E-ISSN: 2707-6644
P-ISSN: 2707-6636
IJCPDM 2022; 3(1): 01-04
Received: 02-11-2021
Accepted: 10-12-2021

Raut AS
Department of Biotechnology, Shivaji University, Kolhapur, Maharashtra, India

Patil AM
Rajlaxmi Foundation's College of Agricultural Biotechnology, Madadgaon, Ahmednagar, Ahmednagar, Maharashtra, India

Shinde SS
Rajlaxmi Foundation's College of Agricultural Biotechnology, Madadgaon, Ahmednagar, Ahmednagar, Maharashtra, India

Sarode DK
Rajlaxmi Foundation's College of Agricultural Biotechnology, Madadgaon, Ahmednagar, Ahmednagar, Maharashtra, India

Shinde NA
Rajlaxmi Foundation's College of Agricultural Biotechnology, Madadgaon, Ahmednagar, Ahmednagar, Maharashtra, India

Corresponding Author:
Patil AM
Rajlaxmi Foundation's College of Agricultural Biotechnology, Madadgaon, Ahmednagar, Ahmednagar, Maharashtra, India

Protein structural databases in computational biology

Raut AS, Patil AM, Shinde SS, Sarode DK and Shinde NA

DOI: https://doi.org/10.33545/27076636.2022.v3.i1a.33

Abstract
Protein databases have becomes a Pivotal part of computational biology. Huge amount of data for protein structure function and particularly sequence are being generated. Searching database is a first step of study to find new protein. We introduce the basics of protein structural bioinformatics. Protein perform most essential biological and chemical function in a cell. They also play important role in structural, enzymatic, transport and regulatory functions. and is determined by their structure. This review covers some basic of protein structure and associated databases and it is more advanced topic of protein structural bioinformatics. This work provides to explore the potential of protein databases on internet.

Keywords: Bioinformatics, biological databases, protein data bank (PDB), homology modeling

Introduction
Protein databases have becomes a pivotal part of computational biology. Huge amount of data for protein structure function and particularly sequence are being generated (Altschul SF, 1997, Arnold 2009) [1, 2]. comparison between protein or protein families provide information about the relationship between protein within genome or across different Species (Apweiler R, 2001, Attwood TK 1999) [3, 4]. The use of database helps to understand the structure and function of protein. Protein comparison through databases allows one to view life as a forest instead of individual trees (Bairoch A, 1993, Bairoch A, 1999) [5, 6]. In addition, secondary databases derived from experimental databases are also widely available. This databases recognize and annotate the data or provide Predictions (Barker WC 1999, Benson DA 1999) [7, 8]. The use of multiple databases often helps Researchers, Understand evolution, structure and function of a protein (Bernstein FC 1977, Bourne P, 1997) [9, 10].
Although some protein databases are widely known, they are far from being fully utilized in the protein Science Community (Contreras-Moreira B, 2010, Corpet F, 1999) [11, 12].
Quantization and quantitative tools are indispensible in modern biology. Most biological research involves application of some type of mathematical, statistical, or computational tools to help synthesize recorded data and integrate various types of information in the process of answering a particular biological question (Etzold T 1999, Finn RD 2010) [13, 14].
Bioinformatics, which will be more clearly defined below, is the discipline of quantitative analysis of information relating to biological macromolecules with the aid of computers (Gao J 2009, Gromiha MM 1999) [15, 16]. A classic example in the history of genetics is by Gregor Mendal and Thomas Morgan, who by simply counting genetic variations of plants and fruit flies (Gupta R, 1999 Heazlewood JL, 2007) [17, 18]. For very sophisticated uses of quantitative tools may Involve using calculus to predict the growth rate of a human population or to establish a kinetic model for enzyme catalysis (Hu ZZ 2004, Huang DW 2009, Kraulis P 1991, Laskowski RA 1997) [19-22]. For very sophisticated uses of quantitative tools, one may find application of the "game theory" to model animal behavior and evolution, or the use of millions of nonlinear partial differential equation to model cardiac blood flow (Liu T 2007, Marchler-Bauer A 1999) [23, 24].

Bioinformatics
It is an interdisciplinary research area at the interface between computer science and biological science. Bioinformatics differs from a related field known as a computational biology. It is limited to sequence, structural and functional analysis of genes and genomes and their corresponding products and is often considered computational molecular biology. However, Computational biology encompasses all biological areas that involve computation.

For example Mathematical modeling of ecosystem, population dynamics, application of game theory in behavioral studies, and Phylogenetic construction using fossil records all employ computational tools, but do not necessarily involve biological macromolecules. It is worth nothing that there are other views of how the two terms relate.

Biological databases

Current biological databases use all three types of databases structures: Flat file format, relational, and object oriented despite the obvious drawbacks of using flat files in database management, many biological database still use this format. The justification for this is that this system involves minimum amount of database design and the search output can be easily understood by working biologists.

Based on the content, biological database can be roughly divided into 3 Categories

1. primary database
2. Secondary database
3. Specialized database

PDB: The protein data bank is a database for three dimensional structural data of large biological molecule such as protein and nucleic acid. (https.rcsb.org.in).

It includes data obtained by x-ray crystallography and nuclear magnetic resonance spectrometry submitted by biologist and biochemist from all over the world Many secondary sources of information are derived from PDB data.

1. **Covalent bond distance and angles:** proteins are compared with standard values from (Nogales-Cadenas R 2009 & Ogata H 1999) [25, 26], nucleic acid bases are compared with standard values from (Orengo CA 1997 & Pierleoni A 2007) [27, 28], Sugar and phosphate are compared with standard values from (Porter CT 2004 & Prilusky J 2011) [29, 30].
2. **Stereo chemical validation:** All chiral centres of proteins and nucleic acid are checked for correct stereochemistry.
3. **Atom Nomenclature:** The Nomenclature of all atoms is checked for compliance with IUPAC standards and is adjusted if Necessary.
4. **Close Contact:** The distance between all atoms within crystal structure and the unique molecule of NMR structure are calculated for crystal structures, contact between symmetry-related molecules are also checked.
5. **Ligand and Atom Nomenclature:** Residue and atom nomenclature is compared against a standard dictionary. For all legands as well as standard residue and bases new legands are added to the dictionary as they are deposited.
6. **Sequence Comparison:** The sequence provided by the depositor is compared with the sequence derived from the co-ordinate records. This information is displayed in the table where any differences or missing residue are annoted. During the annotation process the sequence database reference provided by author are checked for accuracy. If no reference given a BLAST search is used to find the best match. Any conflict between the PDB records and sequence derived from the co-ordinate records is resolved by comparison with various sequence database.

Gen Bank

GenBank is the most complete collection of annotated nucleic acid sequence data for almost every organism. The content includes genomics DNA, mRNA, cDNA, ESTs, high throughput raw sequence data, and sequence polymorphisms. There is also a Genpept database for protein sequences, the majority of which are conceptual translations from DNA sequences, although a small number of the amino acid sequences are derived using peptide sequencing techniques.

Homology modelling

It is also called as comparative modelling. The principle behind is that if two protein share a high enough sequence similarity, they are likely to have very similar 3dimensional structures. If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence. Homology modeling produces an all-atom model based on alignment with template proteins.

The overall homology modelling procedures consist of 6 steps. The first step is template selection, which involves identification of homologous sequences in the protein structure database to be used as templates for modelling. The second step is alignment of the target and template sequences. The third step is to build a framework structure for the target protein consisting of main chain atoms. The fourth step of model building includes the addition and optimization of side chain atoms and loops. The fifth step is to refine and optimize the entire model according to energy criteria. The final step involves evaluating of the overall quality of the model obtained.

Future Aspects

Despite the pitfalls there is no doubt that bioinformatics is a field that holds great potential for Revolutionizing biological research in the coming decades. Currently the field is undergoing major expansion. In addition to providing more reliable and more rigorous computational tool for sequence, structural, and functional analysis, the major challenge for future bioinformatics development is to develop tools for elucidation of the function and interaction of all gene products in a cell. This presents a tremendous challenge because it requires integration of disparate fields of Biological knowledge and a variety complex mathematical and statistical tools. To gain a deeper understanding of cellular Functions, Mathematical model are needed to stimulate a wide variety of Intracellular reaction and interaction at the whole cell level. This molecular simulation of all the cellular process is termed system biology. Achieving this goal will represent a major leap toward fully understanding a living system. That is why the system level simulation and integrations are considered the future of bioinformatics. Modelling such complex networks and making predictions about their behaviour present tremendous challenges and opportunities for Bioinformaticans. The ultimate goal of the endeavour is to transform biology from a qualitative science to a quantitative and predictive science. This is truly an exciting time for bioinformatics.

Summary

Databases are fundamental to modern biological research, especially to Genomic studies. The goal of a biological

database is two fold: information retrieval and knowledge discovery. Electronic databases can be constructed either as flat files, Relational, or object oriented. Flat files are simple text files and lack any form of organization to facilitate information retrieval by computers. Relational databases organize data as tables and search information among tables with shared features. Object-oriented database organize data as objects and associate the objects according to hierarchical relationships, biological databases encompasses all three types based on their content, biological databases are divided into primary, secondary and specialized databases, primary database are simply archive sequence or structure. for example merging redundant sequences into a single entry or store highly redundant sequences into a separate database.

Conclusion

On that basis we can visualize a structure of protein with the help of structural databases, which we can separate with the help of file format and specialised database, primary, secondary, databases, which we can distribute the structural databases on the basis of physiochemical properties.

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al*. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl Acids Res 1997;25:3389-3402. [PMC free article] [PubMed] [Google Scholar]
2. Arnold K, Kiefer F, Kopp J, Battey JN, Podvinec M, Westbrook JD, *et al*. The protein model portal. J Struct Funct Genomics 2009;10:1-8. [PMC free article] [PubMed] [Google Scholar]
3. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M *et al*. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucl Acids Res. 2001;29:37-40. [PMC free article] [PubMed] [Google Scholar]
4. Attwood TK, Flower DR, Lewis AP, Mabey JE, Morgan SR, Scordis P *et al*. PRINTS prepares for the new millennium. Nucl Acids Res. 1999;27:220-225. [PMC free article] [PubMed] [Google Scholar]
5. Bairoch A. The ENZYME data bank. Nucl Acids Res. 1993;21:3155-3156. [PMC free article] [PubMed] [Google Scholar]
6. Bairoch A, Apweiler R. The UniProt protein sequence data bank and its supplement TrEMBL in 1999. Nucl Acids Res 1999;27:49-54. [PMC free article] [PubMed] [Google Scholar]
7. Barker WC, Garavelli JS, McGarvey PB, Marzec CR, Orcutt BC, Srinivasarao GY, *et al*. The PIR-international protein sequence database. Nucl Acids Res. 1999;27:39-42. [PMC free article] [PubMed] [Google Scholar]
8. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, *et al*. Genbank. Nucl Acids Res. 1999;27:12-17. [PMC free article] [PubMed] [Google Scholar]
9. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, *et al*. The protein data bank: A computer based archival file for macromolecular structures. J Mol Biol. 1977;112:535-542. [PubMed] [Google Scholar]
10. Bourne P, Berman H, Watenpaugh K, Westbrook J, Fitzgerald P. The macromolecular crystallographic information file (mmCIF) Methods Enzymol. 1997;277:571-590. [PubMed] [Google Scholar]
11. Contreras-Moreira B. 3D-footprint: a database for the structural analysis of protein-DNA complexes. Nucl Acids Res. 2010;38:D91-97. [PMC free article] [Pub Med] [Google Scholar]
12. Corpet F, Gouzy J, Kahn D. Recent improvements of the ProDom database of protein domain families. Nucl Acids Res 1999;27:263-267. [PMC free article] [PubMed] [Google Scholar]
13. Etzold T, Ulyanov A, Argos P. SRS: Information retrieval system for molecular biology data banks. Methods Enzymol 1996;266:114-128. [PubMed] [Google Scholar]
14. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, *et al*. The Pfam protein families database. Nucl Acids Res 2010;38:D211-22. [PMC free article] [PubMed] [Google Scholar]
15. Gao J, Agrawal GK, Thelen JJ, Xu D. P3DB: a plant protein phosphorylation database. Nucl Acids Res 2009;37:D960-D962. [PMC free article] [PubMed] [Google Scholar]
16. Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Sarai A. Protherm: Thermodynamic database for proteins and mutants. Nucl Acids Res. 1999;27:286-288. [PMC free article] [PubMed] [Google Scholar]
17. Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. Nucl Acids Res. 1999;27:370-372. [PMC free article] [PubMed] [Google Scholar]
18. Heazlewood JL, Verboom RE, Tonti-Filippini J, Small I, Millar AH. SUBA: the Arabidopsis Subcellular Database. Nucl Acids Res. 2007;35:D213-218. [PMC free article] [PubMed] [Google Scholar]
19. Hu ZZ, Mani I, Hermoso V, Liu H, Wu CH. iProLINK: an integrated protein resource for literature mining. Comput Biol Chem. 2004;28:409-416. [PubMed] [Google Scholar]
20. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4:44-57. [PubMed] [Google Scholar]
21. Kraulis P. MOLSCRIPT-a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr. 1991;24:946-950. [Google Scholar]
22. Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDB sum: A web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci. 1997;22:488-490. [PubMed] [Google Scholar]
23. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. Binding DB. a web-accessible database of experimentally determined protein-ligand binding affinities. Nucl Acids Res. 2007;35:D198-201. [PMC free article] [PubMed] [Google Scholar]
24. Marchler-Bauer A, Addess KJ, Chappey C, Geer L, Madej T, Matsuo Y, *et al*. MMDB: Entrez's 3D structure database. Nucl Acids Res. 1999;27:240-243. [PMC free article] [PubMed] [Google Scholar]
25. Nogales-Cadenas R, Abascal F, Díez-Pérez J, Carazo JM, Pascual-Montano A. Centrosome DB: human centrosomal proteins database. Nucl Acids Res.

2009;37:D175-80. [PMC free article] [PubMed] [Google Scholar]
26. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucl Acids Res. 1999;27:29-34. [PMC free article] [PubMed] [Google Scholar]
27. Orengo CA, Michie AD, Jones DT, Swindells MB, Thornton JM. CATH-a hierarchic classification of protein domain structures. Structure 1997;5:1093-1108. [PubMed] [Google Scholar]
28. Pierleoni A, Martelli PL, Fariselli P, Casadio R. eSLDB: eukaryotic subcellular localization database. Nucl Acids Res. 2007;35:D208-212. [PMC free article] [PubMed] [Google Scholar]
29. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucl Acids Res 2004;32:D129-33.
30. Prilusky J, Hodis E, Canner D, Decatur WA, Oberholser K, Martz E, *et al*. Proteopedia: A status report on the collaborative, 3D web-encyclopedia of proteins and other biomolecules. Journal of Structural Biology 2011;175:244-252.