Chapter 3

Tools & Databases
Listed are the resources and introduction for online as well as offline tools, servers and databases, used in current research work

3.1 Online Tools and Servers:

  BLAST (Basic Local Alignment Search Tool) is a similarity search program developed at NCBI. It is available as a free service over the Internet that provides very fast, accurate and sensitive database searching. BLAST uses a heuristic algorithm that seeks local as opposed to global alignment and is therefore able to detect relationships among sequences that share only isolated regions of similarity. I have utilized it for subtractive study.

- TMHMM (version 2.0) ([http://www.cbs.dtu.dk/services/TMHMM-2.0/](http://www.cbs.dtu.dk/services/TMHMM-2.0/))
  TMHMM is a server for prediction of transmembrane helices in proteins provided by Centre of Biological Sequence analysis (CBS). TMHMM allows prediction of batch of protein sequences with maximum limit of 4000 sequences in FASTA format during a single run. TMHMM 2.0 is also available as a stand-alone software package.

  An automatic server developed by Institute of Enzymology at Hungary for predicting transmembrane helices and topology of proteins. The software is based on trained Hidden markove model for prediction and has acceptable accuracy.

- PSORTb (version 3.0.2) ([www.psort.org/psortb/](http://www.psort.org/psortb/))
  A web-based server which handles archaeal sequences as well as Gram-positive and Gram-negative bacterial sequences for bacterial protein subcellular localization prediction. The server is maintained by Brinkman Laboratory, Simon Fraser University, British Columbia, Canada. The prediction was carried
out by setting parameter of organism type to Bacteria, Gram stain to negative, output format to normal and show result to via web.

- **TargetP (version 1.1)** ([http://www.cbs.dtu.dk/services/TargetP/](http://www.cbs.dtu.dk/services/TargetP/))
  A neural network-based tool, provided by Centre of Biological Sequence analysis (CBS) for large-scale subcellular location prediction of newly identified proteins. The prediction was carried out by setting parameter of organism group to non-plant and cutoffs to no cutoffs. TargetP allows prediction of batch of protein sequences with maximum limit of at most 2,000 sequences and 200,000 amino acids per submission; each sequence not more than 4,000 amino acids in FASTA format during a single run.

- **PSLpred** ([http://www.imtech.res.in/raghava/pslpred/](http://www.imtech.res.in/raghava/pslpred/))
  The SVM based server maintained by Centre of Bioinformatics at UCL Global University London. A web server for predicting subcellular localization of gram-negative bacterial proteins with an overall accuracy of 91.2%. The prediction method includes various SVM modules based on different feature of proteins such as Amino acid composition, Dipeptide Composition, Composition of 33 physico-chemical properties and evolutionary information of PSI-BLAST. The predictions are performed by masking sequences with low complexity regions.

- **CPHmodels** ([http://www.cbs.dtu.dk/services/CPHmodels/](http://www.cbs.dtu.dk/services/CPHmodels/))
  CPHmodels 3.2 is a protein homology modeling server, provided by Centre of Biological Sequence analysis (CBS). The template recognition is based on profile-profile alignment guided by secondary structure and exposure predictions. Prediction of only one sequence, with not less than 15 and not more than 4,000 amino acids is allowed per submission.
SMART (http://smart.embl-heidelberg.de/help/smart_about.shtml)
SMART, a Simple Modular Architecture Research Tool allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signalling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Each domain found in a non-redundant protein database as well as search parameters and taxonomic information are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa.

STRING (http://string-db.org/)
STRING is a search Tool for the Retrieval of Interacting Genes/Proteins. The search tool uses STRING database for predicting interactions. STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources namely Genomic Context, High-throughput Experiments, Conserved Coexpression and known literature. STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 5214234 proteins from 1133 organisms.

BHAGEERATH (http://www.scfbio-iitd.res.in/bhageerath/index.jsp)
An Energy Based automated Protein Structure Prediction Server BHAGEERATH is available at Supercomputing Facility for Bioinformatics and Computational biology (SCFBIO), IIT Delhi. The present version of Bhageerath accepts amino acid sequence and secondary structure information to predict 5 candidate structures for the native. It is anticipated that at least one native like
structure (RMSD < 7Å without end loops) is present in the final structures. The server has been validated on 80 small globular proteins.

- ModWeb (https://modbase.compbio.ucsf.edu/modweb/)
  MODWEB is a web server for automated comparative protein structure modeling that relies on MODPIPE. ModPipe is a completely automated software pipeline that can calculate protein structure models for a large number of sequences without manual intervention. MODWEB server is also available at the Protein Model Portal (PMP) as an option of interactive modeling server named Modweb server and in current studies, MODWEB server is used via PMP.

- PatchDock Server http://bioinfo3d.cs.tau.ac.il/PatchDock/patchdock.html
  PatchDock is an algorithm for molecular docking, based on shape complementarity criteria. The input is two molecules of either type proteins or DNA or peptides or drugs. PatchDock algorithm is inspired by object recognition and image segmentation techniques used in Computer Vision. The algorithm has three major stages namely, Molecular Shape Representation; Surface Patch Matching and Filtering and Scoring.

- FireDock Server (http://bioinfo3d.cs.tau.ac.il/FireDock/)
  FireDock is a Fast Interaction REfinement in molecular DOCKing server. The FireDock server addresses the refinement problem of protein-protein docking solutions. The method simultaneously targets the problem of flexibility and scoring of solutions produced by fast rigid-body docking algorithms. Given a set of up to 1000 potential docking candidates, FireDock refines and scores them according to an energy function, spending about 3.5 seconds per candidate solution. To the best of my knowledge, this is the first web server that allows performing large-scale flexible refinement and scoring of docking solutions online.
SAVES Server (http://nihserver.mbi.ucla.edu/SAVES/)
SAVES is a protein Structural Analysis and Verification Server provided by NIH laboratory of Structural Genomics and Proteomics. It has facility to analyze and verify the protein structure with five different software options namely PROCHECK, WHAT_CHECK, ERRAT, PROVE AND VERIFY_3D. Using them protein can be verified for its stereochemical quality, overall structure geometry and its compatibility of an atomic model (3D) with its own amino acid sequence (1D).

3.2 Offline Tools

Swiss-PDBViewer
Deep View - Swiss-PdbViewer (http://us.expasy.org/spdv/) is an application that provides a user friendly interface allow analyzing several proteins at the same time. The proteins can be superimposed in order to deduce structural alignments and compare their active sites or any other relevant parts. Amino acid mutations, Hbonds, angles and distances between atoms are easy to obtain thanks to the intuitive graphic and menu interface. It helps in viewing, manipulating and modeling of multiple molecules in layers. It helps to manipulate interactive Ramachandran-plot for protein modeling. It has an interface to the Swiss Model database for homology modeling from known sequences and the output to Quick draw 3D and POVRay for publication grade imaging.

PyMOL (http://www.pymol.org/)
PyMOL is an open-source, user-sponsored, molecular visualization system created by Warren Lyford DeLano. It is currently commercialized by Schrödinger; Inc. PyMOL can produce high-quality 3D images of small molecules and biological macromolecules, such as proteins. PyMOL is one of a
few open-source visualization tools available for use in structural biology. PyMOL is written in C, C++ and Python computational languages. The Py portion of the software’s name refers to the fact that it extends, and is extensible by the Python programming language.

- **ArgusLab** ([http://www.arguslab.com/arguslab.com/ArgusLab.html](http://www.arguslab.com/arguslab.com/ArgusLab.html))
  We used Argus Lab as a docking tool because it’s allows user to make a different group like Binding site group-where ligand are actually going to bind or user want to bind at that particular site, and a ligand group. These are a much better option for docking a molecule. Argus Lab allow .PDB or .MOL file format for input option. Argus Lab is a very useful tool for anyone working in a research lab, studying or working in the pharmaceutical industry. The program allows you to draw very complex protein configurations, obtain helical chains of amino acids and folded leaves etc. Argus Lab uses a tree system to organize all the elements to add to any structure.

- **Artemis** ([https://www.sanger.ac.uk/resources/software/artemis/](https://www.sanger.ac.uk/resources/software/artemis/))
  Artemis is a free genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation. Artemis is written in Java, and is available for UNIX, Macintosh and Windows systems. It can read EMBL and GENBANK database entries or sequence in FASTA, indexed FASTA or raw format. Other sequence features can be in EMBL, GENBANK or GFF format.

- **Auto Dock Vina** ([http://vina.scripps.edu/](http://vina.scripps.edu/))
  AutoDock Vina is an open-source program for doing molecular docking. It was designed and implemented by Dr. Oleg Trott in the Molecular Graphics Lab at The Scripps Research Institute. AutoDock Vina tends to be faster and accurate
in comparision to AutoDock 4. AutoDock Vina needs to be used in combination of MGL Tools for preparation of macromolecules.


Visual Studio is a comprehensive collection of tools and services to help you to create a wide variety of applications, both for the Microsoft platform and beyond. Visual Studio also connects all of our projects, teams, and stakeholders. Visual Studio allows designing mission-critical .NET applications, writing blazing fast code with C++ AMP, or testing and debugging a cloud-connected HTML/JavaScript application that runs on many devices.

Some important features and advantages of Visual Studio 2010 are:

- Built-in tools for Windows 7, including multi-touch and “ribbon” UI components.
- Rich, new editor with built-in Windows Presentation Foundation (WPF) that you can highly customize to suit how you work.
- Multi monitor support.
- New Quick Search, which helps to find relevant results just by quickly typing the first few letters of any method, class, or property.
- Great support for developing and deploying Microsoft Office 2010, SharePoint 2010, and Windows Azure applications.
- Multi core development support that allows you to parallelize your applications, and a new specialized debugger to help you track the tasks and threads.
- Improvements to the ASP.NET AJAX framework, core JavaScript IntelliSense support, and the inclusion in Visual Studio 2010 of jQuery, the open-source library for DOM interactions.
- Multi targeting/multi framework support.
- Support for developing WPF and Silverlight applications with enhanced drag-and-drop support and data binding. This includes great new
enhancements to the designers, enabling a higher fidelity in rendering your controls, which in turn enables you to discover bugs in rendering before they happen at run time (which is a great improvement from previous versions of Visual Studio). New WPF and Silverlight tools will help you to navigate the visual tree and inspect objects in your rich WPF and Silverlight applications.

- Great support for Team Foundation Server (TFS) 2010 (and previous versions) using Team Explorer. This enables you to use the data and reports that are automatically collected by Visual Studio 2010 and track & analyze the health of your projects with the integrated reports, as well as keep your bugs and tasks up to date.

### 3.3 Databases

A **database** consists of an organized collection of data for one or multiple uses. Among them Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analyses. The extensive data produced by the biological research raise a need for biological databases. They are needed for storing and handling these biological data along with allowing the manipulation and retrieval of data, in order to get related information either from different databases or from different analysis tools. There are two major methods by which the database along the interactive front end could be created. In both the cases the database which is back end could be created by MySQL but the front end depends on the type of scripting we select.
I. Client-side scripting: when we run our application on our computer and it works without the need of server (installed), which otherwise help with the functionality of the script, is called client-side scripting. But it has few limitations like:

- Download time.
- Browser compatibility
- Visible code – since JavaScript and VBScript code is included in the HTML page, and then anyone can see the code by viewing the page source
- Possible security hazards for the client computer

II. Server-side scripting: When the server is being used to process our application. Here the Script or code that is run on the server does not appear on or in the web page that is sent to your browser. The web page that your browser receives is only the RESULT of the script or database code or processing that happened on the server.

Some of its benefits are:

- Programs run on a known computer – not dependent on browser
- Greater flexibility in what scripts can do – e.g. access databases, modify files on server
- Code is not viewable by users

Along with this it has a limitation that our site (interactive front end) on a server which is capable of running your scripts and it should grant us the permission to create your own scripts. These server-side scripting can be done by using various language in combination of MySQL.
- **DEG** ([http://tubic.tju.edu.cn/deg/](http://tubic.tju.edu.cn/deg/))

DEG is Database of essential genes. DEG hosts records of currently available essential genomic elements, such as protein-coding genes and non-coding RNAs, among bacteria, archaea and eukaryotes. Essential genes in a bacterium constitute a minimal genome, forming a set of functional modules, which play key roles in the emerging field, synthetic biology. To perform BLAST search against DEG, users have 4 options Blast with single gene or multiple gene or annotated genome or un-annotated genome.


National Center for Biotechnology Information (NCBI) is a division of the National Library of Medicine (NIM) at the National Institute of Health (NIH), of federal agency of the US government. The Entrez site for PubMed, nucleotide, genome, protein and structure databases are being maintained in the USA. The mission of the NCBI is to ensure that the growing body of information from molecular biology and genome research is placed in the public domain and is accessible freely to all facets of the scientific community in ways that promote scientific progress.

- **PDB** ([http://www.rcsb.org/pdb/](http://www.rcsb.org/pdb/))

PDB is the single International database for the processing and distribution of 3D biological macromolecular data. The data has been determined by X-ray crystallography and NMR. The structures are annotated by RCSB and released as per the depositor’s specifications.
PubChem provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative. PubChem is organized as three linked databases within the NCBI's Entrez information retrieval system. These are PubChem Substance, PubChem Compound, and PubChem Bioassay. PubChem also provides a fast chemical structure similarity search tool.

ExPasy (Expert Protein Analysis System) (www.expasy.org)
ExPasy is a proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures. This database provides integrated access to a variety of databases and molecular biology analytical tools dedicated to understanding proteins. It provides support for several other databases like SWISSPROT, TrEMBL, PROSIT, ENZYME, which can be use for tasks relevant to proteomics (e.g., obtaining protein physical properties) and Performing similarity searches, pattern searches and profile searches, topology prediction, including primary, secondary and tertiary structure analysis.

KEGG (http://www.genome.jp/kegg/)
The KEGG (Kyoto Encyclopedia of Genes and Genomes) is a suite of databases and associated software for understanding and simulating higher-order functional behaviors of the cell or the organism from its genome information. First, KEGG computerizes data and knowledge on protein interaction networks (PATHWAY database) and chemical reactions (LIGAND database) that are responsible for various cellular processes. Second, KEGG attempts to reconstruct protein interaction networks for all organisms whose genomes are completely sequenced (GENES and SSDB databases). Third, KEGG can be utilized as reference knowledge for functional genomics (EXPRESSION database) and proteomics (BRITE database) experiments.
MEROPS (http://merops.sanger.ac.uk/)

The MEROPS database is an information resource for peptidases (also termed proteases, proteinases and proteolytic enzymes) and the proteins that inhibit them. The MEROPS database uses an hierarchical, structure-based classification of the peptidases. In this, each peptidase is assigned to a Family on the basis of statistically significant similarities in amino acid sequence, and families that are thought to be homologous are grouped together in a Clan. There is a Summary page for each family and clan, and these again have indexes to choose from.