Analysis, Detection and Prediction of some of the Structural Motifs in Proteins

Kunchur Guruprasad[1]
[1]*Bioinformatics, Centre for Cellular and Molecular Biology (CCMB), Uppal Road, Hyderabad – 500 007, INDIA.*
Email : guru@ccmb.res.in

ABSTRACT

We are generally interested in the analysis, detection and prediction of structural motifs in proteins, in order to infer compatibility of amino acid sequence to structure in proteins of known three-dimensional structure available in the Protein Data Bank. In this context, we are analyzing some of the well-characterized structural motifs in proteins. We have analyzed simple structural motifs, such as, β-turns and γ-turns by evaluating the statistically significant type-dependent amino acid positional preferences in enlarged representative protein datasets and revised the amino acid preferences. In doing so, we identified a number of "unexpected" isolated β-turns with a proline amino acid residue at the (i+2) position. We extended our study to the identification of multiple turns, continuous turns and to peptides that correspond to the combinations of individual β and γ-turns in proteins and examined the hydrogen-bond interactions likely to stabilize these peptides. This led us to develop a database of structural motifs in proteins (DSMP) that would primarily allow us to make queries based on the various fields in the database for some well-characterized structural motifs, such as, helices, β-strands, turns, β-hairpins, β-α-β, ψ-loops, β-sheets, disulphide bridges. We have recently implemented this information for all entries in the current PDB in a relational database called O-DSMP using Oracle9i that is easy to update and maintain and added few additional structural motifs. We have also developed another relational database corresponding to amino acid sequences and their associated secondary structure for representative proteins in the PDB called PSSARD. This database allows flexible queries to be made on the compatibility of amino acid sequences in the PDB to "user-defined" super-secondary structure conformation and vice-versa. Currently, we have extended this database to include nearly 23,000 protein crystal structures available in the PDB. Further, we have analyzed the 'structural plasticity' associated with the β-propeller structural motif. We have developed a method to automatically detect β-propellers from the PDB codes. We evaluated the accuracy and consistency of predicting β and γ-turns in proteins using the residue-coupled model. I will discuss results of our work and describe databases and software applications that have been developed.

1. INTRODUCTION

One of the interesting problems in structural biology is to understand 'the rules' that determine protein folding - from primary to tertiary structure. Although this remains an unsolved problem yet, much can be learnt from the analysis of a number of protein sequences and their corresponding three-dimensional structures that are deposited in public databank repositories, such as, SWISSPROT [1], EMBL [2] and the Protein Data Bank (PDB) [3]. With increase in the number of protein sequence and structure data resulting from world-wide whole genome sequencing and structural genomics projects, respectively, we felt that it would be useful to have a web-based database relating amino acid sequence to structure "at the level of well-characterized structural motifs in proteins" along with other useful details that can be queried. Few such databases that are available; BIPED [4], SESAM [5], IDITIS [6] either belong to the pre 'world wide web' era or are commercially available or limited only to the recognition of protein secondary structure. We therefore generated a Database of Structural Motifs in Proteins (DSMP) [7]. The data in DSMP was primarily extracted using the PROMOTIF program [8] from the PDB and implemented as a database using the flat-file Sequence Retrieval System (SRS) [9]. The data corresponding to the structural motifs includes; amino acid sequence, position in polypeptide chain, geometrical parameters, classification type, unique code, keywords, resolution of crystal structure and three-dimensional co-ordinates. Using features in SRS, DSMP could be queried to extract information from one or more structural motifs. We have now implemented the data corresponding to all entries in the PDB in a relational database O-DSMP (unpublished).

For the analysis of structural motifs in proteins, we chose 'turns' as they are the simplest and structurally and are often functionally important and present in a number of proteins. Turns are essentially of two types; β-turns and γ-turns. A β-turn consists four consecutive residues defined by positions i, i+1, i+2, i+3 which are not present in a α-helix; the distance between $C_\alpha(i)$ and $C_\alpha(i+3)$ is less than 7.0 Å [10, 11]. The β-turns have further

been classified into nine different types (I, II, VIII, I', II', VIa1, VIa2, VIb, IV) based on dihedral angle values (ϕ,ψ) corresponding to the $(i+1)^{st}$ and $(i+2)^{nd}$ residues in the turn [12]. A γ-turn comprises three consecutive residues at positions i, i+1, i+2 defined by the existence of a main-chain hydrogen bond between CO group of the i^{th} residue and NH group of the $(i+2)^{nd}$ residue [11]. γ-turns are also further classified into two types; classic and inverse, based on dihedral angle values corresponding to the $(i+1)^{st}$ residue [11]. During the last decade or so, the number of β-turns has nearly doubled and the number of γ-turns has increased more than seven times in a representative dataset corresponding to proteins in the PDB. We were therefore interested in the analysis of the amino acid preferences at individual positions in the different turn types [13]. We were also keen to examine whether isolated β-turns with a proline amino acid residue at the "unexpected" (i+2) position may be present in proteins and if so what are the possible interactions that stabilize them [14]. The occurrence of β-turns as multiple turns and for the γ-turns as 'compound gamma-turns' is reported earlier [12, 15]. We intended to provide a simple nomenclature to refer to the multiple turns and to analyze in the enlarged representative protein dataset [16]. This led us to explore whether turns may occur consecutively, i.e., one followed by the other without sharing any residues in common [17]. When we saw a number of such instances in the PDB, we extended our study to the analysis of combinations of turn types in the PDB [18]. All these analyses led us to the development of a database of structural motifs that relate amino acid sequence to structure that also included some of the other well-characterized structural motifs in proteins [7] described in the earlier part of this introduction. The β-propeller structural motif is observed in a number of protein tertiary structures and associated with different functions. We intended to evaluate whether the β-propeller architecture may be associated with 'structural plasticity' [19]. We do not currently have access to databases in the public domain, where given a "user-defined" super-secondary structure conformation corresponding to a protein of known three-dimensional structure in the PDB, it is possible to derive the amino acid sequences that are compatible and vice-versa. We therefore developed the Protein Sequence Structure Analysis Relational Database (PSSARD) [20]. The database would be useful for studies related to protein sequence-structure analysis, prediction, modeling and design. Further, we also intended to develop a method and a computer program that could automatically 'detect' β-propellers from the PDB [21], just like there are methods to detect some of the other structural motifs, such as, β-barrels, greek-keys, and so on. For the prediction of structural motifs we chose the β- and γ-turns [22, 23].

2. METHODS

The representative protein dataset referred here were selected according to the PDB_SELECT program [24] and available at the website http://homepages.fh-giessen.de/~hg12640/pdbselect/recent.pdb_select25. A 25% pairwise sequence identity cut-off value corresponding only to the protein crystal structures in the PDB refined at better than 2.0 Angstrom resolution constituted the non-redundant representative dataset. The β and γ-turns were extracted from the PDB using the PROMOTIF program. The conformational potentials, positional potentials and the turn-type dependent positional potentials for these turns were calculated as described in Hutchinson and Thornton [12]. The potentials were examined for statistical significance by the *d*-test (based on normal distribution) described by Wilmot and Thornton [25]. The hydrogen-bond interactions described in Overington [26] were examined in order to deduce significant interactions for new amino acid preferences. We developed our own programs to identify the multiple turns [16], continuous turns [17] and combinations of turns [18] based on the turns data derived using the PROMOTIF program. The database of structural motifs in proteins (DSMP) [7] contains data primarily extracted using the PROMOTIF program and implemented as a web-based network service using the Sequence Retrieval System (SRS). Using features in SRS, DSMP can be queried to extract information from one or more structural motifs. The PSSARD database was generated by extracting the sequence and secondary structure information by following the 'sequence details' links at the PDB site http://www.rcsb.org. The coil region was represented by the letter 'C' and amino acid residues not determined in the crystal structure by a '-'. A summary of the secondary structure was generated by examining contiguous secondary structure conformation along the protein chain for each PDB entry. The different β-propellers and their types were according to the structural classification of proteins (SCOP) extracted from the website (http://www.scop.mrc.cam.uk). This dataset was used for the analysis of β-propeller 'structural plasticity'.

In order to develop a method to detect β-propellers in the PDB, we have generated a 'β-propeller signature code' for all the known β-propellers. It essentially corresponds to the number of amino acid residues associated with individual β-strands that is associated with blades of the β-propeller and the number of amino acid residues between the β-strands. The β-propeller location in

the protein chain was identified by following the protein chain link in the PDBsum [27] website available at http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin-pdbsum. The PDBsum was also used to extract the number of amino acid residues associated with individual β-strands corresponding to the blades by following the ProMotif link for the β-strands and to derive the β-sheet associated β-strand pattern discussed later. Next the C_α-atom to C_α-atom distances corresponding to amino acid residues at ends of the individual β-strands in blades of the β-propeller was evaluated from coordinates in the PDB. Given a PDB code, the equivalent amino acid sequence is generated. The length of sequence equivalent to a β-propeller signature code is mapped on to the sequence and the distance criterion for the β-strand end amino acid residues in the blades is evaluated by comparing with expected values. If all distances are satisfied then the equivalent region in the query protein corresponds to a β-propeller and its location in the protein is reported along with the other structurally similar β-propellers. Otherwise, the amino acid sequence is slid by one residue at a time and the entire process is repeated until end of the protein chain is reached.

In order to evaluate the 'prediction' of β-turns using the residue-coupled model and to extend it to the prediction of γ-turns, we followed the method described by Chou [28]. We used the original protein dataset by Chou comprising 30 proteins for evaluating the predictions. In our work, we used two additional representative protein datasets that correspond to 425 and 619 protein chains that were selected by using the PDB_SELECT program at different time intervals. Further, in order to evaluate the predictions, we also tested on a jackknife dataset comprising 476 protein chains using potentials derived from the training dataset comprising 425 protein chains. Likewise, for the prediction of γ-turns we used two representative protein datasets comprising 315 and 434 protein chains that were again selected using the PDB_SELECT at different time intervals. By excluding 93 proteins common to these two datasets we generated two mutually exclusive datasets comprising 222 and 341 protein chains that were used to carry out jackknife test for the predictions.

3. RESULTS & DISCUSSION

3.1 Analysis of turns, β-propellers & generation of databases corresponding to some of the well-characterized structural motifs in proteins

Nearly 30% amino acids corresponding to a representative PDB dataset constitute the β-turns in proteins and only 3.4% amino acid residues constitute the γ-turns. A revised set of statistically significant amino acid preferences at individual positions for the different turn types were identified [13]. We further observed where NH(i+3)→CO(i) type hydrogen bond is not observed, such β-turns are possibly stabilized through mainchain – sidechain, sidechain – sidechain or mainchain and sidechain interactions with water molecules in the protein. Further, a number of "isolated" β-turns with a proline at the (i+2) position contrary to expectations were observed with predominant solvent interactions surrounding this proline amino acid residue [14]. We observed that the Ramachandran ϕ,ψ dihedral angle space for proline amino acid residue at the unusual (i+2) position that corresponds to a proline amino acid residue at the commonly observed (i+1) position in an overlapping β-turn is restricted to ϕ=(-112.5, -45.0) and ψ=(-67.5, 45.0). However, for proline at the (i+2) position in an "isolated" β-turn or for a proline residue at the (i+2) position in a multiple β-turn where this proline does not correspond to (i+2) proline of the overlapping turn, there is a distinct additional ϕ,ψ preference in several turns for ϕ=(-67.5, 45.0) and ψ=(112.5, 180.0). We refer to the multiple turns involving a γ- or a β-turn and sharing one, two or three amino acid residues in common as belonging to one of the 4 types; γβ, βγ, γγ or ββ. Also depending upon the number of overlapping residues, the multiple turns are classified as $(\gamma\beta)_{1,2,3}$ or $(\beta\gamma)_{1,2,3}$ [16] whereas for the γγ or ββ multiple turns, we follow the nomenclature proposed earlier [12]. We have identified the statistically significant amino acid preferences at individual positions when turns occur as multiple turns [16]. Peptides of equivalent length show distinct amino acid positional preferences. In multiple ββ turns the NH(i+2)→CO(i) is more predominant. Continuous turns are also observed in protein tertiary structure [17]. The γβ/βγ, γγ and ββ continuous turns represent peptides corresponding to either 7, 6 or 8 amino acid residues in length, respectively, but with varying conformations. The γβ is frequently observed between a coil and strand, βγ between helix and strand, γγ within coil and ββ either between strands or between a strand and coil or within coil. Continuous turns are relatively few compared to the multiple turns, but both may be recognized as components of protein super-secondary structure. The β and γ-turns combine as multiple or continuous turns to form combinations of turns that span large segments of the protein polypeptide chain [18]. Around 475 peptides that correspond to combinations of turns resulted from the analysis of a non-redundant protein dataset comprising 248 high resolution crystal structure data selected from the PDB. For example, the maximum peptide length comprises 15 amino acid residues observed in the electron transport protein

(PDB code:1i80A) between amino acid residue numbers 28 and 42. This peptide represents combinations of only type II β-turns. Peptides corresponding to the combination of turns most frequently observed in the PDB comprise 9 and 10 amino acid residues.

We observed that the β-propeller architecture is characterized by 'structural plasticity' that corresponds to the number of β-strands associated with blades of the β-propeller, the number of amino acid residues associated with equivalent β-strands in the different blades and the presence of α-helices and twisted β-strands. Accordingly, we proposed a β-sheet associated β-strand pattern for the β-propellers. The type 6- and 7-bladed β-propellers known to be associated with sequence and functional diversity is more common and associated with relatively more structural variations as compared to the other β-propeller types [19].

The Database of Structural Motifs in Proteins - DSMP (version 1.0) reported in the year 2000 corresponds to 10,213 PDB entries (PDB release 89). It includes data relevant to helices, turns, β-hairpins, ψ-loops, β-α-β motifs, β-strands, β-sheets and disulphide bridges primarily extracted using PROMOTIF and some of our own computer programs and implemented using the Sequence Retrieval System – SRS. It also contains three-dimensional coordinates corresponding to the structural motifs extracted from the PDB that is useful for graphics visualization. Later on we updated the database to include more recent entries from the PDB and added the multiple turns and continuous turns, β-barrels, coils (unpublished). The data corresponding to each structural motif includes; protein in which the motif is present, a unique code, location in the polypeptide chain, amino acid sequence, geometrical parameters, classification type, keywords and crystal structure resolution. The DSMP can be queried based on any or combination of above fields using features in SRS to extract information from one or more structural motifs. We are implementing the corresponding data for the most recent entries corresponding to crystal structures in the PDB (~23,000 entires) in Oracle9i relational database called O-DSMP that allows flexible queries to be made and is also easy to update and maintain. The homepage (under development) is shown in Figure 1. Amino acid sequences corresponding to the structural motifs that can adopt different conformation or vice-versa can be carried out in addition to querying on other fields described above. In another more recent database, the Protein Sequence Structure Analysis Relational Database – PSSARD that we have implemented in Oracle9i, it is possible to retrieve amino acid sequences from several proteins simultaneously that are compatible to a "user-defined" super-secondary structure or vice-versa. The information content in PSSARD includes the protein description, PDB code, crystal structure resolution, total number of amino acid residues in the protein chain, amino acid sequence, secondary structure conformation and its summary. The database is freely accessible from the website (http://203.200.217.185:8000/rdpssa/index.htm). A sample PSSARD output is shown in Figure 2. We have updated this to include all 23,000 protein PDB entries (unpublished) (http://203.200.217.185:8000/pssafullrd/index.htm). Our work has implications for protein sequence-structure-function analysis, protein structure prediction, modeling and design.

3.2 Method and computer software application for the detection of β-propellers in proteins

Our β-propeller detection (BPD) method [21] automatically detects β-propellers from the PDB(s), identifies the location of the β-propeller in the protein structure, specifies the β-propeller type, the β-sheet associated β-strand pattern and the structurally similar β-propellers observed in other proteins in the PDB. When tested on 21,566 proteins in the PDB, the BPD method was capable of correctly identifying all the 245 known β-propellers described in the structural classification of proteins (SCOP) with less than 0.2% false positives detected. Most β-propellers can be detected by the 56 representative β-propellers identified by the BPD method based on an exhaustive all to all β-propeller search. The BPD method compares with some of the popular web-based programs that can automatically detect 'structural similarities' between the query and target proteins. Our method has the advantage of also being capable of detecting β-propellers associated with 'structural plasticity' and in situations where the target and query proteins differ in amino acid sequence length. The software that can detect β-propellers given a list of PDB codes is available from the authors and the web-based application for detecting the β-propellers given one PDB code at a time is accessible from the website at http://203.200.217.184:8080/cgi-bin/bp_detection.pl. The homepage for this application is shown in Figure 3 and sample output of the results in Figure 4. We are currently extending this work to the 'prediction' of β-propellers from amino acid sequence and also to the 'detection' of β-propellers with β-propeller signature code similar but not identical to those used in the current 'training' dataset.

3.3 Prediction of turns in proteins

We observed that there has been an over prediction of the β-turns using a dataset of 30 proteins and the residue-coupled model reported earlier. Our own work using the residue-coupled model and different protein datasets described in methods suggest that the percentage accuracy of β-turn predictions is ~68% and is also consistent when applied on the other datasets or on a jackknife dataset [22]. Likewise, extending our analysis to the prediction of γ-turns using the residue-coupled model and different representative protein datasets we observed that it is ~57% [23].

REFERENCES

[1] A. Bairoch, B. Boechmann. The SWISS-PROT protein sequence databank. Nucleic Acids Res., 19:2247--2249, 1991.

[2] P. J. Stroehr, G. N. Cameron. The EMBL data library. Nucleic Acids Res., 19:2227--2230, 1991.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. The Protein Data Bank. Nucleic Acids Res., 28:235--242, 2000.

[4] J. M. Thornton, S. P. Gardner. Trends Biochem. Sci., 14:300—304, 1989.

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

[6] S. Gardner, J. Thornton. Iditis: protein structure database. Acta Crystallogr. D Biol. Crystallogr. 54:1071-1077, 1998.

[7] K. Guruprasad, M. S. Prasad, G. R. Kumar. Database of structural motifs in proteins. Bioinformatics, 16(4):372—5, 2000.

[8] E. G. Hutchinson, J. M. Thornton. PROMOTIF—a program to identify and analyze structural motifs in proteins. Protein Sci., 5:212—220, 1996.

[9] T. Etzold, A. Ulyanov, P. Argos. SRS: Information retrieval system for molecular biology data banks. Methods Enzym., 266:114—128, 1996.

[10] J. S. Richardson. The anatomy and taxonomy of protein structure. Adv. Protein Chem., 34:167—339, 1981.

[11] G. D. Rose, L. M. Gierasch, J. A. Smith. Turns in peptides and proteins. Adv. Protein Chem., 37:1—109, 1985.

[12] E. G. Hutchinson, J. M. Thornton. A revised set of potentials for beta-turn formation in proteins. Protein Sci., 3:2207—2216, 1994.

[13] K. Guruprasad, S. Rajkumar. β and γ -turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials. J Biosci., 25(2):143—156, 2000.

[14] K. Guruprasad, M. N. Pavan, S. Rajkumar, S. Swaminathan. Isolated and multiple β -turns with proline in the third position. Current Science, 79(7):992—994, 2000.

[15] E. J. Milner-White. Situations of gamma-turns in proteins. Their relation to alpha-helices, beta-sheets and ligand binding sites. J. Mol. Biol., 216:386—397, 1990.

[16] K. Guruprasad, M. S. Prasad, G. R. Kumar. Analysis of βγ, βγ, γγ, ββ multiple turns in proteins. J Pept Res., 56(4):250—63, 2000.

[17] K. Guruprasad, M. S. Prasad, G. R. Kumar. Analysis of βγ, βγ, γγ, ββ continuous turns in proteins.
J Pept Res., 57(4):292—300, 2001.

[18] K. Guruprasad, M. J. Rao, S. Adindla, L. Guruprasad. Combinations of turns in proteins. J Pept Res., 62(4):167--74, 2003.

[19] K. Guruprasad, P. Dhamayanthi. Structural plasticity associated with the β -propeller architecture. Int J Biol Macromol., 34(1-2):55—61, 2004.

[20] K. Guruprasad, K. Srikanth, A. V. N. Babu. PSSARD: Protein Sequence-Structure Analysis Relational Database. Int J Biol Macromol., (2005, in press)

[21] K. Guruprasad, K. Archana. The automatic detection of known beta-propeller structural motifs from protein tertiary structure. Int J Biol Macromol., (2005, in press)

[22] K. Guruprasad, S. Shukla. Prediction of β -turns from amino acid sequences using the residue-coupled model. J Pept Res., 61(4):159—62, 2003.

[23] K. Guruprasad, S. Shukla, S. Adindla, L. Guruprasad. Prediction of γ -turns from amino acid sequences. J Pept Res., 61(5):243—51, 2003.

[24] U. Hobohm, C. Sander. Enlarged representative set of protein structures. Protein Sci., 3:522—524, 1994.

[25] C. M. Wilmot, J. M. Thornton. Analysis and the prediction of the different types of beta-turn in proteins. J. Mol. Biol., 203:221—232, 1988.

[26] J. Overington, M. S. Johnson, A. Sali, T. L. Blundell. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. R. Soc. London B. Biol. Sci., 241:--132-145, 1990.

[27] R. A. Laskowski. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res., 29:221--222, 2001.

[28] K. C. Chou. Prediction of β-turns. J. Peptide Res., 49:120—144, 1997.

ACKNOWLEDGEMENTS

I would like to thank the DST, DBT and the CSIR, India for supporting my participation at the BIOINFO 2005 meeting. I would also like to thank all past co-workers and my present colleagues, particularly, V. JaiKarthik, N. Kumudini, T. Lavanya, A.V. Naresh Babu, Beena Popuri and S. Savitha for their assistance.

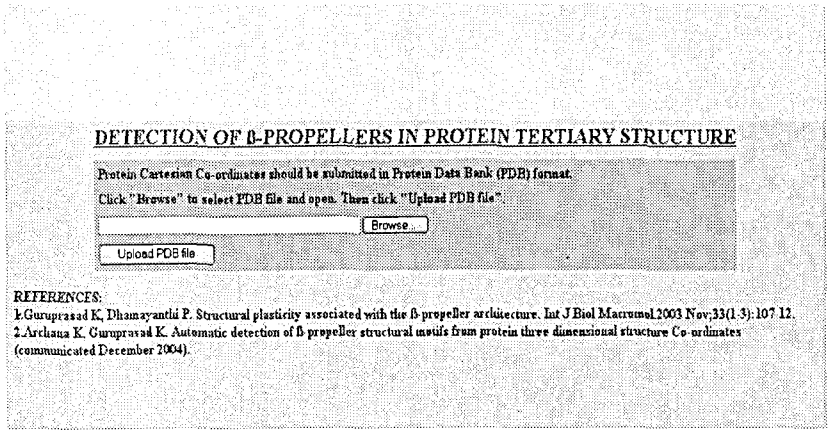

Figure 3: Betapropeller detection hompage
(http://203.200.217.184:8080/cgi-bin/bp_detection.pl)

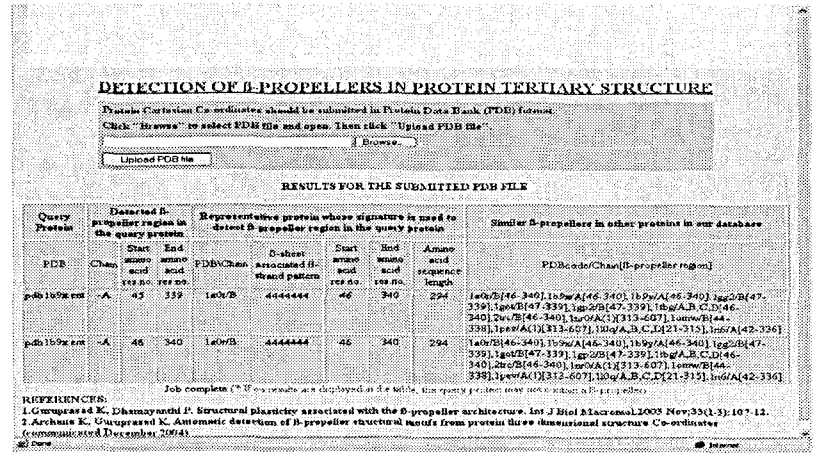

Figure 4: Betapropeller detection results

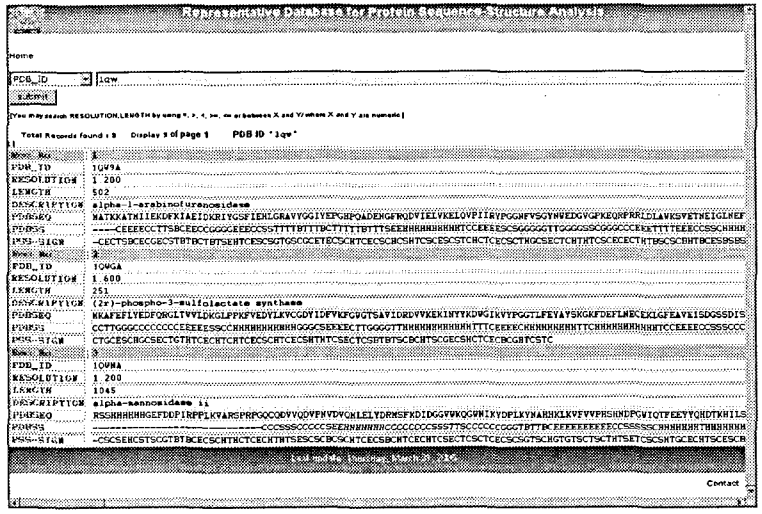

Figure 2: PSSARD
(http://203.200.217.185:8000/rdpssa/index.htm)

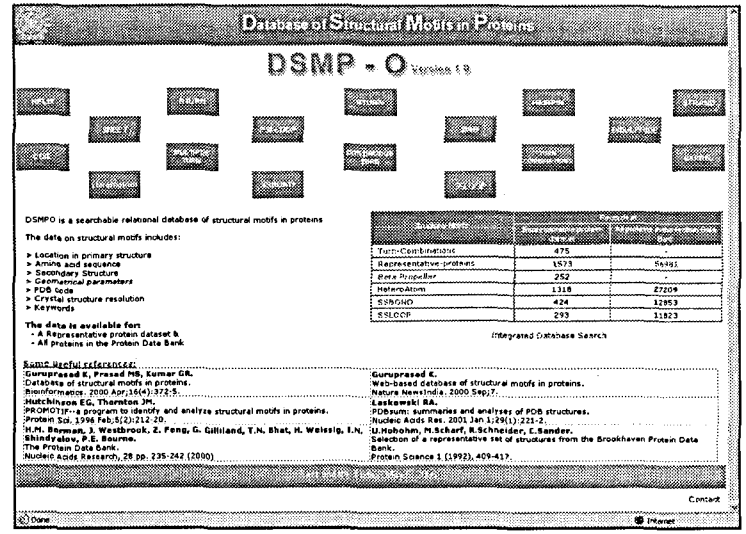

Figure 1: DSMP-O (under development)
(http://203.200.217.185:8000/dsmpo/index.htm)