Databases and Tools
CHAPTER III

DATABASES AND TOOLS

3.1. Sequence Retrieval

3.1.1. Uniprot/Swissprot

It is the central protein repository created through the combination of Swiss-Prot, TrEMBL and PIR-PSD databases (Uniprot, 2010). It is a high-quality, comprehensive and freely accessible database of protein sequences. It comprises of functional information derived from research literature. Importantly, Uniprot/Swissprot is a manually annotated, non-redundant protein sequence database (Bairoch et al., 2004).

URL: http://www.expasy.org/

3.2. Structure Retrieval

3.2.1. Protein Data Bank

The Protein Data Bank (PDB) is a repository for the three-dimensional structural data of macromolecules such as proteins and nucleic acids. All these data are generated through the methods of X-ray crystallography, NMR spectroscopy and electron microscopy and submitted by biologists and biochemists from around the world for free access (Berman, 2008). It is a key resource in field of structural biology. The file formats available here are PDB, mmCIF (macromolecular Crystallographic Information file) and an XML version format, called PDBML (Westbrook et al., 2005). All these coordinates files can be visualized using ICM-Browser (Molsoft), VMD (Humphrey et al., 1996), UCSF Chimera (Pettersen et al., 2004), Rasmol (Sayle and Milner-White, 1995), and Swiss-PDB Viewer (Guex and Peitsch, 1997).

URL: http://www.rcsb.org/pdb/home/home.do

3.2.2. Seaweed database

The Seaweed Metabolite database is a publically accessible database which provides information about bioactive secondary metabolites from brown and red algae. Mainly Red algae of the genus Laurencia (Ceramiales, Rhodomelaceae) are the most prolific producers of secondary metabolites in the marine environment. Secondary
metabolites from these algae are predominantly sesquiterpenes, diterpenes, triterpenes and C15-acetogenins, characterized by the presence of halogen atoms in their chemical structures. The entries of this database are generated through text mining of published articles. There are around 517 compounds reported as on Aug, 2012. Out of these 331 compounds are Lipinski compliant. These compounds cover 37 different species of Laurencia and other genera. (Davis and Vasanthi, 2011).

URL: http://www.swmd.co.in/

3.2.3. Pubchem

PubChem is a composite database of chemical molecules and their activities against biological assays which is maintained by the National Center for Biotechnology Information (NCBI). This freely accessible database has millions of compound structures and descriptive datasets that can be freely downloaded via FTP. PubChem is also reported to harbor annotated small molecules with fewer than 1000 atoms and 1000 bonds. More than 80 database vendors contribute to the growing PubChem database with 31 million entries, with pure and characterized chemical compounds (Bolton et al., 2008).


3.2.4. Chemical Book

Chemical Book is a database with 5455261 different products from 12407 global chemicals suppliers (as on Nov, 2013). The available information includes chemical number and molecular formula. Further, the structures are available in MOL file format. In addition to this, they also provide the end users with chemical compounds based on CAS index, chemical name and company website index apart from browse by category which includes amino acid derivatives, unnatural amino acid derivatives, and boronic acid derivatives.

URL: http://www.chemicalbook.com/

3.3. Modeller9v10

MODELLER tool is used for homology modelling of query protein sequence into 3D structure (Martí-Renom et al., 2000). This software is quite inspired by the technique of nuclear magnetic resonance known as satisfaction of spatial restraints, by which a set
of geometrical criteria are used to create a probability density function for the location of each atom in the protein. This program developed by Sali and Blundel is based on sequence homology between the query and the template sequence. Thus they assist in bridging the gap between the number of sequences and the available 3D structures in the Protein Data Bank. They come with different flavors like basic (single template based), advanced (multiple template based and loop refining), iterative and difficult modelling (Sali and Blundell, 1993).

URL: http://salilab.org/modeller/

3.4. Structure Validation

3.4.1. PROCHECK

It is a suite of programs to check the stereo chemical quality of protein structures (crystallized and modelled) by analyzing residue-by-residue geometry and overall structure geometry with the modelling of 3D structures both locally and globally using various criteria. The global analysis investigates the phi, psi and chi 1 torsion angles and hydrogen bond energies. Whereas, the local analysis includes the irregularity in proline phi angles, peptide bond planarity, disulfide bond lengths, and their chi 3 torsion angles (Laskowski et al., 1993).

URL: http://nihserver.mbi.ucla.edu/SAVES/

3.4.2. ERRAT

ERRAT is a protein structure validation algorithm that is especially designed for the evaluation of the progress of crystallographic model building and refinement. This tool mainly analysis the statistics of non-bonded interactions between different atom types. A single output plot generated through this software gives the value of the error function vs. position of a 9-residue sliding window. By comparison with statistics from highly refined structures, the error values have been calibrated to give confidence limits. This is extremely useful in making decisions about reliability of the experimental and generated model (Colovos and Yeates, 1993).

URL: http://nihserver.mbi.ucla.edu/SAVES/
3.4.3. Verify3D

This derives a “3D-1D” profile based on the local environment of each residue, described by the statistical preferences like the area of buried residues, the fraction of side-chain area that is covered by polar atoms (oxygen and nitrogen), and the local secondary structure (Bowie et al., 1991).

URL: http://nihserver.mbi.ucla.edu/SAVES/

3.4.4. QMEAN (Qualitative Model Energy Analysis)

Generally, the experimental and modelled structure needs to be validated to identify the potential residual clashes. Even though many tools are available both on-line and off-line, still one of the prominent tools is QMEAN (Qualitative Model Energy Analysis) which has a composite scoring function describing the major geometrical aspects of protein structures. Here the local geometry is analyzed by a new kind of torsion angle potential over three consecutive amino acids which are very effective in recognizing the native fold. A secondary structure-specific distance-dependent pairwise residue-level potential is used to assess long-range interactions. Further, solvation potential describes the burial status of the residues (Benkert et al., 2008).

URL: http://swissmodel.expasy.org/qmean/cgi/index.cgi

3.4.5. Protein Structure Analysis (ProSA)

ProSA (Sippl, 1993) identifies potential errors in experimentally determined structures and theoretically modelled structures (Banci et al., 2006; Petrey and Honig, 2005). The generated plot through this tool highlights the potential problems in the protein structures. Basically the z-score indicates the overall model quality. The positive value refers to the problematic part of the structure. A single residue energies based plot usually contains large fluctuations and is of limited value for model evaluation. Thus a 40-residue fragment based average energy was calculated to understand the 3D structures. The 3D coordinates of X-ray, NMR and theoretically modelled structures generally get validated here.

URL: https://prosa.services.came.sbg.ac.at/prosa.php
3.4.6. BLAST (Basic Local Alignment Search Tool)

Basic Local Alignment Search Tool (BLAST) is a heuristic approach based algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search generally helps in identification of orthologous or paralogous sequences from the sequence databases (Genbank CDS translation; Protein Data Bank, Swissprot, Protein Information Resource (PIR), Protein research Foundation database of peptides sequences (PRF)) that resemble the query sequence above a certain threshold value. Different types of BLASTs are available according to the query sequences. The BLAST program was designed by Stephen Altschul for the identification of evolutionarily related proteins from different sequences and structural databases (Altschul et al., 1990).

URL: http://blast.ncbi.nlm.nih.gov/Blast.cgi

3.5. Docking Softwares

3.5.1. Glide

Glide from Schrödinger is designed to assist in high-throughput screening of potential ligands based on binding mode and affinity for a given receptor module. One can compare ligand scores with those of other test ligands or compare ligand geometrics with those of a reference ligand. Additionally Glide is used to generate one or more plausible binding modes for a newly designed ligand. Thus, Glide searches for favorable interactions between ligand molecules and a receptor molecule (a protein) (Friesner et al., 2006).

3.5.2. iGEMDOCK

It is a docking tool which follows generic evolutionary method for computing a ligand conformation and orientation relative to the active site of target protein. iGEMDOCK, generally integrates virtual screening stages from preparations to post-screening analysis. After docking the post-screening is carried out based on the pharmacological interactions from screening compounds without relying on the experimental data of active compounds. In particular, the pharmacological interactions represent conserved interacting residues which generally form the binding pockets with specific physio-chemical properties to play the
essential functions of a target protein. Thus, these pharmacological interactions generated by iGEMDOCK are actively involved in biological functions (Yang and Chen, 2004).

URL: http://gemdock.life.nctu.edu.tw/dock/igemdock.php

3.5.3. HADDOCK

HADDOCK (High Ambiguity Driven protein-protein DOCKing) follows information-driven flexible docking approach for the modeling of bimolecular complexes. HADDOCK identifies itself as the docking program based on predicted protein interface and ambiguous interaction restraints (AIR) which is distinct in comparison with ab-initio docking. HADDOCK can deal with large classes of modelling problems including protein-protein, protein-nucleic acids and protein-ligand complexes (De Vries et al., 2010).

URL: http://haddock.science.uu.nl/services/HADDOCK/haddock.php

3.6. Ligand preparation

3.6.1. Ligprep

In silico based approaches plays a vital role in lead identification which ultimately reduces the cost and time during drug designing. Thus, all the methods require accurate 3D molecular models as a starting point. However, many databases only comprises of 2D structures. Moreover, their conversion into a stable 3D format is quite trivial. Efficient and accurate 2D to 3D conversion is therefore a key precursor to computational analyses. Thus, Ligprep generates accurate, energyminimized 3D molecular structures with no steric clashes. Additionally, they also eliminate the compounds that do not fit into the user specified criteria thus allowing the generation of a completely customized ligand library. As a result, they eliminate bad intra-molecular contacts and optimize bond lengths and angles to produce low-energy structures (Ligprep, 2011).

3.6.2. PRODRG

This server generally considers small molecules (3D ) in PDB/mol/mol2 file format and generates varieties of topologies for use with softwares like GROMACS, WHAT IF, Autodock, and HEX in an energy minimized format (Schüttelkopf and van Aalten, 2004).

URL: http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg/submit.html
3.7. ADME Analysis

3.7.1. Qikprop

It is an accurate, easy-to-use adsorption, distribution, metabolism and excretion (ADME) prediction program designed by Professor William L. Jorgensen. This tool predicts physically significant descriptors and pharmaceutically relevant properties for organic structures, either individually or in batches. They also provide ranges for comparing particular molecules properties with those of 95% of known drugs. Qikprop also flags 30 types of reactive functional groups that may cause false positives in high-throughput screening (HTS) assays (Qikprop, 2011).

3.8 Molecular Simulation

3.8.1. Desmond

Desmond is a software package developed at D. E. Shaw Research laboratory to perform high-speed molecular dynamics simulations of biological systems on conventional commodity clusters. The code uses novel parallel algorithms and numerical techniques to achieve high performance and accuracy on platforms containing a large number of processors, but may also be executed on a single computer (Shaw, 2005).

URL: http://www.deshawresearch.com/resources_desmond.html

3.9. Molecular Visualization

3.9.1. SwissPdbViewer

SwissPdbViewer is an interactive molecular graphics program for viewing and analyzing protein and nucleic acid structures. In combination with Swiss-Model (a server for automated comparative protein modeling maintained at http://www.expasy.org/swissmod) new protein structures can also be modeled (Guex and Peitsch, 1997).

URL: http://spdbv.vital-it.ch/disclaim.html

3.9.10. CHIMERA

UCSF Chimera is an extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, structure building, clash detection,
measurements of distances, angles, surface area and volume apart from molecular dynamics trajectory analysis (Pettersen, et al., 2004).

URL: http://www.cgl.ucsf.edu/chimera/download.html

3.10. Hydrophobicity Analysis

3.10.1. ProtScale

ProtScale allows computing and representing the profile produced by any amino acid scale on a selected protein sequence. The most frequently used scales are the hydrophobicity or hydrophilicity scales and the secondary structure conformational parameters scales, but there are many other scales exist which are based on different chemical and physical properties of the amino acids. This software provides around 50 predefined scales entered from the literature (Gasteiger et al., 2005).

URL: http://web.expasy.org/protscale/

3.11. Amino Acid Substitution Tools

3.11.1. PANTHER (Protein ANalysis THrough Evolutionary Relationships)

PANTHER is an amino acid analysis tool which helps in estimating the likelihood of a particular non-synonymous (amino-acid changing) coding SNP to cause a functional impact on the protein. PANTHER aligns a query sequence to hidden Markov models (HMMs) of protein families and subfamilies in its collection. The probability of a variant being pathogenic is calculated from the variation over each alignment column. It mainly calculates the subPSEC (substitution position-specific evolutionary conservation) score based on an alignment of evolutionarily related proteins (Thomas et al., 2003).

URL: http://www.pantherdb.org/tools/csnpScoreForm.jsp

3.11.2. Polyphen 2 (Polymorphism Phenotyping v2)

PolyPhen2 is fully automated software for the prediction of the possible impact of an amino acid substitution on the structure and function of protein. Such analysis plays a significant role in the interpretation of large datasets of rare genetic variants, which has many applications in modern human genetics research. This software generally extracts various sequence and structure-based features of the substitution site and feeds them to a
probabilistic classifier to identify rare alleles that might cause Mendelian disease (Adzhubei et al., 2010).

URL: http://genetics.bwh.harvard.edu/pph2/

3.11.3. I-Mutant 2.0

I-Mutant2.0 is a support vector machine (SVM)-based tool for the automatic prediction of protein stability changes upon single point mutations. Here the prediction is generally carried out using either the protein sequence or their structure. The training sets for the same are derived from ProTherm which is based on thermodynamic experimental data of free energy changes of protein stability upon mutation under different conditions (Capriotti, 2005).

URL: http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi

3.12. Multiple sequence alignment

3.12.1. ClustalW

Clustal is a popular multiple sequence alignment computer program with three flavors viz ClustalW (command line interface), ClustalX (graphical user interface) and ClustalOmega (for aligning thousands of sequences using multiple processors). (Chenna R, 2003; Larkin et al., 2007; Thompson et al., 1997; Sievers, 2011) This program accepts the sequences in NBRF/PIR, FASTA EMBL/Swiss-Prot, Clustal, GCC/MSF, GCG9 RSF, and GDE file format. They generate multiple alignment through three steps which includes pairwise alignment, guide tree generation and multiple alignment based on the guide tree.