Bioinformation
www.bioinformation.net

open access
Database

ExSer: A standalone tool to mine protein data bank (PDB) for secondary structural elements

Dhandapani Vignesh, Paul Daniel, Natarajan Raja, Ponnusamy Balasubramanian, Loganathan Arul *

Department of Plant Molecular Biology and Biotechnology, Centre for Plant Molecular Biology, Tamil Nadu Agricultural University, Coimbatore - 641 003, India; Loganathan Arul, Email: arulsra@gmail.com, *corresponding author

Received April 10, 2009; Revised February 03, 2010; accepted March 03, 2010; published March 31, 2010

Abstract:
Detailed structural analysis of protein necessitates investigation at primary, secondary and tertiary levels, respectively. Insight into protein secondary structures pave way for understanding the type of secondary structural elements involved (α-helices, β-strands etc.), the amino acid sequence that encode the secondary structural elements, number of residues, length and, percentage composition of the respective elements in the protein. Here we present a standalone tool entitled "*ExSer*" which facilitate an automated extraction of the amino acid sequence that encode for the secondary structural regions of a protein from the protein data bank (PDB) file.

Availability: *ExSer* is freely downloadable from http://code.google.com/p/tool-exser/

Keywords: Protein structure, secondary structural elements, α-helix, β-strand

Background:
Protein structures are determined by experimental methods which include NMR, X-ray and Electronic Microscopy. The experimentally solved protein structures are archived at the publicly available databases such as, Protein Data Bank (PDB) **[1]**. Protein Data Bank (PDB) is an internationally referred protein structure database, it comprises of the atomic coordinates of the three dimensional (3D) structure of proteins **[2]**. The coordinate file is organized in a common file format, called the PDB format **[3]**. Amino acid sequence, secondary structural information and atom positions are organized systematically in the PDB file enabling machine learning **[4]**. Many molecular visualization tools (RasMol, JMol, PyMOL etc.) have been developed based on the PDB format leading to visualization of the 3D structure of proteins from the atomic coordinates. The information pertaining to the fold class at the tertiary level and the arrangement of secondary structural elements (motifs) at the lower levels in the structural hierarchy could directly be inferred from the 3D structure. However, information that are embedded in the PDB file such as the amino acid sequence of the secondary structural elements viz., α-helices and β-strands, is of significant interest to the scientific community. Knowing the amino acid sequence of the secondary structures throw light on the biochemical and structural properties of the residues which inturn governs the formation of α-helices and β-strands in proteins. It furthers understanding on the relationship between proteins and their possible evolutionary origins, the relationship between the amino acid sequence and the tertiary structure and, structure-function analysis. To the best of our understanding, the amino acid sequence is manually extracted based on the identifier(s) that are associated with the secondary structural elements in PDB format. Towards an automated extraction of the amino acid sequence encoding the secondary structural regions, we developed a standalone tool entitled *ExSer* (**Ex**tract the amino acid **Se**quences of secondary structu**r**es). *ExSer* decipher the amino acid sequence encoding the secondary structural elements (helices and strands), number of residues, position and their percentage composition, given the PDB coordinate file as query.

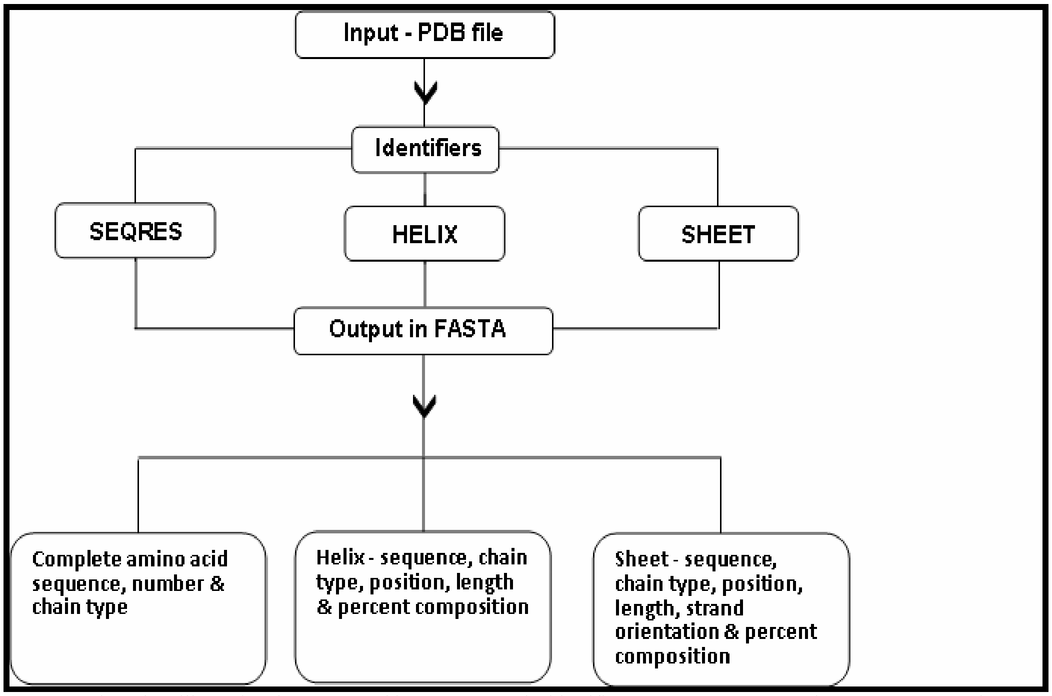
Input - PDB file
Identifiers
SEQRES
HELIX
SHEET
Output in FASTA
Complete amino acid sequence, number & chain type
Helix - sequence, chain type, position, length & percent composition
Sheet - sequence, chain type, position, length, strand orientation & percent composition

Figure 1: Flow chart illustrating the *ExSer* working processes. The identifiers, SEQRES, HELIX and SHEET are as per the PDB format

Implementation:

ExSer was developed using Practical Extraction and Report Language (PERL) **[5]** and Tool kit (Tk) **[6]**. PERL is used for scripting and Tk for designing graphical user interface (GUI). *ExSer* provide the option to either select a PDB file via the browse button or to paste the structural information in the PDB format. Firstly, the tool extract the entire amino acid sequence of the polypeptide chain(s) based on the identifier "SEQRES" defined in the PDB format (latest ver. 3.20) **[3]**. Simultaneously it picks up the start and end positions of the helices and strands using "HELIX" and "SHEET" identifiers mentioned in the PDB file. Finally, based on the above positions the amino acid sequences of helices and strands were extracted from the PDB coordinate file **(Figure 1)**. The *ExSer* output comprise of, the complete amino acid sequence of the polypeptide chain(s) (likely A, B, C etc.) that make the protein structure, the amino acid sequences of secondary structural regions (helices and strands), the number of residues involved, the start and end positions of the secondary structural regions and their percentage composition with respect to the total protein. Sub-classification in the helices such as right-handed alpha, omega, pi, gamma, 3_{10} and left-handed alpha, omega, gamma etc., finds mention in the output. Similarly, the parallel and anti-parallel nature of the strands with respect to the previous strand is indicated. The extracted amino acid sequence finally appears in the FASTA format as output and can be saved in a separate text file, as well.

Caveats and future development:

ExSer had been validated with more than 300 coordinate files downloaded from PDB. Since, turns finds no explicit mention in the PDB format (ver. 3.20), the tool cannot retrieve the amino acid sequence corresponding to the turns in protein structure. This tool can work on all windows operating systems including windows vista. In future, *ExSer* will be available for Linux and Macintosh operating systems.

Acknowledgement:

This tool was developed at the Bioinformatics centre, Tamil Nadu Agricultural University with the financial support of the Department of Biotechnology (DBT), Government of India under the Biotechnology Information System (BTIS) program.

References:

[1] HM. Berman, *et al*., *Nucl. Acids Res.*, (2000) **28:** 235 [PMID: 10592235]
[2] K Henrick, *et al*., *Nucl. Acids Res.*, (2008) **36:** D426 [PMID: 18073189]
[3] http://www.wwpdb.org/documentation/format32/v3.2.html
[4] S. R. Hall, *et al*., *J. Chem. Inf. Comp. Sci.*, (1991) **31**: 326
[5] http://www.perl.org/
[6] http://www.perltk.org/

Edited by P. Kangueane
Citation: Vignesh *et al.* Bioinformation 4(9): 390-391 (2010)