International Journal of Computer Science and Informatics

Volume 2 | Issue 2 Article 15

October 2012

A Need for Development of SDK for Reading PDB File

Nikita V. Mahajan
Dept. of Computer Science & Engineering, G.H. Raisoni College of Engineering, Nagpur, India,
nv.mahajan18@gmail.com

Recommended Citation
Mahajan, Nikita V. (2012) "A Need for Development of SDK for Reading PDB File," *International Journal of Computer Science and Informatics*: Vol. 2 : Iss. 2 , Article 15.
DOI: 10.47893/IJCSI.2012.1081
Available at: https://www.interscience.in/ijcsi/vol2/iss2/15

A Need for Development of SDK for Reading PDB File

Nikita V. Mahajan & L.G.Malik

Dept. of Computer Science & Engineering, G.H. Raisoni College of Engineering, Nagpur, India
E-mail : nv.mahajan18@gmail.com

Abstract **-** Study of protein, predicting its structure to know its function is an important field in bioinformatics. Protein unit that is twenty amino acids have total information for converting linear sequences of amino acid into its unique and globular structures. The data to make protein fold correctly depends on the information present in the input PDB file, but data present in the file is noisy. Thus there is need for developing a customize software development kit or API to read the PDB file in structural manner.

Keywords **-** *Amino Acids, SEQRES, Helix, Beta Sheet, PDB, SDK.*

I. INTRODUCTION

A) Background

Protein has Origin or derived from Greek word PROTEIOS meaning holding first. Protein is made up of 20 linear amino acids chain which is covalently bounded with other amino acids (referred as residues). [3].

TABLE I: ELEMENTS OF PROTEIN

Sr. No.	Element Name	Amount
1	Carbon	50-53%
2	Hydrogen	6-7.3%
3	Oxygen	19-24%
4	Nitrogen	13-19%
5	Sulfur	0-4%

Amino acid is made up of organic compounds which contain two functional groups that are amino group and carboxyl group. Amino group is represented as - NH_3 which is basic in nature. While carboxyl group is acidic in nature and it is represented as, - COOH. Amino acid is also termed as α amino acid if both amino and carboxyl group is attached to some carbon atom. An alpha-amino acid has the generic formula, where R is an organic substituent. This only changes in each amino acid

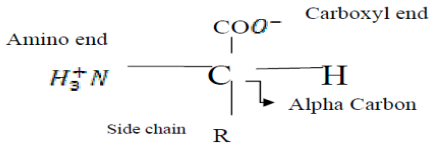

Fig. 1: Amino Acid Structure

Length of these monomers ranges from ten to thousand and they are very important part of mostly biochemical process There are twenty different amino acids; they are represented as three letter codes, such as, Glycine as GLY, Proline as PRO, Alanine by ALA and Cytosine as CYS etc, which are grouped in such a fashion to represent the primary structure of protein.

TABLE II: ELEMENTS OF PROTEIN

Essential	Nonessential
Isoleucine	Alanine
Leucine	Asparagine
Lysine	Aspartic acid
Methionine	Cysteine*
Phenylalanine	Glutamic acid
Threonine	Glutamine*
Tryptophan	Glycine*
Valine	Proline*
	Selenocysteine*
	Serine*
	Tyrosine*
	Arginine*
	Histidine*
	Ornithine*
	Taurine*

This table shows that which amino acid is essential and which are not essential for living.

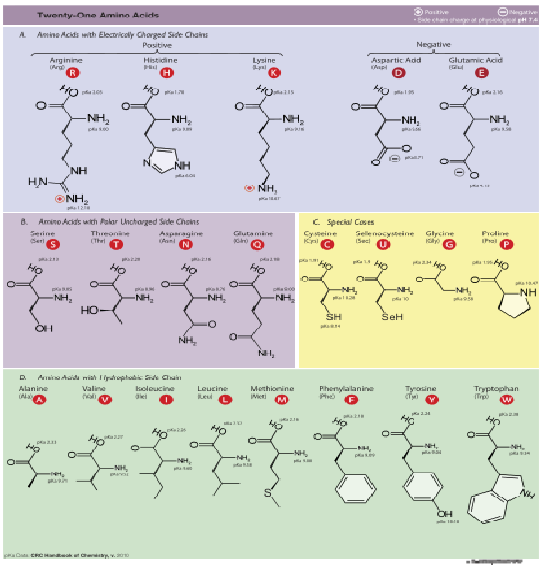

Fig. 2 : Molecular structure of all amino acid [5]

B) PDB File

Drs. Edgar Meyer and Walter Hamilton had founded Protein Data Bank (PDB) in 1971 at Brookhaveni National Laboratory. The Protein Data Bank (PDB) is a repository for the 3-D structural data of large biological molecules, such as proteins and nucleic acids. The data and the three dimensional data is typically obtained by X-ray crystallography or NMR spectroscopy which are submitted by biologists and biochemists from around the world, so that it can be freely accessible to research , scientist and student through Internet via the websites of its member organizations (PDBe, PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB. According to PDB selected list it has non redundant protein structure with sequence identity lower than 25% [2] [9].

C) PDB File Format

The PDB format has 12 sections, in which 46 different fields are represented through the 70 column format.

The different fields are Header, Title, Date, Author, Family to whom the data belong, Remark, Compound, source, Keywords, REVDAT, and JRNL.

The following fields are relevant for model: SEQRES (defines the amino acids sequence of the protein), HELIX and SHEET (identify the amino acids that form these secondary structures), and ATOM (represents the spatial distribution of the atoms of the protein in its native conformation).

- *SEQRES Format*

It defines the sequences of the amino acid of protein in each chain.

TABLE III: SEQRES FORMAT

1-6	Record name	Example
9-10	Serial Number	2
12	Chained	A
14-17	NumRes	23
20-22	Residue name “resName “	MET
24-26	Residue name “resName “	LEU
28-30	Residue name “resName “	ALA
32-34	Residue name “resName “	GLY
36-38	Residue name “resName “	THR
40-42	Residue name “resName “	THY
44 -46	Residue name “resName “	VAL
48 -50	Residue name “resName “	GLN
52 -54	Residue name “resName “	CYS
56 -58	Residue name “resName “	ASN
60 -62	Residue name “resName “	SER
64 -66	Residue name “resName “	PHE
68 – 70	Residue name “resName “	HIS

- *SECONDARY STRUCTURE DATA FORMAT*

1) HELIX Format

Alpha Helix (α) is the common motif in the secondary structure of proteins. It’s a right-handed coiled or spiral conformation. Every backbone N-H group donates a hydrogen bond to the backbone C=O group of the amino acid. One turn of the helix represents 3.6 amino acid residues. A single turn of the a-helix involves 13 atoms from the O to the H of the H bond [10].

TABLE IV: HELIX FORMAT

1-6	Record name	Example
8-10	Serial Number	1
12 – 14	HelixID	HA
16 – 18	ResName	GLY
20	ChainID “	A
22 – 25	SeqNum “	86
26	Insertion code	

28 - 30	EndResName	LEU
32	EndChainID	A
34 – 37	EndSeqNum	94
38	EndICode	
39 – 40	HelixClass	1
41 – 70	String	
72 – 76	Length	9

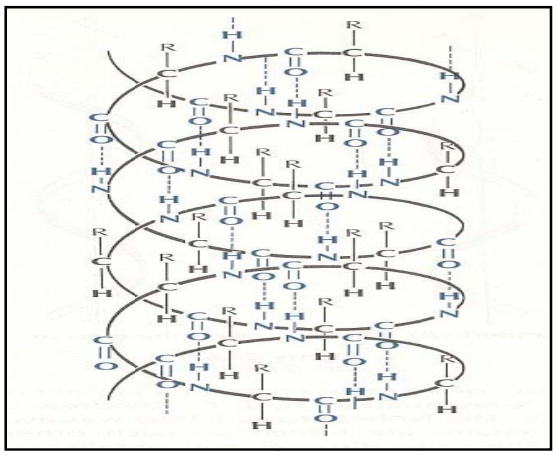

Fig. 3: Helix Structure

The initial residue present in helix sheet is one with n-terminal residue. The helix class present in column 39-40 are classified as

TABLE V: HELIX CLASS

Class Number	Type of helix
1	Right – handed alpha
2	Right handed omega
3	Right handed pi
4	Right handed gamma
5	Right handed 310
6	Left -handed alpha
7	Left- handed omega
8	Left-handed gamma
9	27 ribbon/helix
10	Polypro –line

2) SHEETS

Beta sheets consist of the beta strands connected laterally by at least two or three backbone hydrogen bonds, forming a generally twisted, pleated sheet. In the parallel b-pleated sheet, adjacent chains run in the same direction (N ® C or C ® N). In the anti-parallel b-pleated sheet, adjacent strands run in opposite directions [8] [10].

TABLE VI: BETA SHEET FORMAT

1-6	**Record name**	**Example**
8 – 10	Strand	4
12 – 14	Sheeted	A
15 – 16	NumStrands	5
18 – 20	Initial ResName	THR
22	Initial ChainID	A
23 – 26	Initial residue SeqNum	236
27	ICode	
29 – 31	endResName	THR
33	endChainID	A
34 – 37	endSeqNum	244
38	endICode	
39 – 40	Sense	-1/0
42 – 45	curAtom	N
46 – 48	curResName	ASN
50	curChainId	A
51 – 54	curResSeq	239
55	curICode	
57 - 60	prevAtom	O
61 – 63	prevResName	VAL
65	prevChainId	A
66 – 69	prevResSeq	232
70	prevICode	

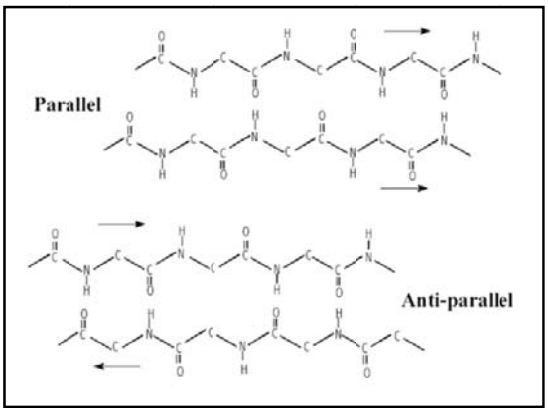

Fig. 4 : Beta Sheet Structure [5]

Column 39-40 that is sense tell that the sheet present will be parallel or anti parallel.

- *ATOM Format*

The ATOM records present the atomic coordinates for standard residues. ATOM records for proteins are listed from amino to carboxyl terminus. ATOM records for proteins are listed from amino to carboxyl terminus

TABLE VII: ATOM FORMAT

1-6	Record name	Example
7-11	Serial Number	145
13-16	Atom name	N/CA/CB
17	Alternate location	
18-20	Resdue name	VAL
22	ChainID	A
23-26	Residue sequence number	26
27	Insertion code	
31-38	x- co-ordinate value	19.45
39-46	y-co-ordinate value	16.54
47-54	z-co-ordinate value	57.87
55-60	Occupancy	1.00
61-66	Temperature factor	11.92
73-78	Segment identifier	
77-78	Element symbol	A1
79-80	Charge on atom	C

II. PROBLEM DEFINITION

As Proteins play a central role in nearly all biological processes at the cellular level; they are responsible for catalysing and regulating biochemical reactions, transporting material and information, and form the basic structures such as skin, hair, and tendon of a protein is determined by its chemical composition of the molecule and by its spatial structure [7].

The folding information process is present totally in linear sequences of amino acids, determines its unique fold, and the geometry of a protein fold largely determines its specific biological function. Predicting the tertiary structure of a protein by using its residue sequence is called the Protein Folding Problem [1] [4] [7].

One problem with protein data is that, it is very noisy complexity in Reading PDB File, As the PDB file contain 12 section each with different column format, it also face lot of complexity such as many references are not properly placed that is some-time the space between the column are more or less, sometime it has some insertion column which is not present in many of the section, Thus for developing the protein related application, totally interlay on the same file as the input.

III. PROPOSED SYSTEM

A) Software Developing Kit

A set of software development application that allow the application to develop certain software packages, framework, and hardware platform and such type of set is called as Software development kit [6].

In the form of some files to interface to a particular programming language or include sophisticated hardware to communicate with a certain embedded system SDK may be an application programming interface (API).

Providers of SDKs for specific systems or subsystems may sometimes substitute a more specific term instead of software. For instance, both Microsoft and Apple provide driver development kits (DDK) for developing device drivers

SDKs also frequently include sample code and supporting technical notes or other supporting documentation to help clarify points from the primary reference material.

For Example in C for graphic program there is Graphic.H header file similarly for mathematical calculation there is Math.H header file, if such header files are not created then reading and calculating all mathematical won't be possible

B) Development Platform

The object-oriented programming is becoming more popular is that software reuse is becoming more important. Developing new systems is expensive, and maintaining them is even more expensive [6]. Development platform for the SDK is vb.net. In this there will be developing the namespaces, developing the customize classes which can be used for different platform. Polymorphism reduces the number of procedures, and thus the size of the program, Class inheritance permits a new version of a program to be built without affecting the old [6].

C) Function Generation

- *Information Extraction Function*

Information Extraction is nothing but the data which are less important than the other data which would be needed for extracting data developing application. Data are such as section containing information related to the input file, Author data who

has extracted the data, date when the file was generated and remark when anything updated.

- *Meta Data Extraction Function*

Meta data are the data which would be needed by any application for studying the evolutionary step. This data is such as residue forming helix, beta and atom present according their section and the last molecule that is TER data. So the generated function will be like get. Helix (), get Sheet ().

- *Manipulations Function*

This function is related to calculation such as get particular co-ordinate data which will give all the co-ordinate value such as x, y, z co-ordinate data, contact map calculation function which will help to directly calculate the contact map from the co-ordinate data, get current cursor function will give the value where point is located. So the generated function will be like get co-ordinate of res.name (), get. Current cursor ()

IV. CONCLUSION

As there are no such software developing kit to solve the problem related to the reading the protein file, this proposed application will be more useful. By developing such kind of platform will help the developers to build the application more rapidly and thus will reduce the time complexity as the software will be ease for developer to use, to build and to plot the PDB file data and finally there will be the leading step for studying the evolutionary process. This is a set of design techniques that will make software more reusable

REFERENCES

[1] Fernanda Hembecker, Heitor Silvério Lopes (2010), "A Molecular Model for Representing Protein Structures and its Application to Protein Folding" IEEE.

[2] Narjes Khatoon Habibi, Mohammad Hossein Saraee (2009), "Protein Contact Map Prediction Based On an Ensemble Learning Method", In: International Conference on Computer Engineering and Technology.

[3] Yosi Shibberu and Allen Holder (2011), "A Sepctral Approach to Protein Structure Alignment". In: IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, NO. 4.

[4] Nitin Gupta, Nitin Mangal and Somenath Biswas (2004), "Evolution and similarity evaluation of Protein Structures in Contact map space

[5] http://www.ics.uci.edu/~baldig/ betasheet_data.html, 2009

[6] http://www.laputan.org/drc.html

[7] Kibeom Park, Michele Vendruscolo, and Eytan Domany (2000), "Toward an Energy Function for the Contact Map Representation of Proteins", In: PROTEINS: Structure, Function, and Genetics 40:237–248

[8] Zafer Aydin, Yucel Altunbasak, and Hakan Erdogan (2011), "Bayesian Models and Algorithms for Protein B-Sheet Prediction".In: Transactions on computational biology and bioinformatics, vol. 8, no. 2

[9] The Protein Data Bank," http:// www.rcsb.org/pdb, 2009

[10] Introduction to protein structure and structural bioinformatics, "secondary-sructure.html"

◈◈◈