Structure databases : CATH and PDB

SOS In Microbiology

M.Sc 2nd Semester

Paper code: 204 (Biostatistics, Computer application and bioinformatics)

Unit 4

INTRODUCTION

- All structural databases have a common ancestor, the Protein Data Bank (PDB), which was established in 1971 and six years later contained 77 atomic co-ordinate entries for 47 macromolecules Over the years, a conspicuous number of homologous databases have evolved from the PDB.

- Many of them concentrate on various classes of structural features, such as protein domains, loops, contact surfaces, quaternary structure, small-molecule ligands, metals and disordered regions.

- Other databases concentrate on biological themes. Databases of membrane proteins or of selected protein families, such as kinases or P450 containing systems, are typical examples of adding biological information to structural data.

CATH

- The CATH database provides hierarchical classification of protein domains based on their folding patterns.

- Domains are obtained from protein structures deposited in the Protein Data Bank and both domain identification and subsequent classification use manual as well as automated procedures.

- The accompanying website (www.cathdb.info) provides an easy-to-use entry to the classification, allowing for both browsing and downloading of data.

- The CATH database is a classification of protein domains (sub-sequences of proteins that may fold, evolve and function independently of the rest of the protein), based not only on sequence information, but also on structural and functional properties.

- CATH offers an important tool to researchers, as proteins with even very little sequence similarity often are both structurally and functionally related.

- The most recent version of CATH (version 3.2.0, released July 2008) contains domains, classified in a hierarchical scheme with four main levels (listed from the top and down) called class (C), architecture (A), topology (T) and homologous superfamily (H) — hence the name CATH.

- At the C-level, domains are grouped according to their secondary structure content into four categories: mainly alpha, mainly beta, mixed alphabeta; and a fourth category which contains domains with only few secondary structures.

- The A-level groups domains according to the general orientations of their secondary structures. At the T-level, the connectivity (ie the order) of the secondary structures is taken into account.
- The grouping of domains at the H-level is based on a combination of both sequence similarity and a measure of structural similarity obtained from the dynamic programming algorithm SSAP.

- The CATH homepag (http://www.cathdb.info/) provides easy access to the CATH classification.

- The first site element contains a quick description of CATH, with a link to a more thorough introduction.

- In addition to the four main levels, CATH comprises five more layers, called S, O, L, I and D. The first four layers group domains according to increasing sequence overlap and similarity.
- (eg. two domains with the same CATHSOLI classification must have 80 per cent overlap, with 100 per cent sequence identity), whereas the D-level assigns a unique identifier to every domain, thus ensuring that no two domains have exactly the same CATHSOLID classification.

Data accessebility-

- Besides a Quick Search box, which facilitates easy searching, links are provided to various other ways of accessing the data:

(1) search by keyword or domain ID;

(2) search using a sequence in FASTA format;

(3) browse the database from the top of the hierarchy;

(4) download datasets.

PDB (Protein Data Bank)

- The Protein Data Bank was established at Brookhaven National Laboratory (BNL; Bernstein et al., 1977) in 1971 as an archive for biological macromolecular crystal structures.
- Nobel prizes have been awarded for the determination and analysis of some of the structures in the PDB.
- It represents one of the earliest community-driven molecular-biology data collections.
- Initial use of the PDB had been limited to a small group of experts involved in structural biology research. Today, depositors to the PDB have expertise in the techniques of X-ray crystal structure determination, NMR, cryo-electron microscopy and theoretical modeling.

- Since October 1998, the PDB has been managed by the three members of the Research Collaborator for Structural Bioinformatics (RCSB) Rutgers, The State University of New Jersey, the San Diego Supercomputer Center at the University of California, San Diego, and the National Institute of Standards and Technology.

Data acquisition and processing

- A key component of the PDB is the effcient capture and curation of the data: data processing.
- In the present system, data (atomic coordinates, structure factors and NMR restraints) may be submitted via e-mail or via the web-based AutoDep Input Tool (ADIT; Westbrook et al., 1998; http://deposit.pdb.org/adit/) developed by the RCSB PDB.

- Each deposition to the PDB is represented by the PDBid - a four- character code of the form nxyz, where n is an integer and x, y and z are alphanumeric characters, e.g. 4hhb.
- After a structure has been deposited using ADIT, a PDBid is automatically and immediately returned to the author. This is the first stage, in which information about the structure is loaded into the internal core database, validated and annotated.
- This step involves using ADIT to help diagnose errors or inconsistencies in the files.
- The completely annotated entry as it will appear in the PDB resource, together with the validation information, is sent back to the depositor. After reviewing the processed file, the author sends any revisions .

- Depending on the nature of these revisions, steps 2 and 3 may be repeated. Once approval is received from the author, the entry and the tables in the internal core data- base are ready for distribution. The schema of this core database is a subset of the conceptual schema specified by the mmCIF dictionary.

Content of the data collected by the PDB-

- All the data collected from depositors by the PDB are considered primary data. Primary data contain, in addition to the coordinates, general information required for all deposited structures and information specific to the method of structure determination.