Chapter – III

Materials and Methods
NCBI- National Center for Biotechnology Information

(http://www.ncbi.nlm.nih.gov/)

The National Center for Biotechnology Information (NCBI) provides a comprehensive website for biologists that includes biology-related databases, and tools for viewing and analyzing the data inherent in the databases. A division of the National Library of Medicine at the National Institutes of Health, NCBI is the agency responsible for creating automated systems for storing and analyzing the rapidly growing profusion of genetic and molecular data. One of the most difficult challenges faced in the field of bioinformatics is how to store, in an easily accessible manner, the overwhelming abundance of new information, including the sequences of entire genomes, the ongoing discoveries of new genes and gene products, and the determinations of their functions and structures. NCBI was established as the government's response to the need for more and better information processing methods to deal with this challenge (Benson et al., 1994).

A relatively good overview of the tools and databases that can be accessed through NCBI is provided in the list along the left border of the home page. Clicking on the link entitled About NCBI produces a second menu containing the topics ‘A Science Primer’, and ‘Databases and Tools’, among others. Many bioinformatics terms defined in this section are clear-cut and basic manner. Selecting ‘Databases and Tools’ from the ‘About NCBI’ webpage menu yields a complete and well-ordered listing of accessible
Materials and Methods

information. This web page containing the databases and tools menu is a good choice for those who are inclined toward bookmarking.

The second item under the ‘Databases and Tools’ menu is ‘Entrez Databases’. Entrez is a search and retrieval system developed by NCBI that is capable of accessing integrated information by searching many of the NCBI databases with just one query. The Nucleotide Databases link under the Databases and Tools menu lists all the sequence databases available through NCBI. These sequence databases contain annotated collections of publicly available DNA, RNA and protein sequences. The evolution of bioinformatics data mining methods has been largely driven by the prodigious amount of sequence information collected by scientists in recent years. New sequences of unknown function can be compared with sequences of well-characterized genes and proteins. Similarities can be identified between the new, unknown sequences and the well-characterized sequences, and used to postulate theories regarding function or structure (Altschul et al., 1990). NCBI advances science and health by providing access to biomedical and genomic information. The fasta sequences of all NF-κB protei...s were retrieved in NCBI and further used for other analysis.

UniProt KB/Swiss-Prot (http://web.expasy.org/)

UniProtKB/Swiss-Prot is the manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB). It has high quality annotated and non-redundant protein sequences, which brings together experimental results, computed features and scientific NF-κB and Proteomics
conclusions. The tool was used for the determination of biological process, domain, cellular component etc. And also for the prediction of functional and post translational modifications of NF-κB in each organism.

**PROTPARAM (http://web.expasy.org/protparam/)**

Protparam computes various physico-chemical properties that can be deduced from a protein sequence. The protein can either be specified as a Swiss-Prot/TrEMBL accession number or ID, or in form of a raw sequence. Providing the accession number of a Swiss-Prot/TrEMBL entry, will prompt to select the portion of the sequence on which one would like to perform the analysis. The choice includes a selection of mature chains or peptides and domains from the Swiss-Prot feature table, as well as the possibility to enter start and end position in two boxes. Otherwise, by default the complete sequence will be analyzed (Edelhoch, 1967).

The parameters computed by ProtParam include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). Molecular weight and theoretical pI are calculated as in Compute pI/Mw. The amino acid and atomic compositions are self-explanatory. Is a tool which allows the computation of various physical and chemical parameters of a given protein (Pace et al., 1995). The
Materials and Methods

determination of molecular weight, theoretical pI, atomic composition, half life etc. of NF-kB were predicted using this tool.

PSORT (http://www.psort.org/)

PSORT.org provides links to the PSORT family of programs for subcellular localization prediction as well as other datasets and resources relevant to localization prediction. There are several programs for the localization prediction. Some of the common one which is used are PSORTb, PSORTdb, PSORTII and WoLF PSORT. They are maintained by PSORTb and PSORTdb are maintained by the Brinkman Laboratory, Simon Fraser University, British Columbia, Canada.

Identification of a protein’s subcellular localization provides valuable clues regarding its biological function. Computational subcellular localization analysis of the growing number of completed genomes or individual proteins allows researchers to screen for vaccine/drug candidates, automatically annotate gene products or select proteins for further study. PSORTdb is a database of subcellular localization for bacteria that contains both information determined through laboratory experimentation (ePSORTdb dataset) and computational predictions (cPSORTdb dataset). The dataset of experimentally verified information (>11000 proteins) was manually curated and represents the largest dataset of its kind. The second component of this database contains computational analyses of proteins deduced from the most recent NCBI dataset of completely sequenced
Materials and Methods

genomes. Analyses are currently calculated using PSORTb, the most precise automated subcellular localization predictor for bacterial proteins (Yu et al., 2011).

WoLF PSORT is an extension of the PSORT II program for protein subcellular location prediction. WoLF PSORT converts protein amino acid sequences into numerical localization features; based on sorting signals, amino acid composition and functional motifs such as DNA-binding motifs. After conversion, a simple k-nearest neighbour classifier is used for prediction. Using html, the evidence for each prediction is shown in two ways: (i) a list of proteins of known localization with the most similar localization features to the query, and (ii) tables with detailed information about individual localization features (Paul et al., 2007). It is a bioinformatics tool used for the prediction of protein localization sites in cells. The tool was used to determine the sub-cellular localization of NF-κB in all the different organisms taken for the study.


PubMed is a free digital database of full-text scientific literature in biomedical and life sciences. It grew from the online Entrez PubMed biomedical literature search system. PubMed Central was developed by the U.S. National Library of Medicine (NLM) as an online archive of biomedical journal articles. The full text of all PubMed Central articles is free to read, with varying provisions for reuse. Some participating publishers delay the release of their articles on PubMed Central for a set time after paper publication. As of
January 2013, the archive contains approximately 2.6 million items, including articles, editorials and letters. It appears to be growing by at least 7% per year.

PubMed comprises over 22 million citations for biomedical literature from MEDLINE, life science journals, and online books. PubMed citations and abstracts include the fields of biomedicine and health, covering portions of the life sciences, behavioral sciences, chemical sciences, and bioengineering. PubMed also provides access to additional relevant web sites and links to the other NCBI molecular biology resources. It is a free resource that is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH). Publishers of journals can submit their citations to NCBI and then provide access to the full-text of articles at journal web sites using LinkOut (PubMed, 2006). PubMed is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The database was used for doing literature survey on the amount of work already done and also for the tools and softwares used in in silico proteomic studies.

**TOPPRED (http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html)**

It is a tool for predicting the topology of bacterial inner membrane proteins. It is based on the basis of hydrophobicity analysis, automatic generation of a set of possible topologies and ranking of these according to the positive-inside rule. TopPred is to compile all existing knowledge about topology in order to permit easy access to prediction of
membrane protein topologies. TopPred II is an improved version of the preceding freeware TOP-PRED. The compiled program is very compact (~90 kbytes) and all the default parameters, scales and texts have been built into resources to allow easy access and ability for permanent modifications by the user. Input sequence files are limited to 2000 amino acids. Sequences can be handled one by one or in groups of up to 20. Parameters and scales can be easily modified through standard dialogs, which also enable one to re-establish the TopPred II default values. Calculation of the hydrophobicity profile, transmembrane segments and topologies can be requested (Gunnar, 1992; Miklos et al., 1997). Is a tool which is used for the examining the number of membrane-spanning segments present in a particular sequence. For the determination of the number of transmembrane site and the level of hydrophobicity, this tool was used.

**InterPro (http://www.ebi.ac.uk/interpro/)**

InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. InterPro combines signatures from multiple, diverse databases into a single searchable resource, reducing redundancy and helping users interpret their sequence analysis results. By uniting the member databases, InterPro capitalises on their individual strengths, producing a powerful diagnostic tool and integrated resource.

NF-κB and Proteomics
Materials and Methods

InterPro is used by research scientists interested in the large-scale analysis of whole proteomes, genomes and metagenomes, as well as researchers seeking to characterise individual protein sequences. Within the EBI, InterPro is used to help annotate protein sequences in UniProtKB. It is also used by the Gene Ontology Annotation group to automatically assign Gene Ontology terms to protein sequences (Hunter et al., 2011).

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combined protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. For the determination of the number of domains, repeats and detailed signature matches this tool was used.

GOR-Garnier-Osguthorpe-Robson (http://gor.bb.iastate.edu/cdm/)

The GOR (Garnier-Osguthorpe-Robson) method uses both information theory and Bayesian statistics for predicting the secondary structure of proteins. Over the years, the method has been improved by including larger databases and more detailed statistics, which account not only for amino acid composition, but also for amino acid pairs and triplets. The most crucial change in the algorithm was the inclusion of evolutionary information using
Materials and Methods

PSI-BLAST to increase the information content for improved discrimination among secondary structures.

The GOR method analyzes sequences to predict alpha helix, beta sheet, turn, or random coil secondary structure at each position based on 17-amino-acid sequence windows. The original description of the method included four scoring matrices of size $17 \times 20$, where the columns correspond to the log-odds score, which reflects the probability of finding a given amino acid at each position in the 17-residue sequence. The four matrices reflect the probabilities of the central, ninth amino acid being in a helical, sheet, turn, or coil conformation. In subsequent revisions to the method, the turn matrix were eliminated due to the high variability of sequences in turn regions. The method was considered as best requiring at least four contiguous residues to score as alpha helices to classify the region as helical, and at least two contiguous residues for a beta sheet (Garnier et al., 1978; Mount, 2004).

It is an information theory-based method for the prediction of secondary structures in proteins. For the secondary structure determination of the NF-κB in selected organisms, the GOR was used.

PDB (http://www.rcsb.org/pdb/)

The Worldwide Protein Data Bank (wwPDB) consists of organizations that act as deposition, data processing and distribution centers for PDB data. Members are: RCSB
Materials and Methods

PDB (USA), PDBe (Europe), PDBj (Japan), and BMRB (USA). The wwPDB’s mission is to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community. PDB is a repository for the 3-D structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organizations (Berman et al., 2003; Berman, 2008).

The PDB is a key resource in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, such as the NIH in the USA, now require scientists to submit their structure data to the PDB. If the contents of the PDB are thought of as primary data, then there are hundreds of derived (i.e., secondary) databases that categorize the data differently. The three dimensional structures of NF-κB were retrieved from PDB.

Clustal Omega (http://www.ebi.ac.uk/Tools/services/web_clustalo)

Clustal Omega is the latest addition to the Clustal family. It offers a significant increase in scalability over previous versions, allowing hundreds of thousands of sequences to be aligned in only a few hours. It will also make use of multiple processors, where present. In addition, the quality of alignments is superior to previous versions, as measured by a range of popular benchmarks. CLUSTAL-OMEGA is a general purpose multiple sequence alignment program for proteins and DNA/RNA. It produces high quality MSAs
and is capable of handling data-sets of hundreds of thousands of sequences in reasonable time.

In default mode, users give a file of sequences to be aligned and these are clustered to produce a guide tree and this is used to guide a ‘progressive alignment’ of the sequences. There are also facilities for aligning existing alignments to each other, aligning a sequence to an alignment and for using a hidden Markov model (HMM) to help guide an alignment of new sequences that are homologous to the sequences used to make the HMM. This latter procedure is referred to as ‘external profile alignment’ or EPA (Fabian et al., 2011).

Clustal-Omega uses HMMs for the alignment engine, based on the HHalign package from Johannes Soeding. Guide trees are made using an enhanced version of mBed which can cluster very large numbers of sequences in less time. Multiple alignment then proceeds by aligning larger and larger alignments using HHalign, following the clustering given by the guide tree. In its current form Clustal-Omega has been extensively tested for protein sequences, DNA/RNA support has been added since version 1.1.0. Clustal-Omega accepts 3 types of sequence input: (i) a sequence file with un-aligned or aligned sequences, (ii) profiles (a multiple alignment in a file) of aligned sequences, (iii) a HMM. Clustal Omega is a new multiple sequence alignment program that uses seeded guide
trees and HMM profile-profile techniques to generate alignments. It produces biologically meaningful multiple sequence alignments of divergent sequences.

**ClustalW2 Phylogeny**

([http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/](http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/))

It is a tool to perform basic phylogenetic analysis on a multiple sequence alignment. Phylogenetics aims to model the substitutions that have occurred over evolutionary time and derive and represent the evolutionary relationships between sequences. There are two ways to use this service at the EBI. The first is interactively (default) and the second is by email. Using it interactively, the user must wait for the results to be displayed in the browser window. The email option means that the results will not be displayed in the browser window but instead a link to the results will be sent by email. The email option is the better one to take when submitting large amounts of data or a job that might take a long time to run. The program accepts protein multiple sequence alignments, in the following multiple sequence formats: Pearson (FASTA), ALN/ClustalW, GCG/MSF and RSF (Larkin et al., 2007; Goujon et al., 2010).

It is a commonly used phylogenetic tree generation method provided by the ClustalW2 program. It accepts the multiple sequence alignment in any supported format and provides the tree in Clustal, Distance Matrix and NEXUS format. The above tool is used to construct rooted and unrooted phylogenetic tree of NF-κB sequences.
**Materials and Methods**


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. The PubChem Substance Database contains descriptions of samples, from a variety of sources, and links to biological screening results that are available in PubChem BioAssay. If the chemical contents of a sample are known, the description includes links to PubChem Compound. The PubChem Compound Database contains validated chemical depiction information provided to describe substances in PubChem Substance. Structures stored within PubChem Compounds are pre-clustered and cross-referenced by identity and similarity groups. The PubChem BioAssay Database contains bioactivity screens of chemical substances described in PubChem Substance. It provides searchable descriptions of each bioassay, including descriptions of the conditions and readouts specific to that screening procedure.

Links from PubChem's chemical structure records to other Entrez databases provide information on biological properties. These include links to PubMed scientific literature and NCBI's protein 3D structure resource. Links to depositor web sites provide further information. A PubChem FTP site, Download Facility, Power User Gateway(PUG), Standardization Service, Score Matrix Service, Structure Clustering, and Deposition Gateway are also available. The structural information of drug compounds from obtained from PubChem.
Materials and Methods

CHEM DRAW ULTRA 6.0

ChemDraw Ultra is a chemical structure drawing software package designed for drawing stereochemically correct structures from chemical names, to get accurate IUPAC names for structures and to estimate NMR spectra from a ChemDraw structure with direct atom to spectral correlation. ChemDraw is a molecule editor developed by the cheminformatics company CambridgeSoft. ChemDraw is, along with Chem3D and ChemFinder, part of the ChemOffice suite of programs and is available for Macintosh and Microsoft Windows. The native file formats for ChemDraw are the binary CDX and the preferred XML based CDXML formats. ChemDraw also can import from, and export to MOL, SDF, and SKC chemical file formats (Mills, 2006).

ChemDraw is the drawing tool of choice for chemists to create publication-ready, scientifically intelligent drawings for use in ELNs, databases and publications and for querying chemical databases. Continually building on 25 years of experience in cheminformatics, ChemDraw is the world’s leading chemical drawing program. Hundreds of thousands of users benefit from its ease of use, high quality output, robust chemical intelligence, and integration with the ChemOffice suite. Chemists who use ChemDraw to predict properties are able to save time and reduce costs by identifying compounds that are likely to have the desired properties before actually synthesizing them. Chemists can also save time and increase data accuracy using ChemDraw to generate spectra, construct correct IUPAC names, and calculate reaction stoichiometry.
Materials and Methods

A powerful set of tools to handle substructural query types (such as variable points of attachment, Rgroups, ring/chain size, atom/bond/ring types, and generic atoms) ensures that compounds are quickly and accurately located by searches, no matter how they are stored in commercial, public or in-house databases (Li et al., 2004).

PRODRG (http://davapc1.bioch.dundee.ac.uk/prodrg/)

PRODRG will take a description of a small molecule (as PDB coordinates / MDL Molfile / SYBYL Mol2 file / text drawing) and from it generate a variety of topologies for use with GROMACS, WHAT IF, Autodock, HEX, CNS, REFMAC5, SHELX, O and other programs, as well as energy-minimized coordinates in a variety of formats. PRODRG takes input from existing coordinates or various two-dimensional formats and automatically generates coordinates and molecular topologies suitable for X-ray refinement of protein-ligand complexes. Energy minimization of dug compounds were done using ProDrG.

Swiss-PdbViewer

Swiss-PdbViewer is an application that provides a user friendly interface allowing to analyze 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, H-bonds, angles and distances between atoms are easy to obtain. Swiss-PdbViewer has been developped since 1994 by Nicolas Guex. Swiss-PdbViewer is tightly linked to SWISS-MODEL, an automated homology modeling server developed
within the Swiss Institute of Bioinformatics (SIB) at the Structural Bioinformatics Group at the Biozentrum in Basel (Johansson et al., 2012).

Working with these two programs greatly reduces the amount of work necessary to generate models, as it is possible to thread a protein primary sequence onto a 3D template and get an immediate feedback of how well the threaded protein will be accepted by the reference structure before submitting a request to build missing loops and refine sidechain packing. Swiss-PdbViewer can also read electron density maps, and provides various tools to build into the density (Guex and Peitsch, 1997). In addition, various modeling tools are integrated and residues can be mutated. Swiss-PdbViewer can also read electron density maps, and provides various tools to build into the density. In addition, various modeling tools are integrated and residues can be mutated. C terminal Oxygen atoms were added to protein structure using the above tool.

**CastP (http://cast. engr. uic. edu.)**

Computed Atlas of Surface Topography of proteins (CASTp) provides an online resource for locating, delineating and measuring concave surface regions on three-dimensional structures of proteins. The server was developed by Joe Dundas and Zheng Ouyang as an update to the CASTp server developed by Andrew Binkowski and Shapor Naghibzadeh, under the guidance of Prof. Jie Liang, who developed the castP program and the original server. The computation is based on the pocket algorithm of the alpha shape theory, and its core is the alpha shape API developed in Edelsbrunner's group and the NCSA.
Materials and Methods

surface regions on three-dimensional structures of proteins. These include pockets located on protein surfaces and voids buried in the interior of proteins. The measurement includes the area and volume of pocket or void by solvent accessible surface model (Richards' surface) and by molecular surface model (Connolly's surface), all calculated analytically. CASTp can be used to study surface features and functional regions of proteins. CASTp includes a graphical user interface, flexible interactive visualization, as well as on-the-fly calculation for user uploaded structures. Binding sites and active sites of proteins and DNAs are often associated with structural pockets and cavities (Joe et al., 2006).

CastP server uses the weighted Delaunay triangulation and the alpha complex for shape measurements. It provides identification and measurements of surface accessible pockets as well as interior inaccessible cavities, for proteins and other molecules. It measures analytically the area and volume of each pocket and cavity, both in solvent accessible surface (SA, Richards' surface) and molecular surface (MS, Connolly's surface). It also measures the number of mouth openings, area of the openings, and circumference of mouth lips, in both SA and MS surfaces for each pocket. The amino acids and atoms present in binding pocket of NF-κB proteins were predicted using this tool.

Ligplot (http://www.ebi.ac.uk)

The LIGPLOT program automatically generates schematic 2-D representations of protein-ligand complexes from standard Protein Data Bank file input. The output is a
Materials and Methods

color, or black-and-white, PostScript file giving a simple and informative representation of the intermolecular interactions and their strengths, including hydrogen bonds, hydrophobic interactions and atom accessibilities. The program is completely general for any ligand and can also be used to show other types of interaction in proteins and nucleic acids. It was designed to facilitate the rapid inspection of many enzyme complexes, but has found many other applications (Wallace et al., 1995).

AutoDock V3.0

AutoDock is a suite of automated docking tools. It is designed to predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure. Current distributions of AutoDock consist of two generations of software: AutoDock 4 and AutoDock Vina. AutoDock was the most cited docking software. It is very fast, provides high quality predictions of ligand conformations, and good correlations between predicted inhibition constants and experimental ones. AutoDock has also been shown to be useful in blind docking, where the location of the binding site is not known. AutoDock has now been distributed to more than 29000 users around the world. It is being used in academic, governmental, non-profit and commercial settings. In January of 2011, a search of the ISI Citation Index showed more than 2700 publications have cited the primary AutoDock methods papers. AutoDock is now distributed under the GPL open source license and is freely available for all to use (Schames et al., 2004).
Materials and Methods

AutoDock consists of two main programs: AutoDock for docking of the ligand to a set of grids describing the target protein; AutoGrid for pre-calculating these grids. AutoDock has an improved version, AutoDock Vina which has an improved local search routine and allows the use of multicore/multi-CPU computer setups. AutoDock was carried out automatically using the Lamarckian Genetic Algorithm. As a result of these calculations the output file of the protein-ligand complex with flexible residues and the ligand located within the binding pocket was obtained. Each structure were scored and ranked by the program by the calculated interaction energy and based on number of hydrogen bonds formed.

TopMatch (http://topmatch.services.came.sbg.ac.at/)

Computational tools for the alignment and superposition of protein structures are essential instruments in structural biology. TopMatch-web provides an easy-to-use interface to a suite of techniques for protein structure alignments called TopMatch. Given a pair of protein structures, TopMatch calculates a list of alignments ordered by structural similarity. The corresponding superpositions can be explored in a 3D molecule viewer which highlights the structurally equivalent parts of the proteins (Sippl and Wiederstein, 2012; Slater et al., 2012).

The sequence alignments resulting from the structure comparison are provided on-line and in PDF format. Coordinates of the input structures after superposition are available for download. TopMatch requires Java to run the Jmol plugin. Due to this platform
Materials and Methods

independent setup. TopMatch should run on most operating systems as well as on all common browsers. TopMatch-web requires the atomic coordinates of the two protein structures that are to be compared (Sippl and Wiederstein, 2008). The NF-κB sequences taken from different mammals were compared for structures using TopMatch and results were deduced based on root mean square deviations obtained between the compared structures.