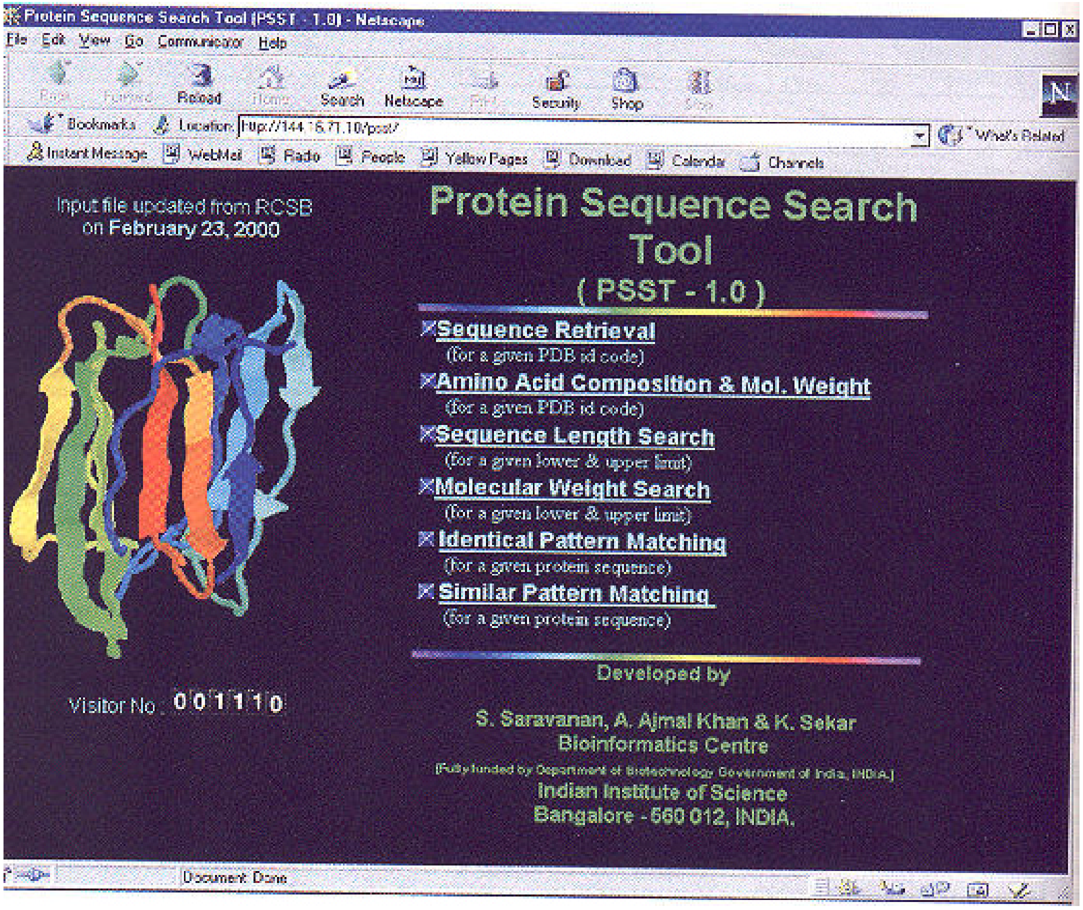
Protein sequence search tool: A web-based interactive search engine

Protein sequence search tool (PSST) is a web-based interactive search tool developed to extract and analyse the protein/nucleic acid sequence and other related information from the database of protein and nucleic acid structures. A search could be performed using all the protein structures or in a selected subset of non-homologous proteins. The basic sequence information is updated every week and hence the results obtained using the search tool are up-to-date at any given time. The package PSST (Version 1.0) is available over the World Wide Web (www) at http://pranag.physics.iisc.ernet.in/psst. The number of available three-dimensional structures of proteins and nucleic acids is increasing rapidly. The three-dimensional structures of nearly 11,800 proteins and nucleic acids are currently available in the public domain of the Protein Data Bank (PDB)[1]. This database also carries the sequence information of all the known three-dimensional structures of proteins and nucleic acids. Prediction of the three-dimensional structure from the amino acid sequence is one of the central problems to be resolved in structural biology. While trying to understand this problem it is important to recognize structural characteristics of sequence patterns. In several DNA, RNA and protein structures similar sequence repeat units are often found, and it is essential to know the correlation with respect to three-dimensional structures. Moreover, identifying the occurrence of functionally or structurally characteristic sequence motifs in various protein or nucleic acid structures is useful. Occurrence of a particular sequence pattern in a protein structure[2] might give insights on the possibility of the associated function and one could fine-tune the three-dimensional structure by means of protein engineering experiments to incorporate the function. Occurrence of sequence repeats in protein structures might suggest symmetries in


Protein Sequence Search Tool (PSST - 1.0) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://144.16.71.10/psst/
What's Related
Instant Message WebMail Radio People Yellow Pages Download Calendar Channels
Document: Done

Input file updated from RCSB on February 23, 2000

Protein Sequence Search Tool (PSST - 1.0)

- Sequence Retrieval (for a given PDB id code)
- Amino Acid Composition & Mol. Weight (for a given PDB id code)
- Sequence Length Search (for a given lower & upper limit)
- Molecular Weight Search (for a given lower & upper limit)
- Identical Pattern Matching (for a given protein sequence)
- Similar Pattern Matching (for a given protein sequence)

Visitor No. 001110

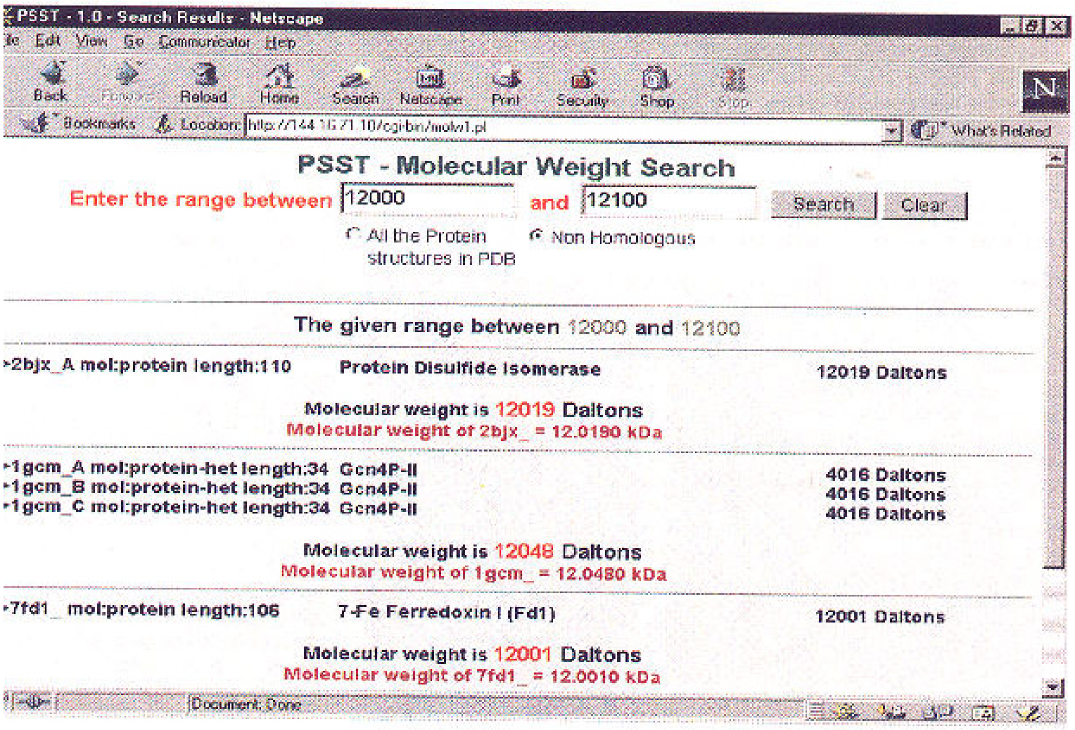
Developed by

S. Saravanan, A. Ajmal Khan & K. Sekar
Bioinformatics Centre
(Fully funded by Department of Biotechnology Government of India, INDIA.)
Indian Institute of Science
Bangalore - 560 012, INDIA.

Figure 1. Web page of PSST along with the available search facilities. Clicking on the appropriate search facility activates the search engine. The molecule on the left side is a lectin from the seeds of jackfruit. The crystal structure[10] of this lectin has four subunits, but only one of them is shown. This picture has been created using the graphics software RASMOL[11].

them[3]. In order to analyse this large volume of database it is essential to have automated computational search tools such as SRS (Sequence Retrieval System)[4]. Towards this effort, several software packages have been developed for identifying the similarity in sequence patterns and their compatibilities with the available databases like SWISS-PROT[5] and PIR (Protein Identification Resource)[6]. To the best of our knowledge, a tool to facilitate search for a well-defined molecular weight and also to perform a search for sequence patterns in a set of non-homologous protein sequences is not yet available. PSST addresses these issues too. The front page of the search tool is shown in Figure 1. PSST offers two different input options for the user to address her/his query. The user has the option of choosing all the protein structures or only using the non-homologous protein structures derived by Hobohm and Sander[7]. In this set of non-homologous protein structures no two proteins have sequence identity of over 30%. At present there are about 1,100 protein chains in the non-homologous category.

The PSST has the following types of search facility: (a) Sequence retrieval in PIR format, (b) Amino acid composition, (c) sequence length search, (d) molecular weight search, (e) identical pattern matching and (f) similar pattern matching. The basic input for (a) and (b) is in a four-letter PDB-id code that usually appears in research papers dealing with the three dimensional macromolecular crystal structures. The search tool displays the corresponding results in an easy and simple format. For example, the amino acid composition search output displays the name of the amino acid (full name, three-letter code and single-letter code), the number of the particular residue and its contribution towards molecular weight. It also displays the percentage content of all the amino acid residues in the input query protein sequence. At the end of the output it displays the total molecular weight of the protein. The search engine requires the lower and upper limits to pick the protein structures, whose lengths satisfy the two-number input (lower and upper bound) criterion for the sequence length search. Similarly, the molecular weight search requires the lower and upper limit values of the molecular weight in Daltons. A sample output of the user-desired lower and upper limit criterion is shown in Figure 2. The amino acid sequence for the proteins and nucleotide sequences for DNA/RNA structures is required as the search input for carrying out identical/similarity pattern matching. Moreover, PSST has an option of the user-desired number of mismatches. For example, in the similarity search the default value for the number of mis-


PSST - 1.0 - Search Results - Netscape
ile Edit View Go Communicator Help
Back
Forward
Reload
Home
Search
Netscape
Print
Security
Shop
Stop
N
Bookmarks
Location: http://144.16.71.10/cgi-bin/molw1.pl
What's Related
Document: Done

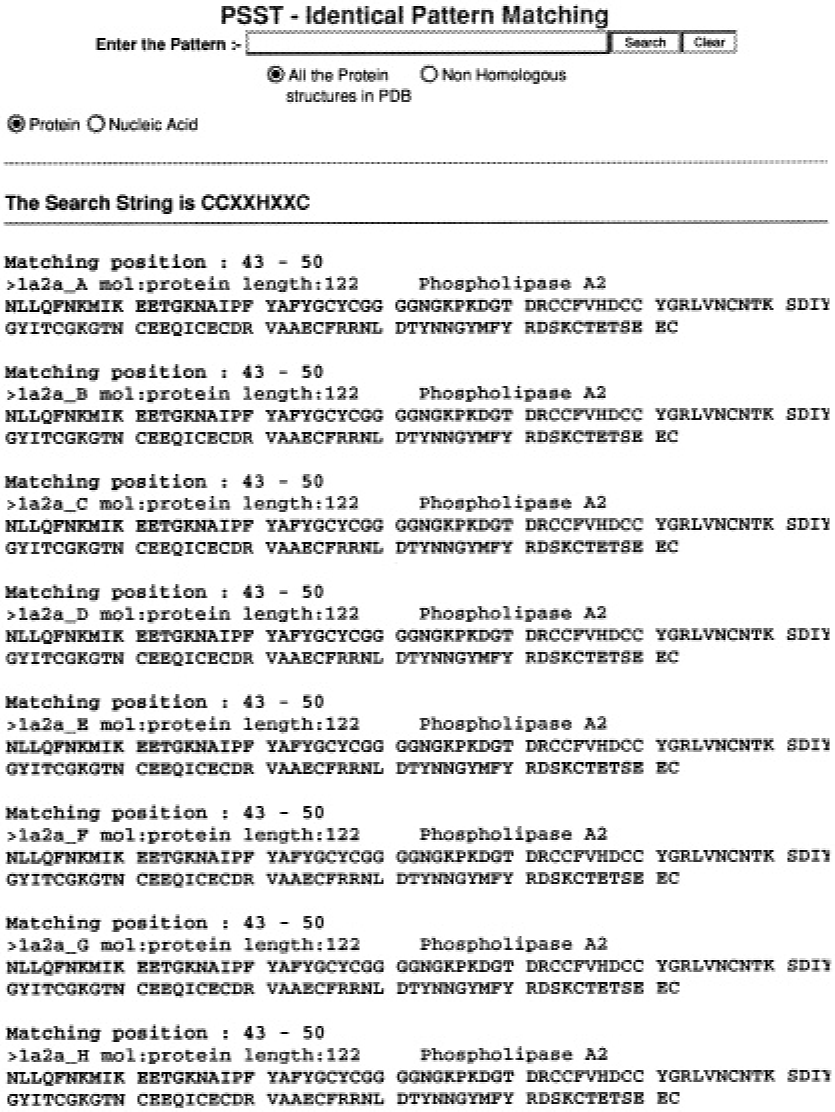
PSST - Molecular Weight Search

Enter the range between 12000 and 12100 Search Clear

○ All the Protein structures in PDB ● Non Homologous

```
                 The given range between 12000 and 12100

>2bjx_A mol:protein length:110       Protein Disulfide Isomerase          12019 Daltons

                    Molecular weight is 12019 Daltons
                  Molecular weight of 2bjx_ = 12.0190 kDa

>1gcm_A mol:protein-het length:34  Gcn4P-II                               4016 Daltons
>1gcm_B mol:protein-het length:34  Gcn4P-II                               4016 Daltons
>1gcm_C mol:protein-het length:34  Gcn4P-II                               4016 Daltons

                    Molecular weight is 12048 Daltons
                  Molecular weight of 1gcm_ = 12.0480 kDa

>7fd1_  mol:protein length:106       7-Fe Ferredoxin I (Fd1)              12001 Daltons

                    Molecular weight is 12001 Daltons
                  Molecular weight of 7fd1_ = 12.0010 kDa
```

Figure 2. Typical output of molecular weight search using the lower and upper limits of the user-desired query from the non-homologous protein sequence database.



The Search String is CCXXHXXC

```
Matching position : 43 - 50
>1a2a_A mol:protein length:122     Phospholipase A2
NLLQFNKMIK EETGKNAIPF YAFYGCYCGG GGNGKPKDGT DRCCFVHDCC YGRLVNCNTK SDIY
GYITCGKGTN CEEQICECDR VAAECFRRNL DTYNNGYMFY RDSKCTETSE EC

Matching position : 43 - 50
>1a2a_B mol:protein length:122     Phospholipase A2
NLLQFNKMIK EETGKNAIPF YAFYGCYCGG GGNGKPKDGT DRCCFVHDCC YGRLVNCNTK SDIY
GYITCGKGTN CEEQICECDR VAAECFRRNL DTYNNGYMFY RDSKCTETSE EC

Matching position : 43 - 50
>1a2a_C mol:protein length:122     Phospholipase A2
NLLQFNKMIK EETGKNAIPF YAFYGCYCGG GGNGKPKDGT DRCCFVHDCC YGRLVNCNTK SDIY
GYITCGKGTN CEEQICECDR VAAECFRRNL DTYNNGYMFY RDSKCTETSE EC

Matching position : 43 - 50
>1a2a_D mol:protein length:122     Phospholipase A2
NLLQFNKMIK EETGKNAIPF YAFYGCYCGG GGNGKPKDGT DRCCFVHDCC YGRLVNCNTK SDIY
GYITCGKGTN CEEQICECDR VAAECFRRNL DTYNNGYMFY RDSKCTETSE EC

Matching position : 43 - 50
>1a2a_E mol:protein length:122     Phospholipase A2
NLLQFNKMIK EETGKNAIPF YAFYGCYCGG GGNGKPKDGT DRCCFVHDCC YGRLVNCNTK SDIY
GYITCGKGTN CEEQICECDR VAAECFRRNL DTYNNGYMFY RDSKCTETSE EC

Matching position : 43 - 50
>1a2a_F mol:protein length:122     Phospholipase A2
NLLQFNKMIK EETGKNAIPF YAFYGCYCGG GGNGKPKDGT DRCCFVHDCC YGRLVNCNTK SDIY
GYITCGKGTN CEEQICECDR VAAECFRRNL DTYNNGYMFY RDSKCTETSE EC

Matching position : 43 - 50
>1a2a_G mol:protein length:122     Phospholipase A2
NLLQFNKMIK EETGKNAIPF YAFYGCYCGG GGNGKPKDGT DRCCFVHDCC YGRLVNCNTK SDIY
GYITCGKGTN CEEQICECDR VAAECFRRNL DTYNNGYMFY RDSKCTETSE EC

Matching position : 43 - 50
>1a2a_H mol:protein length:122     Phospholipase A2
NLLQFNKMIK EETGKNAIPF YAFYGCYCGG GGNGKPKDGT DRCCFVHDCC YGRLVNCNTK SDIY
GYITCGKGTN CEEQICECDR VAAECFRRNL DTYNNGYMFY RDSKCTETSE EC
```

Figure 3. Sample output frame of the consensus pattern search (only first output frame is shown here). The input search string is CCXXHXXC.

matches is 3 and it can be modified while submitting the web form at the user end. The identical and similar patterns blink with a different colour on the screen. The results obtained from the searches correspond to the most recent information available in the PDB[1]. As mentioned earlier, PSST is available on the www and the users can easily submit their queries. In the trial runs, the results appear in about 1–2 min depending on the network speed.

A sample output of the result of a typical search for a sequence pattern in all the proteins is shown in Figure 3. The pattern searched is taken from PROSITE[8] corresponding to a key consensus pattern involved in the function of phospholipase A_2. The pattern is CCXXHXXC (where C is cysteine, X is any amino acid and H is histidine) and histidine is the active site residue. As can be seen from Figure 3 almost all the hits correspond to various structures of phospholipase A_2. This simple search helped in the recognition of all the homologous structures in a family without performing global alignment of complete amino acid sequences. A simple variant of the input sequence pattern could result in the recognition of a non-phospholipase A_2. Depending upon the three-dimensional structure of this hit one might introduce, by site-directed mutagenesis, the residue necessary for the function of phospholipase A_2 in an otherwise non-phospholipase A_2.

The sequence and structure databases are being updated every week from PDB and fully incorporated with the PSST search tool. The search engine is written in Perl, version 5 (ref. 9) and it can be executed on our NT-4.0 Bioinformatics server (a 300 MHz Pentium II processor, 128 Mbytes of main memory). The front end input data part of this tool is written in HTML and JavaScript and allows user-friendly web forms.

1. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. J., *J. Mol. Biol.*, 1997, **112**, 535–542.
2. Kasuya, A. and Thornton, J. M., *J. Mol. Biol.*, 1999, **286**, 1673–1691.
3. McLachlan, A. D. and Stewart, M., *J. Mol. Biol.*, 1976, **103**, 271–298.
4. Etzold, T., Ulyanov, A. and Argos, P., *Methods Enzymol.*, 1996, **266**, 114–128.
5. Bairoch, A. and Apweiler, R., *Nucleic Acids Res.*, 1998, **28**, 38–42.
6. Baker, W. C., Garavelli, J. S., Haft, D. H., Hunt, L. T., Marzec, C. R., Orcutt, B. C., Srinivasarao, G. Y., Yeh, L. S. L., Ledley, R. S., Mewes, H. W., Pfeiffer, F. and Tsugita, A., *Nucleic Acids Res.*, 1998, **28**, 27–32.
7. Hobohm, U. and Sander, C., *Prot. Sci.*, 1994, **3**, 522–524.
8. Bairoch, A., Bucher, P. and Hofmann, K., *Nucleic Acids Res.*, 1997, **25**, 217–221.
9. Wall, L., Christiansen, T. and Schwartz, R. L., *Programming Perl*, Nutshel Handbooks, O'Reilly & Associates Inc., CA, USA, 1996, 2nd edn.
10. Sankaranarayanan, R., Sekar, K., Banerjee, R., Sharma, V. Surolia, A. and Vijayan., M., *Nat. Struct. Biol.*, 1996, **3**, 596–602.
11. Sayle, R. A. and Milner-Whilte, E. J., *Trends Biochem. Sci.*, 1995, **20**, 374–382.

ACKNOWLEDGEMENTS. We gratefully acknowledge the use of the Bioinformatics Centre and the interactive graphics-based molecular modelling facility. These facilities are supported by the Department of Biotechnology (DBT), Government of India.

Received 29 November 1999; revised accepted 12 January 2000

S. SARAVANAN
A. AJMAL KHAN
K. SEKAR*

Bioinformatics Centre,
Raman Building,
Indian Institute of Science,
Bangalore 560 012, India

**Corresponding author*
(e–mail: sekar@physics.iisc.ernet.in)