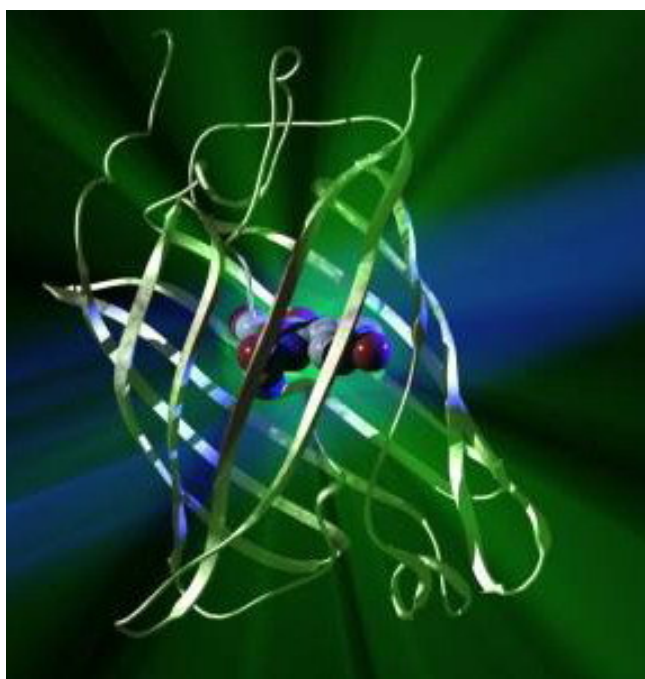




Structural Bioinformatics – Towards Unraveling ‘Molecular Machines’

Muhammad Kamran Haider
kamraider@hotmail.com
B.Sc.(Hons.) Biochemistry
Institute of Biochemistry & Biotechnology,
Punjab University, Lahore, Pakistan



Green Fluorescent Protein – an artistic rendering

Seemingly, simple linear sequences of amino acids – the proteins are even more than ‘Molecular Machines’. Being at the very basis of life, they mediate almost all the processes characteristic to living things. They act as biological catalysts (enzymes), information processors (G-proteins), chemical messengers (hormones), carrier vehicles (globins), force generators (myosin), photosensitive modules (opsins), immune bullets (immunoglobulins) and finally mechanical substructures (collagens). What lies beneath this colossal functional versatility is the unique three-dimensional structure of proteins that confers specific

biological properties and hence functions to them. Structural probe in biochemistry has long sought to unveil the macromolecular gadgetry of biological phenomena. Now, it seems to be accelerated in new dimensions as Structural Bioinformatics continues to mount the arsenal of highly sophisticated computational tools for the avant-grade task of Predictive Structure Elucidation. Biochemical intuition integrated with these computational tools holds great promises in the fields of Novel Protein Engineering, Intelligent Mutagenesis, Rational Drug Design and Structure-oriented Molecular Phylogeny.

Cellular biosynthetic factories, composed of ribosomes, endoplasmic reticulum, nucleic acids & other proteins, manufacture proteins initially as linear polypeptides from DNA-derived genetic information encrypted in mRNA. Once synthesized, these polypeptides decrease their conformational entropy and attain an ordered, well-defined and unique 3-D structure. This process, yet to be known to full scope, is called Protein Folding. The aqueous biological media (cytosol) plays its special role in 3-D rendering by ramming the hydrophobic amino acids into the core of protein structure and bringing hydrophilic amino acids at macromolecular projections. With this 3-D assemblage, different fragments of amino acids within the chain assume different arrangements like helices, pleated sheets and loops, known as secondary structures. These secondary modules give rise to tertiary structure or protein fold which is full 3-D atomic structure of a single polypeptide chain. This structure, called as native conformation, is fully functional and stable and is poised to join the ‘myriad of sorts’ to create mysteries of life.

Proteins are the paragons of structure-functional relationship in biological systems. The structure is key to the understanding of protein function. While at work, these bodily functionaries are even more significant as the ‘molecular disorders’ at their part (due to genetic mutations) lead to diseased conditions. Perhaps, for this reason, proteins have been ideal drug targets and a wide range of pharmaceutical agents act through binding various proteins. Thus, the knowledge of protein structure can aid the process of rational drug design.

Knowledge of protein structure mainly comes from experimental techniques, but our ability to determine structures experimentally cannot keep pace with sequence determination. This is due to the fact that with the large volume of genes being sequenced, and the use of informatics tools to reproduce sequence data, the rate of new protein sequences is growing exponentially. On the other hand experimental structure determination is a slow and labor-intensive process. The two main techniques that it utilizes; X-ray Crystallography and NMR Spectroscopy, are sometimes hampered by problems like non-crystallizability of certain proteins, inadequate atomic resolution and resource limitations.

Structural data streaming through biochemical research are stored in structural databases. Published protein structures are submitted to structural databases e.g., Protein Databank (**PDB**)<www.pdb.org> is the worldwide archive for protein structural data obtained mainly from X-ray crystallography. Molecular Modeling Database (**MMDB**)

www.ncbi.nlm.nih.gov/Structure/MMDB/ is another structural database that provides extensive information on protein sequence and structure.

Similarly, protein sequences are stored in primary sequence databases that are repositories of raw sequence data. For instance, **SWISS-PROT** & **TrEMBL** are the two major databases for the storage of protein sequences. SWISS-PROT www.expasy.ch provides the most up-to-date and extensively annotated information on protein sequences. Both sequence and structure data require standard formats to be used for input to databases and computer applications. Most common sequence formats are NBRF/PIR, FASTA and GDE formats. Each of these formats displays sequence in text form, which is easily readable both by humans and computers. Sequence files, commonly, have any one of the three above-mentioned formats and hence the extensions .seq, .fasta or .gde. These files carry unique code for identification and comments about the source of protein & accession number in the database. The standard file format for protein structural data is PDB. These are text files using a format devised by PDB. Such files have extension .pdb, and contain orthogonal atomic coordinates together with annotations, comments and details.

For the sequence & structure data, which is widely distributed over the WWW, to be useful for scientists, there are several data retrieval tools, provided for the most part, by database curators. **Entrez** www.ncbi.nlm.nih.gov/Entrez/ is an online data retrieval system developed by National Center for Biotechnology Information (NCBI), which integrates information held in all NCBI databases (e.g., **GenBank** & its subsidiaries) and other world topmost biological databases including SWISS-PROT, PDB and MMDB. **ExPASy** (*Expert Protein Analysis System*) is the proteomics server <http://www.expasy.org/swissmod> of the Swiss Institute of Bioinformatics. Besides data retrieval it provides many online tools for analysis of protein sequences and structures. It covers several protein databases such as **SWISS-PROT**, **TrEMBL**, **PROSITE** & **PDB**.

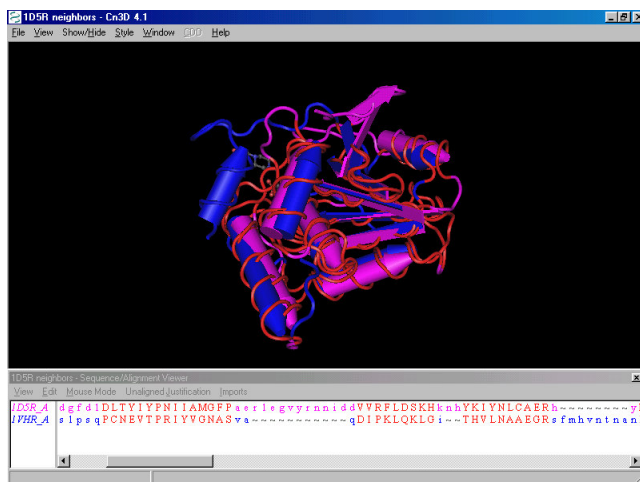


Fig.2. Sequence & Structural Alignment of Human PTEN tumor suppressor protein with its neighbor Dual Specificity Phosphatase.

Sequence & structure data, retrieved through above-mentioned methods, is input to enormous number of bioinformatics softwares, for various types of biomolecular analyses. The huge variety of these softwares accomplishes tasks ranging from simple 3-D structure visualization to advanced structure prediction tools based on statistical thermodynamics and quantum mechanics. **Rasmol**, **Cn3D**, **Chime** are among commonly used programs for visualization and

small-scale manipulation of structural data. Rasmol is a simple standalone application that displays 3-D images whereas Chime has the

same functionality, but it is a browser plug-in to visualize 3-D structures on web pages. Cn3D is another useful visualization tool and what sets Cn3D apart from other software is its ability to correlate structure and sequence information by producing sequence and structure-based alignments. Alignments define relationships between the sequence and structure elements of one biomolecule and those of another, and the degree of similarity reflects the basic data on which we base the biological relationships and further utilize in sequence and structure predictions.

There are several other web-based or locally installed tools used for pivotal computational tasks in structural bioinformatics including physicochemical, thermodynamic, compositional, conformational and functional domain analyses, sequence-based and structure-based, pair-wise and multiple alignments, and finally the structure prediction.

It is useful to categorize structure prediction methods theoretically, either *ab initio* or knowledge-based. *Ab initio* methods attempt to calculate or make a suitable approximation of the minimum thermodynamic energy possessed by a protein when it attains the stable native fold. Proteins, when modeled with enough solvent molecules to be realistic, are huge systems with thousands of atoms and making such detailed calculations is so far an ordeal. For these reasons, *ab initio* methods, despite being more attractive from intellectual viewpoint, are of little use to practicing biochemists. In contrast to *ab initio* methods, knowledge-based methods attempt to predict protein structure from its sequence using information from database of known structures. These methods include comparative modeling, fold recognition, secondary structure prediction. The most accurate and comprehensive structure prediction method is comparative modeling.

The basis of comparative modeling lies in the fact that protein sequences with more than 25% similarity over an alignment of 80 residues or more adopt same basic structure. The

similarity of structures is very high in the so-called core regions, which typically are comprised of a framework of secondary structure elements such as α -helices and β -sheets.

The process of building a comparative model is conceptually straightforward. It begins with the framework construction, which is computed by averaging the position of each atom in the target sequence, based on the location of the corresponding atoms in the template. Following framework generation loops are constructed. Loops form non-conserved surface residues so loop structure prediction is more difficult than backbone prediction. The simplest method makes use of a database of known loop structure from other proteins that are not necessarily similar in sequence. For each loop needed, the database is searched for best possible loop, which fits into the gap of modeled structure. After above steps, the process turns to amino acid side chain positions. For protein side chains there is no structural information available in the templates therefore these cannot be built during the framework generation and must be added later. The number of side chains that need to be built is dictated by the degree of sequence identity between target and template. For this purpose, a table of most probable rotamers for each amino acid side chain is analysed to see if they are acceptable by a van der Waals exclusion test to see if side chains are closer than sterically permissible region. After adding most favoured rotamers, the preliminary structural model for the target sequence is complete. Sometimes, however, it is useful to refine model further, Typically this involves idealisation of bond geometry and removal of unfavorable non-bonded contacts by energy minimization software. Finally, the modeled structure is checked for accuracy from several aspects. As mentioned earlier, the correctness of a model is essentially

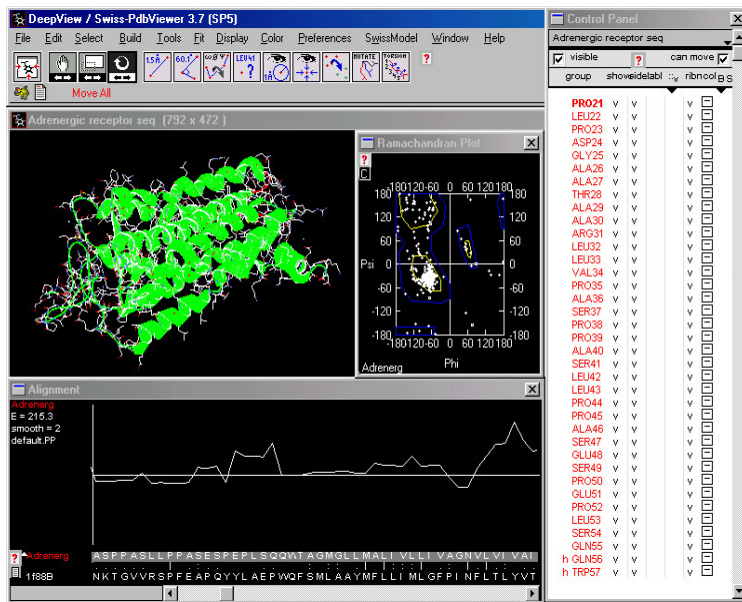


Fig.2. Human β -adrenergic receptor being modeled in SPDBV in combination with Swiss Model server. Depending upon the amount & quality of information, Homology Modeling can generate predictive structures approaching X-ray accuracy.

dictated by the quality of the sequence alignment used to guide the modelling process.

DeepView Swiss-DBViewer (or **SPDBV**) is an interactive molecular graphics program for viewing and analyzing protein and nucleic acid structures. It can also perform Homology Modeling in combination with **Swiss-Model**, a server for automated comparative protein modeling <<http://www.expasy.org/swissmod>>. It can be accessed via ExPASy. A user can generate a modeling project on DeepView and submit it to Swiss Model server. A target sequence can be obtained in many ways. In

order to retrieve the sequence of a human protein whose structure is not experimentally solved we'll have to look first at its corresponding genes. This is done by searching the human genome draft at NCBI, which is a staggeringly vast store of information about the human genome, including over 99% of its sequence. For every gene in human genome, there is a thoroughly annotated page called **Locus Link** page. Along with great variety of resources for the gene, there is NCBI's Reference Sequence that is linked to the Entrez Protein page where sequence of the gene product is found and that is actually the target sequence. This sequence can be saved locally in FASTA format and loaded in DeepView, which initially models the protein as a helix. To find the homologous template on which the target sequence is built to give model, DeepView automatically submits the target sequence to ExPASy site where **BLAST** (a tool for similarity search) is used to search the **ExpPDB** database for appropriate templates. The ExpPDB database is a subset of the PDB, containing all templates available for the Swiss Model server. The template files retrieved from ExpPDB are run on DeepView. The next step is that of adjusting a sequence alignment between the target and the templates. After completing sequence alignment and its manual refinement the project should be saved and submitted to Swiss-Model. Swiss-Model returns modeled 3D structure by e-mail. This structure can be evaluated, realigned and remodeled, if needed.

A main limitation of comparative modeling is the occasional lack of availability of suitable template structure on which the model is to be built. This is because the database of sequences, whose structures are experimentally solved, is quite small. This situation will be changed in coming years as several large-scale projects aimed to create high-throughput structure determination are underway.

With ever refining structural bioinformatics tools fostering the prediction of structures for all protein sequences, especially of complete genomes, in conjunction with experimental work is a realistic goal and when achieved would be a marvel of 21st century science. It would aid in the quest of molecular insight to the life and would reorient the biochemical research and applied spheres of Protein Engineering, Devised Mutagenesis, Organismal Evolution and Structure-based Drug Design.