the FEBS Journal

REVIEW ARTICLE

HuPho: the human phosphatase portal

Susanna Liberti[1], Francesca Sacco[1], Alberto Calderone[1], Livia Perfetto[1], Marta Iannuccelli[1], Simona Panni[1,2], Elena Santonico[1], Anita Palma[1], Aurelio P. Nardozza[1], Luisa Castagnoli[1] and Gianni Cesareni[1,3]

1 Department of Biology, University of Rome 'Tor Vergata', Italy
2 Department of Cell Biology, University of Calabria, Rende, Italy
3 Research Institute 'Fondazione Santa Lucia', Rome, Italy

Keywords
database; phosphatase substrates; phosphatases; phosphorylation; protein interaction

Correspondence
S. Liberti, Department of Biology, University of Rome 'Tor Vergata', Rome, Italy
Fax: +39 062023500
Tel: +39 06725 94315
E-mail: susanna.liberti@gmail.com

(Received 28 March 2012, revised 4 July 2012, accepted 11 July 2012)

doi:10.1111/j.1742-4658.2012.08712.x

Phosphatases and kinases contribute to the regulation of protein phosphorylation homeostasis in the cell. Phosphorylation is a key post-translational modification underlying the regulation of many cellular processes. Thus, a comprehensive picture of phosphatase function and the identification of their target substrates would aid a systematic approach to a mechanistic description of cell signalling. Here we present a website designed to facilitate the retrieval of information about human protein phosphatases. To this end we developed a search engine to recover and integrate information annotated in several publicly available web resources. In addition we present a text-mining-assisted annotation effort aimed at extracting phosphatase related data reported in the scientific literature. The HuPho (human phosphatases) website can be accessed at http://hupho.uniroma2.it.

Introduction

Protein phosphorylation is a key post-translational modification that is regulated by the competing activities of protein kinases and phosphatases [1]. Substrate phosphorylation/dephosphorylation often acts as a switch to activate key regulatory proteins, governing signal propagation. Thus it is not surprising that disease conditions often correlate with alteration of the cell phosphorylation profile as a consequence of a perturbation of kinase and/or phosphatase activities [2–4]. Even though historically kinases have received more attention, recent large experimental efforts have contributed to accumulate a wealth of data on phosphatase function [5–7]. The information is growing to a pace that even domain experts find it difficult to keep track of the available data. In addition, the phosphatase field is not any longer a specialized field and biologists in general are often faced with the problem of retrieving information about a specific phosphatase found to be implicated in the biological problem of their interest. These considerations highlight the merits of resources that capture the information dispersed in the literature and present it to end-users in an organized and user-friendly format.

A number of public databases and web resources have been developed in order to address specific needs in this research area. The Protein Phosphatase Database (http://www.phosphatase.biochem.vt.edu) organizes information retrieved from the primary scientific literature, focusing on and limiting its effort to prokaryotic phosphatases [8]. PhosphaBase is a database that organizes phosphatase sequence information retrieved from UniprotKB [9]. The Protein Tyrosine

Abbreviations
CDC14, cell division cycle 14 phosphatase; DSP, dual specificity protein phosphatase; HAD, haloacid dehalogenase; LP, lipid phosphatase; LPT, lipid phospho-transferase; MTMR, myotubularin; NUDT, nucleoside-diphosphate-linked moiety X; PPI, protein–protein interaction; PPM, metal-dependent protein phosphatase; PPP, phosphoprotein protein phosphatase; PTP, protein tyrosine phosphatase.

Phosphatases website (http://ptp.cshl.edu/) is a compendium of protein tyrosine phosphatases (PTPs) that integrates sequence and structure information with cellular and biological function [10].

Despite being useful resources, none of them aims at the completeness that would meet the needs of the scientific community. In addition, similarly to many such initiatives, they face the problem of securing stable assets for maintaining the information updated.

With the project presented here we intend to address this community need by developing an online resource to assist the work of those scientists interested in studying human protein phosphatases or of those who would like to rapidly recover information on a specific phosphatase they have stumbled on. This new web portal is named 'HuPho' (human phosphatase) and it is accessible at http://hupho.uniroma2.it.

The aim of the project is to provide the scientific community with a resource that can be easily queried to obtain structural and functional information about human protein phosphatases. Most of the information is dynamically retrieved by querying the web services of publicly available repositories. This information is integrated with data specifically curated by our group and stored in our internal database. Finally, the information is organized to allow the user to browse the knowledge base through a single pane of glass.

Results and Discussion

The HuPho resource

The HuPho project addresses the need for recovering functional information about protein phosphatases. We started by retrieving and classifying human proteins containing a phosphatase domain according to sequence similarity and substrate preference (pSer/pThr, pTyr, lipids, carbohydrates etc.; see Methods) of their phosphatase domains. Next, we identified relevant web resources that contain literature-curated data accessible through web services. Finally, we identified missing information and we undertook a text-mining-assisted curation effort to complement the data automatically recovered from the publicly available resources. The result is the most comprehensive resource to date for extracting published information about phosphatase function.

Despite our effort, we cannot claim completeness. However, the website is designed to facilitate the addition of new phosphatases or extra annotation. In addition much of the information is not statically stored in an informatics warehouse but rather retrieved dynamically from primary databases. Thus the quality and coverage of the information that can be retrieved from HuPho mirrors the quality and coverage of the primary databases. Users are encouraged to contact the authors to point out missing phosphatases or missing substrates and/or interactions.

Protein phosphatase classification

In order to achieve a genome-wide perspective of the phosphatase enzymes encoded in the human genome, we performed PSI-BLAST searches using the sequences of different families of phosphatase domains [11] as a seed. The result is a compendium of 199 human proteins containing a phosphatase domain, subdivided into six distinct functional and structural superfamilies: protein tyrosine phosphatases (PTPs, 108 members), metal-dependent protein phosphatases (PPMs, 13 members), phosphoprotein protein phosphatases (PPPs, 15 members), lipid phosphatases (LPs, 37 members), haloacid dehalogenases (HADs, 21 members) and nucleoside-diphosphate-linked moiety X (NUDT, five members). HuPho also contains information about 116 regulatory subunits. Phosphatases within the six major superfamilies have been further classified based on additional structural and functional information. The details are described in the Methods section and can be downloaded from the 'Research & Tools' drop-down menu of the HuPho website. This classification is colour coded in the interactive cake graph that can be accessed from the portal home page (Fig. 1).

Website

The HuPho home page (http://hupho.uniroma2.it) offers immediate access to two search options: a free text search field and an interactive cake graph whose coloured sectors represent the different superfamilies and subfamilies.

By using the free text search field, the search can be performed by typing names, protein names, UniprotKB AC, synonyms or keywords. HuPho will retrieve from its database the protein whose description matches the entered string and will display the matching proteins as suggestions in real time. The protein of interest can be selected by clicking on one of the suggestions. The portal can also be interrogated by browsing the interactive cake graph where phosphatases are clustered according to family membership and coloured according to functional classification (Fig. 1). By moving the mouse over the cake sectors the user can highlight and select the family of interest. By clicking a relevant sector the members of the superfamily are listed and the user can easily select the phosphatase of interest.

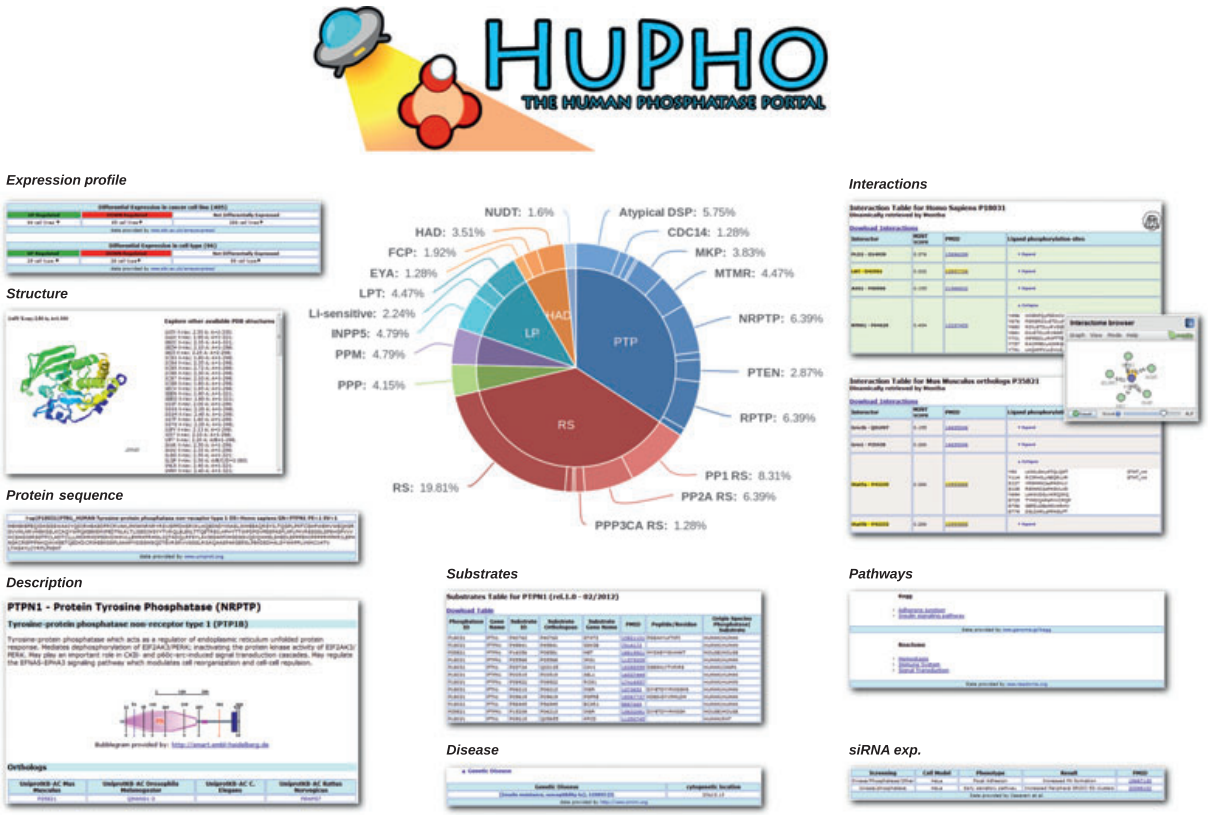

Fig. 1. The HuPho website. The figure offers an overview of HuPho functionalities. In addition to a 'standard text search' phosphatase information can be accessed via an interactive cake graph, shown at the centre of the figure. Sectors of the cake graph are colour coded and represent the phosphatase superfamilies and subfamilies. A variety of web pages, described in the main text and displaying different functional and structural information, can be reached via nine different tabs.

Once a protein is selected, the application retrieves data from nine curated 'static' tables and five different web resources, using publicly available web services. Thanks to this strategy, the latter type of data are constantly aligned with the latest release of the remote data source so that the user is provided with the most up-to-date information without extra workload for the HuPho database curators. The retrieved data are organized in sections and displayed in tabs.

Description tab

The page that can be accessed by clicking the 'Description' tab displays basic information about the queried phosphatase. This includes protein and gene names and aliases, as retrieved from the UniprotKB resource, together with information about superfamily and subfamily classification as described in Methods. In addition, we capture from the SMART database a graphical representation of domain composition. Finally, the identifiers of orthologous phosphatases in four model organisms (*Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster* and *Caenorhabditis elegans*) are reported in a separate table.

Protein Sequence tab

The sequence page dynamically retrieves from the UniprotKB database the sequence of the query protein. The amino acid sequence of the orthologues in four model organisms is also shown. In addition the 'align' button links to a UniprotKB web page that performs a CLUSTALW alignment between the orthologous proteins. Thus, conserved residues can be easily identified and visualized.

Structure tab

From the 'Structure' tab the user can reach a page where an integrated JMOL applet permits visual inspection and interaction with the 3D structures of the queried phosphatase. The atomic coordinates are

retrieved from the RCSB repository [11] and displayed with JMOL, an open-source Java viewer for displaying chemical structures in 3D (http://jmol.sourceforge.net). The RCSB repository currently offers 713 structures of phosphatase proteins and regulatory subunits. In the HuPho database 197 phosphatases (49%) have at least one structure, with PTP1B alone having 118 different ones. The structures are evenly distributed among the six families. PTP, HAD and NUDT have the highest coverage, 58%, 62% and 60% respectively, while PPM and regulatory subunits are the least represented, 20% and 25%.

Interaction tab

Protein–protein interaction information is a source of crucial importance as it allows the interpretation of protein functions in the context of the intricate interaction web inside the cell. Placing phosphatases in such an interaction web would help to infer the function of poorly characterized phosphatases and to identify potential substrates [12–14]. In order to offer a proteome-wide perspective of the phosphatase interactome, we have embarked on an extensive text-mining-assisted literature curation effort to extend phosphatase interaction information that was not yet covered by protein–protein interaction (PPI) databases. Interaction evidence captured by expert curators was annotated in the protein interaction database MINT [15] according to the rapid curation standard [16]. This data set was next integrated with protein interaction information from three additional major PPI databases, IntAct, BioGRID and DIP [17–19]. These databases are part of the PSIMEx consortium [20] and adopt a common data model and common controlled vocabularies, thus facilitating data integration. Duplicated entries were merged and redundant interactions have been removed.

As a result, from the HuPho website it is possible to explore experimental evidence from 718 scientific articles reporting $\sim$ 4600 experiments supporting protein interactions where at least one of the partners is a phosphatase. Since some interactions are supported by more than one piece of evidence, the actual number of non-redundant interactions is smaller, $\sim$ 2500 at the time of writing this paper. Moreover, 199 phosphatases have at least one reported ligand, while 53 have none. Interaction evidence is fairly evenly distributed in the four PSIMEx resources suggesting a substantial lack of overlap among the data curated by each database.

The PPP and PTP families, as well as the regulatory subunits, are well connected in the human phosphatase network – more than 80% have known interactors – while LPs and HADs are still poorly characterized and only approximately 60% of them are integrated in the network.

Although the vast majority of the information derives from experiments with human proteins, HuPho also captures interaction evidence obtained in *M. musculus* and *R. norvegicus* model organisms. To further increase the coverage of the PPI network, these interactions are then mapped to the corresponding human orthologues and presented to the user in organism-specific tables. Each row of the table corresponds to one interaction partner and is linked to the PubMed abstract of the paper reporting the experimental evidence. Most importantly, each interaction is assigned a score [15] that estimates the degree of confidence, given the supporting experimental evidence. The last column of the table lists all the sites in the partner proteins that are annotated as phosphorylated in the Phosphosite database [21]. This information may help for inferring dephosphorylation sites in ligand proteins. In addition curators of DIP, MINT and IntAct also capture interactions that lead to dephosphorylation of the ligand when proved to be direct by *in vitro* experiments. These interactions are highlighted in yellow in the table. The interaction information is also offered to the user via an interactive graphical applet that permits the phosphatase interactions in the context of the human interactome to be explored.

Substrate tab

Protein interaction databases by default do not capture information on enzymatic interactions unless these are demonstrated to be direct by *in vitro* experiments. However, a variety of *in vivo* and *ex vivo* experiments provide useful information about proteins whose phosphorylation levels depend on the activity of a specific phosphatase, although they do not offer definitive evidence of direct dephosphorylation. This type of information was captured from the literature by an additional curation effort. Since the evidence cannot be annotated in IMEx interaction databases, we created a static table containing evidence of dephosphorylation reactions. In this table we also integrate information deriving from experiments in model organisms mapped onto the corresponding human orthologues. Whenever available, the sequence context of the target phosphorylated residue is also shown. Each row in the table is linked to the PubMed entry reporting the corresponding evidence. In addition, we concisely classified evidence as '*in vitro*', '*in vivo*' or 'trapping', according to the experimental background

in which the dephosphorylation was demonstrated. Experiments performed with purified proteins demonstrating a direct dephosphorylation are classified as '*in vitro*'. Dephosphorylation characterized in a cellular context underlines their functional relevance but does not distinguish between direct and indirect. These are classified as '*in vivo*'. Finally, experiments that take advantage of the substrate trapping mutant technique [22] are indicated as 'trapping'.

Expression Profile tab

The page accessed from the 'Expression Profile' tab is subdivided into three subsections: 'Tissue Expression', 'Cell Expression' and 'Subcellular Localization'. Each of the subsections shows the expression profile of the selected gene or protein from a different perspective. In 'Tissue Expression' data annotated from UniprotKB curators are displayed and hyperlinked to the relevant supporting evidence. 'Cellular Expression' links to experimental evidence extracted from ArrayExpress [23], a database that stores and integrates a large number of DNA array experiments. The retrieved data are organized by listing the cancer cell lines and the cell types where the phosphatase is upregulated or downregulated [24]. Finally the 'Subcellular Localization' subsection displays information on the subcellular localization of the selected protein, as annotated in the Human Protein Atlas website [25].

Disease tab

The page that can be accessed via the 'Disease' tab provides information about the role of the selected protein in both cancer and genetic diseases. The first section displays information curated by our group and stored in a static table that classifies phosphatases into oncogenes and tumour suppressors. In order to obtain a comprehensive picture of phosphatase involvement in cancer, we first used automated text-mining strategies to automatically retrieve abstracts containing both a phosphatase gene name and the words 'cancer' or 'tumour'. Next, we identified relevant papers by reading each of the automatically retrieved abstracts. Finally, from each paper and for each phosphatase gene, we have annotated the PubMed identifier of the paper and we have extracted information regarding its role in cancer development and progression. By this procedure 32 genes are classified as oncogenes and 41 as tumour suppressors while 14 have oncogene or tumour suppressor function depending on context. This classification is hyperlinked to the relevant bibliography used for the classification. The second section displays the list of genetic diseases that are associated with the queried phosphatase in the OMIM database [26]. Convenient hyperlinks permit the annotated information in the primary OMIM database to be navigated.

Pathways tab

In all, 125 KEGG pathways [27] and 75 Reactome pathways [28] include information on the activity of at least one phosphatase or a regulatory subunit. The list of pathways that are modulated by the query protein can be explored by clicking the 'Pathways' tab. The information is hyperlinked to the pathway page in the primary databases.

'siRNA experiment' tab

In the 'siRNA experiment' tab, it is possible to access functional information from siRNA screenings in human cell lines. Data were retrieved both from the recently developed database GenomeRNAi [29] and from additional data curated by our group. Thus, for 172 of 313 phosphatases and regulatory subunits the phenotype observed upon gene downregulation was annotated together with the cell type where the phenotype has been observed. The data are hyperlinked to the relevant bibliographic reference.

Conclusions

HuPho is a web tool designed to provide convenient access to comprehensive information about phosphatase enzymes. This is achieved via a combination of methods that interrogate web services of established databases and integrate the extracted information with additional published data specifically curated for this project by our group. Via the web portal the user can browse the information presented in nine different web pages or download the data for local use.

By this effort most phosphatases have at least one annotation, an interaction, a substrate, a functional annotation by siRNA knockdown experiments or an association to a disease. Large-scale efforts are under way and the available information is likely to increase dramatically soon. We intend HuPho to become a community resource and we encourage users to point out missing information and to submit their new data at the time of publication. Finally we want to stress once more that most of the information that can be extracted by interrogating HuPho is retrieved dynamically from established specialized public repositories. The completeness or incompleteness of the

HuPho resource mirrors the coverage of phosphatase information in these public resources. The community of phosphatase scientists can contribute to the value of HuPho by ensuring that the results of their work are annotated by the specialized public resources at the time of publication.

Methods

Website

The HuPho web resource is entirely developed in PHP, with the use of Ajax and jQuery for the interactive cake graph. An underlying Postgres database stores the data provided by our group as well as a static copy of some of the dynamically accessed data. The static copy of the data is returned to the user in case of network problems that inhibit the use of the remote web services such as, for instance, timeouts of the remote service. The website structure is fully portable and could be easily adapted to any list of proteins; it is stored on a Linux server, running APACHE2, PHP 5 and POSTGRES 8. The integration of data from several web resources was achieved by using the UniProtKB AC of the protein as main identifier. To retrieve data from databases that use different identifiers, gene name approved symbols or EnsembleGN, we used the dictionary implemented in the UniprotKB ID mapping tool.

Phosphatase classification

The table that can be downloaded from the 'Phosphatase Classification' item in the 'Resources & Tools' menu arranges the six major phosphatase superfamilies into more specific subfamilies according to the criteria described below.

PTP superfamily

The tyrosine phosphatase superfamily forms the largest and most complex group and consists of 108 enzymes, all characterized by the active-site signature motif 'HCX_5R', also dubbed 'PTP signature'. The residues in this peptide motif play important roles in the catalytic mechanism [30]. The PTP superfamily can be further subdivided into three different subfamilies. The first subfamily consists of the 'classical' tyrosine phosphatases, which have strictly tyrosine specificity (41 members), and the dual specificity protein phosphatases (DSPs) (VH1-like), which dephosphorylate both Ser/Thr and Tyr residues (61 members). The second subfamily consists of a unique tyrosine-specific low molecular weight phosphatase, whose origin appears to be more ancient than classical PTP. In the third subfamily, we include three tyrosine/threonine specific phosphatases (CDC25) that most probably evolved from a bacterial rhodanese-like enzyme [31,32]. The classical PTPs consist of 41 phosphatases, which can be classified into transmembrane, receptor-like enzymes and the intracellular, non-receptor-like phosphatases. In the human genome the receptor-like enzymes are represented by 21 proteins, while the cytosolic PTPs are encoded by 20 genes [32]. The VH1-like family consists of 63 members, which can be further subdivided into seven distinct subfamilies: mitogen activated kinase phosphatases, atypical DSPs, slingshot homologues, phosphatases of regenerating liver, cell division cycle 14 (CDC14) phosphatases, PTEN phosphatases and myotubularins (MTMRs) [31]. Eleven genes encode the mitogen activated kinase phosphatases, which specifically dephosphorylate the threonine and tyrosine residues of the activation loop contained in the MAPKs ERK, JNK and p38 [33,34]. These phosphatases possess a CH2 region and a number of MAPK targeting motifs, such as the kinase interacting motif (KIM), which are completely absent in the 19 atypical DSPs, whose substrate specificity is different and little characterized. Atypical DSPs, slingshot homologues and the three phosphatases of regenerating liver are poorly characterized enzymes [31]. CDC14 phosphatases are regulators of cell cycle and mitosis exit [35]. Finally the last two subgroups, the PTENs (nine genes) and MTMRs (14 genes), target both phosphorylated lipids and, albeit with a lower K_{cat}, phosphorylated tyrosines, serines and threonines. MTMR phosphatases mainly dephosphorylate phosphatidylinositol-3-phosphate on internal cell membranes; PTENs target phosphatidylinositol-3,4,5-trisphosphate at the plasma membrane [36]. Remarkably, five members of MTMR phosphatases (MTMR9, MTMR10, MTMR11, SBF1 and SBF2) have inactivating mutations within the catalytic site.

PPP superfamily

In contrast to the PTP family, PPP phosphatases catalyse a bimetal (Fe^{3+} and Zn^{2+}) dependent dephosphorylation of protein substrates [37]. The PPP superfamily consists of 13 members, whose catalytic activity, subcellular localization and substrate activity is determined by their interaction with a large number of regulatory subunits and adaptors [38].

PPM superfamily

Similarly to PPP phosphatases, PPMs dephosphorylate phosphorylated serine and threonine residues. Their catalytic domain is composed of about 250 amino acids and some of the invariant residues bind divalent metal ions of Mn^{2+}. The alignment of prokaryotic and eukaryotic PPM catalytic domains enabled the identification of 11 short signature motifs, only six of which are conserved in human PPMs. These include FX_3DGH, GDSR, DHK, EX_2RI, IGD and DGxW, where the conserved aspartic acid plays a key role in binding the divalent metal ions at the catalytic centre [39].

HAD superfamily

The HAD superfamily is a distinct and heterogeneous group that can be subdivided into three different subfamilies: FCP, EYA and HAD phosphatases. The only known function of the 11 members of the FCP subfamily is the dephosphorylation of the carboxy terminal domain of RNA polymerase II [37].

The EYA subgroup has been recently discovered and consists of four enzymes, which share a high sequence homology. Interestingly EYAs have a HAD-like phosphatase catalytic domain and they are able to dephosphorylate both phosphotyrosine and phosphoserine residues [31], although the latter activity may be in question. The elucidation of the mechanisms underlying specificity of substrate selection of this enzyme subfamily requires further analysis [40,41]. Similarly, HAD phosphatases are a heterogeneous and poorly characterized subfamily, whose functional role and substrate specificity is still unknown.

LP superfamily

The LP superfamily can be subdivided into three different subfamilies: lipid phospho-transferases (LPTs), Li-sensitive phosphatases and INPP5s. LPTs dephosphorylate both lipid phosphomonoesters and diacylglycerol-pyrophosphate. The LPT subfamily consists of 15 members whose catalytic domain is characterized by the following signature motifs: SRH and SRX_5HHX_2D, where histidine and aspartic acid residues act as nucleophiles during catalysis [42]. The Li-sensitive phosphatase subfamily consists of seven enzymes, which are completely inhibited by sub-millimolar concentrations of lithium [43]. The representative members of this enzyme subfamily are inositol polyphosphate 1-phosphatases (INPP1) and inositol monophosphate phosphatases, which both catalyze the removal of 1-phosphate from inositol. Inositol can be targeted also by the inositol 5 phosphatases subfamily (INPP5), which is specifically involved in the dephosphorylation of the following four substrates: Ins-1,4,5-P3, Ins-1,3,4,5-P4, and the lipids PtdIns-4,5-P2 and PtdIns-3,4,5-P3 [44].

NUDT superfamily

NUDT phosphatases are a heterogeneous group of enzymes that hydrolyse the pyrophosphate linkage in a variety of nucleoside triphosphates [45].

Phosphatase regulatory subunits

Regulatory subunits are very diversified in structure and domain composition. In addition it is more difficult to give a clear functional definition. We have not made an attempt to redefine this protein class but rather we have accepted the annotation made by the HUGO Gene Nomenclature Committee. To the proteins annotated in this database as phosphatase regulatory subunits we have added 11 proteins that are described in the literature to modulate phosphatase activity: O43423, O59626, P39687, Q92688, Q9BTT0, O75167, Q01105, Q96KR7, Q9C0D0, Q99653, Q9Y6j0.

Protein interaction curation

PubMed was searched for abstracts containing a phosphatase name or synonym together with a second protein name (putative partner) and a word indicating an interaction, either physical or enzymatic. This approach was combined with a systematic search in the Human Protein Reference Database for interaction entries containing at least one phosphatase name [46]. This strategy allowed us to collect $\sim$ 700 PubMed abstracts (PMIDs) that were read by expert curators to filter out irrelevant publications. The 240 PMIDs that passed this screening have been annotated in the MINT database [15] according to the rapid curation standard, established by the PSI-MI initiative. The information annotated in the Human Protein Reference Database was also re-curated to match the PSI-MI standards.

Three interaction types were distinguished depending on the type of experimental evidence: (a) 'physical interaction' when the experimental evidence supported the existence of a complex; (b) 'enzymatic interaction' (dephosphorylation), if the paper provides evidence that the phosphatase was directly dephosphorylating its substrate; (c) 'genetic interaction' whenever the experiment demonstrates that perturbing a phosphatase activity would affect the target substrate phosphorylation level but no evidence of direct dephosphorylation is provided. The rapid curation standard captures from any experimental evidence the following information: (a) the identity of the two interacting proteins by mapping their partners to the UniProtKB ID; (b) the experimental method; (c) the PubMed identifier of the paper reporting the experimental evidence. This curation effort resulted in the annotation of 240 new articles.

Orthology mapping

To identify orthologous phosphatases in different taxa, human phosphatases and regulatory subunits were used for a search in the Ensemble-BioMart web service. Each human sequence was automatically aligned against the proteome of *M. musculus*, *R. norvegicus*, *D. melanogaster* and *C. elegans* by the TreeBeST method [47]. For each human gene, the corresponding orthologue was annotated.

Retrieval of data from ArrayExpress

Among the different parameters that can be chosen to query the ArrayExpress web service, we decided to use

'cancer cell line' and 'human cell type'. As a result in the 'Cellular Expression' subsection it is possible to visualize, for the phosphatase of interest and for each cancer cell line or human cell type, the experiments in the ArrayExpress repository that have reported the gene to be upregulated or downregulated. Parameters are referred to in ArrayExpress as EFO terms (Experimental Factor Ontology) [24].

Acknowledgement

This work was supported by Telethon (GGP09243), the Italian Association for Cancer Research (AIRC), the FIRB project Oncodiet and the FP7 projects Affinomics and PSIMEX.

References

1 Manning G, Whyte DB, Martinez R, Hunter T & Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* **298**, 1912–1934.
2 Easty D, Gallagher W & Bennett DC (2006) Protein tyrosine phosphatases, new targets for cancer therapy. *Curr Cancer Drug Targets* **6**, 519–532.
3 Gee CE & Mansuy IM (2005) Protein phosphatases and their potential implications in neuroprotective processes. *Cell Mol Life Sci* **62**, 1120–1130.
4 Tonks NK (2006) Protein tyrosine phosphatases: from genes, to function, to disease. *Nat Rev Mol Cell Biol* **7**, 833–846.
5 MacKeigan JP, Murphy LO & Blenis J (2005) Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. *Nat Cell Biol* **7**, 591–600.
6 Goudreault M, D'Ambrosio LM, Kean MJ, Mullin MJ, Larsen BG, Sanchez A, Chaudhry S, Chen GI, Sicheri F, Nesvizhskii AI *et al.* (2009) A PP2A phosphatase high density interaction network identifies a novel striatin-interacting phosphatase and kinase complex linked to the cerebral cavernous malformation 3 (CCM3) protein. *Mol Cell Proteomics* **8**, 157–171.
7 Barr AJ, Ugochukwu E, Lee WH, King ONF, Filippakopoulos P, Alfano I, Savitsky P, Burgess-Brown NA, Müller S & Knapp S (2009) Large-scale structural analysis of the classical human protein tyrosine phosphatome. *Cell* **136**, 352–363.
8 Kennelly PJ (2001) Protein phosphatases – a phylogenetic perspective. *Chem Rev* **101**, 2291–2312.
9 Wolstencroft KJ, Stevens R, Tabernero L & Brass A (2005) PhosphaBase: an ontology-driven database resource for protein phosphatases. *Proteins* **58**, 290–294.
10 Andersen JN, Del Vecchio RL, Kannan N, Gergel J, Neuwald AF & Tonks NK (2005) Computational analysis of protein tyrosine phosphatases: practical guide to bioinformatics and data resources. *Methods* **35**, 90–114.
11 Deshpande N, Addess KJ, Bluhm WF, Merino-Ott JC, Townsend-Merino W, Zhang Q, Knezevich C, Xie L, Chen L, Feng Z *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res* **33**, D233–D237.
12 Sacco F, Perfetto L, Castagnoli L & Cesareni G (2012) The human phosphatase interactome: an intricate family portrait. *FEBS Lett* **586**, 2732–2739.
13 Ferrari E, Tinti M, Costa S, Corallino S, Nardozza AP, Chatraryamontri A, Ceol A, Cesareni G & Castagnoli L (2011) Identification of new substrates of the protein-tyrosine phosphatase PTP1B by Bayesian integration of proteome evidence. *J Biol Chem* **286**, 4173–4185.
14 Sacco F, Tinti M, Palma A, Ferrari E, Nardozza AP, Hooft van Huijsduijnen R, Takahashi T, Castagnoli L & Cesareni G (2009) Tumor suppressor density-enhanced phosphatase-1 (DEP-1) inhibits the RAS pathway by direct dephosphorylation of ERK1/2 kinases. *J Biol Chem* **284**, 22048–22058.
15 Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* **40**, D857–D861.
16 Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C *et al.* (2004) The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. *Nat Biotechnol* **22**, 177–183.
17 Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32**, D452–D455.
18 Stark C, Breitkreutz BJ, Chatr Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* **39**, D698–D704.
19 Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU & Eisenberg D (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res* **32**, D449–D451.
20 Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman F, Cesareni G *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods* **9**, 345–350.
21 Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V & Sullivan M (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* **40**, D261–D270.

22 Flint AJ, Tiganis T, Barford D & Tonks NK (1997) Development of 'substrate-trapping' mutants to identify physiological substrates of protein tyrosine phosphatases. *Proc Natl Acad Sci USA* **94**, 1680–1685.
23 Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, Contrino S, Vilo J, Abeygunawardena N, Mukherjee G, Holloway E *et al.* (2003) ArrayExpress: a public database of gene expression data at EBI. *C R Biol* **326**, 1075–1078.
24 Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A & Parkinson H (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118.
25 Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S *et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* **28**, 1248–1250.
26 Hamosh A, Scott AF, Amberger JS, Bocchini CA & McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514–D517.
27 Kanehisa M, Goto S, Sato Y, Furumichi M & Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109–D114.
28 Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* **37**, D619–D622.
29 Gilsdorf M, Horn T, Arziman Z, Pelz O, Kiner E & Boutros M (2010) GenomeRNAi: a database for cell-based RNAi phenotypes. 2009 update. *Nucleic Acids Res* **38**, D448–D452.
30 Barford D, Jia Z & Tonks NK (1995) Protein tyrosine phosphatases take off. *Nat Struct Biol* **2**, 1043–1053.
31 Alonso A, Sasin J, Bottini N, Friedberg I, Osterman A, Godzik A, Hunter T, Dixon J & Mustelin T (2004) Protein tyrosine phosphatases in the human genome. *Cell* **117**, 699–711.
32 Andersen JN, Jansen PG, Echwald SM, Mortensen OH, Fukada T, Del Vecchio R, Tonks NK & Moller NP (2004) A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage. *FASEB J* **18**, 8–30.
33 Saxena M & Mustelin T (2000) Extracellular signals and scores of phosphatases: all roads lead to MAP kinase. *Semin Immunol* **12**, 387–396.
34 Keyse SM (1998) Protein phosphatases and the regulation of MAP kinase activity. *Semin Cell Dev Biol* **9**, 143–152.
35 Visintin R, Craig K, Hwang ES, Prinz S, Tyers M & Amon A (1998) The phosphatase Cdc14 triggers mitotic exit by reversal of Cdk-dependent phosphorylation. *Mol Cell* **2**, 709–718.
36 Wishart MJ & Dixon JE (2002) PTEN and myotubularin phosphatases: from 3-phosphoinositide dephosphorylation to disease. *Trends Cell Biol* **12**, 579–585.
37 Cohen P (2004) Overview of protein serine/threonine phosphatases. *Protein Phosphatases* **5**, 1–20.
38 Shi Y (2009) Serine/threonine phosphatases: mechanism through structure. *Cell* **139**, 468–484.
39 Bork P, Brown NP, Hegyi H & Schultz J (1996) The protein phosphatase 2C (PP2C) superfamily: detection of bacterial homologues. *Protein Sci* **5**, 1421–1425.
40 Tootle TL, Silver SJ, Davies EL, Newman V, Latek RR, Mills IA, Selengut JD, Parlikar BE & Rebay I (2003) The transcription factor Eyes absent is a protein tyrosine phosphatase. *Nature* **426**, 299–302.
41 Rayapureddi JP, Kattamuri C, Steinmetz BD, Frankfort BJ, Ostrin EJ, Mardon G & Hegde RS (2003) Eyes absent represents a class of protein tyrosine phosphatases. *Nature* **426**, 295–298.
42 Sigal YJ, McDermott MI & Morris AJ (2005) Integral membrane lipid phosphatases/phosphotransferases: common structure and diverse functions. *Biochem J* **387**, 281–293.
43 Patel S, Yenush L, Rodríguez PL, Serrano R & Blundell TL (2002) Crystal structure of an enzyme displaying both inositol-polyphosphate-1-phosphatase and 3′-phosphoadenosine-5′-phosphate phosphatase activities: a novel target of lithium therapy. *J Mol Biol* **315**, 677–685.
44 Majerus PW, Zou J, Marjanovic J, Kisseleva MV & Wilson MP (2008) The role of inositol signaling in the control of apoptosis. *Adv Enzyme Regul* **48**, 10–17.
45 McLennan AG, Cartwright JL & Gasmi L (2000) The human NUDT family of nucleotide hydrolases. Enzymes of diverse substrate specificity. *Adv Exp Med Biol* **486**, 115–118.
46 Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al.* (2009) Human Protein Reference Database – 2009 update. *Nucleic Acids Res* **37**, D767–D772.
47 Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R & Birney E (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327–335.