CHAPTER 18

CONSTRUCTION OF BIOLOGICAL DATABASES: A CASE STUDY ON THE PROTEIN PHOSPHATASE DATABASE (PPDB)

Prasanna R. Kolatkar

Genome Institute of Singapore
kolatkarp@gis.a-star.edu.sg

Kui Lin

Beijing Normal University
linkui@bnu.edu.cn

Biological data is being created at ever-increasing rates as different high-throughput technologies are implemented for a wide variety of discovery platforms. It is crucial for researchers to be able to not only access this information but also to integrate it well and synthesize new holistic ideas about various topics. A key ingredient in this process of data-driven knowledge-based discovery is the availability of databases that are user-friendly, that contain integrated information, and that are efficient at storage and retrieval of data.

Implementations of integrated databases include GenBank,[77] SWISS-PROT,[41] InterPro,[29] PIR,[903] *etc*. No single one of these databases contains all the information one might need to understand a specific topic. So unique databases are required to provide researchers better access to specific information. Databases can be built using a variety of tools, techniques, and approaches;[416, 756, 897] but there are some methods that are extremely powerful for management of large amounts of data and that can also integrate this information.

We have created several purpose-built integrated databases. We describe one of them—the Protein Phosphatase DataBase (PPDB)—in this chapter. PPDB has been constructed using a data integration and analysis system called Kleisli. Kleisli can model complex biological data and their relationships, and integrate information from distributed and heterogeneous data resources.[162, 189, 894]

ORGANIZATION.

Section 1. The importance of protein phosphatases in signal transduction in eukaryotic cells is briefly explained. The lack of a well-integrated protein phosphatase database is noted, motivating the construction of PPDB.

Section 2. An overview of the architecture of PPDB and its underlying workflow is then given.

Section 3. This is followed by a brief introduction to the data integration tool and the object representation used in constructing PPDB.

Sections 4–9. The next several sections present in greater detail the types of information provided in PPDB for protein phosphatases. The types of information include protein and DNA sequences, structure, biological function related information, disease related information, and so on, as well as the classification tree of protein phosphatases.

Sections 10–15. After that, the details of data collection, integration, and update are described. The techniques and rules used in PPDB for the automatic classification of protein phosphatases into one of 8 categories are also discussed.

1. Biological Background

Most cellular processes are extremely complex and require tightly controlled signal transduction events. One of the fundamental mechanisms that a cell utilizes to control its biological processes is via protein phosphorylation and dephosphorylation reactions that are catalyzed by protein kinases and protein phosphatases.[834, 936] This type of reversible covalent modification of proteins is of crucial importance for modulation of the cell's biochemical pathways.

Protein kinases catalyze the transfer of g-phosphate of a nucleotide triphosphate—usually ATP—to an acceptor residue in the substrate protein, while protein phosphatases do the reverse job of removing the phosphate. It was originally thought that protein kinases were the key enzymes controlling the phosphorylation. In fact, the protein phosphatases are also composed of a large family of enzymes that parallel protein kinases in terms of structural diversity and complexity. It is now clear that the protein phosphatases and kinases play equally important roles in signal transduction in eukaryotic cells as the Yin and Yang of protein phosphorylation and cell signaling function.

Papers published over the last several years document crucial physiological roles for protein phosphatase in a variety of mammalian tissues and cells.[32, 200, 323] However, to date, there are no unique, well-integrated protein phosphatase databases available that a researcher can access to locate a large amount of related information for his favorite phosphatase. We thus have implemented the Protein Phosphatase DataBase (PPDB) not only to integrate related information concerning protein phosphatases but also to classify and organize the data as hierarchical classes based on the specified biochemical, genetic, and biological characteristics. We believe that such data organization is helpful for researchers.

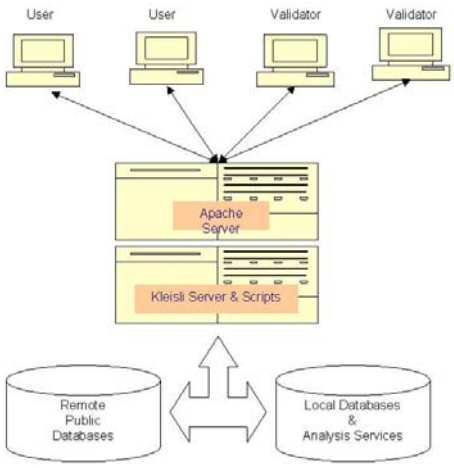


Fig. 1. The 3-tier architecture of PPDB.

2. Overview of Architecture and Workflow of PPDB

In PPDB ver. 1.0, we focus only on human phosphatases. An overview of the 3-tier architecture of the database is shown in Figure 1. Users can access the database via the Internet by using their browsers. Figure 8 gives the general workflow of how the PPDB is created using the Kleisli data integration and analysis system.[162, 189, 894]

At the time of writing (October 2001), there are 70 phosphatases, which are clustered into the 8 categories shown in Figure 2. Currently, we mostly focus on the classification of tyrosine phosphatases, which function Of this type of phosphatases, 4 subcategories—tyrosine specific phosphatase, dual specific phosphatase, low molecular weight phosphatase, and anti phosphatase—are devised to classify them. The tyrosine specific phopahatase subclass is in turn divided into 2 subclasses, termed receptor-like group[458] and cytosolic phosphatase.[200, 871] At the same time, we also provide two additional subcategories, “non-protein” and “not-sure” for storing those phosphatases that are not protein phosphatases or currently have unknown function.

In the following sections, we describe individual data types, methods for data collection and integration, automatic phosphatase classification schema, expertise

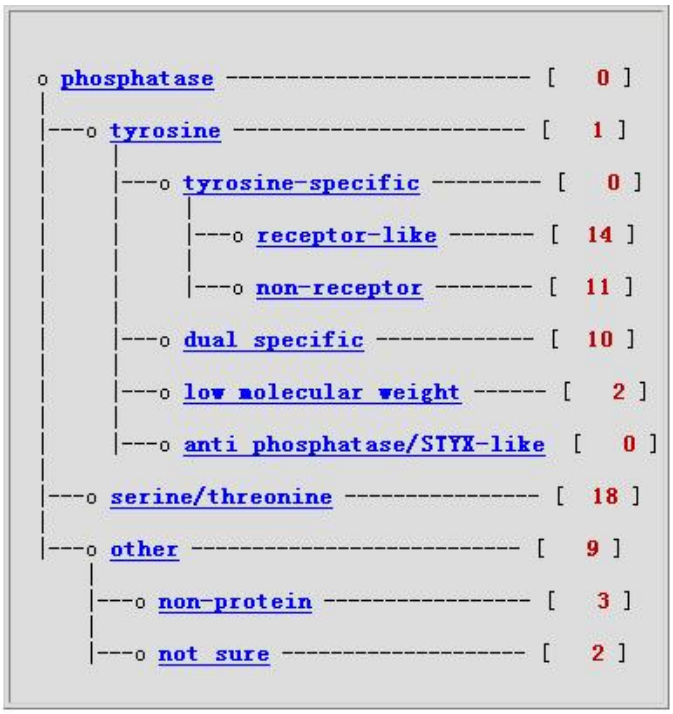


Fig. 2. The classification and organization tree for phosphatases in PPDB. The number of entries in various subcategories is given in brackets.

for validation, data storage, content update, and visualization in detail.

3. Data Integration Tool and Object Representation

PPDB has been implemented using a powerful data integration and analysis system called Kleisli. Kleisli can facilitate efficient building of a prototype for complex biological data objects and their relationships, as well as integrating information from distributed and heterogeneous data resources.[162,189,894]

There are 3 types of object containers in Kleisli,[894] *viz.* set, bag, and list. Objects in a container can be either homogeneous or heterogeneous. In the latter, objects are tagged by different labels using variant data type. Record objects are the main data type to describe entities in the real world. Similar to the general records in relational databases, the records consist of several labeled fields as well. However, each field of a record can contain any type of data within Kleisli—for instance, it can be a bag, a list, a set, a record, *etc.* Hence, the type of data modeling

facilitated by Kleisli is more powerful than other methods.[897]

Kleisli is equiped with many relevant built-in operators and functions. In particular, many existing bioinformatics tools, algorithms, and databases are integrated into Kleisli in a transparent way.[162, 894] Kleisli is hence not only a powerful data integration system but also a data analysis system. It is highly suitable for addressing bioinformatics problems.[146–148, 162, 189, 509, 895, 896]

In PPDB, each phosphatase object is presented as a complex data object—*e.g.*, Figures 3 and 4—in the data exchange format of Kleisli.[895] All the same type of objects are usually organized into a set or a list of Kleisli complex objects. Most of the objects are presented as Kleisli records in a set or a list. However, the fields of a record can also be any of types of Kleisli objects—*i.e.*, basic types such as string and number; or complex types such as record, set, or list. Thus, data objects are allowed to be deeply nested to reflect the natural complex structure of real-life biological data.

For example, the field `#organism` of a protein sequence in Figure 3, which is a complex record, includes the field `#lineage`, which is a list of taxonomy. Another object is the classification tree object in Figure 4, which is a recursive record to represent a hierarchical category data structure. In PPDB, the core data object is the tree object by which all of the other data objects are connected together, such as protein and DNA sequences, literature abstracts, related information from OMIM,[318] ENZYME,[40] and PDB[880] databases.

```
(#uid: 4096844,
 #title: "protein tyrosine phosphatase Cr1PTPase precursor",
 #accession: "4096844",
 #common: "human",
 #organism: (#genus: "Homo", #species: "sapiens",
             #lineage: ["Eukaryota", "Metazoa", ...]),
 #feature: {
   (#name: "Protein", #start: 0, #end: 656,
    #anno: [ (#anno_name: "product",
              #descr: "protein tyrosine phosphatase ...")]),
   (#name: "CDS", #start: 0, #end: 656,
    #anno: [ (#anno_name: "gene",
              #descr: "Ch-1PTPase alpha"),
             (#anno_name: "note",
              #descr: "receptor-type protein tyrosine ..."),
             ...]), ...}, ...)
```

Fig. 3. A snapshot of phosphatase Cr1PTPase curated in PPDB in the Kleisli data exchange format. This figure only shows part of the whole protein information.

```
(#name: "phosphatase",
 #classes: [
    (#name: "tyrosine",
     #classes: [
        (#name: "tyrosine-specific",
         #classes: [
            (#name: "receptor-like", #classes:  []),
            (#name: "non-receptor", #classes:  []) ]),
        (#name: "dual specific", #classes:  []),
        (#name: "low molecular weight", #classes: []),
        (#name: "anti phosphatase/STYX-like",#classes:[])]),
    (#name: "serine/threonine", #classes: []),
    (#name: "other",
     #classes: [
        (#name: "non-protein", #classes: []),
        (#name: "not sure", #classes: []) ]) ])
```

Fig. 4. The classification tree object in Kleisli data exchange format.

4. Protein and DNA Sequences

Figure 3 shows a Kleisli complex data object of the phosphatase protein called "Cr1PTPase precursor" within the PPDB. The object is transformed from NCBI's GenPept[78] database of translated protein sequences from GenBank. We also have curated related DNA sequence information from GenBank,[77] and each DNA sequence object has a similar data structure like the protein object. A sequence object is a record, and consists of several fields. Some of them are simple and basic data types, such as unique identifier (number), title (string), and the amino acid sequence (long string). Some are complex data types, for example, the field `#feature` is a set of records of annotated information which is the most important part in any type of sequence database. Researchers usually extract useful information from this field for each protein, DNA, EST, SNP, *etc*. For example, we use the annotated information in this field to extract the gene name for each phosphatase-related DNA sequence; see Figure 5.

5. Structure

When a new phosphatase is validated and curated into PPDB, the system automatically parses the potentially related Protein Data Bank (PDB)[475, 880] identifier for the phosphatase. If there is a related entry found, the system downloads and

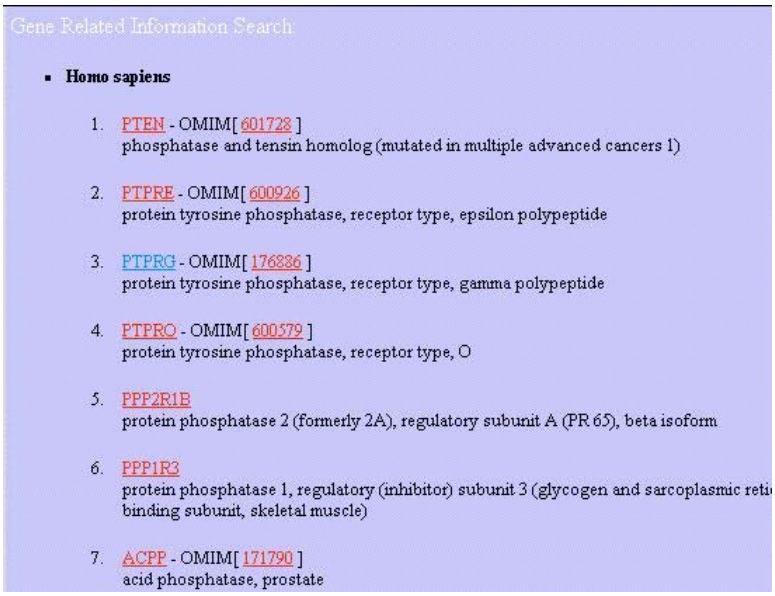


Fig. 5. A view of the list of gene names in PPDB.

integrates that part of the information from PDB for the respective phosphatase. The information is represented and visualized in HTML format. Figure 6 shows the visualization of 3D structure information that is extracted from PDB for the phosphatase PTPmu/RPTPmu[360] in PPDB.

6. Biological Function and Related Information

In addition to the classification tree, the most interesting and important information in PPDB is the related biological function information linked to each phosphatase. This includes structural features, gene expression properties, catalytic properties, substrates, interacting proteins, cell effects, disease linkage, and knockout/transgenics properties, *etc.* Figure 7 lists the related biological, biochemical, and genetic information of the phosphatase PTPmu/RPTPmu in PPDB. This information is retrieved by the phosphatase expert and integrated into PPDB by the subsystem called InfoCollection which is validated by the expert via a user-friendly web interface.

PDB: 1RPM

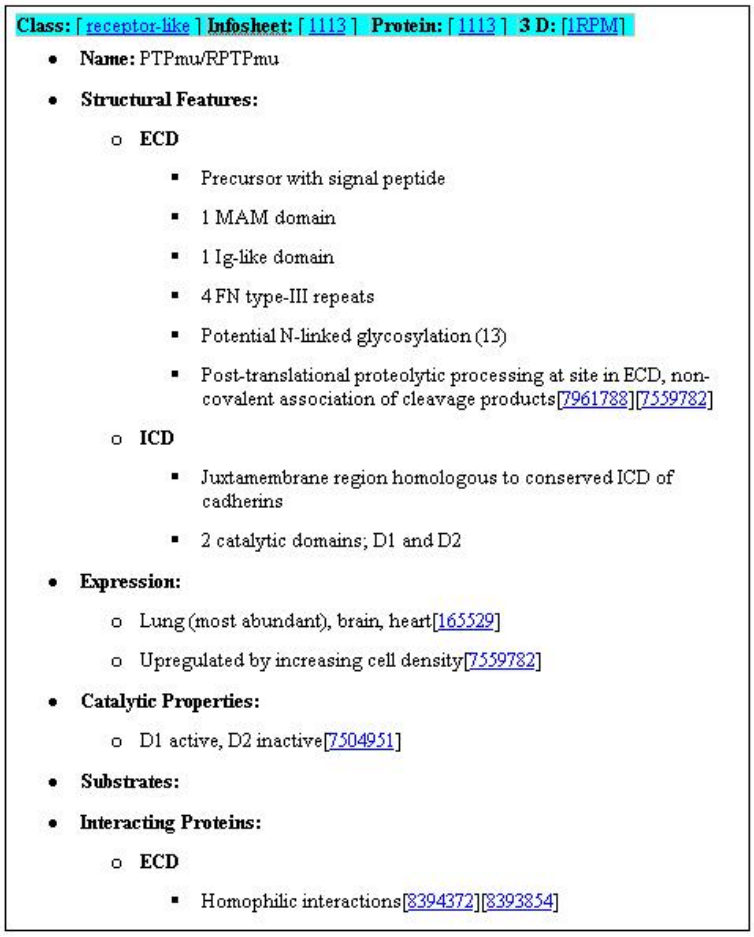
- **Compound:**
  1. **Molecule:** Receptor Protein Tyrosine Phosphatase Mu
     - **Domain:**
     - **Fragment:** Cytosolic Membrane Proximal Catalytic Domain
- **Classification:** Receptor
- **Resolution:** 2.3[Å]
- **R value:** 0.195
- **Space group:** P 1 21 1
- **Atoms:** 4838
- **Unit cell:**

a[Å]	b[Å]	c[Å]	alpha[¯]	beta[¯]	gamma[¯]
95.54	36.29	95.38	90.0	95.66	90.0

Fig. 6. The 3D information of phosphatase PTPmu/RPTPmu.

7. Other Data Objects

There are other data objects integrated into PPDB, such as phosphatase related gene information, associated MEDLINE abstracts, and relevant disease information with links to the OMIM database.[318] These diverse types of information are used to classify phosphatases automatically according to pre-defined classification rules. The hotlinks to the corresponding databases are given in the presentation of query results, *e.g.*, Figure 5.



Class: [receptor-like] **Infosheet:** [1113] **Protein:** [1113] **3 D:** [1RPM]

- **Name:** PTPmu/RPTPmu
- **Structural Features:**
  - **ECD**
    - Precursor with signal peptide
    - 1 MAM domain
    - 1 Ig-like domain
    - 4 FN type-III repeats
    - Potential N-linked glycosylation (13)
    - Post-translational proteolytic processing at site in ECD, non-covalent association of cleavage products[7961788][7559782]
  - **ICD**
    - Juxtamembrane region homologous to conserved ICD of cadherins
    - 2 catalytic domains; D1 and D2
- **Expression:**
  - Lung (most abundant), brain, heart[165529]
  - Upregulated by increasing cell density[7559782]
- **Catalytic Properties:**
  - D1 active, D2 inactive[7504951]
- **Substrates:**
- **Interacting Proteins:**
  - **ECD**
    - Homophilic interactions[8394372][8393854]

Fig. 7. The biological function information of phosphatase PTPmu/RPTPmu.

8. Classification Tree

The most important Kleisli data object in PPDB is the classification tree depicted in Figure 12. It integrates all the related information in PPDB together. It is also used to explore and visualize the whole database by allowing access to all relevant information. In Kleisli, the classification tree is a hierarchical object, as shown in Figure 4. We curate the data by automatic classification of newly deposited and unclassified phosphatases by integrating different types of information linked to

the specified phosphatase, and displaying the related web links. The tree is used to visualize the query results as well.

9. Data Integration and Classification

We employ Kleisli to collect phosphatase information from several public protein databases, such as NCBI Entrez GenPept,[78,758] GenBank,[77] PubMed, SWISS-PROT,[41] PDB,[880] *etc.* Each newly collected phosphatase is classified automatically into an existing category of the classification tree according to its related evidence. The evidence considered include sequence, annotation features, related PubMed abstracts, and specific domain knowledge.

The automatically classified result and the supporting evidence are then displayed to allow phosphatase experts to validate the findings. All consensus phosphatases are integrated into PPDB with their related data, which includes genomic information, pathway information, related disease information along with associated mutation data, and protein function information.

10. Data Collection

Although more complex searching mehcanisms can be applied, we decide to implement simple keyword searching using, for example, "phosphatase" to search the NCBI GenPept database via the "Title" field. Subsequent to the search, information pertaining to all the hits is downloaded and saved as a local dataset. This local dataset can be updated automatically based on the classification results and the phosphatase entries in our database can be curated and validated.

In addition to protein sequence data, we also need to collect those DNA and PubMed items that are directly linked to the phosphatase proteins in our database. For example, for PPDB ver. 1.0, we download about 2,600 protein records from GenPept and save them as RawPhosphatases, Figure 8. We filter out those entries whose titles include words like "similar" or "predict." We therefore obtain a relatively smaller data set as the input for the automatic classification subsystem.

11. Phosphatase Classification Strategy

Phosphatases that utilize phosphoproteins as substrates have been divided into two major categories based on their substrate specificity:[834] protein-tyrosine phosphatases (PTPs) and protein serine/threonine phosphatases. PTPs can then be further subdivided into four subclasses. The PPDB categorizes the PTPs according to their biological functions. Figure 2 shows the current classification tree of phosphatases used in PPDB.

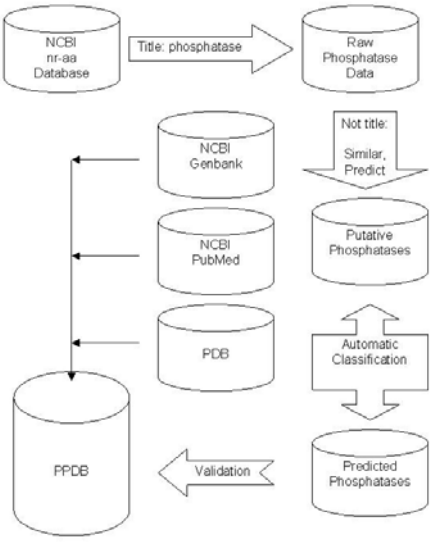


Fig. 8. The processes of data collection and integration of PPDB.

Currently, we use 3 simple approaches for automatically classifying the input entries into their specific subgroups in the classification tree. These approaches are:

- keyword searching,
- specified sequence motif matching, and
- prosite database searching

For example, we utilize the following methods to distinguish the PTPs:

(1) Most (if not all) of the Y-specific PTPs have the active site motif [VI]HCXAGXXR[TS].[201,806] Here "X" stands for any amino acid, and "[VI]" ("[TS]") stands for either the amino acid "V" ("T") or the amino acid "I" ("S").
(2) The Y-specific PTPs tend to have other regions of conservation in the catalytic domains near the N-terminus, which often has the sequence KNRY;[396] and further along the sequence are the residue stretches of DYINAS or XYINAX; and finally the WPD motif.
(3) If the protein sequence has a transmembrane region, it is classified as a receptor-like PTP (RPTP). To date, RPTPs are all Y-specific.

(4) If the protein sequence has two tandem catalytic domains, it is a RPTP (Y-specific).
(5) It is likely to be a true PTP if it has the minimal conserved active site motif CXXXXXR;[163] and unlikely to be a true PTP if the C or the R is substituted.
(6) The dual specificity PTPs have less well conserved sequence patterns throughout the catalytic domain and only the CXXXXXR motif is usually conserved.

12. Automatic Classification and Validation

According to the predefined classification rules of Section 11, we set the classification conditions for each subclass in the tree first and use the whole tree to discriminate each new input protein by evaluating the possibilities of all subclasses to which it may belong. Thus, we simply pick the most likely result as the predicted result, and publish the predicted subclass along with the whole tree in HTML format together with various pieces of evidence—*e.g.*, motifs along the amino acid sequence, keywords found, and so on—to allow phosphatase experts to validate the choice manually. These experts can put the protein into the correct class by just clicking a button on the web interface. The system then collects the validated result returned through the web interface and saves it as the trusted classification result.

13. Data Integration and Updates

After facilitating the validation, the system prompts the expert to supply specific biological information such as structural features, expression information, catalytic properties, substrates, interacting proteins, regulation, cell effects, disease linkages, crystallographic information, knockout/transgenics; see Figure 7. It subsequently combines all collected information in the database. In addition to this information, we also collect the related 3D structure information (Figure 10) from PDB,[880] disease linkage information (Figure 5) from OMIM,[318] and pathway information from ENZYME[40] (Figure 9 and 10).

To keep PPDB updated, we adopt three different strategies. The simplest strategy is to keep as many symbolic links to the specified databases as possible. For example, we use hotlinks for OMIM,[318] GenBank,[77] *etc*. The second strategy is to only re-create the webpages for the newly collected information of relevant phosphatases. That is, to only update existing phosphatases' with additional information, such as new literature links, instead of re-creating the whole database. The third strategy is the integration of new phosphates into PPDB in the future. This step is not trivial and we need to keep the internal unique identifiers for each phosphatase to preserve data integrity.

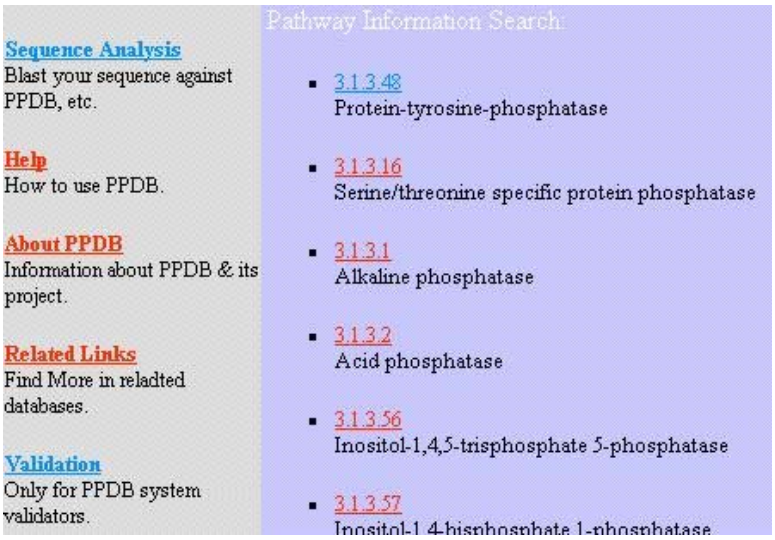


Fig. 9. A display of the list of EC numbers of enzymes in PPDB.

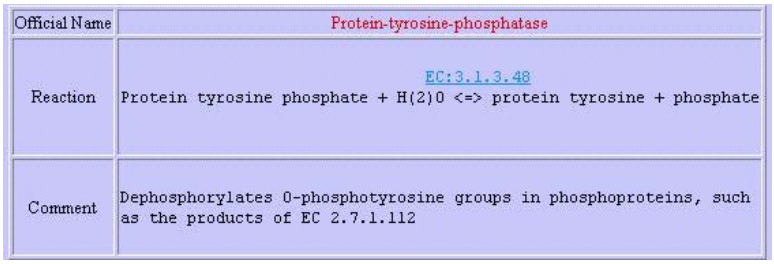


Fig. 10. The visualization of the protein tyrosine phosphatase pathway.

14. Data Publication and Internet Access

In PPDB ver. 1.0, there are 70 phosphatases, which are clustered into 8 categories, as shown in Figure 2. One phosphatase, however, belongs to two separate categories corresponding to different functions when it is found at different locations within a cell. We provide additional subcategories for storing phosphatases that currently have unknown function. As new phosphatase information is discovered and their biochemical features become available, we will classify them with into the correct groups in the classification tree.

Our classification system is flexible and we can easily add new subcategories into the existing node of the classification tree. Although we currently concentrate on tyrosine phosphatases, we will classify serine/threonine phosphatases into their functional subcategories as well.

Using the classification scheme for phosphatases, PPDB provides researchers more powerful and systematic searches. A vast amount of related information is displayed together, providing a comprehensive but compact interface allowing researchers to manage data and facilitate use of more detailed and specific information.

PPDB is accessible online through the web interface at `http://www.gis.a-star.edu.sg/PPDB`, depicted in Figure 11. There are several different ways to query the PPDB. The tree exploration allows users to step through the classification tree and easily find interesting specific information. Users can click on any category and obtain information from for all members of that specific phosphatase category. Figure 12 demonstrates the result of tree exploration of the tyrosine subclass.

For general search functions, PPDB ver. 1.0 provides 3 different methods, *viz.* gene-based search, pathway-based search, and keyword-based search. These methods are described briefly below.

- Gene-based search. PPDB provides a list of the existing gene names of the curated phosphatases. It allows users to pull out all phosphatases related to each gene. See Figure 5.
- Pathway-based search. PPDB gives out a list of known and curated enzymatic reactions. It allows users to explore the interesting reactions. See Figure 9.
- Keyword-based search. Users can input any interesting keywords to search the PPDB and find relevant information in the database. The search function of PPDB currently simply implements the title content of related data objects.

Like many genomic databases, PPDB also provides additional tools to allow users to do sequence search and analysis as well. For example, a user can submit a protein sequence to BLAST search[24] against the PPDB protein dataset. The system returns to the user the phosphatase homologs in their classified categories. PPDB also provides a localized ProfileScan 2.0 tool[361] for domain hunting to perform sequence domain analysis.

15. Future Development

PPDB curate phosphatases and their related information from diverse public resources, especially for structural, pathway, and functional information. We plan to

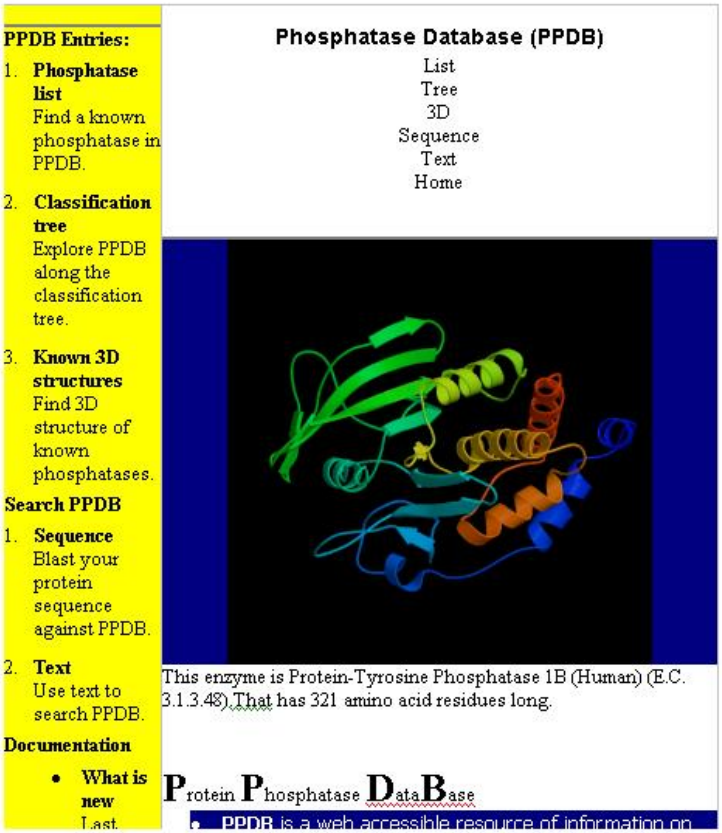
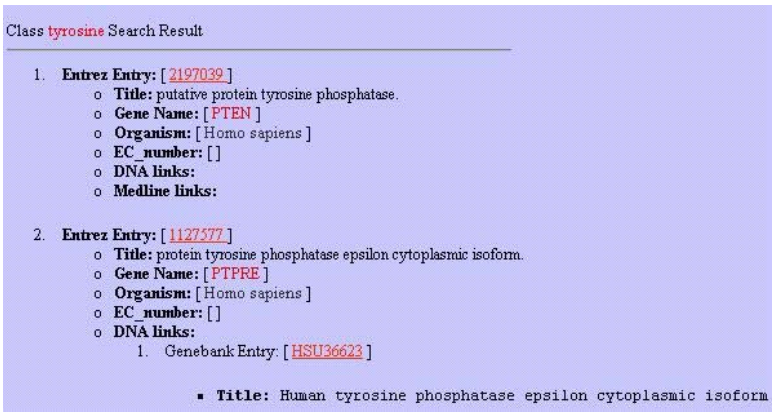


Fig. 11. A snapshot of the home page of PPDB.

include phosphatases from other important model species, such as *S. cerevisiae*, *E. coli*, *etc*. As the content increases, we also plan to do comprehensive domain comparisons within one species and cross species to better understand their biochemical functions in the cell.

Acknowledgement

We thanks Catherine J. Pallen and Kah-Leong Lim for helpful discussions and suggestions.


Class tyrosine Search Result

1. **Entrez Entry:** [2197039]
   - **Title:** putative protein tyrosine phosphatase.
   - **Gene Name:** [PTEN]
   - **Organism:** [Homo sapiens]
   - **EC_number:** [ ]
   - **DNA links:**
   - **Medline links:**

2. **Entrez Entry:** [1127577]
   - **Title:** protein tyrosine phosphatase epsilon cytoplasmic isoform.
   - **Gene Name:** [PTPRE]
   - **Organism:** [Homo sapiens]
   - **EC_number:** [ ]
   - **DNA links:**
     1. Genebank Entry: [HSU36623]

        ▪ **Title:** Human tyrosine phosphatase epsilon cytoplasmic isoform

Fig. 12. The result of tree exploration of tyrosine subclass.