WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Review: Protein Structure Visualization Tools and ORF Scanning Tools

Rohit[1], Ruchika[1], Sanjana[1], Shivam[1], Shivani[1], Rachna Yadav[2], Anita Grewal[1*]

[1] Department of Biotechnology, UIET, Kurukshetra University Kurukshetra 136119, Haryana
[2] Department of Biotechnology, Indira Gandhi University, Meerpur, Rewari, Haryana.
Email*: apunia2015@kuk.ac.in

1. Introduction:

1.1 Protein structure visualization tools

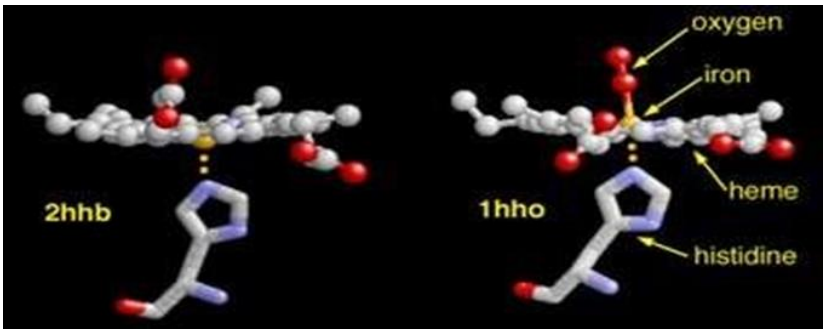
Prior to the invention of computer visualisation software, molecular structures were represented physically using wire, rod, and spherical models made of metal. Programmes were created to visualise and manipulate three-dimensional structures as computer hardware, software, and graphics technology advanced. The computer images assist in analysing and comparing protein structures to determine how proteins operate. The scientists use molecular visualisation to bioengineer protein molecules. This branch of bioinformatics gives bio scientists a full-fledged, scientific excitement because to its user-friendly graphic interface. Molecular visualisation instruments: Many software programmes, both free and paid, are available to visualise the biomolecules.. The most commonly used free software are:

i) Rasmol

Rasmol is a Protein structure visualization tool.This site was established in mid-September 2000 to provide a home for developers of Open-Source versions of RasMol. RasMol is an important scientific tool for visualisation of molecules created by Roger Sayle in 1992. RasMol is used by hundreds of thousands of users world-wide to view macromolecules and to prepare publication-quality images.RasMol is a molecular graphics program proposed for the visualization of proteins, nucleic acids and small molecules.The program is aimed at display, teaching and generation of publication quality images. The program reads in molecular coordinate files and interactively displays the molecule on the screen in a variety of representations and colour schemes.RasMol runs on wide range of architectures and operating systems including Microsoft Windows, Apple Macintosh, UNIX and VMS systems.

Characteristics-

1. The ability to automatically mark non bonded atoms in wireframe and stick displays.
2. The ability to report coordinates.
3. Additions to the list of pre-defined colours.
4. Improved accuracy of coordinates in pseudo-PDB output.
5. Updating the picture title with the PDB ID code and EXPDTA information, so models will be clearly distinguished from experimental data.
6. Introduction of a multilingual structure for RasMol.
7. Population of messages and menu lists for English and Spanish.
8. Correction of coordinate handling for Mol2 and XYZ coordinates
9. An attempt to fix some of the chirality reversals in some of the output modes.



Link for RasMol-: http://www.openrasmol.org

RasMol Features:

The program consists of two windows:

a) For the command line

b) For providing the graphics.

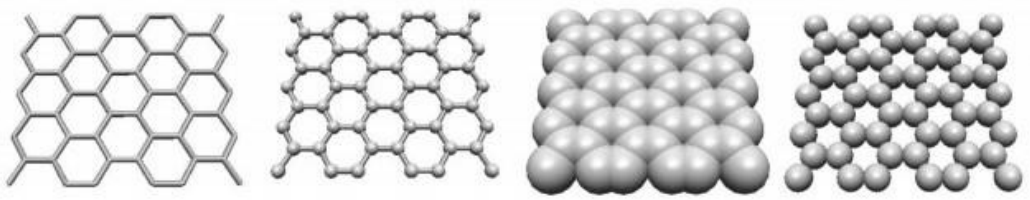
1. Input file format: The input file can be in PDB format and can be downloaded from the PDB structure database. Protein Data Bank (PDB) files can be downloaded for visualization from members of the Worldwide Protein Data Bank (wwPDB). These have been uploaded by researchers who have characterized the structure of molecules usually by X-ray crystallography, protein NMR spectroscopy, or cryo-electron microscopy.

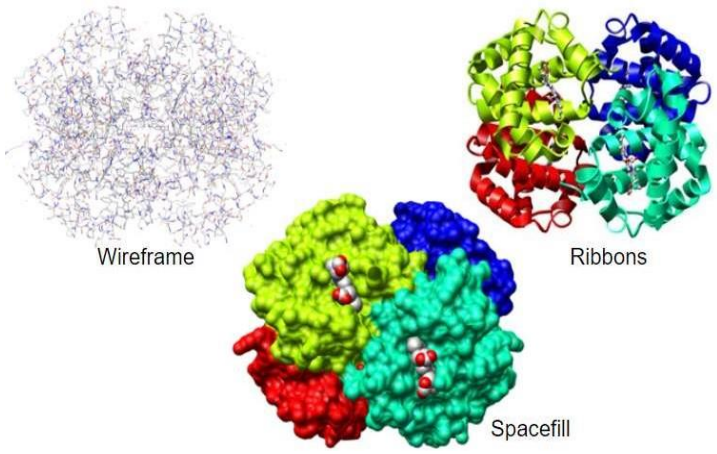
2. Display: There are different ways of displaying coordinates. These include: wireframe, sticks, spacefill, strands & cartoons.

A **textfile** that includes **atomic coordinates**, observed sidechain rotamers, secondary structure assignments, atomic connectivity, ...

record type	atom number	atom	amino acid	chain ID	residue number	coordinates x	y	z	occupancy	temperature factor	element name
ATOM	1	N	MET	D	1	14.322	20.430	-2.337	1.00	17.78	N
ATOM	2	CA	MET	D	1	14.423	20.285	-0.855	1.00	18.66	C
ATOM	3	C	MET	D	1	15.153	21.479	-0.242	1.00	18.46	C
ATOM	4	O	MET	D	1	15.811	22.241	-0.941	1.00	18.84	O
ATOM	5	CB	MET	D	1	15.068	18.970	-0.457	1.00	20.20	C
ATOM	6	CG	MET	D	1	16.569	18.895	-0.674	1.00	20.60	C
ATOM	7	SD	MET	D	1	17.240	17.319	-0.103	1.00	22.81	S
ATOM	8	CE	MET	D	1	16.378	16.194	-1.196	1.00	13.23	C
ATOM	9	N	LEU	D	2	14.983	21.653	1.071	1.00	18.40	N
ATOM	10	CA	LEU	D	2	15.568	22.825	1.718	1.00	19.14	C
ATOM	11	C	LEU	D	2	17.093	22.722	1.765	1.00	18.53	C
ATOM	12	O	LEU	D	2	17.655	21.647	1.945	1.00	19.07	O
ATOM	13	CB	LEU	D	2	15.025	23.078	3.121	1.00	21.35	C
ATOM	14	CG	LEU	D	2	15.438	24.404	3.773	1.00	22.45	C
ATOM	15	CD1	LEU	D	2	14.856	25.606	3.049	1.00	23.53	C
ATOM	16	CD2	LEU	D	2	15.042	24.430	5.244	1.00	23.83	C

The initial image is shown as a „wire" model. -from the Display menu one can choose other visualisation styles such as spacefill", „stick", „ball and stick" as well as the visually most attractive, ribbon" and ‚cartoon" models.In the last two styles, alpha helices are rendered as helical ribbons and beta structures as flat arrows pointing in the direction of the polypeptide chain.





3. Colour:

The atoms of the model can be coloured by the standard CPK (named after Corey, Pauling and Koltun) To colour by atom type: Colours/CPK

Carbon: grey Hydrogen: white Oxygen: red Nitrogen: blue Sulphur: yellow Iron: yellow

The protein can be coloured based on polypeptide chains, the chemical property of the amino acids.

To colour by the protein-secondary structure: Colours/Structure

-helices: magenta

-sheets: yellow -turns: pale blue

all other residues: white

The structure can be cut in the z-dimension

The left and right mouse buttons can be used to rotate the protein along the „x" and „y" axes.

By clicking any part of the structure, the residue number of the given chain and the particular atom will be shown in the command window.One can select a chain, a particular residue or segment of the chain by using the "select" command. "Select 25A" in the command window means that the 25th residue of chain A will be selected.The name of the residue can also be used. A segment of a chain can also be selected: for example, "select 1-33" means that the first 33 residues of the chain will be selected. If the structure contains a ligand (coenzyme, substrate, metal ion etc.) besides the polypeptide chain, it can be selected by the "hetero" command or by its name (e.g., "ca" refers to a Ca2+ ion).The background of the image can be set by the "background colour" syntax. If we want to remove part of the structure, it can be done by using the "restrict" command (e.g. "restrict 1-56" will remove the rendering of the chain from residue 57 to the C-terminal end). One can save the modified structure (e.g. for later manipulation) by the "write script" command and a file name. The finished structure can be saved in common graphics file formats (gif, jpeg, etc.). The "help" menu can explain many additional commands that can be used to manipulate the structure.

ii) Swiss-PDB Viewer

SWISS-MODEL is a structural bioinformatics web-server dedicated to homology modelling of protein 3D structures. Swiss-PdbViewer has been developed since 1994 by Nicolas Guex.Swiss-PDB Viewer is tightly linked to SWISS-MODEL, an automated homology modelling server developed within the Swiss Institute of Bioinformatics (SIB) at the Structural Bioinformatics Group at the Biozentrum in Basel.Swiss-PDB Viewer is an application that provides a user-friendly interface allowing to analyse several proteins.The proteins can be superimposed in order to deduce structural alignments.It is used to compare their active sites or any other relevant parts.Amino acid mutations, H-bonds, angles and distances between atoms are easy to visualize.Swiss-PDB Viewer can also read electron density maps, and provides various tools to build into the density.Various modelling tools are integrated and residues can be mutated.Homology modelling is currently the most accurate method to generate reliable three- dimensional protein structure models.Homology (or comparative) modelling methods make use of experimental protein structures ("templates") to build models for evolutionary related proteins ("targets").

Today, SWISS-MODEL consists of three tightly integrated components:

The SWISS-MODEL pipeline – a suite of software tools and databases for automated protein structure modelling SWISS-MODEL pipeline comprises the four main steps that are involved in building a homology model of a given protein structure:

a) Identification of structural template(s). BLAST and HHblits are used to identify templates. The templates are stored in the SWISS-MODEL Template Library (SMTL), which is derived from PDB.

b) Alignment of target sequence and template structure(s).

c) Model building and energy minimization. SWISS-MODEL implements a rigid fragment assembly approach for modelling.

d) Assessment of the model's quality using QMEAN, a statistical potential of mean force.

The SWISS-MODEL Workspace(A web-based graphical user workbench) In this mode the input is a project file that can be generated by the DeepView (Swiss PDB Viewer) visualization and structural analysis tool, to allow the user to examine and manipulate the target-template alignment in its structural context.The SWISS-MODEL Repository The SWISS-MODEL Repository provides access to an up-to-date collection of annotated three-dimensional protein models for a set of model organisms of high general interest. SWISS-MODEL Repository is integrated with several external resources, such as UniProt, InterPro, STRING, and Nature PSI SBKB.

Uses:

1. To find hydrogen bonds within proteins and between proteins and ligands.
2. To view several protein structures simultaneously and superimpose them to align their structures and sequences.
3. To examine electron-density maps from crystallographic structure determination
4. To judge the quality of maps and models, and to identify many common problems in protein models.
5. It computes electrostatic potentials and molecular surfaces, and carries out energy minimization.

New developments of the SWISS-MODEL expert system feature

a) automated modelling of homo-oligomeric assemblies

b) modelling of essential metal ions and biologically relevant ligands in protein structures

c) local (per-residue) model reliability estimates based on the QMEAN local score function

d) mapping of UniProt features to models.

1.2 Open Reading Frame Scanning Tools

In molecular genetics, an open reading frame (ORF) is the part of a reading frame that contains no stop codons. The transcription termination pause site is located after the ORF, beyond the translation stop codon, because if transcription were to cease before the stop codon, an incomplete protein would be made during translation.Normally, inserts which interrupt the reading frame of a subsequent region after the start codon cause frame shift mutation of the sequence and dislocate the sequences for stop codons.

Significance:

One common use of open reading frames is as one piece of evidence to assist in gene prediction. Long ORFs are often used, along with other evidence, to initially identify candidate protein coding regions in a DNA sequence.The presence of an ORF does not necessarily mean that the region is ever translated. For example, in a randomly generated DNA sequence with an equal percentage of each nucleotide, a stop-codon would be expected once every 21 codons.A simple gene prediction algorithm for prokaryotes might look for a start codon followed by an open reading frame that is long enough to encode a typical protein, where the codon usage of that region matches the frequency characteristic for the given organism's coding regions. By itself even a long open reading frame is not conclusive evidence for the presence of a gene.

Example:

If a portion of a genome has been sequenced (e.g., 5'-ATCTAAAATGGGTGCC-3'), ORFs can be located by examining each of the three possible reading frames on each strand. In this sequence two out of three possible reading frames are entirely open, meaning that they do not contain a stop codon:

1. ...A TCT AAA ATG GGT GCC...
2. ...AT CTA AAA TGG GTG CC...
3. ...ATC TAA AAT GGG TGC C...

Possible stop codons in DNA are "TGA", "TAA" and "TAG". Thus, the last reading frame in this example contains a stop codon (TAA), unlike the first two.

1.3 Open Reading Finding Tool:

i) ORF Finder:

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

ii) ORF Investigator:

ORF Investigator is a program which not only gives information about the coding and noncoding sequences but also can perform pairwise global alignment of different gene/DNA regions sequences. The tool efficiently finds the ORFs for corresponding amino acid sequences and converts them into their single letter amino acid code, and provides their locations in the sequence. The pairwise global alignment between the sequences makes it convenient to detect the different mutations, including single nucleotide polymorphism.

Needleman and Wunsch algorithms are used for the gene alignment. The ORF Investigator is written in the portable Perl programming language, and is therefore available to users of all common operating systems.

iii) ORF Predictor:

ORF Predictor is a web server designed for identifying protein-coding regions in expressed sequence tag (EST)-derived sequences. For query sequences with a hit in BLASTX, the program predicts the coding regions based on the translation reading frames identified in BLASTX alignments, otherwise, it predicts the most probable coding region based on the intrinsic signals of the query sequences. The output is the predicted peptide sequences in the FASTA format, and a definition line that includes the query ID, the translation reading frame and the nucleotide positions where the coding region begins and ends. ORF Predictor facilitates the annotation of EST-derived sequences, particularly, for large-scale EST projects.

REFERENCES:

1. The Deep Learning Revolution" by Terrence J. Sejnowski.
2. bioinfo.vanderbilt.edu/zhanglab/lectures/AB2011Lecture07.pdf
3. ORF Finder (bioinformatics.org)