

Towards Understanding the Protein Structure and its Problems: Selecting Data Sources and Data Representation

Jamaludin Sallim¹, Roslina Abdul Hamid², Rosni Abdullah³, Ahamad Tajudin Abdul Khader⁴

^{1,3,4} School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia
jamals@cs.usm.my, rosni@cs.usm.my, tajudin@cs.usm.my

² Faculty of Computer Systems and Software Engineering,
University Malaysia Pahang, 25000, Kuantan, Pahang, Malaysia
roslina@ump.edu.my

Abstract - This paper aims to provide a basic idea for bioinformatics researchers in gearing their works on protein structure. We will discuss generally on the protein structure, accessing the data and represent them with using selected visualization tools. The major problems of protein structure analysis also being highlighted.

Keywords - protein structure, structural problems, data access, data representation.

INTRODUCTION

Research in protein structure and solving its problems is currently major issues in bioinformatics. The rapid growth in protein data continues and the size and complexity of individual protein structure databases and the explosion of new databases reflected the growth of protein data. For new comers in bioinformatics research, to select the protein data from respective database and to represent them in order to analyze by using computational method can be considered as a major challenge. Since data is important entity in research, it is essential that bioinformatics researchers make a strong attempt to select and present the data. For this paper, we try to guide new bioinformatics researches to select the protein structure data and represent them based on the requirement of computational methodology, which they are going to use. There are many databases were developed for depositing protein structure and one of them and the most well known is Protein Data Bank (PDB), Protein Structure Database (PSdb, which derived from PDB) and EMBL-Dali (which compare protein structures in 3D shapes in PDB). PDB has evolved into a sophisticated resource that contains not only the results of structure analyses but information about experimental procedures, derived information about structures and links to related data sources [1]. In this single repository database, the process of depositing protein data experimentally starts by X-ray crystallography, nuclear magnetic resonance (NMR), neutron diffraction and most

recently, cryo-electron microscopy [2]. (We will describe more detail on PDB in the next section). Figure 1 illustrates the scientific and computational approach in analyzing protein structure.

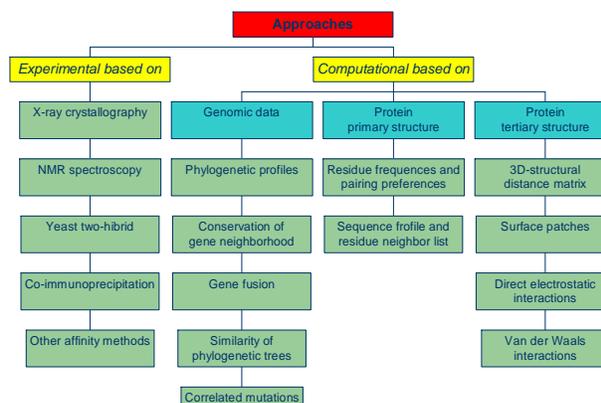


Fig. 1. Protein Structure Analysis Approach

PROTEIN STRUCTURE IN GENERAL

Proteins are large polymers, such as enzymes, that are actively involved in the biological reactions of living organisms and play a structural role in the cell [3]. For this section, we follow precisely the scientific definition from wikipedia.org and based on our observation, all definitions for *Amino Acids*, *Protein Secondary Structure*, *Protein Tertiary Structure* and *Protein Quaternary Structure* are taken from published books, journals and papers. Since our background are computer scientist, we only cite the most important information that related to our research works.

Amino Acids

In chemistry, an amino acid is any molecule that contains both amine and carboxyl functional groups. In biochemistry, this term refers to alpha amino acids. These

alpha amino acids are components of proteins. In proteins, amino acids are joined together in a chain by peptide bonds between their amino and carboxylate groups. An amino acid residue is one amino acid that is joined to another by a peptide bond. Each different protein has a unique sequence of amino acid residues, this is its primary structure. Just as the letters of the alphabet can be combined to form an almost endless variety of words, amino acids can be linked in varying sequences to form a huge variety of proteins. Amino acids are the basic structural building units of proteins. They form short polymer chains called peptides or longer chains either called polypeptides or proteins [wikipedia].

Protein Primary Structure

In biochemistry, the primary structure of a biological molecule is the exact specification of its atomic composition and the chemical bonds connecting those atoms (including stereochemistry). Primary structure is sometimes mistakenly termed *primary sequence*, but there is no such term, as well as no parallel concept of secondary or tertiary sequence [wikipedia].

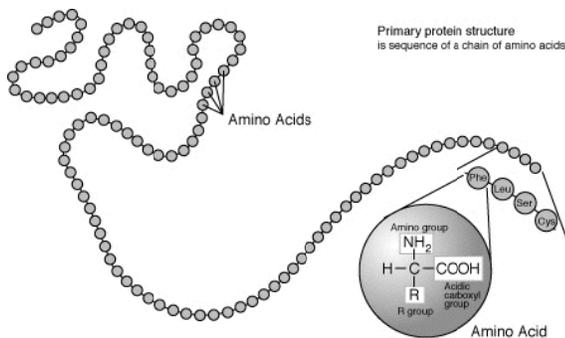


Fig. 2. A protein primary structure is a chain of amino acids (source: Wikipedia).

Protein Secondary Structure

In biochemistry and structural biology, secondary structure is the general three-dimensional form of *local segments* of biopolymers such as proteins and nucleic acids (DNA/RNA). It does not, however, describe specific atomic positions in three-dimensional space, which are considered to be tertiary structure. Secondary structure is formally defined by the hydrogen bonds of the biopolymer, as observed in an atomic-resolution structure. In proteins, the secondary structure is defined by patterns of hydrogen bonds between backbone amide groups (sidechain-mainchain and sidechain-sidechain hydrogen bonds are irrelevant). In nucleic acids, the secondary structure is defined by the hydrogen bonding between the nitrogenous bases. Secondary structure was introduced by Kaj Ulrik Linderstrom-Lang in the 1952 Lane medical lectures at Stanford [wikipedia].

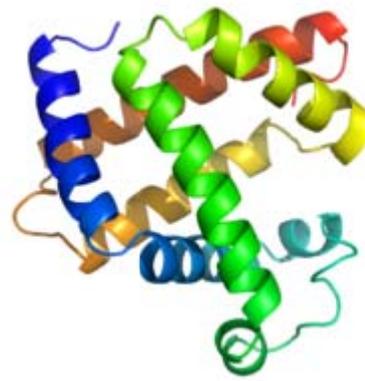


Fig. 3. A representation of the 3D structure of the myoglobin protein. (source: wikipedia)

Protein Tertiary Structure

In biochemistry, the tertiary structure of a protein is its overall shape, also known as its fold. Protein molecules are linear chains of amino acids that typically assume a specific three-dimensional structure in which they perform their biological function. The study of protein tertiary structure is known as structural biology.

Relationship to Primary Sequence

Tertiary structure is considered to be largely determined by the protein's primary sequence, or the sequence of amino acids of which it is composed. Efforts to predict tertiary structure from the primary sequence are known generally as protein structure prediction. However, the environment in which a protein is synthesized and allowed to fold are significant determinants of its final shape and are usually not directly taken into account by current prediction methods [wikipedia].

Protein Quaternary Structure

In biochemistry, quaternary structure is the arrangement of multiple folded protein molecules in a multi-subunit complex. Many proteins are actually assemblies of more than one polypeptide chain, which in the context of the larger assemblage are known as protein subunits.

PROTEIN STRUCTURE PROBLEMS

Most of bioinformatics researchers focus on protein analysis from the stand-point of its sequence and structure. For this paper, we focus on protein structure because scientifically proven that protein structure can provide more information if compare with sequence [5]. The major problems in analyzing protein structures are prediction or folding, classification and protein-protein interaction. Protein structure prediction or folding refers to the computational methods that have been developed to predict protein secondary structure with a reasonable degree of accuracy. Currently, there are predictions

methods for predicting tertiary structure, but the accuracy of such methods is highly dependent on whether or not the protein in question is related in sequence to any members of the existing library of known protein structure. [Protein Folding and Secondary Structure Prediction]. For protein structure classification, it refers to process of classifying protein domains based on their tertiary structure that can provide a valuable resource, which can be used to understand protein function and evolutionary relationships [6]. The third problem, protein-protein interactions refer to the association of protein molecules and the study of these associations from the perspective of biochemistry, signal transduction and networks. The interactions between proteins are important for many biological functions [wikipedia]. The following figures (Figure 4,5 & 6) illustrate the general overview of three problems mentioned.

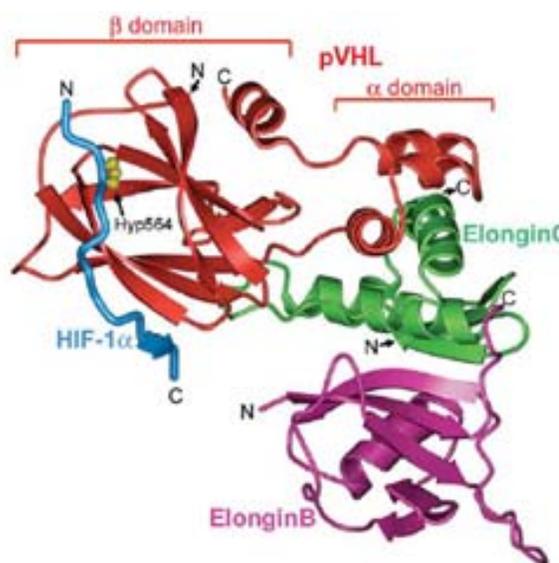


Fig. 6. Protein-Protein Structure Interaction

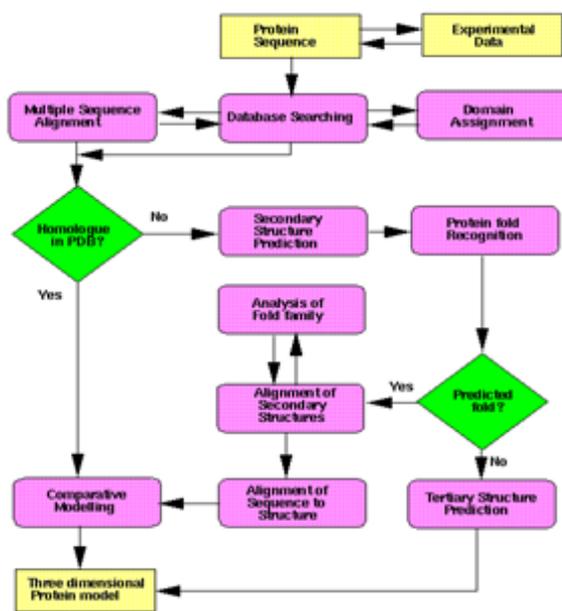


Fig. 4. Protein Structure Analysis Approach

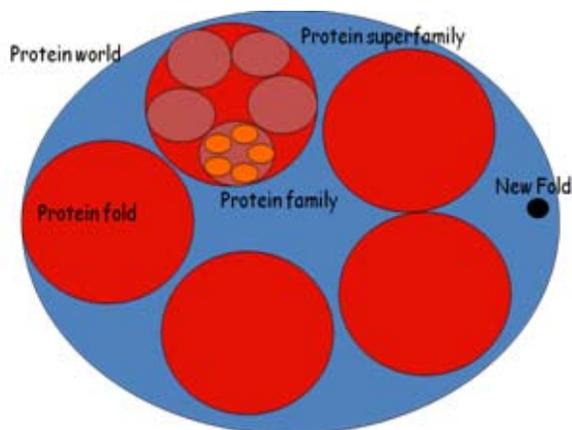


Fig. 5. Protein Structure Classification

PROTEIN STRUCTURE DATA SOURCE

An mentioned, we survey on three structural databases, Protein Data Bank (PDB), Protein Structure Database (PSdb) and EMBL-Dali. PSdb is derived from PDB while EMBL-Dali compare protein structures in 3D shapes in PDB. The reason we select these database is that they have a similarity in data type, format and justification.

Protein Data Bank (PDB)

The Protein Data Bank (PDB), free access database, is a repository for 3-D structural data of proteins and nucleic acids. This data, typically obtained by X-ray crystallography or NMR spectroscopy, is submitted by biologists and biochemists from around the world, is released into the public domain. Protein Data Bank (PDB) format is a standard for files containing atomic coordinates. Structures deposited in the Protein Data Bank are written in this standardized format. The complete PDB file specification provides for a wealth of information, including authors, literature references, and the identification of substructures such as disulfide bonds, helices, sheets, and active sites.

ATOM 1	N	VAL	1	-13.090	1.966	9.741	1.00	0.00
ATOM 2	CA	VAL	1	-12.852	3.121	8.892	1.00	0.00
ATOM 3	C	VAL	1	-13.047	4.399	9.711	1.00	0.00
ATOM 4	O	VAL	1	-12.143	5.228	9.800	1.00	0.00
ATOM 5	CB	VAL	1	-13.753	3.058	7.658	1.00	0.00
ATOM 6	CG1	VAL	1	-13.930	4.446	7.036	1.00	0.00
ATOM 7	CG2	VAL	1	-13.208	2.063	6.631	1.00	0.00
ATOM 8	H	VAL	1	-13.919	1.449	9.527	1.00	0.00
ATOM 9	HA	VAL	1	-11.816	3.075	8.557	1.00	0.00
ATOM 10	HB	VAL	1	-14.734	2.707	7.977	1.00	0.00
ATOM 11	1HG1	VAL	1	-13.951	4.357	5.950	1.00	0.00
ATOM 12	2HG1	VAL	1	-14.866	4.883	7.384	1.00	0.00
ATOM 13	3HG1	VAL	1	-13.098	5.085	7.333	1.00	0.00

Fig. 7: Protein Tertiary Structure Data in PDB.

Protein Structure Database (PSdb)

The Protein Structure Database (PSdb), a new protein database that relates secondary, supersecondary and tertiary information to the primary structure. The data for each protein is supplied on a residue by residue basis and encoded in a series of flat ASCII files. Relationships between the various levels of structure (primary, secondary, tertiary) can be investigated visually using PSdbView, a graphical tool provided to view the information within the PSdb [7].

Atom	Hybrid	Radius[1]	Example	Connolly[2]	Eisenberg[3]	Sanders[4,5]
C	sp3	1.893	Methyl,Methylene	1.90	1.90	1.87/1.80
C	sp2	1.874	CarbonylCarbon,Aromatic	1.90	1.90	1.80/1.80
C	sp1	1.812	Alkyne	1.90	1.90	/1.80
N	sp3	1.734	Amine,ammonium	1.82	1.70	/1.80
N	sp2	1.668	Amide	1.82	1.70	1.65/1.80
O	sp3	1.566	Ether,Alcohol	1.65	1.40	/1.80
O	sp2	1.549	CarbonylOxygen	1.65	1.40	1.40/1.80
F	sp3	1.420	Alkylfluorides	1.55	1.42	/1.80
Cl	sp3	1.934	Alkylchlorides	2.05	1.934	/1.80
Si	sp3	2.004	Silicates	1.80	1.80	/1.80
P	sp3	2.051	Phosphates	2.10	1.80	/1.80
S	sp3	2.016	Thiol	2.10	1.80	/1.80

Fig. 8: Example of Protein Structure in PSdb.

EMBL-Dali

The Dali server is a network service for comparing protein structures in 3D. When the user submit the coordinates of a query protein structure and Dali compares them against those in the PDB. A multiple alignment of structural neighbours is emailed back to the user. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences [8].

237127	329	332	102	105
237127	325	328	96	99
237127	305	316	69	80
237127	266	282	48	64
237127	257	265	30	38
237127	243	253	18	28
237127	233	242	7	16
237127	8	12	1	5

Fig. 9. Example of Protein Structure Pairwise Alignment in EMBL-Dali

PROTEIN STRUCTURE DATA REPRESENTATION

Diplaris et. al. (2005) quoted that a crucial issue in bioinformatics is structural biology, especially to represent the structure of several biological

macromolecules. 3D protein structure knowledge can be considered a strong weapon in combating many diseases, since most of them are caused by malfunctions of the proteins involved in several functions of the human cells. [9]. Even though there so many tools have been developed recently, researchers must select the most appropriate tools and methods in representing protein, not only to view them but give the related information which are needed. We study three visualization tools; Rasmol, VMD and PSdbview to visualize the data and to get the related information.

Rasmol

RasMol is a molecular graphics program intended for the visualization of proteins, nucleic acids and small molecules. The program is aimed at display, teaching and generation of publication quality images. RasMol runs on wide range of architectures and operating systems. Figure 10 illustrate viewing protein 3D structure using Rasmol.

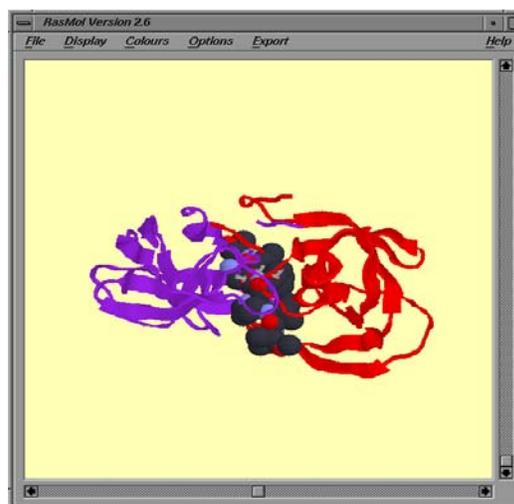


Fig. 10. Viewing protein 3D structure using Rasmol

PSdb View

PSdbView is a graphical tool provided to view the information and investigate visually on relationships between the various levels of structure (primary, secondary, tertiary) within the PSdb. It allows for side by side comparison of residue based data and includes a variety of standard mechanisms for visualizing protein data [10]. Figure 11 illustrate the basic information screen of PSdb View.

Visual Molecular Dynamics (VMD)

VMD is a tool designed for the visualization and analysis of biological systems such as proteins, nucleic acids etc. VMD can read standard Protein Data Bank (PDB) files and display the contained structure. In particular, VMD can act as a graphical front end for an external MD program by displaying and animating a

molecule undergoing simulation on a remote computer [11]. Figure 12 illustrate the sample of protein structure viewed by using VMD.

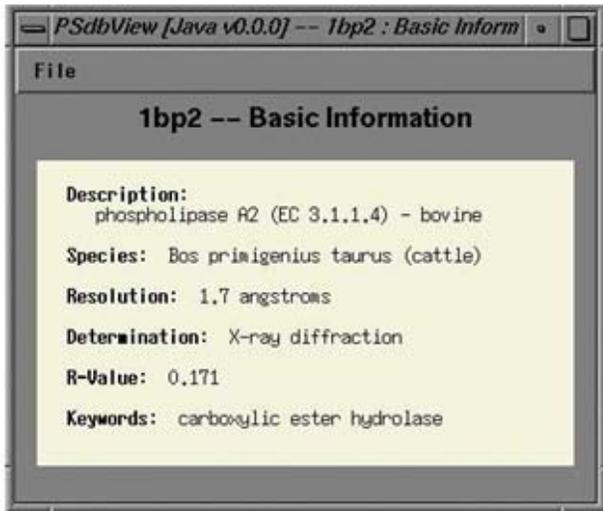


Fig. 11. PSdb View: Basic Info Screen

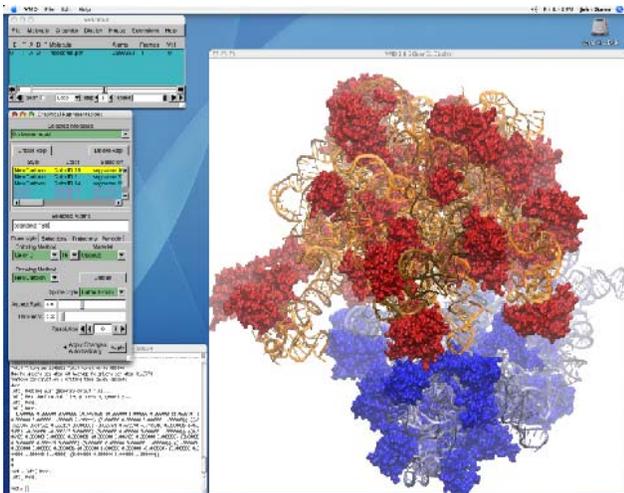


Fig. 12. Using VMD To View Protein Structure

We only give a general idea on these tools and technically we can't conclude or suggest with tool is the most appropriate because each tool has it's own features and was designed to meet certain purpose. It is up to researcher to select the tool, because they are the people who really know their data sets and the methods requirement. For us, this among the best tool to visualize protein structure in PDB and can assist us to analyze the data with regard to our methodology requirement.

SUMMARY

In this background research works, we introduced the protein structure data and its problem. We also describe the data representation by using three visualizing tools. We hope that with this elementary introduction to protein structure, it will give a basic idea to the new

bioinformatics researchers to start their work on protein structure, which we believe that there are so many unresolved issues arises with regards to the biological data growth.

REFERENCES

- [1] Berman H.M, Bourne P.E, and Westbrook J., "The Protein Data Bank: A Case Study in Management of Community Data," *Current Proteomics*, Bentham Science Publishers Volume 1, Number 1, January 2004 , pp. 49-57(9)
- [2] Philip E. B., John D.W. and Helen M. B. "The Protein Data Bank and lessons in data management," *Briefings in Bioinformatics* 2004 Mar;5(1):23-30.
- [3] <http://www.ebi.ac.uk/2can/tutorials/structure/struct2.html>
- [4] Scientific definition from <http://wikipedia.org>
- [5] Scott F. Smith,. "Protein family classification using structural and sequence information," in *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*.
- [6] Tim J.P., Alexey G. M, Steven E. B. and Cyrus C "SCOP: a Structural Classification of Proteins database," *Nucleic Acids Res.*, 25:236-- 239, 1997.
- [7] <http://www.psc.edu/~deerfiel/PSdb/>
- [8] <http://www.ebi.ac.uk/dali/>
- [9] Diplaris S., Tsumakias G., Mitkas P. and Vlahavas I, "Protein Classification with Multiple Algorithms", *10th Panhellenic Conference on Informatics (PCI 2005)*, P. Bozanis and E.N. Houstis (Eds.), Springer-Verlag, LNCS 3746, pp. 448-456, Volos, Greece, 11-13 November, 2005.
- [10] <http://www.psc.edu/~deerfiel/PSdb/PSdbPaper/>
- [11] <http://www.ks.uiuc.edu/Research/vmd/>