Pak. J. Agri. Sci., Vol. 53(1), 129-134; 2016
ISSN (Print) 0552-9034, ISSN (Online) 2076-0906
DOI: 10.21162/PAKJAS/16.5127
http://www.pakjas.com.pk

RPDB: A Relational Databank of Protein Structures

Naeem Aslam[1,6,*], Asif Nadeem[1], Masroor Ellahi Babar[2], Muhammad Tariq Pervez[3], Maryam Javed[1], Muhammad Aslam[4],Muhammad Wasim[1] and Wasim Shehzad[1]

[1]Institute of Biochemistry and Biotechnology, University of Veterinary and Animal Sciences, Lahore, Pakistan; [2]Department of Bioinformatics and Computational Biology, Virtual University of Pakistan; [3]Department of Computer Science, Virtual University of Pakistan; [4]Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan; [5]Department Molecular Biology, Virtual University of Pakistan; [6]Department of Computer Science, NFC Institute of Engineering & Technology, Multan, Pakistan.
[*]Corresponding author's e-mail: naeemiet@gmail.com

Proteins are the building blocks of cells of living creatures of all kingdoms. Due to pivotal role of protein structures in understanding evolutionary relationships and inferring their functions, the need to have a publically available databank was recognized and this results in establishment of protein data bank. This is a repository of protein structures in form of text files. Understanding and accessing protein structures in plain text files is very cumbersome job and a little bit difficult to understand the data as well. Several struggles have been made to transform protein structures of protein data bank in relational databases but all available relational databases of protein structures till now either do not provide complete information about the protein structures or they are not convenient to. Relational Protein Databank, presented in this paper, is a repository of all protein structures of protein databank stored in tables. A web interface has been provided to access various elements of protein structures in much more user friendly way. The structure of relational protein database has made data selection, insertion, deletion and editing very easy. RPDB is developed in MySQL which is an open source relational database management system.
Keywords: Lignment, protein structure comparison, PDB, database search, DBMS

INTRODUCTION

Proteins are the building blocks of all cells and perform all types of functions ranging from dynamic process of life maintenance, reproduction, defense, and replication. Due to pivotal role of protein structures in understanding evolutionary relationships and inferring their functions, the need to have a publically available databank was recognized and this results in establishment of Protein Data Bank (PDB) (Bernstein *et al*., 1978). Current version of PDB is termed as Worldwide Protein Data Bank (wwPDB) (Berman *et al*., 2007) and it comprises protein structures generated by NMR spectroscopy and X-ray crystallography. In the current era, a large number of protein structures are being determined by electron microscopy and added in PDB (Lee *et al*., 2012).

One of early databases was BIPED (Thornton and Gardner, 1989) that had data of protein secondary structure location and solvent accessibilities. SESAM (Huysmans et. al., 1991) was developed using the SYBASE (MacNicol and French, 2004) package. It was a relational database and allowed to incorporate raw data on protein structure and sequence ligands. Protein secondary structure (PSS) (Suzuki *et al*., 1991) was developed to associate protein sequence database with the atomic co-ordinates of the PDB. One of the drawbacks of these databases was that due to non-availability of World Wide Web at those times they were not accessible easily. Iditis (Gardner and Thornton, 1998) was a good commercial relational database for protein structures that allowed a wide range of queries to be used to fetch data. PACSY (Protein structure and Chemical Shift NMR spectroscopY) (Lee *et al*., 2012) is a relational database that incorporates information from PDB, the Biological Magnetic Resonance Data Bank (BMRB) (Markley *et al*., 2008) and the Structural Classification of Proteins (SCOP) database. It allows a user to get 3-D coordinates and chemical shifts of atoms along with derived information such as solvent accessible surface areas, torsion angles. It consists of 7 table types connected to one another for coherence.

The size of PDB is increasing at an enormous rate. The protein structures are distributed in text form with a specific format, one text file for every protein structure. The individual protein structures can be investigated in detail, however, this pattern does not allow to examine a protein structure on record level to analyze specific features of interest. PDB allows a user to search protein structures by using limited parameters such as protein name, function, source and simple experimental information. Secondly, searching a protein structure based on some specific parameter or comparing a protein structure with all structures in PDB is very time consuming task. Several relational databases have been developed to make searching and comparison of protein structures efficient and user friendly. However, some of the databases have either

provided limited range of searching criteria or the interface provided by them is not user friendly.

Relational Protein Data Bank (RPDB) is developed in MySQL. It comprises of several tables which have relationships to each other for selection, insertion, deleting and editing protein structures very easily. All protein structures in PDB have been transformed and stored in RPDB. RPDB provides a web interface for searching and comparing protein structures. The interface is designed using hypertext markup language (HTML), cascading style sheet (CSS). Business logic is written in Java programming language. It has made convenient to retrieve and analyze data of a single protein. For example, the user can obtain header data, the journal in which the protein structure was published, the atom coordinates, sequence residues, helix/ beta sheets, the chains and any other information of a particular structure. RPDB also allows the user to compare various components of one protein structure to the other protein structure. For example, a user can compare secondary structure elements, atom coordinates or sequence residues of two given protein structures. Similarly, all these activities can be performed while comparing a single protein structure against multiple structures.

MATERIALS AND METHODS

***Relational Protein Databank (RPDB)*:** The Relation Protein Databank is a database, which was used to store PDB data in the form of relations. Each PDB file is stored in different entities of RPDB with their desired attributes. MySQL relational database management system was used to create and maintain RPDB. This RDBMS is an open source management system and stores data in the form of intersection of rows and columns. The database system has provided the support of Structured Query Language (SQL) for communication between application program and RDBMS.

***Methodology of downloading PDB files and inserting data in RPDB*:** The whole process of downloading PDB files and inserting records of a PDB file in RPDB (Fig. 1).

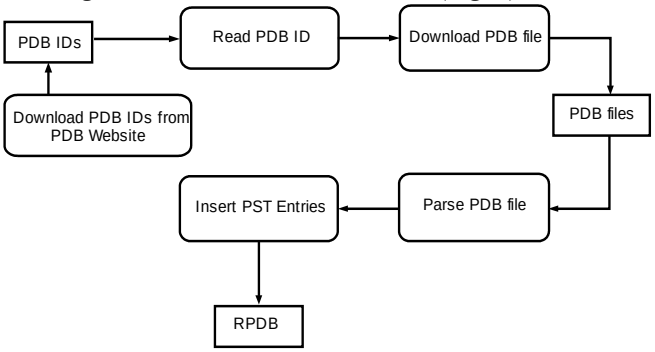

Figure 1. Methodology for Downloading PDB files and inserting data in RPDB.

***Downloading PDB Files*:** PDB IDs were downloaded from the website of PDB and saved in an excel file. These IDs were used to download PDB files using a software tool written in Java programming language (Fig. 1).

***Inserting protein structures into RPDB*:** A parser written in Java programming language was used to extract all records of a PDB file and insert them in RPDB.

***Computer languages*:** RPDB was written using java servlets, Java Sever Pages (JSPs), JavaScript and HTML. JSP was used to implement presentation layer. JavaBeans were used to implement business logic. Java servlets were used to write server side script. JavaScript was used to write client side script and verify data provided by user. HTML was used to present data and contents to the user in an attractive manner.

***Development environment*:** NetBeans IDE 7.4 was used as an integrative developing environment to write desktop as well as web based versions of SuitePST.

RESULTS

RPDB is a relational protein databank developed in MySQL. All PDB files available in PDB were parsed and records of PDB files were stored into several related tables. These tables have entity and referential integrity constraints which ensure data integrity. The minimal data redundancy is achieved using normalization technique up to 3rd normal form. This methodology has made retrieval and analysis of data very efficient.

***RPDB architecture*:** Since RPDB is a web application so its architecture can be described through three tier architecture as shown Fig. 2.

This tier architecture is composed of three layers: (1) presentation layer displays results and graphs and tables on the user screen, (2) business logic layer performs computations and analyze the data and send it to the presentation layer, (3) data layer has all pdb files downloaded by the PDB files downloader**.**

***Entity relational model (ER-Model)*:** The ER-model (Fig. 3) was developed to describe the entities of RPDB, their attributes and relationship among them. The structural constraints i.e., cardinality and participation were used to maintain entity and referential integrity. The 'primary key' and 'not null' constraints were made available on entity level and foreign key constraint was made available between entities. The domain values of attributes were selected according to data requirement.

***Main features provided by RPDB*:** Main interface of RPDB allows a user to perform following three activities.

- To find information of a single protein structure
- To compare two protein structures
- To compare multiple protein structures with a given protein structure.
- Computing details of master record of a single protein structure

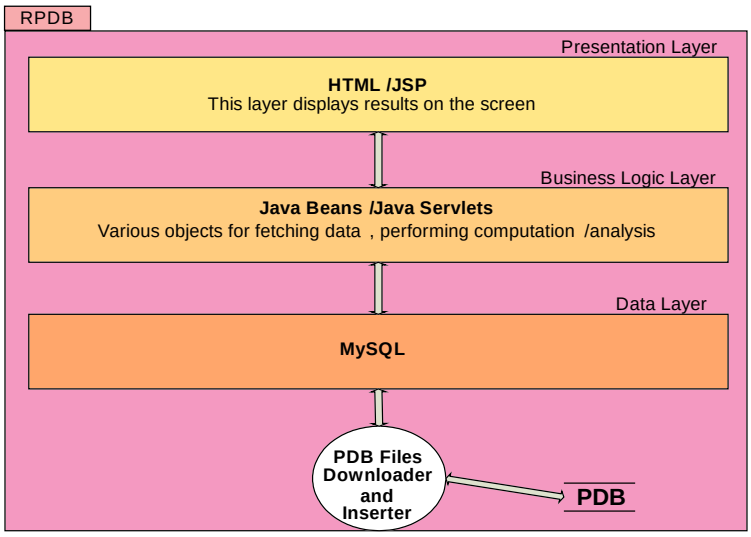

Figure 2. Three tier architecture of RPDB.

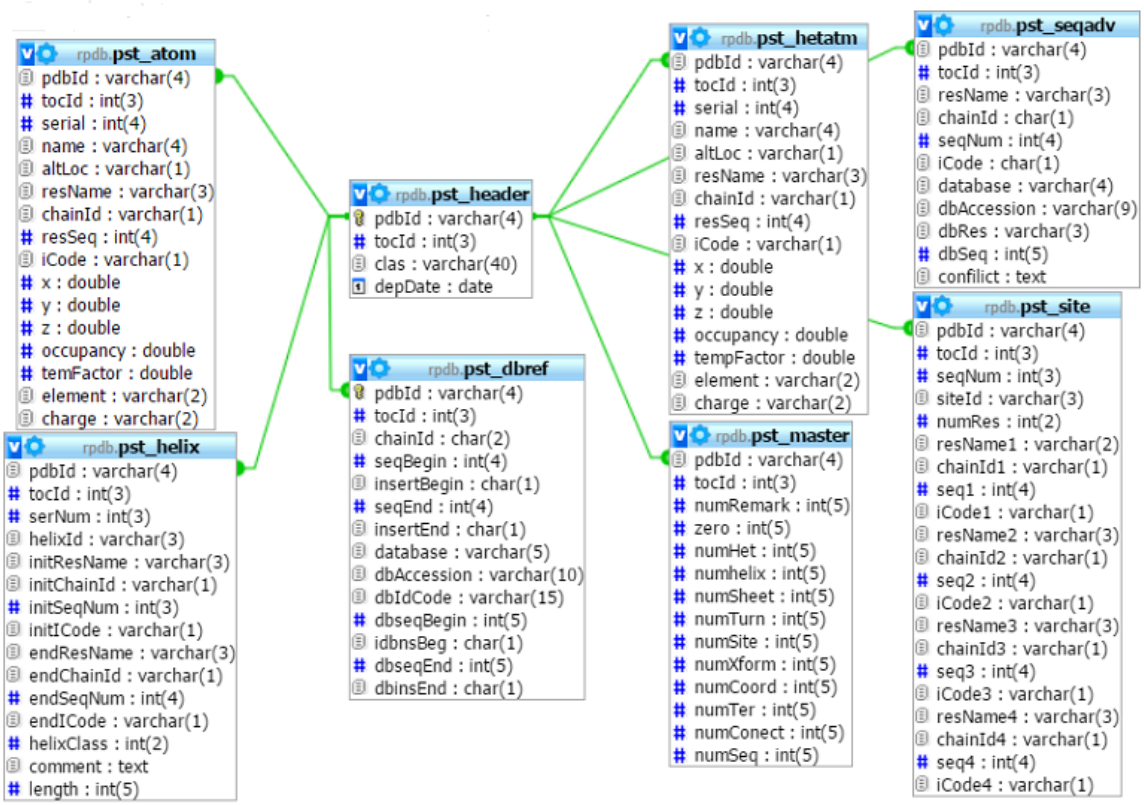

Figure 3. ER-Model used to design database of RPDB.

- Computing Helix information of a single protein structure
- Computing HET/HETATM information of a single protein structure
- Computation of Site details of a protein structure
- Computing Sheet information of a single protein structure
- Computing and comparing master information of two protein structures

When user selects any one of these options, interface of RPDB loads the respective web page and allows the user to perform a number of activities relevant to the selected option.

***Comparison with other relational databases*:** Like other relational databases for protein structures, RPDB provides a

Table 1. Comparison of RPDB with other relational database.

Parameters/RDBs	Iditis	PACSY	PSSARD	RPDB
Year of Published	1998	2012	2005	2015
DBMS Used	Own developed	MySQL		MySQL
Relations	9	6		42
Programming Language	Unspecified	C++/PASCAL	PHP	Java
Open Source	No	No	Yes	Yes
Dedicated for PDB	Ys	No	Yes	Yes
User Friendly Level	Medium	Medium	Very low (Multiple steps for first process)	Very High
Record based search	Yes	Yes (Limited)	Yes (Limited)	Yes
Comparison of MPS	No	No	No	Yes
Searching Levels	1	1	1	3
Information retrieval of Single Protein	Yes	No	No	Yes
Support of No. Records to Search	8	6	6	14
Support of Relational Operators	Yes (Limited)	Yes (Limited)	Yes (Limited)	Yes
Criteria to search PS	8	7	5	16

lot of features and criteria to fetch information about a protein structure or compare multiple protein structures. It outperforms other databases in several ways such as

- It allows a user to find out comprehensive information about a single protein using a number of criteria,
- It supports search of protein structures using large number of records.
- It allows searching protein structures using nested search criteria.
- It provides a user friendly interface for usage of relational operators to make a search query comprehensive.
- It allows a user to compare a given protein structure with the whole database using a number of criteria.
- It allows to compare two given protein structures based on various criteria.

Comparison of RPDB with other relational databases is summarized in Table 1.

DISCUSSION

Proteins are the building blocks of all cells in all living creatures of all kingdoms. Their functions can be examined through their structures (Whisstock and Lesk, 2003; Prabhavathi *et al*., 2011; Bordoli *et al*., 2009). Applications of protein structures can be observed in a number of research studies such as designing site-directed mutagenesis experiments, virtual screening, rationalizing the effects of variations in sequence (Hillisch *et al*., 2004; Kopp and Schwede, 2004; Marti-Renom *et al*., 2000; Peitsch, 2002). They perform all types of functions ranging from dynamic process of life maintenance, reproduction, defense, and replication. Due to pivotal role of protein structures in understanding evolutionary relationships and inferring their functions, the need to have a publically available databank was recognized and this results in establishment of Protein Data Bank (PDB) (Bernstein *et al*., 1978). Current version of PDB is termed as Worldwide Protein Data Bank (wwwPDB) (Berman *et al*., 2007) and it comprises protein structures generated by NMR spectroscopy and X-ray crystallography. In the current era, a large number of protein structures are being determined by electron microscopy and added in PDB (Lee *et al*., 2012). RPDB is a relational database of protein structures available in PDB (Bernstein *et al*., 1978). This database was developed in MySQL (Kofler, 2001) which is an open source (West and Gallagher, 2006) relational database management system. All records of protein structures downloaded from PDB were downloaded and saved in relational protein database using a program written in java programming language. Saving of protein structures in a lot of relations has made the search and retrieval of information very efficient and user friendly. Table of content is the most important relation in the database. It is used to manage other relations. As a part of RPDB, web interface and web server were also developed using JSP and HTML.

Other relational databases such as Protein structure and chemical shift NMR spectroscopy (Lee *et al*., 2012) is a relational database management system that extracts and integrates information from three data sources. It does not provide the support to analyze protein structures record by record and (Mobilio, 2010) presented three databases i.e. Protein Relational Database, Matrix Metalloproteinase Knowledge Base and Kinase Knowledge Base. They extract data from protein databank. They made it easy to find atom-atom distances among protein and ligand. Ring controids, centoroid-atom and centroid-centroid distances are also provided by this database. However, this database does not support comparison of two or multiple protein structures based on various elements such as HET Atom, secondary structure elements, atom co-coordinates sequence residues.

Islam and Sternberg (1989) developed a relation database using ORACLE database management system. It provides information about several aspects of protein structure. The database stores information about coordinates set, various chains in protein, ligands and amino acid residues, salt bridges, atomic co-ordinates, hydrogen bonds, disulphide bridges and close tertiary contacts. However, this database does not allow the user to make comprehensive comparison of the protein structures. Secondly, currently, no web interface for this database exists or maintained. Chebrek *et al*. (2014) developed data repository for polyproline helix II secondary structure element.

RPDB allows comparing two protein structures based on a number of criteria similarly to the search of a single protein. A user can also compare a particular protein structure against the whole database. Web interface of RPDB allows a user to get protein structures using sixteen criteria such as the user can obtain protein structures with a specific HELIX class, specific HELIX length, parallel/anti-parallel strands, specific experimental techniques, specific Organism Name, specific number of Residues in SITE record, links to specific database, conflicts of specific class, specific resolution, specific number of amino acids, specific R-Value, specific number of models, JRNL reference, specific number of strands, having specific percentage of different amino acids, specific number of chains. It also computes various types of statistics to analyze a single protein, to compare two proteins and make comparison of a protein to multiple proteins. A user is allowed to find any type of information of a single given protein. Several records of two protein structures can be compared and analyzed. Similarly, records of a query protein structures can be compared with given multiple protein structures or the whole database.

RPDB outperforms other relational databases in several ways such as now a user can find out well organized information about a single protein using a number of criteria, search protein structures using large number of records, use nested search criteria, build a query using relational operators, compare a given protein structure with the whole database or compare two protein structures using a number of criteria. One of the core features of RPDB is that normalization process was followed to reduce redundancy and keep data in the well managed form. This feature helped a user to get information about a particular part of a given protein. Front end interfaces of other relational databases provide limited features to get information about a protein structures. Some of them involve multiple steps (Gardner and Thornton, 1998) to run a simple query. Some of them (Lee *et al*., 2012) lack the ability of generating queries using nested criteria.

***Conclusion*:** RPDB is relational database of protein structures which allows the user to search and analyze elements of a single protein structure, two protein structures and comparing a protein with a number of provided protein molecules or the database. It is very useful for structure biologists to infer structure and function of an unknown protein and finding the evolutionary relationships among protein molecules. A single protein can be analyzed by a number of techniques. A user can find master record, HET groups with their group names, formulas and atoms, secondary structure elements, sequence residues and atom co-ordinates. A user can compare and analyze two proteins using common records such as sites of proteins, HELIX, Sheets. Web server of RPDB allows to compare a query protein with multiple comma separated proteins or the whole database. The main feature of RPDB is that data is already parsed and available in form of records which can be accessed directly without traversing other elements of a protein structure.

***Availability*:** http://www.suitepst.com/SuitePST/

REFERENCES

Berman, H., K. Henrick, H. Nakamura and J.L. Markley. 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucl. Aci. Res. 35: 301-303.

Bernstein, F.C., T.F. Koetzel, G.J.B. Williams, E.F. Jr. Meyer, M. Brice, J.R. Rogers, O. Kennard, T. Shimanouchi and M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macro-molecular studies. J. Mol. Biol. 112:535–542.

Bordoli, L., F. Kiefer, K. Arnold, P. Benkert, J. Battey and T. Schwede. 2009. Protein structure homology modeling using SWISS-MODEL workspace. N. Protocols 4: 1-13.

Chebrek, R., S. Leonard, A.G. de-Brevern and J.C. Gelly. 2014. PolyprOnline: polyproline helix II and secondary structure assignment database. Database 2014: 1-8.

Gardner, S. and J. Thornton. 1998. IDITIS: Protein structure database. Acta Cryst. D. 54: 1071-1077.

Hillisch, A., L.F. Pineda and R. Hilgenfeld. 2004. Utility of homology models in the drug discovery process. Drug Discov. Today 9: 659-669.

Huysmans, M., J. Richelle and S.J. Wodak. 1991. SESAM: a relational database for structure and sequence of macromolecules. Proteins. 11: 59-76.

Islam, S.A. and M.J. Sternberg. 1989. A relational database of protein structures designed for flexible enquiries about conformation. Protein Eng. 2: 431-442.

Kofler, M. 2001. What Is MySQL? Apress.

Kopp, J. and T. Schwede. 2004. Automated protein structure homology modeling: a progress report. Pharmacogenomics 5: 405-416.

Lee, W., W. Yu, S. Kim, I. Chang, W. Lee and J.L. Markley. 2012. PACSY, a relational database management system for protein structure and chemical shift analysis. J. Biomol. NMR 54: 169-179.

MacNicol, R. and B. French. 2004. Sybase IQ multiplex-designed for analytics. Proceedings of the Thirtieth

International Conference on Very large Databases 31 AugUST 2004. VLDB Endowment; pp.1227-1230.

Markley, J.L., E.L. Ulrich, H.M. Berman, K. Henrick, H. Nakamura and H. Akutsu. 2008. BioMagResBank as a partner in the Worldwide Protein Data Bank: new policies affecting biomolecular NMR depositions. J. Biomol. NMR 40: 153-155.

Marti-Renom, M.A., A.C. Stuart, A. Fiser, R. Sanchez, F. Melo and A. Sali. 2000. Comparative protein structure modeling of genes and genomes. Ann. Rev. Bioph. Biom. 29: 291-325.

Mobilio, D., G. Walker, N. Brooijmans, R. Nilakantan, R.A. Denny, J. DeJoannis and C. Humblet. 2010. A Protein Relational Database and Protein Family Knowledge Bases to Facilitate Structure-Based Design Analyses. Chem. Biol. Drug. Des.76: 142-153.

Peitsch, M.C. 2002. About the use of protein models. Bioinformatics 18:934–938.

Prabhavathi, M., K. Ashokkumar, N. Geetha and D.K.M. Saradha. 2011. Homology modeling and structure prediction of thioredoxin protein in wheat (Triticum aestivum L.). J. Biosciences 1: 20-32.

Suzuki, H., A.S. Kolaskar, S.L. Samuel, J. Otsuka and A. Tsugita. 1991. A protein secondary structure database (PSS). Protein Seq. Data Anal. 4: 97-104.

Thornton, J.M. and S.P. Gardner. 1989. Protein motifs and data-base searching. Tre. Biochem. Sci. 14:300-304.

West, J. and S. Gallagher. 2006. Challenges of open innovation: the paradox of firm investment in open source software. R&d Management 36:319-331.

Whisstock, J.C. and A.M. Lesk. 2003. Prediction of protein function from sequence and structure. Q. Rev. Biophys. 36: 307-340.