# iProt - A Data Warehouse for Protein Database

M. I. Jaya, Z. Zainol and N.H. Ahamed Hassain Malim

*School of Computer Sciences, Universiti Sains Malaysia, 11800, Penang, Malaysia*

An integrated database is a vital approach to help biologist to analyze protein data from heterogeneous formats and resources. The iProt (Integrated Protein Data Warehouse) is a pool of protein dataset that provide a comprehensive protein sequence, 3D structure, enzymatic reaction, gene description, and taxonomical data from five different protein databases; Swiss-Prot, Protein Data Bank (PDB), ENZYME, NCBI Taxonomy and Gene Ontology (GO). The iProt database is design based on relational model which eliminate problems of data organization and data access experienced in the legacy, flat file method. More than 1,029,511 protein data from each database was combined into iProt's database using warehouse approach. Each data is grouped according to its function and each group is maintain in a respective table which linked by cross-reference entry from each databases. As a result, the iProt greatly shortened user search time and additionally improved data consistencies. A complete summary report including sequence features, protein function, enzymatic reaction and three-dimension structure display is provided via our user friendly web application. The implementation of this warehouse give biologist full control of protein data and extend the capabilities to mine protein of interest. Furthermore, it enabled biologist to fully analyze related protein data trough a single query. The iProt is implemented on a MySQL database in a local UNIX server.

## 1. Introduction

The development of complete bioinformatics database is crucial to help biologist fully store and explore valuable proteins data. Therefore, hundreds of newly biological databases were created and open to the community through out these years. Each was design for different purposed and supports different needs among biologist. As a result, biologist task has become very time consuming and complicated. The reason is heterogeneity of available data sources, format and size. Currently, more than a million protein data are available in the internet and these data is scattered according to its type. PDB itself contain more than 18,000 structures and their sizes grow exponentially every year(1). Searching task is more complicated now as a single answer to a biological problem may require access to many cross-referenced databases. Manual query to multiple databases can create bottlenecks, which then reduce the effectiveness and process speed. The new approach to overcome this problem is to integrate these databases into a single database, which then will highly reduce the queries overlap(2). As a result, new drugs discoveries process will be speed up. Interoperation among multiple databases in one local server is a vital salvation to the problems of searching difficulties and data heterogeneity(2). The main problem that happens today is the heterogeneity of proteins format and data locations. As consequences, data stored may be defined in multi-standard format and contained many error and data inconsistencies. To solve those problems, researchers need a local platform that contains most of data they used so that they are free to edit and filter the inconsistencies occurred. Moreover, they need a uniform format to store those data in a standard nomenclature that they can understand(3).

Today, to solve a complex and important problem, biologist need to manually scan different databases and relate each cross-referenced record before a complete knowledge is discovered. This is very time consuming to manually search and download the records. Biologist may need extra time to study different format and nomenclature for each database. Another problem is the usage of flat-file format in public databases. This format brings a lot of difficulties to bioinformatics scientist as they are written in a standard text format. The major challenge in bioinformatics is to facilitate biologist to view relationship among protein, structure, function and pathways in a single query. This requires a standard format for each database and automated cross-referencing among them. Warehouse approach solves this problem by storing multiple databases in a single local RDBMS server with standard format. A framework which contains multiple set of database that can help biologist to answer different problem in automated environment is proposed in this paper. This eliminates the used of flat file which is difficult and limited usage. In future, a lot of protein data analysis tool can be built on top of this project. More over, it may provide a platform that satisfied the better usage of proteins data. In this paper, we proposed the used of data warehouse concept by integrating PDB, SWISSPROT, NCBI Taxonomy, Gene Ontology (GO) and ENZYME database into our local warehouse using RDBMS platform. We also developed a web based application to be integrated with our warehouse. iProt is equipped with warehouse searching capabilities, descriptive protein report and JMol , a 3D protein structure modelling tool

## 2. Related work

COLUMBA (3) is an integrated database of protein structure and annotations. It comprises of 12 different databases including PDB, KEGG, Swiss-Prot, CATH, SCOP, the Gene Ontology , ENZYME and etc. COLUMBA features protein structure data, structural and sequence based classification scheme, functional annotation, secondary structure element and metabolic pathways participation in a single warehouse. COLUMBA works by dividing PDB file into a compound. Each compound then relate to function, protein-fold classification and annotation group. Each group contain databases which upgrade the information available in PDB file. COLUMBA gets protein domain information from SCOP and CATH in protein-fold classification. Sometimes, links from PDB file to other database such as Swiss-Prot are needed so that link to NCBI Taxonomy database is possible. PostGreSQL are used as database management system. BioPhyton and Perl parser were use to parse flat file data into relational database format. Integration between protein data in COLUMBA is done trough multidimensional data integration and star schema. COLUMBA provided XML format download and also JMol, an on-line viewer to visualize molecular structure.

Where as BioMolQuest(4) integrates four databases into their data warehouse; PDB, Swiss-Prot, ENZYME and CATH. Using these four databases, structural information, functional annotation and domain classification can be served to biologist under one roof. BioMolQuest is built in Perl scripting language and MySQL as their database management system. The main idea in BioMolQuest is to do cross-referencing among all databases included. This can fasten the query processing. Query is passed sequentially trough these databases. Problem with BioMolQuest is, it response slowly to a large queries. This may due to the design problem as many queries are cross-linked to each other. As consequent, query may become too complex and the processes become slow. If too much table involve in one queries then the processing time are too long. To avoid this, table and database schema must be redesign so that one query doesn't involve too much cross-referencing. On the other hand, iProClass (5) is an integrated database which linked 50 databases into its warehouse. It provides wide range of features including protein sequence, function and pathways, protein modification, gene and structural data. This warehouse employed in Oracle database and present protein view in two summary reports. Protein sequence report derived protein data and second report, which called superfamily report, which present PIR superfamily information of queried protein. This warehouse combines both local storage warehouse and hypertext navigation methods. This approach fit most of biologist needs as it provides latest update data.

## 3. Methodology

We gather data of protein sequence, 3D structure, enzymatic reaction, gene description, and taxonomical information into iProt. iProt currently integrates five databases, which are Swiss-Prot, Protein Data Bank (PDB), ENZYME, Gene Ontology (GO), and NCBI Taxonomy. iProt integration approach is base on the hybrid of Biowarehouse and OpenMMS schema represented in Relational Database Management System (RDBMS) platform. All databases that we integrated in iProt are cross-linked trough cross-reference record.

### 3.1 System Overview

As shown in Figure 1, user need to type keyword such as protein name or sequence and sent query trough iProt web base search application to the local server. This will generate summary report that consists of database information associated to queried protein. In general, all queries were centralizing on Swiss-Prot cross-reference data before being passed to any related entry. Structural data were invoked and pass to user for viewing purposed when related PDB ID for the queried protein is found in Swiss-Prot CrossReference table. An output file in XYZ format is then created and stored in the local server.

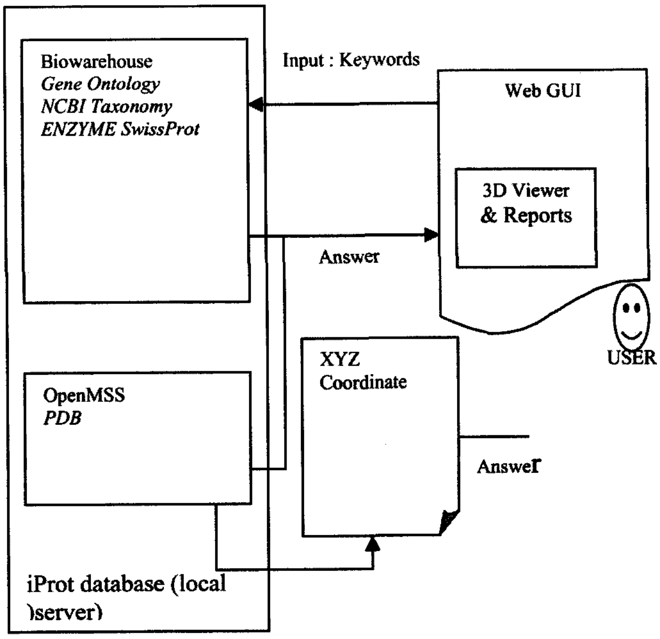


Fig. 1. System Model

### 3.2 Database Integration Phase

Database integration phase can be divided into two layer, which is data conversion layer, and data integration layer. These two layers are responsible to convert and integrate protein data from flat file into RDBMS format. In data conversion layer, the origin of used data is scattered and available in different format. Most of them stored in flat file format which only suitable for simple and small data scheme. This format is no longer fit the growth of protein data available and queries to them will heavily depend on the parsing algorithm complexity. iProt integrates multiple protein databases in RDBMS format and reduce the complexities of query process. Input file in flat file, XML, and mmCIF file format was converted into RDBMS format using parser program that dedicated to each database. For the data integration layer, each database contain in iProt is given a unique Db_id and each data such as protein sequence and XYZ coordinate has a unique Warehouse_ID that is associated to Db_id. Cross-linking records between Db_id and Warehouse_ID is stored. As shown in figure 2, all protein data is stored in a table that corresponding to their function. For example, protein sequence and protein name data is stored in table protein. Cross-linking information is use to link all related data in iProt database. This collection of related protein data is then uses to generate a protein summary.

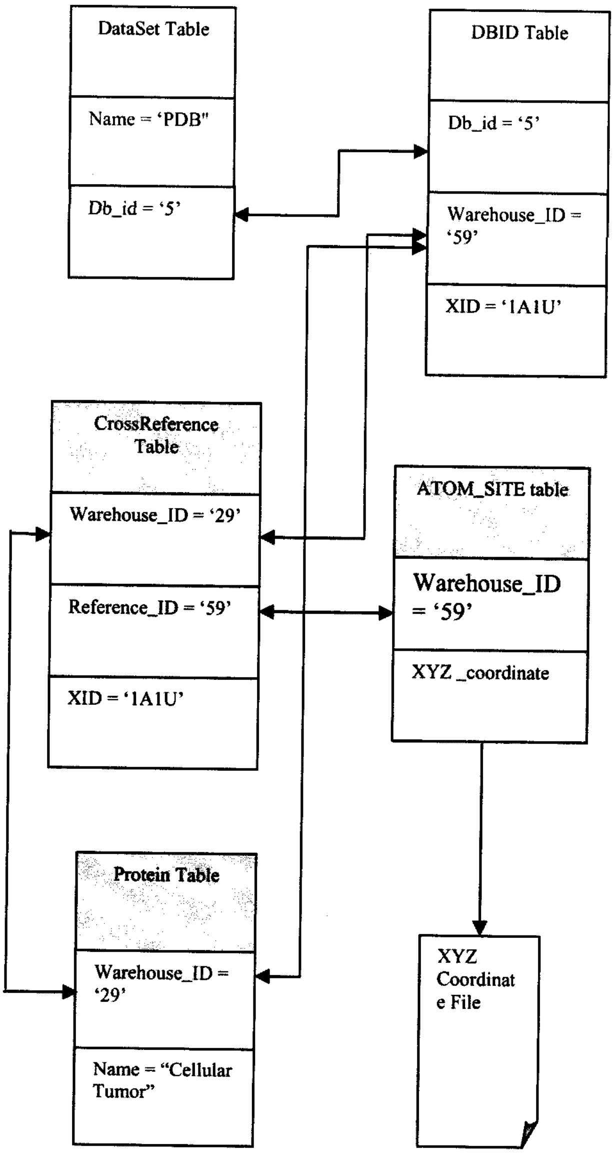


Fig. 2. Db_id, Warehouse_ID, Reference_ID, and XID Relationship in iProt Database.

In Figure 2, we show how query of protein structure from keyword, protein name = "*Cellular Tumor*" is passed in iProt. The server will search in Protein Table where Name = "*Cellular Tumor*" and get corresponding Warehouse_ID. Warehouse_ID = '*29*' will be pass to CrossReference Table to collect all related Reference_ID. Then, query will be pass to DBID Table to filter corresponding Reference_ID where Db_id = '5' (as PDB database in DataSet Table is refer in value '5'). ATOM_SITE used filtered Reference_ID and search for XYZ_coordinate record where Warehouse_ID = Reference_ID. XYZ coordinate then will be written into output file. iProt stored all data in each source databases into corresponding warehouse table. We try to avoid from deleting any record, as we want to preserve data consistency and accuracy in iProt.

Data update and data cleaning method can be done on selected database by selecting all Warehouse_ID where Db_id is equal to database Db_id value in DataSet Table. This avoids us from deleting all records if update of Swiss-Prot database is available in public server. Furthermore, iProt does not need to be redesign if new data added to server.

### 3.3 Database Design Phase

iProt database is designed in our local server as shown in Figure 3. This design is base on the hybrid of Biowarehouse and OpenMMS schema. iProt's database contained more than 1,029,511 data in 201 tables.

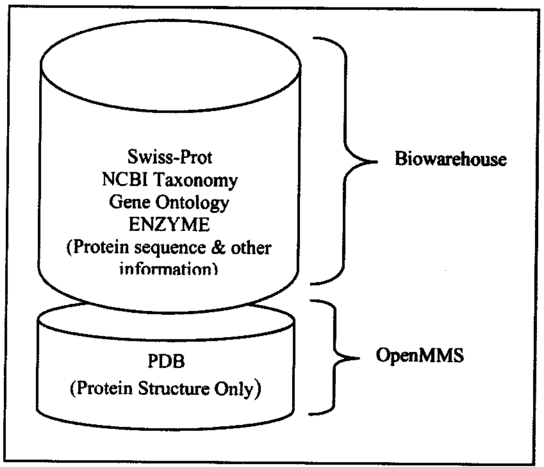


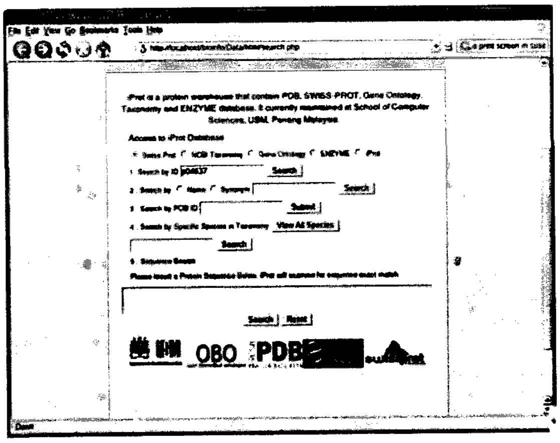
Fig. 3. Overview Database Design

### 3.4 Application Layer

iProt is implemented in a user-friendly web based application. This application enable user to query protein data on their web browser. This application has four major functions, which are Protein searching capabilities with specific keywords, data insertion form, 3D structure modeler and complete reports, which summarize data from all databases in iProt. For protein search features, user has an option to conduct search base on Protein name, Synonym name, Original database ID, Species Sequence and Structure ID. iProt provide a form for new data insertion. User need to specified protein data such as protein name, sequence, species, and related entry to the found protein.

## 4. Implementation and Results

We implemented iProt system in a SUSE LINUX system with 40GB hard disk capacity. We used MySQL version 4.1.11, PHP and apache server. Both, database integration and design phase is implemented on *localhost* server. As describe in methodology part, a parser will run in iProt's server environment to convert *flat file* into RDBMS format before copy the data into iProt's MySQL database. A PHP script then run and invokes these data using embedded SQL command based on keyword given by user. The results then return by server to the application layer. iProt's web application generate a full report using the result they received. User can query to iProt database by sending keyword to iProt server. Keyword such as original database ID, synonym, scientific name, species name and protein sequence is type by user to generate summary report of related protein as shown in Figure 4.


File Edit View Go Bookmarks Tools Help
http://localhost/bioinfo/Data/html/search.php
Go print screen in suse

iProt is a protein warehouse that contain PDB, SWISS-PROT, Gene Ontology, Taxonomy and ENZYME database. It currently maintained at School of Computer Sciences, USM, Penang Malaysia.

Access to iProt Database

● Swiss Prot ○ NCBI Taxonomy ○ Gene Ontology ○ ENZYME ○ iProt

1 . Search by ID p04637 [Search]

2 . Search by ○ Name ○ Synonym [Search]

3 . Search by PCB ID [Submit]

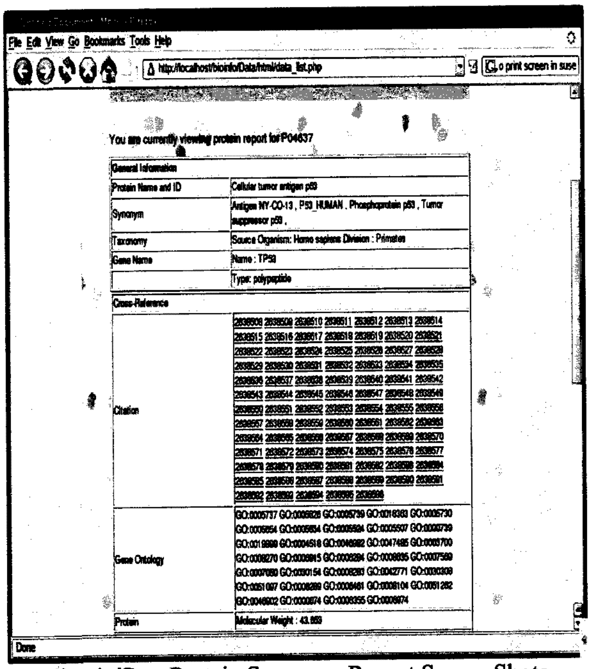
4 . Search by Specific Species in Taxonomy [View All Species]

[Search]

5 . Sequence Search

Please insert a Protein Sequence Below. iProt will scan for sequence exact match

[Search] [Reset]


File Edit View Go Bookmarks Tools Help
http://localhost/bioinfo/Data/html/data_list.php
Go print screen in suse

You are currently viewing protein report for P04637

| General Information | |
|---|---|
| Protein Name and ID | Cellular tumor antigen p53 |
| Synonym | Antigen NY-CO-13 , P53_HUMAN , Phosphoprotein p53 , Tumor suppressor p53 , |
| Taxonomy | Souce Organism: Homo sapiens Division : Primates |
| Gene Name | Name : TP53 |
| | Type: polypeptide |

| Cross-Reference | |
|---|---|
| Citation | 2638508 2638509 2638510 2638511 2638512 2638513 2638514 2638515 2638516 2638517 2638518 2638519 2638520 2638521 2638522 2638523 2638524 2638525 2638526 2638527 2638528 2638529 2638530 2638531 2638532 2638533 2638534 2638535 2638536 2638537 2638538 2638539 2638540 2638541 2638542 2638543 2638544 2638545 2638546 2638547 2638548 2638549 2638550 2638551 2638552 2638553 2638554 2638555 2638556 2638557 2638558 2638559 2638560 2638561 2638562 2638563 2638564 2638565 2638566 2638567 2638568 2638569 2638570 2638571 2638572 2638573 2638574 2638575 2638576 2638577 2638578 2638579 2638580 2638581 2638582 2638583 2638584 2638585 2638586 2638587 2638588 2638589 2638590 2638591 2638592 2638593 2638594 2638595 2638596 |
| Gene Ontology | GO:0005737 GO:0006826 GO:0005739 GO:0016363 GO:0005730 GO:0005654 GO:0005634 GO:0005524 GO:0005507 GO:0000739 GO:0019899 GO:0004518 GO:0046982 GO:0047485 GO:0003700 GO:0008270 GO:0006915 GO:0006284 GO:0008635 GO:0007569 GO:0007050 GO:0030154 GO:0008283 GO:0042771 GO:0030308 GO:0051097 GO:0006289 GO:0006461 GO:0008104 GO:0051262 GO:0046902 GO:0000074 GO:0006355 GO:0006974 |
| Protein | Molecular Weight : 43.653 |

Fig. 4. iProt Protein Summary Report Screen Shots

Summary report is created when information about user query are found in iProt database. Report organized by section such as general information about protein and cross-reference record of that protein of interest. Figure 4 give an example of the output from input, Swiss-Prot ID = 'P04637';

## 5. Conclusion and Future Work

This project has successfully integrates Swiss-Prot, ENZYME, PDB, NCBI Taxonomy and Gene Ontology database into iProt. A web-based application, which used to assist biologist query to iProt server was also successfully being implemented. This application featured a protein search capabilities, JMol 3D structure modeler, and a comprehensive protein report. We believe iProt would greatly benefit to biological research. iProt has also eliminated several problems occurred in flat file protein database. We plan to improve iProt by implement sequence and structure similarity algorithm in its searching function. This will give user more accurate result in percentage of similarities occurred in protein. More databases also should be added into iProt's database to cover wider area in biological sciences. We also suggest automatic update to be implemented in iProt's future work. Update should be run automatically each time public server announced the update to maintain data consistencies and accuracy. We believe that, iProt is a good basis to develop more protein analysis application in future

## References

(1) Francois B and Peer K. 2003. A Computational Biology Database Digest: Data, Data Analysis and Data Management. ACM Digital Library, Distributed and Parallel Databases. Pg 7-42

(2) Mark G. 2005. Integrative Databases Analysis in Structural Genomics. Nature Structural Biology. PubMed.

(3) Silke T. kristian R, Heiko M. 2005. Columba: An Integrated Database of Proteins, Structures and Annotations. BMC Bioinformatics (6).

(4) Yury V.B, Jeffrey S. 2001. BioMolQuest: Integrated Database-based Retrieval of Protein Structural and Functional Information. Bioinformatics 17 (5); pg 468 - 478.

(5) Hongzhan H, Winona C.B, Yongxing C, Cathy H.W. 2005. iProClass: An Integrated Database of Protein Family, Function and Structure Information. Nucleic Acids Research 31 (1): pg 390 - 392.