CHAPTER 1

Introduction to Bioinformatics Tools in Genomic Data Analysis and Rational Drug Design
1.1 Bioinformatics tools in genomic data analysis

The information from the completed genome sequence projects stored in the databanks such as EMBL, GenBank for nucleotides and SWISSPROT, UNIPROT, NRDB for proteins are freely available to the public. Millions of sequences are available in these databanks that provide basic information about the respective proteins. Characterization of whole genomes is important to understand the structural and functional principles of living organisms. Whole genome comparisons provide clues on evolutionary relationships (Griffiths et al., 1999).

The wealth of sequence information brought about by the genome sequencing projects has led to the discovery of several computational tools, which enables the researchers to analyze the genes and proteins in whole genomes. These computational methods have been developed to solve the biological problems, using nucleotide and amino acid sequences and other related information.

Though DNA is the genetic material, it does not carry out the processes of life. This genetic code is transcribed and translated in the synthesis of protein molecules, which are present as the structures and molecular machines that make the cell function. Proteins contribute to almost all the events in the cells of a living organism. The polypeptide chain of a protein folds into a specific 3-D structure, which governs its function. Recent developments in the techniques of structure determination at atomic resolution, X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy, have enhanced the quality and speed of structural studies (Zhang & Kim, 2003). Nevertheless, current statistics still show that the known protein sequences vastly outnumber the available protein structures (48,778) deposited in protein data bank (PDB) so far. This is due to the inability to express, purify and crystallize some proteins as well as the intrinsic limitations of the structure determination techniques.
Chapter 1

It becomes a challenge for the researchers to annotate this huge genomic data. About 40-50% of proteins in each genome are novel and are not biochemically and structurally characterized. Experimental characterization of each sequence is however time consuming. Therefore, adding value to the structure and function of these novel proteins by means of comparative studies, using computational tools is one of the challenges to the researchers worldwide. Sophisticated mathematical, statistical and computational techniques are developed to handle, analyze and add value to this flood of data. These studies have become one of the frontier areas of research in modern biology.

1.1.1 Nucleotide and protein databases:

Nucleotide sequence databases were first assembled at Los Alamos National Laboratory (LANL), by Walter Goad and colleagues in the GenBank database and at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. GenBank is now under the auspices of the National Center for Biotechnology Information (NCBI) (http://ncbi.nlm.nih.gov). The EMBL Data Library was founded in 1980 (http://www.ebi.ac.uk). The EMBL maintains DNA and protein sequence databases. In 1984 the DNA DataBank of Japan (DDBJ) came into existence (http://ddbj.nig.ac.jp). GenBank, EMBL and DDBJ have now formed the International Nucleotide Sequence Database Collaboration (http://www.ncbi.nlm.nih.gov/collab), which acts to facilitate exchange of data on a daily basis. Translated nucleotide sequence information is included in the Protein Information Resource (PIR) database at the National Biomedical Research Foundation in Washington, DC. GenBank entries provide a large amount of information describing the entry of each sequence. SwissProt is a protein sequence database and it is similar to the EMBL format. It contains more information about
the physical and biochemical properties of the protein. Researchers are encouraged to submit their newly obtained sequences directly to the various types of nucleotide or protein databases. UniProt is a comprehensive resource for protein sequence and annotation data. UniProt is a result of collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the PIR. UniProt is the central hub for the collection of functional information on proteins, with accurate, consistent, and rich annotation. In addition to capturing the core data, each UniProt entry contains the amino acid sequence, protein name or description, taxonomic data and citation information. The “nr” database is the largest nucleotide database available through NCBI. It includes all GenBank, RefSeq Nucleotides, EMBL, DDBJ and PDB sequences.

The format of a database entry is such that each sequence file contains the information about the assigned accession number, source organism, function of the sequence, literature references, location of mRNAs, coding regions, positions of important mutations and sequence.

1.1.2 Sequence analysis tools:
Database searching

Comparison of a sequence with entries in a database is required to identify similar sequences that share homology. This can be done at both nucleotide and protein level. After proper validation of the results, multiple sequence alignments of these related sequences can built using consensus sequences of protein families that help in the identification of domains, motifs or functional sites. Detection of sequence similarity among different proteins has led to the classification of proteins on the basis of structure and function. It has been observed that most often similar sequences share similar structure and function. In addition, database searches are also used as primary requirement in identifying a structural homolog for an
unknown sequence. The most widely used programs for database searching are BLAST and FASTA.

1.1.2.1 Basic Local Alignment Search Tool (BLAST):

The BLAST program is used to identify sequence similar homologs from nucleotide or protein databases. The program takes a query sequence and searches it against the database selected by the user. It aligns the query sequence against every subject sequence in the database and the results are reported in the form of a ranked list followed by a series of individual sequence alignments, plus various statistics and score parameters. Every hit in that list is assigned with a similarity score S. Further, this score is analyzed to calculate the extent of such matching to occur by chance. For that purpose E-value is calculated for every hit. BLAST program finds regions of local similarity and calculates the statistical significance of matches (Altschul et al., 1990) (http://www.ebi.ac.uk/blast2) and (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi/).

The BLAST program first dissects the query sequence into words of length k (3 for proteins and 11 for nucleotides). These words are searched against the database for matches, and scores are assigned with either BLOSUM (Henikoff & Henikoff, 1992) or PAM (Dayhoff, 1978) scoring matrices. Word hits that score more than T (neighborhood word score threshold) are extended in both directions to generate an alignment between segment pairs. The "T" parameter dictates the speed and sensitivity of the search. The extension process is stopped when the scores drop from its maximum achieved score and the segment pairs are referred to high scoring pairs (HSP). The next step is to determine those HSPs of sequences, which have score greater than a cut off score (S). S is determined empirically by examining a range of scores found by comparing random sequences and by choosing a value that is significantly greater. BLAST determines the statistical
significance of HSPs and generates sequence hits in the descending order of E (expectation value) and P (probability score) values. E and P values are different ways of representing the significance of the alignment. These values are the number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The highly significant E or P values will be those close to 0 and lower values. BLAST also filters the low-complexity regions. Filtering is done by SEG and XNU filters and applied to the query sequence alone to make the search focus on more important parts of the sequence. These regions are marked with X in protein sequences and N in nucleotide sequences and are then ignored by BLAST.

The BLASTP offers various user defined options. A choice can be made on database to be searched. Based on the requirement, a user can switch to PDB or SWISSPROT database or a specific organism. Other options include selection of matrices, filters, adjustment of sensitivity and number of alignments etc. The default parameters for BLASTP include BLOSUM62 scoring matrix, a value of 11 is assigned for gap opening and a value of 1 for gap extension.

BLAST uses Smith-Waterman dynamic programming algorithm (Smith & Waterman, 1981a, 1981b). It detects local as well as global alignments using a heuristic approach. The exhaustive Smith-Waterman approach is too slow for searching large genomic databases such as GenBank. Therefore, the BLAST algorithm uses a heuristic approach that is slightly less accurate than Smith-Waterman but over 50 times faster. There are five different BLAST programs, which can be distinguished by the type of the query sequence (DNA or protein) and the type of the subject database.
BLASTP-compares an amino acid query sequence against a protein sequence database.
BLASTN-compares a nucleotide query sequence against a nucleotide sequence database.
BLASTX-compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
TBLASTN-compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).
TBLASTX-compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

1.1.2.2 Position-Specific Iterative BLAST (PSI-BLAST):

PSI-BLAST program is used for finding distant relatives of a protein. The program makes a list of all closely related proteins. These proteins are then combined into a "profile" that is a sort of average sequence. A query against the protein database is then run using this profile, and a larger group of proteins are found. This larger group is used to construct another profile, and the process is repeated (Altschul et al., 1997) till one finds all related proteins in the database. This method is more reliable and used in several other programs such as PSI-PRED, PHD- secondary structure prediction methods. By including related proteins in the search, PSI-BLAST is much more sensitive in picking up distantly related proteins than using the standard protein-protein BLAST (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi/).
1.1.2.3 Pattern Hit Initiated BLAST (PHI-BLAST):

PHI-BLAST is a search program that combines matching of regular expressions with local alignments surrounding the match. The calculation of local alignments is done using a method very similar to gapped BLAST (Zhang, 1998). The most important features of the program have been incorporated into the BLAST framework partially for user convenience and partly so that PHI-BLAST may be combined seamlessly with PSI-BLAST. PHI-BLAST is most preferred to search for pattern occurrences because it filters out those cases where the pattern occurrence is probably random and not indicative of homology. PHI-BLAST may be preferable to other types of BLAST programs because it is faster and allows the user to express a rigid pattern occurrence requirement. PHI-BLAST uses Baeza-Yates and Gonnet (Baeza, 1992), Wu and Manber, 1992 algorithm, permits simple patterns to be represented in a single computer word and matches to be found very efficiently. PHI-BLAST was specifically designed to combine pattern search with the search for statistically significant sequence similarity (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi/).

1.1.2.4 FASTA:

The sequence similarity searching against nucleotide and protein databases are also carried out using the FASTA program. Pearson and Lipman (Pearson & Lipman, 1988) developed this program to achieve good sensitivity for similarity searching at high speed. This is achieved by performing optimized searches for local alignments using a substitution matrix. The search algorithm FASTA proceeds through four steps in determining a score for pair-wise similarity. FASTA searches for the matching sequence patterns called k-tup. Using k-tup FASTA builds a local alignment, scores this alignment and generates a list of sequences similar to a query sequence in the descending order. The high speed of
this program is achieved by using the observed pattern of word hits to identify potential matches before attempting to carry out the more time consuming optimized search. The speed and sensitivity is controlled by the k-tup parameter, which specifies the size of the word. Increasing the value of k-tup decreases the number of background hits. Not every word hit is investigated but instead initially looks for segments containing several nearby hits. This performs a database scan for similarity in a short time, so as to make such scans routinely possible (http://www.ebi.ac.uk/fasta33/).

1.1.3 Multiple sequence alignment:

In bioinformatics, a sequence alignment is the way of arranging the primary sequences of DNA, RNA or proteins in order to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. A multiple sequence alignment arranges more than two sequences such that residues with common structural positions or ancestral residues are aligned in the same column. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutation, and gaps as insertion or deletion mutations that are introduced in one or both lineages in the time since they diverged from one another. In protein sequence alignment, the degree of similarity between amino acid occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineage. The most similar regions in the multiple sequence alignment may represent structural domains or regions of functional importance. Multiple sequence alignments often provide an understanding of evolutionary history of sequences. If the sequences in the alignment are very well conserved, then it implies that these sequences are recently derived from a common ancestor sequence. Conversely, a group of poorly aligned sequences share a more complex
and distant evolutionary relationship. Multiple sequence alignments of related sequences can build consensus sequences of known families, domains, motifs or sites. These are useful in predicting the function and structure of proteins, and also in identifying new members of protein families. Combining these predictions with primary biochemical data can provide valuable insights into protein structure and function.

1.1.3.1 T-Coffee:

T-Coffee is a multiple sequence alignment program, which pre-processes a dataset of all pair-wise alignments between the sequences. This provides us with a library of alignment information that can be used to guide the progressive alignment. Intermediate alignments are then based not only on the sequences to be aligned next but also on how all of the sequences align with each other. This alignment information can be derived from heterogeneous sources such as a mixture of alignment programs and/or structure superposition (Notredame et al., 2000). T-Coffee will compare all sequences two by two, producing a global alignment and a series of local alignments. The program will then combine all these alignments into a multiple alignment. The main characteristic of T-Coffee is that it allows one to combine results obtained with several alignment methods (http://www.ebi.ac.uk/t-coffee/).

1.1.3.2 CLUSTALW:

CLUSTALW is a fully automated program for global multiple alignment of nucleotide and protein sequences. This is very useful in designing experiments to test the function of specific proteins, in predicting the function and structure of proteins and in identifying new members of protein families. CLUSTALW
generates multiple sequence alignments of divergent sequences, and a phylogenetic
tree based on a multiple alignment of sequences. It can manipulate existing
alignments and carry out profile analysis (Thompson et al., 1994). The majority of
the automated multiple sequence alignments are based on the progressive approach
of the Feng and Doolittle (Feng & Doolittle, 1987). CLUSTALW, developed by
Thompson et al., 1994, incorporated a number of improvements to the alignment
algorithm, including sequence weighting, position-specific gap penalties and the
choice of a suitable residue comparison matrix at each stage in the multiple
alignments (http://www.ebi.ac.uk/Tools/clustalw2/).

CLUSTALW produces biologically meaningful multiple sequence
alignments of divergent sequences. It calculates the best match for the selected
sequences, and aligns them up so that the identities, similarities and differences can
be seen. The alignment in CLUSTALW is achieved via three steps: 1) pair-wise
alignment, 2) guide-tree generation and 3) progressive alignment. Evolutionary
relationships can be observed in a diagrammatic form by viewing Cladograms or
Phylograms which will be discussed in detail under the section, “Phylogenetic
analysis”.

In CLUSTALW alignment, scores can be calculated by two methods, slow /
accurate and fast / approximate, that use dynamic programming (Smith &
Waterman, 1981a; 1981b) and Wilbur and Lipman methods (Wilbur & Lipman,
1983) respectively. CLUSTALW provides several options, such as use of slow or
fast pair-wise alignments, nucleotide or protein sequences, protein weight matrix,
gap open, gap extension, end gaps and gap distances. The default parameters for
protein sequences are: Protein Gap Extension Penalty = 0.2; Protein matrix =
Gonnet; Protein ENDGAP = -1; Protein GAPDIST = 4.
1.1.3.3 Phylogenetic analysis:

Phylogenetic analysis of a family of proteins or nucleic acids is the determination of how the family might have been derived during evolution. When the sequences found in two different organisms are similar, then they are likely to have been common ancestor. Phylogenetic analysis is an important area of sequence analysis. There are three main steps in phylogenetic analysis, these are 1) Multiple sequence alignment, 2) Distance calculation and 3) Tree construction. Multiple sequence alignment method first aligns the most closely related sequences and then sequentially adds more distantly related sequences or sets of sequences to these initial alignments. After obtaining the multiple sequence alignment each column is assumed to correspond to an individual site that has been evolving according to the observed sequence variation in the column. Distance methods build trees by grouping them according to their overall similarity. After calculating the distance, one can cluster the data together in a tree. Using the multiple sequence alignment method CLUSTALW, from the input sequences, the program calculates the pair-wise alignments and degree of similarity between all the pairs followed by the calculation of distance. Distance is commonly calculated by number of mismatches in the non-gapped positions between the two sequences. This value is divided with the number of non-gapped pairs. Thus, a distance matrix is generated for all the sequence pairs. Using the distance matrix and neighbor-joining method, CLUSTALW constructs the similarity tree. The root is placed in the middle of the longest chain of consecutive edges. For generating the phylogenetic trees one can use Bootstrapping method to obtain support values for each cluster. Pair-wise distances can be determined with protein parsimony method (Felsenstein, 1996). Representations of the calculated trees can be constructed using TreeView (Page, 1996) (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).
Chapter 1

1.1.4 Motif/pattern:

A sequence motif is a short conserved region found in a number of related protein sequences. Motifs often correspond to core structural and functional elements of the proteins. Their conserved nature allows them to be used to diagnose family membership and predict function. Genome sequencing provides the basis for a systematic analysis of all motifs that are present in a particular organism. Protein sequences can be searched for known motifs in databases such as PROSITE (http://www.expasy.org/prosite/), ProDom (http://prodom.prabi.fr/prodom/current/html/form.php), Pfam (http://www.sanger.ac.uk/Software/Pfam/) and PRINTS (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS). These pattern and profile searches constitute an important resource for the classification of majority of the newly appearing protein sequences into known families.

1.1.5 Protein families and protein domains:

A protein family is a group of evolutionarily related proteins, and is often nearly synonymous with gene family. Proteins in a family descend from a common ancestor and typically have similar 3-D structure, function and significant sequence similarity. Many proteins comprise multiple independent structural and functional units or domains. Domains are structural and functional units that have specific biochemical activities. Due to evolutionary shuffling, different domains in a protein have evolved independently. A brief description of protein domains is discussed in this section.

1.1.5.1 Simple Modular Architecture Research Tool (SMART):

The SMART is an online resource (http://smart.embl.de/) used for protein domain identification and the analysis of protein domain architectures (Schultz et al., 1998; Letunic et al., 2006). SMART offers a high level of sensitivity and
specificity coupled with ease of use. It contains several unique aspects, including automatic seed alignment generation, detection of repeated motifs or domains and a protocol for combining domain predictions from homologous subfamilies. Visualization tools have been developed to allow analysis of gene intron-exon structure within the context of protein domain structure, and to align these displays to provide schematic comparisons of orthologous genes, or multiple transcripts from the same gene. It also allows batch retrieval of multiple entries.

1.1.5.2 INTERPRO:

INTERPRO is a database of protein families, domains and functional sites (Mulder et al., 2005). It can be applied to predict the function and structure of unknown protein sequences (Zdobnov & Apweiler, 2001). INTERPRO provides an integrated view of the commonly used signature databases such as Pfam, PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, and PANTHER. Signatures are manually integrated into INTERPRO entries that are curated to provide biological and functional information. INTERPRO covers over 78% of all proteins in the SWISSPROT and TrEMBL. The database is available for text and sequence based searches via a web server (http://www.ebi.ac.uk/InterProScan/).

1.1.5.3 PFAM:

PFAM is a comprehensive collection of protein domains and families and helps in the genome annotation (Bateman et al., 2004). Each family in PFAM is represented by multiple sequence alignments and Hidden Markov Model (HMM) profile and can be used to view the domain organisation of proteins.

Structural data has been utilised to ensure that families in PFAM correspond to structural domains, and to improve domain based annotation. Predictions of
non-domain regions are also included. In addition to secondary structure, PFAM multiple sequence alignments now contain active site residues highlighted. New search tools, including taxonomy search and domain query, greatly add to the functionality and usability of the PFAM resource. Apart from the well known annotated domains, PFAM also provides the information of functionally uncharacterized families, known as Domains of Unknown Function (DUFs) and Uncharacterized Protein Families (UPFs). DUFs are families that have been created by PFAM and UPFs are those created by SWISSPROT and added to PFAM database (http://pfam.janelia.org/search).
1.2 Drug design

The numbers of disease target proteins are expected to increase considerably due to the completion of many genome sequencing projects. With the arrival of high-throughput protein purification and 3-D structure determination methods, the importance of computational strategies to focus drug discovery efforts will be more useful and will facilitate the rapid development of novel therapies. Computational models provide an effective path to test hypotheses regarding mechanisms that simulate the behavior of biological systems at all levels, including molecular and cellular systems, and can provide models to such hypotheses.

*In silico* design of potential drugs for a given protein target can involve change of existing lead compounds, or *de novo* design of entirely new compounds. A lead compound is described as a compound that binds to the target protein and inhibits the activity of protein at a certain level. Lead compounds can be found through experimental high-throughput screening as well as virtual screening. Combinatorial libraries of analogs can be computationally designed and screened and a subset of identified compounds can be synthesized for validation.

Use of computational techniques in drug discovery and development process is highly attractive due to the ease of implementation and reliability. Different nomenclature is being applied to this area, including computer-aided drug design (CADD), computational drug design, computer-aided molecular design (CAMD), computer-aided molecular modeling (CAMM), rational drug design, *in silico* drug design, computer-aided rational drug design.
1.2.1 Computer-Aided Drug Design (CADD):

Computational chemistry methods were applied to aid the understanding of theoretical chemistry and pharmaceutical drug discovery. A variety of techniques for similarity searching aid scientists to find potentially active molecules and docking methods are helpful to model the binding of these molecules to desired protein targets. This application, known as CADD or molecular modeling, rapidly became an essential part of modern drug discovery, and thus the pharmaceutical industry became a strong supporter of the field (David & Gary, 2006). It was familiar since the 1960s, that computer-based methods can be useful in the discovery of new lead molecules and can eliminate chemical synthesis and screening of many irrelevant compounds. The rapid advance of computer technology and the development of new modeling software have made CADD an increasingly useful tool in drug design. An ideal computational method for lead molecule discovery should be able to generate structurally diverse lead molecules rapidly and should give an estimate of the binding affinities that would correlate with experimental values and ideally go on to become viable drugs (Mohan et al., 2005).

1.2.2 Rational Drug Design (RDD):

The advent of molecular biology, coupled with advances in screening and synthetic chemistry technologies, has allowed a combination of both knowledge around the receptor and random screening to be used for drug discovery. Rational drug design is a process used in the biopharmaceutical industry to discover and develop new lead molecules. RDD uses a variety of computational methods to identify novel compounds, design compounds for selectivity, efficacy and safety, and thus develop compounds into clinical trial candidates. These methods fall into several natural categories such as, structure-based drug design, ligand-based drug
design, *de novo* design and homology modeling depending on how much information is available about drug targets and potential drug compounds (Jurgen, 2000). Advances in molecular biology, protein crystallography and computational chemistry since the 1980s have greatly aided the RDD paradigms. Figure 1.1 shows a flow chart that describes various approaches that enable RDD.

**Figure 1.1**: A Flow Chart Indicating Various Approaches to Rational Drug Design (Adopted and modified from ref. Parrill & Reddy 1999).
1.2.3 Computer-Aided Molecular Modeling (CAMM):

Identification of lead molecules with selective bioactivity, whether intended as potential therapeutics or as tools for experimental research, is essential in medicine and the life sciences. The discovery of a new drug to combat a disease takes years to decade and costs are too high (DiMasi \textit{et al.}, 2003; Dickson & Gagnon, 2004). CAMM represents a potentially useful tool for this purpose, and it has made great advances into improving the odds of finding bioactive lead molecules (Burley & Park, 2005, Blundell \textit{et al.}, 2002). CAMM is a truly successful tool and is provided in an easily available and usable format, for the benefit of the wider scientific community. The state of the art CAMM can be divided into two broad categories: ligand-based drug design and structure-based drug design depending upon the availability of 3-D structure of the target protein.

1.2.3.1 Ligand-Based Drug Design (LBDD):

Many receptors are not readily amenable to structure-based drug design. For example, many important receptors are membrane-bound proteins, which are often difficult to crystallize. In such cases, a lead compound or active ligand must be found, and then the structure of the ligand guides the drug design process in the LBDD. Structure-based drug design studies assemble information from already existing lead molecules or drugs that are active against the target biological molecule of interest. Based on the known information, a set of rules are framed to design either a new ligand or modify an existing ligand in order to improve its biological activity.

1.2.3.1.1 Quantitative Structure-Activity Relationship (QSAR):

The QSAR paradigm has evolved over the last hundred years to embody many quantitative approaches to structure-property correlations in physical organic
chemistry, biochemistry and molecular design. QSAR represents an attempt to correlate structural or property descriptors of compounds with activities. These physicochemical descriptors, which include parameters to account for lipophilicity, hydrophobicity, topology, electronic properties, and steric effects, are determined empirically by computational methods. The QSAR models are useful for various purposes including the prediction of activities of untested molecules.

Early QSAR methods related biological activity to the presence (or absence) of functional groups in a series of structurally related compounds (Free-Wilson model). Later the concept of quantitative correlation of physicochemical properties of molecules with their biological activities termed as QSAR was initiated by Corwin Hansch and coworkers during early 1960. Several 3D-QSAR modeling approaches have emerged in 1980s such as, active analog approach, molecular shape analysis, distance geometry and CoMFA. QSARs attempt to correlate physical and chemical properties of molecules to their biological activities. This can be achieved by simply using easily calculable descriptors (for example, molecular weight, number of rotatable bonds, LogP) and simple statistical methods such as multiple linear regression to build a model which describes the activity of the dataset and predict the activities for untested sets of compounds. These types of descriptors are simple to calculate and allow a relatively fast analysis, but often fail to take into account the 3-D nature of chemical structures (which obviously play a major role in ligand-receptor binding, and hence activity). 3D-QSAR uses probe-based sampling within a molecular lattice to determine 3-D properties of molecules (particularly steric and electrostatic values) and can then correlate these 3-D descriptors with biological activity. Hopfinger et al., 1997, introduced a fourth dimension to the 3D-QSAR modeling and termed it as 4D-QSAR analysis (Hopfinger et al., 1997).
Hologram QSAR (HQSAR) is a relatively new technique, which does not require any physicochemical descriptors or 3-D structure to generate structure-activity models (Naumann & Lowis, 1997). It needs only 2-D structures and activity as input. HQSAR converts the molecules of a dataset into counts of their constituent fragments. These fragment counts are then related to biological data using partial least square analysis. HQSAR is a rapid, highly predictive QSAR technique. Results reported earlier show that HQSAR can readily produce highly predictive QSAR models over a wide variety of datasets.

1.2.3.1.2 Pharmacophore:

Pharmacophore generation is another method for LBDD. A pharmacophore is the spatial arrangement of key chemical features that are recognized by a receptor and are thus responsible for ligand-receptor binding (Gund & Güner, 2000). A pharmacophore is the ensemble of steric and electronic features that are necessary to ensure the optimal supramolecular interactions with a specific biological target (protein or DNA) structure and to trigger (or to block) its biological response. Pharmacophore models are constructed based on molecules of known biological activity and refined as more data is acquired in an iterative process. Alternatively, a pharmacophore can also be generated from the receptor structure. These models can be used for optimizing known ligands or for screening databases to find potential novel lead molecules suitable for further development (Renner et al., 2004, Singh et al., 2002).

1.2.3.2 Structure-Based Drug Design (SBDD):

In SBDD, the 3-D structure of a receptor (drug target) interacting with small molecules is used to guide drug discovery. The active site of a receptor is the area into which a chemical or biological molecule binds in order to initiate a
biochemical reaction. SBDD aims to create a molecule that will bind to the active
site of a targeted receptor, thereby preventing the normal chemical reaction and
ultimately halting the progression of the disease. Much of the work in the drug
design is now based on the structure of the target and virtual screening of libraries.
Captopril is the first success drug that came from the SBDD (Cushman et al.,
1977). After that several drugs came to the market, Carbonic anhydrase I and II
targeted, Dorzolamide for glaucoma (Tsukamoto & Larsson, 2004), Bcr-Abl kinase
targeted Imatinib for cancer (Druker et al., 1996) and HIV protease targeted
Lopinavir, Indinavir, Nelfinavir, Saquinavir and Ritonavir for AIDS (Verbesselt et
al., 2007). Various approaches used for the SBDD are as follows.

1.2.3.2.1 Docking:

Protein-ligand docking is a molecular modeling technique. The goal of
protein-ligand docking is to predict the position and orientation of a ligand when it
is bound to a receptor. Docking is frequently used to predict the binding orientation
of small molecule drug candidates to their protein targets in order to be able to in
turn predict the affinity and activity of the small molecule. Hence docking plays an
important role in the RDD studies (Kitchen et al., 2004). Theoretically, docking is
an energy optimization process concerned with the search of the lowest free energy
binding mode of the ligand within the receptor binding site. In addition, protein
flexibility is computationally expensive; therefore many of the existing docking
programs treat the protein either as a rigid structure or allow flexibility only to the
protein side chain functional groups. A good docking method places the ligand
appropriately in the active site and then estimates the forces involved in the
receptor-ligand recognition (electrostatic, van der Waals and hydrogen bonding).
Chapter 1

Table 1.1 lists some of the existing docking methodologies and the strategies they use.

**Table 1.1**: List of Some of the Available Docking Methodologies and Their Strategies.

<table>
<thead>
<tr>
<th>Searching Algorithm</th>
<th>Brief Description of Methodology</th>
<th>Examples of software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monte Carlo (MC)</td>
<td>Stochastic method of generating conformations. Selection based on Metropolis criterion</td>
<td>Ligand Fit</td>
</tr>
<tr>
<td>Simulated Annealing (SA)</td>
<td>Random thermal motions are induced, through high temperatures, to explore the local search space. System is driven to a minimum energy conformation by decreasing temperature. SA usually combined with MC.</td>
<td>MC-DOCK, AutoDock</td>
</tr>
<tr>
<td>Genetic Algorithm</td>
<td>Based on Darwin principles of evolution. ‘Chromosome’ encoding model parameters (like torsion angles) are varied stochastically. Populations are generated through genetic operations (crossover, mutation, migration). The fittest survives in the population.</td>
<td>GOLD, AutoDock</td>
</tr>
<tr>
<td>Matching Methods</td>
<td>Based on clique detection technique from graph theory. Ligand atoms are matched to the complimentary atoms in the receptor</td>
<td>FLOG, DOCK</td>
</tr>
<tr>
<td>Simulation Methods</td>
<td>Molecular dynamics simulations are used to generate conformations</td>
<td>DOCK, AutoDock, FlexX</td>
</tr>
<tr>
<td>QM-Polarized Ligand Docking algorithm</td>
<td>Quantum mechanic simulations are used to generate conformations</td>
<td>Glide</td>
</tr>
</tbody>
</table>
1.2.3.2.1 Conformation generation:

Drug molecules (ligands) usually bind to protein at a cavity of the receptor, which is called the binding site. It is usually assumed that the geometric constraints are the main determinants in this process. The energetic factors are also important, since molecules in nature are usually found in their low energy conformation. Ligand molecules have many degrees of freedom due to the rotatable bonds. Proteins, however, are much bigger molecules, with several hundreds of atoms. Given a ligand and a protein, finding whether they will bind to each other, and if they do bind their configuration in bound state is a difficult problem to predict as they involve many degrees of freedom of rotation. The conformations sought should each be with energy, less than a given threshold, and possibly with spatial features at specific positions in 3-D space.

1.2.3.2.1.2 Scoring:

The process of evaluating the particular conformation of molecule when bound to protein uses a number of descriptive features such as, number of intermolecular interactions including hydrogen bonds, hydrophobic contacts and van der Waals energy. Scoring function used in docking is a mathematical function whose values are proportional to the binding affinities of the lead molecules. A good scoring function should be able to give reliable estimates of binding affinities of structurally diverse lead molecules for different protein targets while considering the thermodynamic aspects of binding (Ajay & Murko, 1995).

Essentially, three types or classes of scoring functions are currently applied. Force field based empirical and knowledge-based scoring functions (Kitchen et al., 2004). 1) Force field based methods are first principle methods that use force field parameters to score the van der Waals and electrostatic interactions between
receptor and ligand. The score includes receptor–ligand interaction energy and internal ligand energy. These methods do not require calibration or training with experimental binding data. 2) Empirical scoring functions are regression based functions derived from a large sample of crystal structures with known affinities for the bound ligands. These functions reflect a best fit with respect to the training set used, but rarely achieve generality. 3) Knowledge-based scoring functions are designed to reproduce experimental structures rather than binding energies. It evaluates the frequencies of particular type of interaction, the mutual distance between particular types of atoms across the interface, in databases of protein–ligand complexes.

1.2.3.2.2 De Novo ligand design:

De novo design uses structural information to develop a molecule that can fit into the active site by consecutively adding or joining molecular fragments instead of using libraries of existing compounds (Honma, 2003). Structure sampling is carried out by different methods such as: linking, growing, lattice-based sampling, random structure mutation, transitions driven by molecular dynamics simulations, and graph-based sampling. Apart from these, the ligand can also be built from recombination of bioactive conformations of known ligands for a particular target. Recombination is carried out by overlaying the known ligands and swapping the fragments of different ligands. This procedure is carried out recursively, so that the compounds that emerge from recombination are added to the pool of known active molecules and participate in subsequent cycles of recombination. The largest advantage of de novo design is its ability to develop novel scaffolds utilizing the whole chemical space (Schneider & Fechner, 2005). However, this method also suffers limitations such as: 1) synthetic feasibility is not
considered while constructing structures, and 2) the prediction of binding affinities for the designed structures is not accurate.

1.2.3.2.3 Virtual Screening (VS):

There is a growing pressure on the pharmaceutical industry to reduce the cost of drugs and the time taken to market them. The large number of targets made available in the last decade has created a new area for technologies that can rapidly identify quality lead candidates. VS is one such technology that is gaining increasing importance in the drug discovery process. VS is a reliable and inexpensive method currently being employed as a complementary approach to high-throughput screening. VS can be adopted irrespective of the structural information of the target receptor. In the absence of structural data of the receptor, VS using pharmacophore-based search is a major in silico tool. However, when the structure of the receptor is available, VS using both pharmacophore-based and docking techniques can be employed. VS is used as an initial screen for large databases to reduce the number of compounds that are to be screened experimentally (Lyne, 2002). VS protocols include ligand-based screens such as: 1-D filters (e.g. molecular weight), 2-D filters (similarity, substructure fingerprints) and 3-D filters (3-D pharmacophore, 3-D shape matching) and docking based on structure-based screening methods (Sirois et al., 2004). The potential sources of error contributing to the identification of false positives and false negatives in VS include: 1) approximations in the scoring functions employed; 2) improper solvation terms; 3) neglect of protein flexibility and; 4) poor assessment of the protonation states of active site residues or ligands (Lyne et al., 2004). Significant improvements in VS have been made by consensus scoring (Bissantz et al., 2000) of multiple scoring functions and by clustering docking poses, from multiple docking tools before scoring (Paul & Rognan, 2002).
A more sophisticated approach in the current drug design process is “Docking based virtual screening” that allows the user to quickly screen large databases of potential drugs and score the ligand-protein interactions. Docking based virtual screening typically involves fast docking of a large number of chemical compounds against a protein binding site. But the accuracy of these screening approaches are underpinned by the molecular-docking methods, which in turn, depend on the computational algorithms for conformational sampling and scoring of different ligand binding conformations.

1.2.4 Protein structure and small molecule databases:

One of the most important and difficult problems in molecular biology is the protein folding problem. The structure-function relationship in proteins is directly concerned with correlating the 3-D structure of a protein to the primary sequence. The richest source of information about protein structure is the Protein Data Bank (PDB). The development of chemoinformatics has been hampered by the lack of large, publicly available, comprehensive repositories of molecules, in particular small molecules. They can be used as combinatorial building blocks for chemical synthesis (Schreiber, 2000), as molecular probes for perturbing and analyzing biological systems in chemical genomics and systems biology (Stockwell, 2004).

1.2.4.1 Protein Data Bank (PDB):

The PDB is a collection of individual "flat" text files, each of which contains the 3-D co-ordinates of one of the several thousands of protein structures determined by various experimental techniques such as X-ray crystallography and NMR. A variety of information associated with each structure is available through the RCSB PDB, including sequence details, atomic co-ordinates, crystallization
conditions, bound cofactors, metal ions or inhibitors, 3-D structural neighbors computed using various methods, derived geometric data, structure factors, 3-D images and a variety of links to other resources. Information about DNA and RNA structures is also available in these databases. Protein Data Bank maintained by the Rutgers, the State University of New Jersey, San Diego Supercomputer Center (SDSC), Skaggs School of Pharmacy and Pharmaceutical Sciences. It is available to researchers worldwide via the website www.rcsb.org/ (Berman et al., 2000). Till the year 2008, 48,778 experimentally determined structures have been deposited from scientists all over the world.

1.2.4.2 Small molecule databases:

Small molecules mainly comprise atoms such as carbon, hydrogen, nitrogen, oxygen, sulfur, halogens and phosphorus. These molecules play a fundamental role in organic chemistry and biology. There are various types of small molecule databases such as NCI (http://dtp.nci.nih.gov/), CSD (CCDC) (http://relibase.ccdc.cam.ac.uk/), Maybridge and Derwent (Accelrys), ACD (http://www.chemweb.com/databases/) and ChemBank (http://chembank.broad.harvard.edu/). All these databases contain millions of compounds and have useful information for each molecule, including its, structural, physical, chemical and biological properties. These databases are of great value for the screening, design and discovery of useful compounds.

1.2.5 The Lipinski rule of 5:

Experimental and computational approaches to estimate the solubility and permeability of small molecules for the drug discovery and development are described in the Lipinski “the rule of 5”. In the discovery process ‘the rule of 5’ predicts that, the likelihood of poor absorption or permeation of the molecule is
Chapter 1

greater when there are more than, 1) 5 H-bond donors (expressed as the sum of OHs and NHs) 2) 10 H-bond acceptors (expressed as the sum of Os and Ns) 3) The molecular weight is greater than 500, 4) The calculated Log P is greater than 5 and 5) The calculated Moriguchi octanol-water partition coefficient (MlogP) is greater than 4.15.

The problems and methods introduced in this chapter have been instrumental in the advance of our understanding of protein function, organization and structure. These computational analyses are aimed at speeding up the process for identification of protein structure, function and drug design. The computational tools aimed at analyzing the protein data are useful in supporting and explaining the experimental findings, assisting in the design of experiments and creating hypotheses.
1.3 References


Chapter 1


