PSSD: Protein Secondary Structure Database

HAMID NIKBAKHT, PAYAM MAHMOUDIAN, BAHRAM GOLIAEI[*]
[*]Institute of Biochemistry and Biophysics
University of Tehran
Tehran, P. O. Box: 13145-1384
IRAN

Abstract: Protein Secondary Structure Database (PSSD) is a database that incorporates sequences of secondary structure elements of all proteins which their three dimensional structures are defined by experimental methods such as NMR-Spectroscopy or X-Ray Crystallography and their structural data exists in Brookhaven protein databank. Dictionary of Secondary Structure of Proteins (DSSP) criteria have been used to define both ends of each structural element. At present PSSD includes 290,709 alpha helices, 418,362 beta strands, 571,176 turns and 118,109 helices 3(10) of 21,347 proteins. The following information is given for each entry: (i) PSSD Unique ID, (ii) Description, (iii), Organism source, (iv) Author(s), (v) PDB code, (vi) Cross references to PDB, DSSP and Swiss-Prot databanks, (vii) Sequence of secondary structure element, (viii) number of starting and ending amino acids of each element in its corresponding protein chain, (ix) length of element, (x) the number of the element in its regarding protein chain. A user friendly interface is developed for doing search in database using different combinations of fields mentioned above. Facilities provided in this database allow structure-sequence analysis studies faster, more reliable and suitable. Now, the database is located on IBB Bioinformatics Center (IBC) server. The interface can be accessed via: http://www.ibc.ut.ac.ir/pssd/.

Keywords: Database, Protein Secondary Structure, PhP, MySQL, Linux, Bioinformatics.

1 Introduction

Molecular sequence data is now growing at an accelerating pace due to technical improvements in sequencing technology. Advance databases are needed to facilitate the retrieval of relevant information from the huge amount of sequence data available in primary sequence databases [1, 2].

The secondary structure of proteins is essential to the biological function of the protein. Many attempts have been made at describing the various structural and conformational properties of the secondary structural elements of proteins [3, 4]. Analysis of sequence of whole proteins or secondary structural elements has provided valuable information regarding prediction of structural or functional characteristics of proteins [5- 8].

The sequences from all proteins which their structure are determined and represented in the protein data bank (PDB) [9] are compiled in the NRL-3D sequence database [10]. Although NRL-3D database still provides the most reliable means for searching the sequences actually represented the PDB files, there are disagreements regarding the assignment of zonal amino acids in the secondary structures such as alpha helices. The dictionary of the secondary structure of proteins (DSSP) is frequently used as an objective and uniform automatic criteria of assignment of elements of the secondary structure of proteins [11].

Various specialized databases have provided access to sequences of the secondary structural elements of proteins [12-14]. These databases use filtering criteria for excluding sequences with various degrees of similarity from their databases. The Protein Secondary Structure Database (PSSD) has been developed to represent an exhaustive collection of secondary structural elements of all proteins whose structure has been represented in PDB. We have used DSSP criteria for assigning the exact definition of sequence elements of the secondary structures. The database contains information about more than 22000 proteins which their structure has been solved. The web-based querying interface allows rapid and friendly retrieval of information from PSSD.

2 Format of the PSSD database

The database consists of a single flat file that contains information about each secondary structure element such as: PSSD unique ID, sequence, starting and ending amino acid number of each element, the number of each element in its regarding protein chain, a brief description about protein chain, cross references to PDB, DSSP and SWISS-PROT databases and some other information such as authority information and etc. This flat file has a standard unique format that is computer readable. Each line starts with a two letter ID line following with three white spaces; each element in this flat file is separated from others by specific sign”//”. Programmers can write scripts to parse this flat file and get various results regarding to their aim.

3 Fields of each entry

ID (PSSD unique ID): A unique numerical code that is given to each record in database. This unique code has some information in itself. The code is a combination of some properties of element such as PDB code, chain name, structure type and starting residue number.

DR: Cross references to other databases such as PDB and SWISS-PROT.

CH (Chain): The protein chain name regarding the secondary structure element.

ST (Structure): The type of secondary structure.

RN (Residue number): The starting and ending amino acid numbers in its regarding protein chain.

DS (Description): A brief description about each protein chain that is a combination of information given in compound and header in DSSP files.

DT: Date of structural information deposition in PDB

OS (Organism Source): The scientific name of source organism of protein.

AU (Author(s)): Authority information of protein.

SQ (Sequence): The sequence of each entry. The sequence is given in a standard format (10 residues in each column and 60 in each line). Some information about the length of element and the number of element in its regarding chain is given to.

//: The sign each entry is separated from others.

This is an example of a part of flat file that belongs to third alpha helix of Apolipoprotein A- I chain A. (Figure 1)

4 Whole Procedure

We used PDB database [9] as a primary resource that contains experimental information of 3D structure of proteins. DSSP [11] definitions was used to define both start and end of each secondary structure. A program was developed in C++ called ***SSF*** (Secondary Structure Finder) for extracting information from each DSSP flat file and adding the extracted information at the end of flat file (PSSD.DAT flat file). A shell program ran SSF through DSSP files. This step was repeated till the last DSSP file.

```
//
ID   1av1HA72
DR   PDB: 1av1; SWISS-PROT: APA1_HUMAN
CH   A
ST   Alpha-Helix
RN   72-140
DS   LIPID TRANSPORT, APOLIPOPROTEIN A-I;
DT   23-SEP-97
OS   HOMO SAPIENS;
AU   D.W.BORHANI, D.P.ROGERS, J.A.ENGLER, C.G.BROUILLETTE
SQ   SEQUENCE 69 Amino acids; Alpha-Helix No. 3;
     WDNLEKETEG LRQEMSKDLE EVKAKVQPYL DDFQKKWQEE MELYRQKVEP LRAELQEGAR
     QKLHELQEK
```

Figure 1: A sample entry of the database

5 SSF Algorithm

SSF has three modules; two of them are responsible to read and parse each DSSP file and the third is responsible to find the cross reference of each chain in SWISS-PROT database. In according to parse DSSP files and extract information through them, we read each DSSP file in a 2D array (each line as a record of array). We found a constant and unique pattern which addressed us the location of information in DSSP files. An algorithm was developed for recognizing this pattern through each record of 2D array and extracting the desired information. Using this algorithm, we could extract any type of information through DSSP files (or any other flat file) files such as header information and others. Figure 2 shows the whole algorithm of parsing DSSP and updating the flat file.

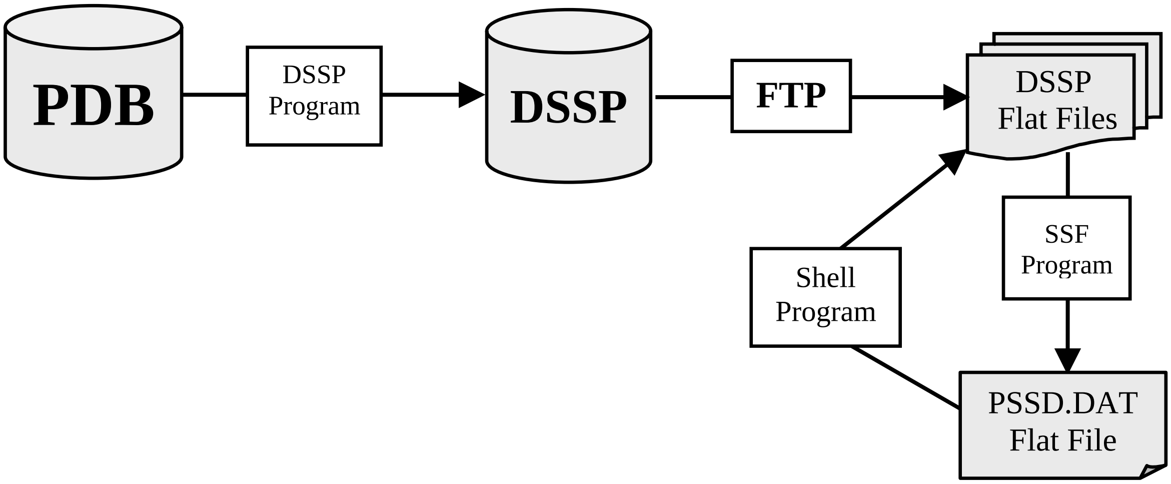

Figure 2: The whole procedure of making PSSD.DAT flat file

6 PSSD Interface

We chose reliable and fast 'MySQL' (a free open source Data Base Management System) for data storage system. Data stored in PSSD.DAT flat file was loaded into a designed table called ‘pssdtable’ using MySQL. The provided PHP web-interface does search on this table using MySQL. PHP server scripting technology lets us to connect to the mentioned DBMS and generate customized data-bounded HTML pages according to the user's requests. Figure 3 shows this procedure.

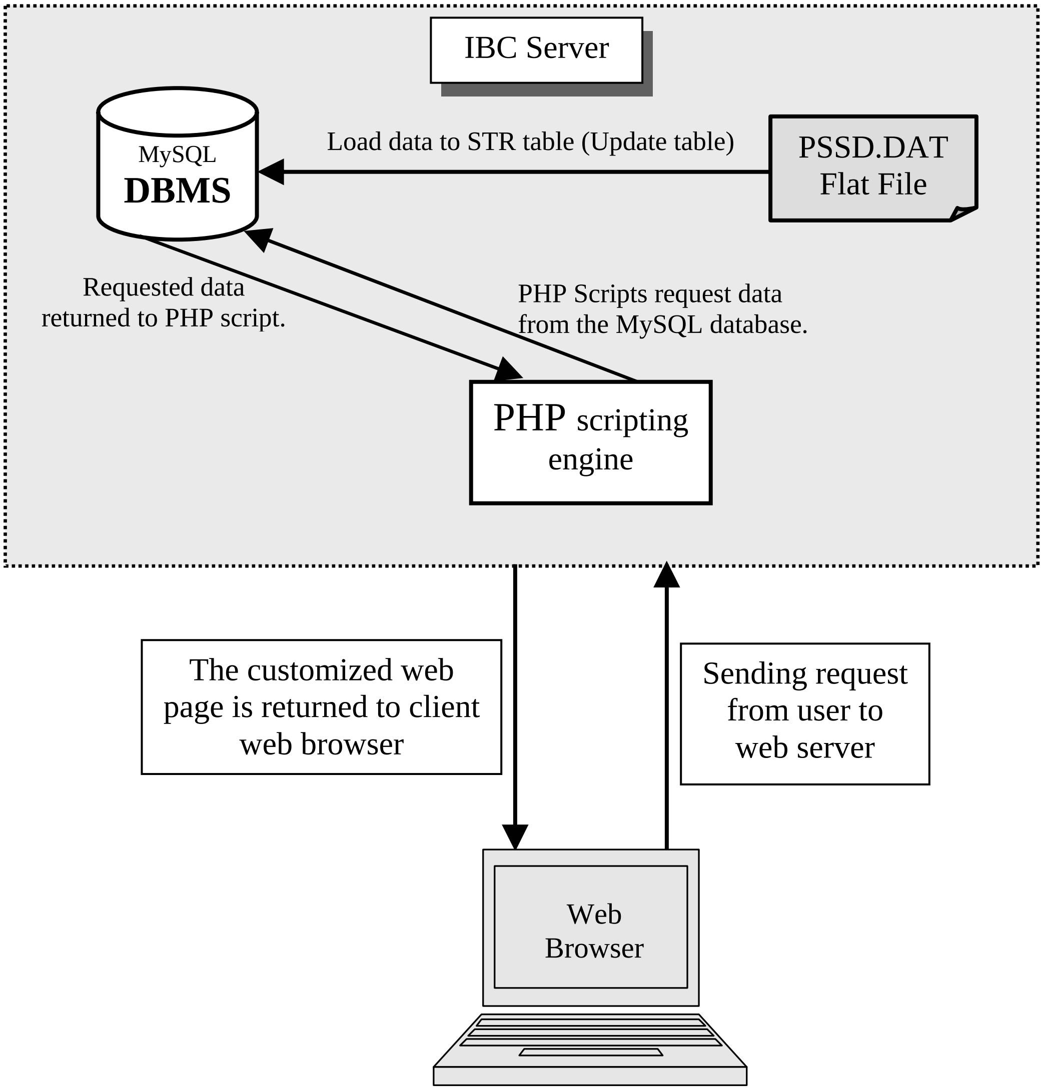

Figure 3: Figure3- The procedure of doing search using PSSD interface

7 Access to the PSSD interface

The PSSD interface is located in IBB Bioinformatics Center (IBC) server and can be accessed via:

http://www.ibc.ut.ac.ir/pssd/

Users can simply do search on database via this interface and get results in any desired format.

8 Search in database

The provided web-interface has several fields that users can use one or a combination of them to do search through database. It’s not necessary to fill each field completely, users can fill a part of desired query and search engine would give back all entries which contained given query. One of aspects of this interface is that users can select various formats for outputs. These formats are: a) a table that contains information in brief that has two links in each row. The first one refers to a page which has a complete information of entry with links to cross references to PDB, DSSP and SWISS-PROT, the second one refers to an index page of project directory. Users can choose a project name and a description that can be used for next refers. The interface makes a directory named by project name and all of the result files will be saved in this directory. b) A tab limited flat file that contains information in a tab limited format. In this format, each entry is given in a single row and all fields are separated by a tab. c) A chunked flat file, In this format information of each entry is fully discussed and entries are separated from each other with a “ // ” sign. These two last formats are very suitable for programmers who want to parse information of secondary structures. d) A FastA sequence format, in this format information of PSSD unique ID, PDB code, structure type and some others are given. This format is very suitable for analysis like multiple alignment among sequences and etc. e) The last format is raw sequence format, by choosing this option; users can achieve sequences of all entries without any other information. This format is very suitable for those who want to do statistical analysis in this area.

9 Database statistics

This database covering the secondary structural information of 21,347 proteins that includes 1,398,356 records consisting 290,709 alpha helices, 418,362 beta strands, 571,176 turns and 118,109 helices 3(10). The size of entire database is about 286.636 MB. An index has been made for table STR for faster search and the final size of database tables became 398.606 MB.

10 References:

[1] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler (2004) GenBank: update. *Nucl. Acids. Res.*, 32, D23-D26

[2] Rolf Apweiler, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O’Donovan, Nicole Redaschi, and Lai-Su L. Yeh. (2004) UniProt: the Universal Protein knowledgebase. *Nucl. Acids. Res.* , 32, D115-D119

[3] Vinayagam, J. Shi, G. Pugalenthi, B. Meenakshi, T.L. Blundell, and R. Sowdhamini DDBASE 2.0: updated domain database with improved identification of structural domains *Bioinformatics*, 19, 1760-1764

[4] Guruprasad K, Prasad MS, Kumar GR. (2001 Apr) Analysis of gammabeta, betagamma, gammagamma, betabeta continuous turns in proteins., *J Pept Res*. ,57(4), 292-300.

[5] Richardson JS, Richardson DC (1988) Amino acid preferences for specific locations at the ends of alpha helices.*Science*. Jun 17, 240(4859), 1648-52.

[6] Steward RE, Thornton JM. (2002 Aug) Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins*., 1;48(2),178-91.

[7] Goliaei B, Minuchehr Z. (2003 Feb) Exceptional pairs of amino acid neighbors in alpha-helices.*FEBS Lett*. 27;537(1-3), 121-7.

[8] Baker EN, Arcus VL, Lott JS. (2003) Protein structure prediction and analysis as a tool for functional genomics. *Appl Bioinformatics*., 2(3 Suppl),S3-10

[9] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne (2000) The Protein Data Bank *Nucl. Acids. Res*., 28, 235-242

[10] Pattabiraman N, Namboodiri K, Lowrey A, Gaber BP. (1990 Oct) NRL-3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment. *Protein Seq Data Anal*., 3(5), 387-405.

[11] Kabsch W & Sander C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22, 2577-2637

[12] Kunchur Guruprasad, Maheshuni S. Prasad, and Gundu R. Kumar (2000) Database of Structural Motifs in Proteins *Bioinformatics*, 16, 372-375

[13] S. A. Fernando, P. Selvarani, Soma Das, Ch. Kiran Kumar, Sukanta Mondal, S. Ramakumar, and K. Sekar (2004) THGS: a web-based database of Transmembrane Helices in Genome Sequences *Nucl. Acids. Res*., 32, D125-D128

[14] V. Shanthi, P. Selvarani, Ch. Kiran Kumar, C. S. Mohire, and K. Sekar (2003)

SSEP: secondary structural elements of proteins *Nucl. Acids. Res*., 31, 3404-3405