and organized.

PhosphaBase is the first database resource to focus on the phosphatase domain, so we hope it will be useful for the phosphatase community. However, the scope of this project could be greatly extended in the future. The properties used in the ontology to describe phosphatases are not properties that are unique to the phosphatase family. They are essentially the same properties that would be used to describe any other enzyme. Consequently, the architecture and organization of PhosphaBase could be used as a basis for developing similar resources for other enzyme families. Since phosphatases and kinases are so often involved in the control of the same biological processes, the protein kinase family would be the natural choice for the expansion of the system.

Many of the properties in the PhosphaBase ontology are not only unique to enzymes but also describe the generic properties of proteins. For this reason, it should also be possible to extend the model even further than enzymes and produce a generic ontology model for capturing protein family data. We have recently been addressing this issue. The benefits gained from using an ontology-driven system

could be advantageous to other research communities working on specific protein families.

REFERENCES

1. Smith CM, Shindyalov IN, Veretnik S, Gribskov M, Taylor SS, Ten Eyck LF, Bourne PE. The protein kinase resource. TIBS 1997;22:444–446.
2. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 2003;31:374–378.
3. Horn F, Vriend G, Cohen FE. Collecting and harvesting biological data: The GPCRDB & NucleaRDB databases. Nucleic Acids Res 2001;29:346–349.
4. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
5. Andersen JN, Jansen PG, Echwald SM, Mortensen OH, Fukada T, Del Vecchio R, Tonks NK, Moller NP. A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage. FASEB J 2004;18:8–30.
6. Cohen P. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. Eur J Biochem 2001;268:5001–5010.
7. Hunter T. Signaling—2000 and beyond: a comprehensive review on phosphorylation and dephosphorylation. Cell 2000;100:113–127.
8. Wishart MJ, Dixon JE. PTEN and myotubularin phosphatases: from 3-phosphoinositide dephosphorylation to disease. Trends Cell Biol 2002;12:579–585.
9. Barford D, Das AK, Egloff MP. The structure and mechanism of protein phosphatases: insights into catalysis and regulation. Annu Rev Biophys Biomol Struct 1998;27:133–164.
10. Cohen PT. Novel protein serine/threonine phosphatases: variety is the spice of life [Review]. Trends Biochem Sci 1997;22:245–251.
11. Schonthal AH. Role of serine/threonine protein phosphatase 2A in cancer. Cancer Lett 2001;170:1–13.
12. Zhang ZY. Protein tyrosine phosphatases: prospects for therapeutics. Curr Opin Chem Biol 2001;5:416–423.
13. Eng C. PTEN: one gene, many syndromes. Hum Mutat 2003;22: 183–198.
14. Wu C, Sun M, Liu L, Zhou GW. The function of the protein tyrosine phosphatase SHP-1 in cancer. Gene 2003;306:1–12.
15. Prevost GP, Brezak MC, Goubin F, Mondesert O, Galcera MO, Quaranta M, Alby F, Lavergne O, Ducommun B. Inhibitors of the CDC25 phosphatases. Drug News Perspect 2003;16:637–648.
16. Ukkola O, Santaniemi M. Protein tyrosine phosphatase 1B: a new target for the treatment of obesity and associated co-morbidities. J Intern Med 2002;251:467–475.
17. Tian Q, Wang J. Role of serine/threonine protein phosphatase in Alzheimer's disease. Neurosignals 2002;11:262–269.
18. Sontag E. Protein phosphatase 2A: the Trojan Horse of cellular signaling. Cell Signal 2001;13:7–16.
19. Tonks NK. PTP1B: from the sidelines to the front lines! FEBS Lett 2003;546:140–148.
20. Dutta AS, Garner A. The pharmaceutical industry and research in 2002 and beyond. Drug News Perspect 2003;16:637–648.
21. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298:1912–1934.
22. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein Knowledgebase. Nucleic Acids Res 2004;32:D115–D119.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research 2000;28:235–242.
24. Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 2001;29:137–140.
25. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32:D258–D261.
26. Bechhofer S, Horrocks I, Goble C, Stevens R. OilEd: a reason-able ontology editor for the Semantic Web. Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence, LNAI 2001;2174:396–408.
27. Horrocks I. DAML+OIL: a reason-able web ontology language. Proc of EDBT 2002:2–13.
28. Andersen JN, Mortensen OH, Peters GH, Drake PG, Iversen LF, Olsen OH, Jansen PG, Andersen HS, Tonks NK, Moller NP. Structural and evolutionary relationships among protein tyrosine phosphatase domains. Mol Cell Biol 2001;21:7117–7136.
29. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJA, Vaughan R, Zdobnov EM. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 2003;31:315–318.
30. Horrocks I. Using an expressive description logic: FaCT or fiction? In: Cohn AG, Schubert L, Shapiro SC, editors. Principles of knowledge representation and reasoning: Proceedings of the 6th International Conference (KR '98). San Francisco: Morgan Kaufmann; 1998. p 636–647.
31. McKusick VA. Mendelian inheritance in man: a catalog of human genes and genetic disorders. Baltimore: Johns Hopkins University Press; 1998.