ELSEVIER

International Journal of Biological Macromolecules 36 (2005) 259–262

INTERNATIONAL JOURNAL OF
Biological
Macromolecules
STRUCTURE, FUNCTION AND INTERACTIONS

www.elsevier.com/locate/ijbiomac

Short communication

PSSARD: Protein sequence-structure analysis relational database

Kunchur Guruprasad *, K. Srikanth, A.V.N. Babu

Centre for Cellular and Molecular Biology (CCMB), Uppal Road, Hyderabad 500007, Andhra Pradesh, India

Received 13 April 2005; received in revised form 14 June 2005; accepted 14 June 2005
Available online 27 July 2005

Abstract

We have implemented a relational database comprising a representative dataset of amino acid sequences and their associated secondary structure. The representative amino acid sequences were selected according to the PDB_SELECT program by choosing proteins corresponding to protein crystal structure data deposited in the protein data bank that share less than 25% overall pair-wise sequence identity. The secondary structure was extracted from the protein data bank website. The information content in the database 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 mentioned below and is useful to query on any of the above fields. The database is particularly useful to quickly retrieve amino acid sequences that are compatible to any super-secondary structure conformation from several proteins simultaneously.

Keywords: Protein sequence-structure analysis; Representative protein datasets; Protein structure prediction; Protein modelling; Protein design; Protein fold

1. Introduction

One of the interesting problems in structural biology is to understand ‘the rules’ that determine the folding of protein from primary structure into three-dimensional 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 [1] that are deposited in public protein sequence databank repositories, such as SWISSPROT [2] and EMBL [3] and the protein structure databank, protein data bank (PDB) [4,5]. The need to establish derived databases from primary databanks described above became necessary when several protein sequences and structure motifs were recognized that were common to many different protein families for their use in modeling and prediction [6,7]. One of the early such database that could be queried was BIPED [6] that incorporated information related to the protein secondary structure location and solvent accessibilities. Later, SESAM [8], a relational database, was implemented under the SYBASE commercial package. SESAM allows full integration of raw data on protein structure, sequence, ligands and heterogroups, obtained from the PDB with pure sequence information available from SWISS-PROT. It contains descriptions of structural and topological properties as well as survey results. It includes a molecular dictionary and complete description of geometric properties and energy parameters used for validating structural data and flagging errors and incomplete PDB entries and for modeling and conformational energy calculations. A protein secondary structure database (PSS) was designed to correlate the protein sequence database with the atomic co-ordinates of the PDB [9]. However, these databases belonged to the pre-‘world wide web’ era and could not easily be accessed or were limited to the recognition of secondary structure, or were only commercially available, such as the IDITIS database [10] that was designed to allow range of queries to be performed on all available protein structures. With increase in the number of protein sequences and structure data resulting from world wide whole genome sequencing and structural genomic projects, respectively, we felt that it would be useful

* Corresponding author. Tel.: +91 40 2719 2779; fax: +91 40 2716 0591/0311.
E-mail address: guru@ccmb.res.in (K. Guruprasad).
URL: http://203.200.217.185:8000/rdpssa/index.htm.

0141-8130/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijbiomac.2005.06.005

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. We, therefore, generated the database of structural motifs in proteins (DSMP) [11]. The DSMP contained data related to the protein secondary structures: helices, β-strands and turns, and some well-characterized protein structural motifs: β-hairpins, β–α–β, ψ-loops, β-sheets and disulphide bridges that could automatically be extracted from the PDB primarily based on the PROMOTIF program [12] and some of our own computer programs and implemented as a network service using the sequence retrieval system (SRS) [13]. 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 that may be useful for sequence-structure analysis, prediction modeling or design. Subsequently, we have extracted the multiple turns [14], continuous turns [15], combinations of turns [16], β-propellers [17] and several other structural motifs from proteins of known three-dimensional structure and implemented these in the database that will be communicated separately.

The knowledge of amino acid sequences that can adopt a repertoire of secondary or super-secondary structure motifs in proteins would aid recognition of structure from sequence and thus be useful to protein structure prediction, modeling and design. Currently, we do not 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, it is possible to derive the corresponding amino acid sequences from several proteins with identical super-secondary structure conformation. Or alternatively, given an amino acid sequence and to be able to detect the corresponding secondary or super-secondary structure conformation(s) and their locations along the protein polypeptide chain, we believe that data storage and information retrieval at the levels of both whole protein sequences and structural motifs is both important in order to obtain useful protein structural bio-informatics. For instance, with the DSMP database [11], it should be possible to identify the specific well-characterized structural motifs available therein, whereas from the database (PSSARD) presented in this work, it should be possible to query on any type of "user-defined" super-secondary structural conformation. We have, therefore, generated a relational database that is based on the primary data in the PDB and the corresponding amino acid sequences available at the website (http://www.rcsb.org/pdb) and we have included further useful details corresponding to each protein PDB entry. The web-based database presents a simple and easy to use facility that is aimed to help researchers to make flexible queries using either amino acid sequences or secondary/"user-defined" super-secondary structure conformation and retrieve appropriate information. The database would be useful for studies related to protein sequence-structure analysis, prediction, modeling and design.

2. Materials and methods

The PDB_SELECT program [18] provides a list of representative protein chains that share less than 25% overall pair-wise sequence identity for proteins of known three-dimensional structure. This list, available at the website http://homepages.fh-giessen.de/~hg12640/pdbselect/recent.pdb_select25, was used to obtain the relevant protein three-dimensional structure co-ordinates from the protein data bank. Only, protein structures determined by X-ray crystallography technique were used in the present analysis. The protein description, crystal structure resolution and the total number of amino acid residues in the protein chain were according to PDB_SELECT. The amino acid sequence and the secondary structure conformation corresponding to individual amino acid residues in the protein chain were extracted from the PDB site by following the 'sequence details' link under the 'structure explorer' for the individual PDB files. The sequence of each chain is represented in the standard single-letter amino acid code format. This also includes amino acid residues that may have been truncated in the sample preparation or which were otherwise invisible in the electron density map. The secondary structure is calculated and described according to an implementation of the method of Kabsch and Sander [19]. The secondary structure assignments are: H, helix; B, residue in isolated beta bridge; E, extended beta strand; G, 3_{10} helix; I, pi helix; T, hydrogen-bonded turn; S, bend. We wrote our own computer programs to determine the amino acid residues that have not been defined in the crystal structure by comparing with the corresponding complete amino acid sequence and to represent the secondary structure corresponding to such amino acid residues by a '-' against the PDBSS field in the database. The secondary structure corresponding to amino acid residues in 'coil' conformation is represented by the letter 'C'. We also generated a "protein secondary structure signature" designated as PSS-SIGN in PSSARD that corresponds to the protein secondary structures in the individual protein chains. The PSS-SIGN represents a contiguous stretch of amino acid residues in a particular secondary structure conformation (according to PDBSS) by its equivalent single letter. For instance, the PSS-SIGN for the protein defensin HNP-3 A-chain (PDB code: 1DFNA) with the secondary structure, i.e. PDBSS = 'CCEEESSCCCTTCCBCEEEEETTEEEEEEC' is 'CESCTCBCETEC'. The data for all representative proteins corresponding to the PDB code, protein name, crystal structure resolution, number of amino acid residues in the protein chain, amino acid sequence, secondary structure and secondary structure signature are incorporated into ORACLE 9i relational database. The web-based database application is developed using JSP and HTML. The retrieval