PTP-central: A Comprehensive Resource of Protein Tyrosine Phosphatases in Eukaryotic Genomes

Teri Hatzihristidis^{1,2,#}, Shaq Liu^{3,#}, Leszek Prysycz⁴, Andrew P. Hutchins³,
Toni Gabaldón⁴, Michel L. Tremblay^{1,2,*} and Diego Miranda-Saavedra^{3,*}

¹Goodman Cancer Research Center, McGill University, 1160 Pine Avenue, Montreal H3A 1A3, QC, Canada.

²Department of Biochemistry, McGill University, Montreal, QC, Canada.

³Bioinformatics and Genomics Laboratory, World Premier International (WPI), Immunology Frontier Research Center (IFReC), Osaka University, 3-1 Yamadaoka, Suita, 565-0871, Osaka, Japan.

⁴Centre for Genomic Regulation (CRG), and UPF. Dr. Aiguader, 88, 08003 Barcelona, Spain.

[#]These authors contributed equally.

* Authors for correspondence: MLT (michel.tremblay@mcgill.ca; Tel. +1 514 398 7290; Fax: +1 514 398 6769); DMS (diego@ifrec.osaka-u.ac.jp, Tel. +81 6 6879 4269; Fax. +81 6 6879 4272).

Running Title: PTP-central, a comprehensive resource of PTPs in eukaryotes.

Abstract

Reversible tyrosine phosphorylation is a fundamental signaling mechanism controlling a diversity of cellular processes. Whereas protein tyrosine kinases have long been implicated in many diseases, aberrant protein tyrosine phosphatase (PTP) activity is increasingly being associated with a wide spectrum of conditions too. PTPs are now regarded as key regulators of biochemical processes instead of simple “off” switches operating in tyrosine kinase signaling pathways. Despite the central importance that PTPs play in the cell’s biochemistry, the tyrosine phosphatomes of most species remain uncharted. Here we present a highly sensitive and specific sequence-based method for the automatic classification of PTPs. As proof of principle we re-annotated the human tyrosine phosphatome, and discovered four new PTP genes that had not been reported before. Our method and the predicted tyrosine phosphatomes of 65 eukaryotic genomes are accessible online through the user-friendly PTP-central resource (<http://www.PTP-central.org/>), where users can also submit their own sequences for prediction. PTP-central is a comprehensive and continually developing resource that currently integrates the predicted tyrosine phosphatomes with structural data and genetic association disease studies, as well as homology relationships. PTP-central thus fills an important void for the systematic study of PTPs, both in model organisms and from an evolutionary perspective.

Keywords: tyrosine phosphatase, tyrosine phosphorylation, database, HMM, sequence analysis, signal transduction.

1. Introduction

The reversible phosphorylation of proteins as carried out by protein kinases and phosphatases is one of the most widespread mechanisms for controlling cellular functions [1]: cells can quickly respond to intracellular and extracellular cues by altering the phosphorylation status of target proteins with the effect of increasing or decreasing their biological activity, modifying their sub-cellular localisation, or affecting protein stability and protein-protein interactions [2]. Reversible protein phosphorylation is thus a simple and flexible regulatory system that has been positively selected for in evolution as a general mechanism of cellular control. The first serine/threonine protein kinase ('Phosphorylase Kinase') was reported by Fischer and Krebs as early as 1955 [3, 4]. It took another 25 years to realize that v-Src (encoded by the Rous sarcoma virus) was a protein kinase [5] that phosphorylates tyrosine residues (a PTK, [6]). On the other hand, the first serine/threonine protein phosphatases were discovered during the late 1970's and early 1980's [7], and the first tyrosine-specific phosphatase (PTP1B) in 1988 [8].

Tyrosine phosphorylation in metazoans is a fundamental signaling mechanism controlling a plethora of processes ranging from development to cellular shape and motility, transcriptional regulation, and proliferation vs. differentiation decisions. Not surprisingly, the abnormal regulation of tyrosine phosphorylation on target proteins is responsible for a wide spectrum of human conditions, including diabetes, obesity, cancer and inflammatory diseases. Many diseases have been associated to PTKs as well as to protein tyrosine phosphatase (PTP) over-expression and deficiencies [9]. Historically, research on PTKs has advanced at a faster rate than investigations into PTPs. Not only were PTKs identified nearly a decade earlier than PTPs, but also the intrinsic difficulties of investigating the "disappearance" of a phosphate moiety as opposed to the appearance of the radioactive phosphate represented a major burden for the PTP field. Yet the use of both embryonic gene targeting models, and the knockdown technologies of siRNA and shRNA have validated the importance of PTP activities in a great number of signaling pathways. Moreover, major advances have been made with the development of substrate trapping techniques where specific mutations of the PTP catalytic domains allow the purification, detection and identification of their physiological substrates.

Whereas genome-wide catalogs of protein kinases have been available since the dawn of the genomics era (e.g. the *S. cerevisiae* kinome was published in 1997 [10]), the systematic identification of tyrosine phosphatomes has lagged behind. Previous work by Andersen et al on classical (tyrosine-specific) PTPs has provided fundamental insights into the structure of the tyrosine phosphatase catalytic domain, a ~280 amino acid region containing 10 discrete and highly conserved motifs [11]. This was followed by the characterization of the tyrosine-specific PTPs in human [12], and a three-way comparison including human, fly and worm [13]. These efforts were later extended by a thorough description of the entire human phosphatome (including tyrosine-specific and non tyrosine-specific PTPs) [14], and a recent review of the functions of human PTPs alongside the development of the PTP field from its origins [15]. Publicly available resources for PTPs include the 'PTP Database' (<http://ptp.cshl.edu/>) and PhosphoregDB (<http://phosphoreg.imb.uq.edu.au/home.shtml>). The PTP Database provides access to tyrosine-specific PTP sequences, multiple sequence alignments, phylogenetic trees, structures and links to human diseases. Although a fundamental source of

knowledge in the PTP field, the sequences stored in the PTP Database are derived from computational analyses done nearly 10 years ago, with the further limitation that it is not a continually updated or interactive database, or is linked to modern databases, but instead consists of downloadable files mostly in the form of text files, PDFs and Excel tables [13]. PhosphoregDB, on the other hand, is limited to providing information on mouse and human protein kinases and phosphatases, including protein sequences and interacting partners, some pathway information, tissue-specific distribution (drawn from the mouse GNF expression atlas), and a systematic subcellular localization screen done in HeLa cells [16]. The Human Phosphatase (HuPho) web portal was recently launched with a focus on human phosphatases only, including serine, threonine and tyrosine phosphatases, as well as hydrolases [17], and information is provided on expression profiles, substrates, interactions and structures. However, since HuPho focuses on more than just PTPs, their proper identification or classification is often sacrificed (for instance, EyA phosphatases are categorized only as haloacid dehalogenases and not PTPs). Finally, an important resource that focuses on human phosphatases is the ‘human DEPhOsphorylation Database’ (DEPOD), where the authors, among other important analyses, integrate substrate information with cellular localization, coexpression data and pathway information to uncover important enzyme-substrate relationships [18].

Despite the central importance that PTPs play in the regulation of cellular biochemistry and the recent advances in high-throughput sequencing, the tyrosine phosphatomes of most genomes remain uncharacterized. We present a highly sensitive and specific sequence analysis method for the automatic classification of PTPs. Our method (‘Y-Phosphatomer’) relies on a collection of publicly available protein domain models to represent the diversity of the PTP sequence universe. Upon evaluation Y-Phosphatomer gave perfect coverage and class-level classification rates on characterized PTPs, suggesting its broad utility for the genome-wide characterization of tyrosine phosphatomes. We scanned 65 distinct eukaryotic genomes and describe their PTP complements and the evolutionary distribution of the various PTP classes. Finally we present PTP-central, a comprehensive resource of tyrosine phosphatomes in multiple genomes, where users can: (i) access our PTP sequence predictions and perform multiple sequence and phylogenetic analyses online; (ii) analyze the orthology and paralogy relationships of PTP sequences across 65 eukaryotic genomes; (iii) access an up-to-date database of PTP structures as well as a curated set of genetic studies where PTPs have been implicated; and (iv) also submit their own sequences for prediction.

2. Material and methods

2.1 Current classification of protein tyrosine phosphatases (PTPs)

The PTP superfamily has been divided into 4 distinct classes that differ both in their catalytic mechanisms and phosphatase catalytic domain sequences [14]. Class I are cysteine-based PTPs including the classical tyrosine-specific phosphatases (both receptor and cytoplasmic), and the dual-specificity phosphatases (DSPs, or VH1-like). DSPs represent the most promiscuous group of PTPs in terms of substrate specificity, with some members dephosphorylating mRNAs while other enzymes dephosphorylate lipids. Class II PTPs are a small but evolutionarily conserved class of PTPs with only one member in human (ACP1); they are also found in bacteria and are structurally related to bacterial arsenate reductases. Class III PTPs, like the Class I and II members, are also cysteine-based enzymes displaying specificity towards phosphotyrosine and phosphothreonine residues. The human enzymes (CDC25A, CDC25B and CDC25C) control cell cycle progression by dephosphorylating cyclin-dependent kinases. Despite sharing a cysteine-based catalytic mechanism, Class I, II, and III PTPs are believed to have evolved independently. A fourth class of PTPs displays an aspartic acid-based catalytic mechanism with dependence on a cation, and is represented by the developmentally important EyA ('Eyes Absent') genes, of which only 1 member is found in the fruit fly versus 4 genes in mouse and human.

2.2 Systematic prediction of PTPs

We present a general method for the automatic prediction and class-level classification of eukaryotic PTPs called Y-Phosphatomer. Y-Phosphatomer relies on a specific combination of publicly available protein domain models (mostly in the form of profile hidden Markov models, or HMMs) diagnostic for the various PTP families. Briefly, we built Y-Phosphatomer by characterizing the specific combination of protein domain signatures from the smallest number of protein domain databases that allows the identification and classification of a curated set of human PTPs into their correct classes, without cross-hitting other classes (Figure 1).

<Figure 1>

2.2.1 Working data sets and identification of protein domain models specific to PTPs

In 2004 Alonso et al reported the PTP complement of the human genome following comprehensive searches in public databases for genes encoding PTPs and PTP-like genes [14]. A total of 107 human genes were found to encode PTPs belonging to 4 distinct classes: Class I (receptor PTPs, cytoplasmic PTPs and dual-specificity phosphatases), Class II (low molecular weight phosphatases), Class III (CDC25), and aspartic acid (Asp)-based phosphatases (EyA homologues). We mapped these 107 human genes onto the Ensembl database (release 67), leading to the removal of MTMR15 and PTPRVP (a pseudogene, originally annotated as PTPRV). We also expanded gene PTPN20 into 3 distinct genes (PTPN20A, PTPN20B and PTPN20C), and merged DUSP13A and DUSP13B into the DUSP13 gene. This brings the updated list of human PTPs to 105 protein-coding genes, including 20

receptor PTPs, 19 non-receptor PTPs, 58 DSPs, 1 LMWP, 3 CDC25 homologues and 4 EyA homologues (Supplementary Table 1).

The proteins encoded by the 105 human PTP genes were retrieved from the Ensembl database and analyzed with a local installation of InterProScan (release 30.0) run with default parameters [19]. This was done to identify protein domain models diagnostic for the enzymes in each PTP family. InterProScan is the functional implementation of the InterPro database, a comprehensive and integrated resource of protein domains, regions and active sites that is widely employed in the automatic annotation of proteins in genome sequencing projects [19]. The overall strategy of the method is summarized in Figure 1. Briefly, first we identified a combination of 14 protein domain models (mostly HMMs) that allows the identification of the human PTP data set and its discrimination into classes (Supplementary Table 2). We found that it was possible to combine a reduced set of protein domain models from the Pfam [20], PRINTS [21], SMART [22], SUPERFAMILY [23] and TIGRFAMs [24] databases to identify all human PTPs with perfect coverage. An advantage of InterPro is that it combines various member databases, each covering a portion of the total protein space. Therefore, by combining a set of protein domain models from a variety of sources we take advantage of the individual strengths of InterPro's member databases. The performance of this specific collection of 14 models (the 'Y-Phosphatome' library) was subsequently evaluated on two distinct data sets, and then applied to annotate the tyrosine phosphatomes of 65 eukaryotic genomes.

2.2.2 Evaluation of Y-Phosphatome

First we tested the Y-Phosphatome library for its ability to identify and classify a collection of PTP sequences from the 'PTP Database' (<http://ptp.cshl.edu/>), an important resource of tyrosine-specific PTPs. Specifically we analyzed the larger of the data sets, which includes both vertebrate and non-vertebrate tyrosine-specific PTP sequences derived from an original PSI-BLAST search, followed by expert inspection, dated August 2004 [13]. This collection of sequences comprises a set of 601 PTP domains from 61 species and 5 phyla. The protein identifiers of these sequences mapped to 6 distinct databases (DDBJ, EMBL, GenBank, PIR, RefSeq and SWISS-PROT). To perform a uniform updating of this 2004 data set, all identifiers were mapped onto the UniProt database, producing a non-redundant list of 383 UniProt proteins. This set of proteins was run through Y-Phosphatome, and in 100% of the cases the sequences were identified as tyrosine-specific PTPs. This result, although promising (coverage and correct classification rate of 100%) is limited to tyrosine-specific PTPs, the subject of study of the PTP Database.

In a second exercise we tested the Y-Phosphatome library on a high-quality set of proteins from the UniProt database, the largest public resource on protein sequences. Specifically we selected those UniProt proteins (release of 16 May, 2012) that were annotated as having PTP catalytic activity according to the UniProt-GO database [25]. The proteins had to be annotated with any of the following experimental codes: IDA ('Inferred from Direct Assay'), IPI ('Inferred from Physical Interaction'), IMP ('Inferred from Mutant Phenotype'), IGI ('Inferred from Genetic Interaction') and EXP ('Inferred from Experiment'), plus any of the Gene Ontology Molecular Function identifiers listed in Supplementary Table 3. This selection resulted in a list of 124 proteins from a diversity of species

including vertebrates (human, mouse, rat, zebrafish), invertebrates (fruit fly), fungi (budding and fission yeasts, *Candida albicans*), the model organisms *Arabidopsis thaliana* and *Dictyostelium discoideum*, and several bacteria. When applying our classification criteria, 28/124 proteins were not classified as protein tyrosine phosphatases, but a close inspection showed that these proteins had been incorrectly annotated as PTPs, or are fragments of real PTPs. Inspection of the 96 remaining proteins that were classified as PTPs showed that all were correctly assigned to their annotated classes. The evaluation tests on this set of UniProt proteins, as well as on the sequences of the PTP Database, suggest that Y-Phosphatomer can retrieve PTPs and classify them into their correct classes with perfect coverage and classification rates (Table 1).

2.3 PTP-central: contents and sequence data sources

The Y-Phosphatomer library was used to scan the predicted protein sets of 65 completely sequenced and published eukaryotic genomes. These genomes may be classified into four of the five eukaryotic supergroups (unikonts, excavates, plants and chromalveolates) following the classification scheme of Keeling and co-workers [26]. Supplementary Table 4 lists the scientific and common names of the 65 species, plus the databases where the predicted proteins were downloaded from (including version and/or data set release date), as well as their predicted tyrosine phosphatomes split by PTP class. PTP-central is stored as a MySQL relational database (<http://www.mysql.com/>) and the server is implemented as a set of Perl and Python scripts running under Apache (<http://www.apache.org/>). PTP-central is freely available at <http://www.PTP-central.org>.

2.3.1 Genetic association studies

Positive associations between PTPs and genetic studies were obtained from the Genetic Association Database (GAD) of the NIH [27]. The GAD is a comprehensive, standardized and curated collection of published genetic association studies of many common disease types. The primary focus of GAD is the archiving and summarization of information from complex genetic diseases (both from candidate gene studies and genome-wide association studies) rather than rare Mendelian disorders. The GAD release analysed (rel. 27 October 2012) features 130 653 entries from 56 397 distinct publications, and which are broadly classified into 19 distinct disease classes, including cancer, immunological, infectious, reproductive, metabolic, cardiovascular, neurological, hematological, developmental and mitochondrial diseases.

2.3.2 Three-dimensional structures of PTPs

PTP structures were retrieved from the Protein Data Bank (<http://www.rcsb.org/>, 6 November 2012), a curated and annotated repository of experimentally determined structures of proteins, nucleic acids and complex assemblies [28]. The identification of PTP structures was done by a combination of sequence-based and specific keyword searches. The initial set of results was manually inspected to remove false-positive hits, resulting in a compilation of 339 structures. The vast majority of the structures in PTP-central comprise PTP catalytic domains, although a few are of accessory PTP domains that play important roles in PTP function.

2.3.3 Orthology and paralogy relationships

Orthologs and paralogs of PTPs were fetched from MetaPhOrs [29] (<http://orthology.phylomedb.org/>, as of February 2013). MetaPhOrs is a public repository of orthologs and paralogs derived from phylogenetic trees deposited in five databases (PhylomeDB [30], Ensembl [31], EggNOG [32], Hogenom [33] and TreeFAM [34]). MetaPhOrs contains predictions for over 13 million proteins from 1963 genomes (representing ~1.4 billion homologous relationships), and is therefore the most comprehensive database of homology relationships currently available. Most importantly, MetaPhOrs also provides quality metrics for each prediction, the most important of which is the Consistency Score (CS). The CS represents the ratio between the number of trees informing about a specific orthology relationship for a pair of proteins and the total number of trees. The CS ranges from 0 (no trees predicting orthology) to 1 (all trees predicting orthology), so the closer the CS value to 1, the more robust the prediction is. We applied a CS cutoff value of 0.5, meaning that at least one-half of all trees available for a given pair of proteins do predict an orthology relationship. Any pair of proteins with a CS<0.5 is annotated as paralogous. MetaPhOrs also describes the type of orthologous relationships (one-to-one, one-to-many or many-to-many) and lists eventual co-orthologs. In total, we describe 79 355 orthologous relationships and 640 099 paralogous relationships among the PTPs predicted in the 65 eukaryotic genomes.

3. Results and Discussion

3.1 Y-Phosphatomer: a novel method for the automatic classification of PTPs

We have developed a sequence-based classification method called Y-Phosphatomer to identify PTP sequences and classify them into specific classes automatically (Figure 1). Y-Phosphatomer works by combining 14 specific protein domain models (mostly HMMs) diagnostic for the various classes of human PTP sequences and drawn from five protein domain databases (Pfam, PRINTS, SMART, SUPERFAMILY and TIGRFAMs). The protein models characteristic of each PTP class do not overlap with those of other classes, thus preventing in practice the cross-annotation of sequences to more than one PTP class. When evaluated, Y-Phosphatomer was able to retrieve all tyrosine-specific PTP sequences (n=383) from the expertly annotated PTP Database (<http://ptp.cshl.edu/>) and classify them as tyrosine-specific PTPs automatically. In a second exercise using the experimentally validated and manually annotated set of PTP sequences from the UniProt database, Y-Phosphatomer managed to retrieve all real, full-length PTP sequences and assign them to their correct families, reporting a coverage of 100% and a mis-classification rate of zero on the class level. The UniProt data set was limited in size (n=124) albeit phylogenetically diverse. Y-Phosphatomer relies on the recommended “trusted cut-offs” of the distinct protein models as provided in InterProScan, which are thought to report relevant matches [19]. **Other than this, no test for false positives was implemented, as no gold standard exists yet for the correct classification of PTPs.**

Andersen and colleagues previously reported a set of guidelines for the computational identification of tyrosine-specific PTPs using web-based resources, including BLAST, PSI-BLAST and keyword searches [13]. Despite the usefulness of this approach, carefully selected collections of protein

models describing specific protein families have been shown to be more sensitive and specific than general tools for the database search and automatic classification of protein sequences, as we have previously shown for protein kinases [35, 36] and ubiquitinating and deubiquitinating enzymes [37]. The protein models combined into the Y-Phosphatome library were specifically built to represent well-studied PTP classes and families, and as a result Y-Phosphatome managed to retrieve all experimentally validated PTPs from organisms representing a diversity of phylogenetic lineages. Thus the high sensitivity and specificity of our tool enables the genome-wide identification and class-level classification of PTPs in a reproducible and automatic manner.

3.2 Reannotation of the human tyrosine phosphatome

Y-Phosphatome was used to scan the predicted proteins of the human genome for PTPs (Ensembl release 67) and found 4 new genes encoding PTPs that were not described in the Alonso et al paper of the human tyrosine phosphatome [14]. These include the DSPs DUPD1 (ENSG00000188716), PTPMT1 (ENSG00000110536), TNS3 (ENSG00000136205) and the tyrosine-specific phosphatase KIAA1274 (PALD1) (ENSG00000107719). Although these genes were not reported as part of the original human tyrosine phosphatome [14], all four have been subsequently discovered and studied independently in recent years. For example, DUPD1 was initially identified as a new phosphatase binding to the intracellular domain of the short isoform of the prolactin receptor and is thought to dephosphorylate MAPK, a protein kinase essential for normal follicular development [38]. PTPMT1 was originally found to locate exclusively to the inner membrane of mitochondria, and its knockdown in a pancreatic cell line was shown to alter the mitochondria's phosphoprotein profile as well as to up-regulate insulin secretion and ATP production [39, 40]. Besides being a potential drug target in type II diabetes, PTPMT1 has recently been found to have an essential role in the biogenesis of cardiolipin (a phospholipid of the inner mitochondrial membrane required for mitochondrial metabolism), and possibly in cardiolipin deficiency diseases too [41, 42]. Tensin-3 (TNS3) is a negative regulator of cell migration that was initially identified in the epidermal growth factor signaling cascade [43]. The genetic deletion of Tensin-3 showed that the gene is essential for the correct development and function of the small intestine, lung and bone [44]. Recently, the tyrosine kinase Src was found to phosphorylate the SH2 domain of Tensin-3 with the effect of promoting the gene's oncogenic function [45]. Finally, the important regulatory phosphatase KIAA1274 (also called PALD1) was initially identified as a negative regulator of insulin signaling in a genome-wide functional screen [46], and its mouse ortholog (Paladin) was identified in another screen for angiogenesis regulators. Paladin is dynamically expressed throughout the vasculature, mainly in endothelial cells in embryonic stages, and in arterial smooth muscle cells in the adult [47]. In summary, the updated count of the human tyrosine phosphatome includes 109 PTP genes (20 receptor PTPs, 20 non-receptor PTPs, 61 DSPs, 1 LMWP, 3 CDC25 homologues, and 4 EyA homologues, Supplementary Table 1) encoding 537 distinct PTP protein sequences (Supplementary Table 4). Although these four new PTP genes were characterised by other researchers following the original publication of the human tyrosine phosphatome [14], their identification proves the usefulness of Y-Phosphatome for the genome-wide characterization of phosphatomes.

3.3 The PTP complements of 65 eukaryotic genomes

The perfect coverage and classification rates reported in the tests on the Y-Phosphatome library indicate that the accurate identification of tyrosine phosphatomes from genomic data sets is feasible. We scanned 65 eukaryotic genomes for PTPs, including species belonging to four of the five eukaryotic supergroups [26]. These are unikonts (33 species), excavates (6 species), plants (10 species) and chromalveolates (16 species), which collectively display a vast range of genome sizes and environmental adaptations, including parasites of great importance such as *Trypanosoma* and *Leishmania* species. If we accept the classification as presented here, our analysis predicted a total of 4605 PTP sequences in 65 eukaryotic genomes classified into the PTP, DSP, LMWP, CDC25 and EyA classes. The genomes analyzed harbor somewhere between 1 PTP enzyme, as is the case of the intracellular parasite *Encaphalitozoon cuniculi* (which also has an extremely economical kinome [48]), to 537 PTPs (human). Given the comparatively large number of unikont genomes, it is not surprising that these species dominate the database in terms of PTP sequences (3916 or 85%), followed by plants (339 or 7.4%), chromalveolates (183 or 4.0%) and excavates (167 or 3.6%). Irrespective of the phylogenetic origin of the sequences, PTP-central is mostly populated by tyrosine-specific PTPs (2258 or 49%) and DSPs (1998 or 43.4%), followed by EyAs (139 or 3.0%), CDC25s (119 or 2.6%) and LMWPs (91 or 2.0%) (Figure 2A). For most species a linear correlation exists between the number of proteins encoded by a genome and its tyrosine phosphatome. However, this correlation does not appear to hold true for the plant genomes whose predicted tyrosine phosphatome sizes remain nearly identical despite large increases in genome sizes (Figure 2B). PTPs and DSPs are the only enzyme classes that are universally present in all eukaryotic supergroups. CDC25 enzymes are reported in all eukaryotic supergroups (but not in all genomes), and are particularly prominent in mammals maybe because of their complex mechanisms of cell cycle regulation. CDC25 enzymes appear to be absent from most excavates, plants and chromalveolates, but not all, suggesting alternative cell cycle regulation strategies (for instance, as previously reported in the Apicomplexa [49, 50]). Likewise, LMWP enzymes are most prominent in mammalian genomes and discretely distributed among all other genomes from all four eukaryotic supergroups, with notable losses as in the Apicomplexa, none of which appear to harbor any LMWPs. Finally, Eye Absent (EyA) homologs are only predicted in true-tissue metazoans and streptophyte plants (1-3 members per genome), and are absent from all other species surveyed (Figure 2C and Supplementary Table 4).

<Figure 2>

3.4 PTP-central: contents and functionality

Our method and predictions are accessible online via the PTP-central resource (<http://www.PTP-central.org/>). PTP-central harbors the precomputed tyrosine phosphatomes of the 65 eukaryotic genomes (n=4605 proteins) (Figure 2A, 2B and 2C, and Supplementary Table 4), as well as the PTP structures retrieved from the PDB (Figure 2D), the majority of which are tyrosine-specific

phosphatases of human origin; 138 distinct positive disease associations drawn from the Genetic Association Database (GAD) [27], and all the homology relationships detected among PTPs across the 65 eukaryotes.

The PTP-central website is a user-friendly and comprehensive resource, featuring essential background information on PTPs and their functions, PTP-related diseases, and the various structural classes as well as the implications for inhibitor design. The database is particularly easy to navigate, as illustrated in the example provided in Figure 3: at the top are displayed the distinct search options of the database (“PHOSPHATOMES”, “DISEASES”, “PTP STRUCTURES” and “PTP HOMOLOGS”), as well as a “PEPTIDE SCAN” interface where users can submit their protein sequences for prediction. As an exercise, let’s say that we wish to investigate the dual-specificity phosphatases that are common to both human and the Placozoan *Trichoplax adhaerens* (likely the simplest extant multicellular metazoan). Ultimately this comparison should be informative about the DSPs that are present in both organisms and which therefore are likely to represent an early core of animal dual-specificity phosphatases. In the “SEARCH DB -> PHOSPHATOMES” interface we select the DSPs present in human and the placozoan (Figure 3A), resulting in the tabulated output shown in Figure 3B: besides the peptide and gene IDs, their classification into one of the PTP classes, gene names and description, the database generates a protein architecture mini-plot (which can be enlarged by clicking on it) that provides a quick way of visually inspecting various protein architectures. The gene names include those provided in the original Ensembl annotation, as well as all the aliases available in the Entrez Gene database [51]. The inclusion of gene aliases makes the keyword-based search for PTPs far more comprehensive and flexible (Figure 3A). Finally, for human genes PTP-central provides a link to the corresponding entry in the OMIM (Online Mendelian Inheritance in Man) database, the main repository of genetic information on Mendelian genetic disorders. OMIM provides extensive expert-based historical annotation on each gene’s identification, functions and involvement in specific diseases [52]. Although the style of the entries in the OMIM database is rather narrative, the entries are typically of very high quality and significance. The user is given the option to download the datasets, either in the “text-only” version (fully annotated protein sequences in Fasta format), or also including the protein architecture plots, as well as in a “tab-separated” format (for uploading to Excel). The “Start Jalview” button launches a Java applet of the multiple sequence analysis tool Jalview [53]. Jalview allows the user to edit and color the sequences by conservation, protein secondary structural properties, or amino acid chemical characteristics and perform on-the-fly calculations of phylogenetic trees (Neighbor-Joining and average distance). The full Jalview application can be launched (“File->View in Full Application”) to access various multiple alignment algorithms (MAFFT, Muscle, ClustalW, T-Coffee and Probcons), secondary structure prediction methods (Jnet), and display structures with Jmol. Additionally, multiple alignments in Jalview can be exported to TOPALi v2 [54] via a synchronized interface where more sophisticated phylogenetic methods are readily available.

<Figure 3>

The “SEARCH DB -> DISEASES” interface features the positive disease associations with human PTPs that are found in the Genetic Association Database (GAD) [27]. We found 138 positive associations between 26 PTP genes (of the PTP and DSP classes) and 14 different types of disease, which suggests that PTP genes are often associated with more than one disease type. The most frequent type of disease is immunological (>50%), followed by metabolic disorders (~14%) (Table 2). PTPN22 is by far the PTP gene associated with the largest number of conditions (54 diseases of the immune, infectious and metabolic types), followed by PTPN11 (14 diseases) and PTPN1 (11 diseases) (Table 3). PTP-central provides links between each PTP gene and its associated diseases to the original publications and the GAD, OMIM and Ensembl databases.

4. Conclusions

The main conclusions of our work are:

- (i) We developed a highly sensitive and specific classification method (‘Y-Phosphatomer’) for the automatic classification of protein sequences into the various PTP classes. Our method relies on a specific collection of protein domain models drawn from five distinct InterPro member databases. Upon evaluation, Y-Phosphatomer reported perfect coverage and classification rates on curated sets of PTPs.
- (ii) As proof of principle we re-annotated the human tyrosine phosphatome, and discovered four new PTP genes that had not been reported in an early important publication of the human PTP repertoire [14].
- (iii) Y-Phosphatomer was used to scan 65 eukaryotic genomes and we report that the sizes of tyrosine phosphatomes vary extensively, and that only tyrosine-specific and dual-specificity phosphatases are universally found in all eukaryotic supergroups.
- (iv) Our method Y-Phosphatomer and the predicted tyrosine phosphatomes are accessible online through the PTP-central resource (<http://www.PTP-central.org/>), which also integrates PTP structural information from the Protein Data Bank, as well as genetic association studies from the Genetic Association Database (GAD) and pre-calculated homology relationships among the 65 tyrosine phosphatomes.
- (v) PTP-central is a comprehensive and continually developing resource that integrates a diversity of data sets essential for the study of PTPs, both in model organisms as well as from an evolutionary perspective. Future plans for PTP-central include the integration of “omics” data sets from various model organisms, the addition of new genome-wide association studies as well as disease-specific data, and the implementation of algorithms for investigating the contents of the resource from multiple angles.

Figure legends

Figure 1. Flow diagram of the Y-Phosphatomer method for the automatic classification of PTPs.

The human set of PTPs was analyzed with a local installation of InterProScan run with default parameters (steps 1 and 2). This analysis determined that the PTPs from the four classes could be unequivocally distinguished by a specific combination of protein domain models, mostly in the form of HMMs (step 3 and 4). We took advantage of this property to build the Y-Phosphatomer library consisting of only 14 protein models (step 5). The Y-Phosphatomer library was evaluated on two data sets and uniformly reported perfect coverage, and a mis-classification rate of zero on the PTP class level (step 6). Therefore, Y-Phosphatomer is a robust library that can be applied to the genome-wide annotation of PTPs (step 7).

Figure 2. Overview of the contents of PTP-central. (A) In the 65 eukaryotic genomes analyzed, nearly 50% of the 4605 PTP sequences are tyrosine-specific phosphatases, closely followed by dual-specificity phosphatases (~43%). LMWP/CDC25/EyA class phosphatases comprise less than 3% each of the entire dataset. (B) For most species a linear correlation exists between a species' tyrosine phosphatome and the total number of proteins encoded in the genome. Plants are the exception, where the sizes of tyrosine phosphatomes remain relatively constant despite large increases in genome sizes. (C) Only tyrosine-specific and dual-specificity phosphatases are universally present in all eukaryotic supergroups surveyed. (D) PTP-central contains a manually curated data set of 339 structures, the vast majority of which are human tyrosine-specific phosphatases from crystallographic studies.

Figure 3. Snapshot of the PTP-central web interface. PTP-central features a user-friendly and flexible interface that allows the user to search (i) the predicted PTPs of 65 eukaryotic genomes (in any combination of PTP class and species), (ii) the genetic diseases associated with human PTPs as reported in the Genetic Association Database (GAD) of the NIH, (iii) a curated set of PTP structures drawn from the PDB, and (iv) the homology relationships among the PTPs of all 65 genomes. Additionally, the user can submit protein sequences for scanning with the Y-Phosphatomer library.

Tables

Table 1. The Y-Phosphatome classification of experimentally characterized PTP sequences retrieved from the UniProt database.

PTP Class	Group	Proteins	Group-level classification accuracy on the data set covered (%)
Class I	PTP	66	66 (100%)
	DSP	13	13 (100%)
Class II	LMWP	3	3 (100%)
Class III	CDC25	0	N/A
Asp-based PTPs	EYA	14	14 (100%)
Total		96	96 (100%)

Table 2. Disease types associated with 26 distinct human PTP genes as reported in the Genetic Association Database (GAD) of the NIH. The most frequent type of disease associated with human PTPs is immunological, and many PTPs are associated with more than one disease type.

Type of disease	Representation among PTPs with a positive disease association
Immune	52.9%
Metabolic	13.8%
Cancer	6.5%
Developmental	5.8%
Cardiovascular Neurological Psychiatric Chem. Dependency Infection Reproduction Hematological Vision Aging Renal	All below 3%

Table 3. The 26 human PTP genes and their associated diseases as reported in the Genetic Association Database (GAD) of the NIH and available through PTP-central.

Gene	PTP Class	Number of diseases	Disease types
PTPN1	PTP	11	Cardiovascular / Immune / Metabolic
PTPN2	PTP	10	Immune
PTPN9	PTP	1	Developmental
PTPN11	PTP	14	Cancer / Cardiovascular / Developmental / Hematological / Immune / Metabolic / Other
PTPN12	PTP	1	Vision
PTPN22	PTP	54	Immune / Infection / Metabolic
PTPRC	PTP	7	Immune / Infection / Other
PTPRD	PTP	6	Chem. Dependency / Metabolic / Neurological / Psychiatric
PTPRE	PTP	1	Other
PTPRF	PTP	1	Metabolic
PTPRG	PTP	2	Cardiovascular / Psychiatric
PTPRJ	PTP	1	Cancer
PTPRK	PTP	1	Immune
PTPRT	PTP	1	Chem. Dependency
TNS1	PTP	1	Cardiovascular
DUSP6	DSP	1	Aging
DUSP12	DSP	1	Metabolic
DUSP13	DSP	1	Renal
MTM1	DSP	1	Other
MTMR2	DSP	1	Other
MTMR3	DSP	1	Immune
MTMR11	DSP	1	Developmental
PTEN	DSP	7	Cancer / Chem. Dependency / Metabolic
ACPI	LMWP	10	Developmental / Immune / Metabolic / Psychiatric / Reproduction
EYA1	EYA	1	Neurological
EYA4	EYA	1	Other

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science (JSPS) through the WPI-IFReC Research Program and a Kakenhi grant; the Kishimoto Foundation; the ETHZ-JST Japanese-Swiss Cooperative Program (to DMS); and a Jeanne and Jean-Louis Lévesque Chair in Cancer Research and a Canadian Cancer Society Research Institute operating grant #700922 (to MLT).

We would like to thank Ms. Mineko Tanimoto for secretarial support.

References

- [1] P. Cohen, *Nat Cell Biol*, 4 (2002) E127-130.
- [2] P. Cohen, *Trends Biochem Sci*, 25 (2000) 596-601.
- [3] E.H. Fischer, E.G. Krebs, *J Biol Chem*, 216 (1955) 121-132.
- [4] E.G. Krebs, E.H. Fischer, *Biochim Biophys Acta*, 20 (1956) 150-157.
- [5] M.S. Collett, R.L. Erikson, *Proc Natl Acad Sci U S A*, 75 (1978) 2021-2024.
- [6] T. Hunter, B.M. Sefton, *Proc Natl Acad Sci U S A*, 77 (1980) 1311-1315.
- [7] T.S. Ingebritsen, P. Cohen, *Eur J Biochem*, 132 (1983) 255-261.
- [8] N.K. Tonks, C.D. Diltz, E.H. Fischer, *J Biol Chem*, 263 (1988) 6722-6730.
- [9] W.J. Hendriks, A. Elson, S. Harroch, R. Pulido, A. Stoker, J. den Hertog, *FEBS J*, (2012).
- [10] T. Hunter, G.D. Plowman, *Trends Biochem Sci*, 22 (1997) 18-22.
- [11] J.N. Andersen, O.H. Mortensen, G.H. Peters, P.G. Drake, L.F. Iversen, O.H. Olsen, P.G. Jansen, H.S. Andersen, N.K. Tonks, N.P. Moller, *Mol Cell Biol*, 21 (2001) 7117-7136.
- [12] J.N. Andersen, P.G. Jansen, S.M. Echwald, O.H. Mortensen, T. Fukada, R. Del Vecchio, N.K. Tonks, N.P. Moller, *FASEB J*, 18 (2004) 8-30.
- [13] J.N. Andersen, R.L. Del Vecchio, N. Kannan, J. Gergel, A.F. Neuwald, N.K. Tonks, *Methods*, 35 (2005) 90-114.
- [14] A. Alonso, J. Sasin, N. Bottini, I. Friedberg, A. Osterman, A. Godzik, T. Hunter, J. Dixon, T. Mustelin, *Cell*, 117 (2004) 699-711.
- [15] N.K. Tonks, *FEBS J*, (2012).
- [16] A.R. Forrest, D.F. Taylor, J.L. Fink, M.M. Gongora, C. Flegg, R.D. Teasdale, H. Suzuki, M. Kanamori, C. Kai, Y. Hayashizaki, S.M. Grimmond, *BMC Bioinformatics*, 7 (2006) 82.
- [17] S. Liberti, F. Sacco, A. Calderone, L. Perfetto, M. Iannuccelli, S. Panni, E. Santonico, A. Palma, A.P. Nardoza, L. Castagnoli, G. Cesareni, *FEBS J*, 280 (2013) 379-387.
- [18] X. Li, M. Wilmanns, J. Thornton, M. Kohn, *Sci Signal*, 6 (2013) rs10.
- [19] E.M. Zdobnov, R. Apweiler, *Bioinformatics*, 17 (2001) 847-848.
- [20] M. Punta, P.C. Coghill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn, *Nucleic Acids Res*, 40 (2012) D290-301.
- [21] T.K. Attwood, P. Bradley, D.R. Flower, A. Gaulton, N. Maudling, A.L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, C. Zygouri, *Nucleic Acids Res*, 31 (2003) 400-402.
- [22] I. Letunic, T. Doerks, P. Bork, *Nucleic Acids Res*, 40 (2012) D302-305.
- [23] D. Wilson, R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia, J. Gough, *Nucleic Acids Res*, 37 (2009) D380-386.
- [24] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T.K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coghill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R.D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Mueller, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A.F. Quinn, C. Rivoire, A. Sangrador-Vegas, J.D.

- Selengut, C.J. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajananathan, P.D. Thomas, C.H. Wu, C. Yeats, S.Y. Yong, *Nucleic Acids Res*, 40 (2012) D306-312.
- [25] E.C. Dimmer, R.P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M.J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M.C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy, I. Cusin, P. Duek-Roggli, I. Xenarios, R. Apweiler, *Nucleic Acids Res*, 40 (2012) D565-570.
- [26] P.J. Keeling, G. Burger, D.G. Durnford, B.F. Lang, R.W. Lee, R.E. Pearlman, A.J. Roger, M.W. Gray, *Trends Ecol Evol*, 20 (2005) 670-676.
- [27] K.G. Becker, K.C. Barnes, T.J. Bright, S.A. Wang, *Nat Genet*, 36 (2004) 431-432.
- [28] P.W. Rose, C. Bi, W.F. Bluhm, C.H. Christie, D. Dimitropoulos, S. Dutta, R.K. Green, D.S. Goodsell, A. Prlic, M. Quesada, G.B. Quinn, A.G. Ramos, J.D. Westbrook, J. Young, C. Zardecki, H.M. Berman, P.E. Bourne, *Nucleic Acids Res*, 41 (2013) D475-482.
- [29] L.P. Pryszcz, J. Huerta-Cepas, T. Gabaldon, *Nucleic Acids Res*, 39 (2011) e32.
- [30] J. Huerta-Cepas, S. Capella-Gutierrez, L.P. Pryszcz, I. Denisov, D. Kormes, M. Marcet-Houben, T. Gabaldon, *Nucleic Acids Res*, 39 (2011) D556-560.
- [31] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A.K. Kahari, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H.S. Riat, G.R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y.A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S.P. Wilder, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernandez-Suarez, J. Harrow, J. Herrero, T.J. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, S.M. Searle, *Nucleic Acids Res*, 40 (2012) D84-90.
- [32] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L.J. Jensen, C. von Mering, P. Bork, *Nucleic Acids Res*, 40 (2012) D284-289.
- [33] S. Penel, A.M. Arigon, J.F. Dufayard, A.S. Sertier, V. Daubin, L. Duret, M. Gouy, G. Perriere, *BMC Bioinformatics*, 10 Suppl 6 (2009) S3.
- [34] J. Ruan, H. Li, Z. Chen, A. Coghlan, L.J. Coin, Y. Guo, J.K. Heriche, Y. Hu, K. Kristiansen, R. Li, T. Liu, A. Moses, J. Qin, S. Vang, A.J. Vilella, A. Ureta-Vidal, L. Bolund, J. Wang, R. Durbin, *Nucleic Acids Res*, 36 (2008) D735-740.
- [35] D. Miranda-Saavedra, G.J. Barton, *Proteins*, 68 (2007) 893-914.
- [36] D.M. Martin, D. Miranda-Saavedra, G.J. Barton, *Nucleic Acids Res*, 37 (2009) D244-250.
- [37] A.P. Hutchins, S. Liu, D. Diez, D. Miranda-Saavedra, *Mol Biol Evol*, (2013).
- [38] Y.S. Devi, A.M. Seibold, A. Shehu, E. Maizels, J. Halperin, J. Le, N. Binart, L. Bao, G. Gibori, *J Biol Chem*, 286 (2011) 7609-7618.

- [39] D.J. Pagliarini, S.E. Wiley, M.E. Kimple, J.R. Dixon, P. Kelly, C.A. Worby, P.J. Casey, J.E. Dixon, *Mol Cell*, 19 (2005) 197-207.
- [40] Y. Boisclair, M.L. Tremblay, *Mol Cell*, 19 (2005) 291-292.
- [41] J. Zhang, Z. Guan, A.N. Murphy, S.E. Wiley, G.A. Perkins, C.A. Worby, J.L. Engel, P. Heacock, O.K. Nguyen, J.H. Wang, C.R. Raetz, W. Dowhan, J.E. Dixon, *Cell Metab*, 13 (2011) 690-700.
- [42] K. El-Kouhen, M.L. Tremblay, *Cell Metab*, 13 (2011) 615-617.
- [43] Y. Cui, Y.C. Liao, S.H. Lo, *Mol Cancer Res*, 2 (2004) 225-232.
- [44] M.K. Chiang, Y.C. Liao, Y. Kuwabara, S.H. Lo, *Dev Biol*, 279 (2005) 368-377.
- [45] X. Qian, G. Li, W.C. Vass, A. Papageorge, R.C. Walker, L. Asnaghi, P.J. Steinbach, G. Tosato, K. Hunter, D.R. Lowy, *Cancer Cell*, 16 (2009) 246-258.
- [46] S.M. Huang, M.K. Hancock, J.L. Pitman, A.P. Orth, N. Gekakis, *PLoS One*, 4 (2009) e6871.
- [47] E. Wallgard, A. Nitzsche, J. Larsson, X. Guo, L.C. Dieterich, A. Dimberg, T. Olofsson, F.C. Ponten, T. Makinen, M. Kalen, M. Hellstrom, *Dev Dyn*, 241 (2012) 770-786.
- [48] D. Miranda-Saavedra, M.J. Stark, J.C. Packer, C.P. Vivares, C. Doerig, G.J. Barton, *BMC Genomics*, 8 (2007) 309.
- [49] L. Reininger, J.M. Wilkes, H. Bourgade, D. Miranda-Saavedra, C. Doerig, *Mol Microbiol*, 79 (2011) 205-221.
- [50] D. Miranda-Saavedra, T. Gabaldon, G.J. Barton, G. Langsley, C. Doerig, *Microbes Infect*, 14 (2012) 796-810.
- [51] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, *Nucleic Acids Res*, 39 (2011) D52-57.
- [52] A. Hamosh, A.F. Scott, J. Amberger, C. Bocchini, D. Valle, V.A. McKusick, *Nucleic Acids Res*, 30 (2002) 52-55.
- [53] A.M. Waterhouse, J.B. Procter, D.M. Martin, M. Clamp, G.J. Barton, *Bioinformatics*, 25 (2009) 1189-1191.
- [54] I. Milne, D. Lindner, M. Bayer, D. Husmeier, G. McGuire, D.F. Marshall, F. Wright, *Bioinformatics*, 25 (2009) 126-127.

Figure 1

Figure 2

A

B

C

UNIKONTS

	PTP	DSP	CDC25	LMWP	EyA
Amoebozoa	●	●	●	●	○
Choanozoa	●	●	●	○	○
Fungi	●	●	●	●	○
Metazoans (Placozoa)	●	●	●	●	○
Metazoans (True tissues)	●	●	●	●	●

EXCAVATES

Euglenozoa	●	●	○	●	○
Metamonada	●	●	●	●	○

PLANTS

Streptophytes	●	●	○	●	●
Green algae	●	●	●	●	○
Red algae	●	●	○	●	○

CHROMALVEOLATES

Alveolates	●	●	●	●	○
Apicomplexa	●	●	○	○	○

D

Figure 3