CHAPTER 3
MATERIALS AND METHODS

The overall objective of the study was to carry out molecular modeling activities on the selected therapeutic targets belonging to cancer and infection. It was accomplished by using computational approach involving homology modeling, molecular docking, fragment based studies, pharmacophore modeling and quantitative structure activity relationship. Several online resources as well as licensed software were used for the implementation of the computational approach.

3.1 Requirements

3.1.1 Hardware and software

Majorly, all the computational studies were carried out on Intel Xeon dual processor based HP Workstation designed for computer aided engineering with following features:

i. Central Processing Unit (CPU): 2.0 GHz (4MB L2 Cache) 64 bit Dual Core Intel Xeon 5130 Processor.

ii. Chipset and Motherboard: Intel 5000X Chipset on Intel Motherboard.

iii. Memory: 4GB DDR-2 Synch DRAM PC2 – 5300 (667MHz) FB.

iv. Hard Disk Drive: 2 x 73 GB SAS HDD at 15000 rpm.

v. Storage controller: Integrated ISI 1068 SAS Controller RAID, 0, 1 & External Connector and also integrated SATA Controller.

vi. Graphics Card: NVIDIA Quadro Fx3500 PCI Express (256 MB) for better graphics,

along with standard keyboard, mouse, PCI slots, Networking features, Bays, Mini tower cabinet, DVD RW, Linux Operating System (64-Bit) and other standard certifications and application software.

In addition, the following servers were also used:-

42
a) SGI Fuel workstation with single 700Mhz MIPS R16000A processor; 4MB L2 cache; Advanced VPro V10 Graphics with 64-Bit IRIX Operating System.

b) Wipro Netpower 7225 series Server with Intel DP Xeon 3.2 GHz, 512KB L2 Cache; 256 MB DDRAM, 10/100 Ethernet; 52*CDR, 40 GB HDD with Windows XP professional.

The major licensed software used for carrying out the computational studies are as under:

i) **Schrodinger suite**

The following software modules of Schrodinger suite were installed on the HP Intel Xeon dual processor based Workstations for carrying out the molecular modeling studies:


With the help of the above mentioned modules, it was possible to carry out small molecular modeling and simulations (conformation generation & clustering, property generation & filtering, molecular mechanics & dynamics and quantum mechanics), macromolecular modeling and simulations (protein crystal structure refinement, protein modeling & bioinformatics, molecular mechanics & dynamics, monte carlo simulations and quantum mechanics/molecular mechanics), lead discovery (cheminformatics, ligand based discovery, fragment based discovery and structure drug discovery), lead optimization (cheminformatics, 2D/3D QSAR, combinatorial chemistry, fragment based design, ligand based design, structure based design and absolute and relative binding affinity prediction), visualization and automation (molecular visualization, workflow and medicinal chemistry application).
ii) **Molecular Operating Environment Software**

The software was installed on Wipro Netpower 7225 and it has the following features: Molecular modeling and simulation (molecular mechanics and dynamics, molecule alignment, molecular builders, implicit solvent electrostatics etc.), bioinformatics (protein structure database, fold identification, conserved features, multiple alignment, homology search etc.), cheminformatics (molecular databases, 3D pharmacophore search, molecular descriptors, high throughput conformational search etc.), protein modeling (multiple alignment, structure analysis, structure prediction, protein mechanics and dynamics etc.), combinatorial chemistry (combinatorial library enumeration, combinatorial library design, molecular descriptors etc.), docking and pharmacophore (ligand-receptor docking, active site detection, multi-fragment and 3D pharmacophore search).

iii) **Insight-II and Cerius2 software**

These software modules are from Accelrys (formerly known as Molecule Simulation Inc. (MSI)) and were installed on SGI Fuel workstation. Insight-II is a comprehensive graphic molecular modeling program. In conjunction with the molecular mechanics/dynamics program such as Discover, or CHARMM Insight II program can be used to build and manipulate virtually any class of molecule or molecular system. In conjunction with other MSI products, you can study molecular properties.

Cerius2 is another module from Accelrys which can be used for research in Life and Material Sciences. The key features of Cerius2 are Molecular visualization, Structure editing and optimization, crystal and amorphous structure building, powder diffraction studies, molecular simulation, property calculations and morphology predictions.
3.1.2 Databases

All the 3D crystal structural information about the target proteins was obtained from the Protein Data Bank (PDB) (Berman et al., 2000). The amino acid sequence detail of the target proteins, wherever the 3D structure was not available, was obtained from National Center for Biotechnology Information (NCBI) database in order to arrive at an in silico derived 3D structure. A brief overview about the databases in bioinformatics and the Entrez search system is given in Figure 1. The Entrez Global Query Cross-Database Search System is a powerful federated web portal that allows users to search many discrete health sciences databases at the NCBI website (Figure 3.1). All the molecular modeling studies were carried out using the licensed software MOE, Accelrys Cerius2 & Insight-II and Schrodinger suite, and the homology modeling was carried out using several online servers such as LOMETS (Wu et al., 2007), PHYRE (Kelley et al., 2009), MODELLER (Sali et al., 1993) etc. The structure of the ligands was drawn on Maestro 9.1 & 9.2 version of Schrodinger suite. The databases used frequently in bioinformatics is tabulated along with their links and salient features in Table 3.1
Figure 3.1. Classification of Bioinformatics databases based on Entrez
Table 3.1. Frequently used databases in bioinformatics and their links and salient features

<table>
<thead>
<tr>
<th>Databases</th>
<th>Salient features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary nucleotide sequence</strong></td>
<td></td>
</tr>
<tr>
<td>Genebank</td>
<td>Provides nucleotide sequence database maintained by the National Center for Biotechnology Information (NCBI), USA</td>
</tr>
<tr>
<td><strong>Other nucleotide sequence</strong></td>
<td></td>
</tr>
<tr>
<td>Unigene</td>
<td>The nucleotide sequences of genbank in the form of clusters representing genes are available</td>
</tr>
<tr>
<td>Genome Biology</td>
<td>The information about the complete genome is available</td>
</tr>
<tr>
<td><strong>Protein sequence database</strong></td>
<td></td>
</tr>
<tr>
<td>SWISS PROT</td>
<td>Provides the description of the structure of a protein, its domain’s structure, post translational modifications, variants etc.</td>
</tr>
<tr>
<td><strong>Other databases</strong></td>
<td></td>
</tr>
<tr>
<td>PDB</td>
<td>This is the primary database for 3D structures of biological macromolecules</td>
</tr>
<tr>
<td>KEGG</td>
<td>The Kyoto Encyclopedia of Genes and Genomes is with latest computerized information on biomolecules and cell biology. KeGG provides details on information pathways interacting molecules and the connecting links with genes.</td>
</tr>
</tbody>
</table>
3.2 Homology Modeling

Homology modeling aims to build 3D protein structure models using the experimentally determined structure of related family members as templates (Bordoli et al., 2009) as shown in Figure 3.2.

Figure 3.2. Flowchart depicting the homology modeling technique

- **Target Sequence**: Check for completeness, species, polymorphism
- **Identify Templates**: Use consensus methods
- **Secondary structure prediction**: Match secondary structure prediction
- **Multiple Sequence**: Use more than one method to align and check for errors
- **Model Building**: Construct backbone and Adjust sidechain
- **Model Loops**: Build more than one model
- **Refine Model**: Use reliable programs and limit to less than 15 amino acids
- **Optimize and Validate**: Make sure to cap the ends of proteins, missing regions, choose proper constraints and force fields
- **Use experimental data as restraints**
To elucidate further, homology modeling technique works as under: the initial step involves the alignment of the target and the template sequences followed by the secondary structure determination of the target sequence and its alignment with the secondary structure of the target sequence (the 3D structure of the target is already known). Subsequently, after adjusting the secondary structure of the target and the template, backbone of the target is constructed and finally the side chains are incorporated. If required Further refinement of the arrived structure is carried out based on revised secondary structure alignment, if required. The process is pictorially depicted in Figure 3.3.
In order to get a reliable model, wherever the 3D crystal structure of the selected target was unknown, several online servers were used for *in silico* modeling of these proteins (*Table 3.2*). The template for modeling was selected based on the structure and function of the template protein. The developed protein models were first prepared and minimized using Protein Preparation Wizard of the Schrodinger Modeling Software Package. During minimization the root mean square deviation (rmsd) value was set to 0.3Å, as any deviation below this is considered to be negligible.

All the protein models were then checked for Ramachandran Core value, disallowed residues and bad contacts (*Ramachandran et al., 1963*) using Procheck module of the SAVS server (*http://nihserver.mbi.ucla.edu/SAVES*). SAVS is Structural Analysis and Verification Server available from National Institute of Health (NIH) for sterochemical analysis of protein models. The description of modules within the SAVS server is given in *Table 3.3*.
Table 3.2. Online servers used for structure prediction and their URLs

<table>
<thead>
<tr>
<th>S.No</th>
<th>Name of the Server</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>SWISS-MODEL</td>
<td>swissmodel.expasy.org</td>
</tr>
<tr>
<td>2.</td>
<td>CPHmodels-3.0 Server</td>
<td><a href="http://www.cbs.dtu.dk/services/CPHmodels">www.cbs.dtu.dk/services/CPHmodels</a></td>
</tr>
<tr>
<td>3.</td>
<td>HOMER SERVER</td>
<td><a href="http://protein.criibi.unipd.it/homer">http://protein.criibi.unipd.it/homer</a></td>
</tr>
<tr>
<td>4.</td>
<td>3D-Jigsaw</td>
<td>bmm.cancerresearchuk.org/~3djigsaw</td>
</tr>
<tr>
<td>7.</td>
<td>GENO3D</td>
<td>geno3d-pbil.ibcp.fr</td>
</tr>
<tr>
<td>8.</td>
<td>HHpred</td>
<td>toolkit.tuebingen.mpg.de/hhpred</td>
</tr>
<tr>
<td>9.</td>
<td>CHOYCE</td>
<td>choyce.ismb.lon.ac.uk</td>
</tr>
<tr>
<td>10.</td>
<td>CASPR</td>
<td><a href="http://www.igs.cnrs-mrs.fr/Caspr">www.igs.cnrs-mrs.fr/Caspr</a></td>
</tr>
<tr>
<td>11.</td>
<td>Phyre</td>
<td><a href="http://www.sbg.bio.ic.ac.uk/~phyre">www.sbg.bio.ic.ac.uk/~phyre</a></td>
</tr>
<tr>
<td>12.</td>
<td>I-TASSER</td>
<td>zhanglab.ccmb.med.umich.edu/I-TASSER</td>
</tr>
<tr>
<td>13.</td>
<td>PS2 server</td>
<td>ps2.life.nctu.edu.tw</td>
</tr>
<tr>
<td>14.</td>
<td>@TOME-2</td>
<td>atome.cbs.cnrs.fr</td>
</tr>
<tr>
<td>15.</td>
<td>LOMETS</td>
<td>zhanglab.ccmb.med.umich.edu/LOMETS</td>
</tr>
<tr>
<td>17.</td>
<td>MUSTER</td>
<td>zhanglab.ccmb.med.umich.edu/MUSTER</td>
</tr>
<tr>
<td>18.</td>
<td>M4T server 3.0</td>
<td>manaslu.aecom.yu.edu/M4T</td>
</tr>
<tr>
<td>19.</td>
<td>MMM server</td>
<td>manaslu.aecom.yu.edu/MMM</td>
</tr>
<tr>
<td>20.</td>
<td>FFAS03</td>
<td>ffas.ljcrf.edu</td>
</tr>
<tr>
<td>21.</td>
<td>PROTEUS 2.0</td>
<td><a href="http://www.proteus2.ca">www.proteus2.ca</a></td>
</tr>
<tr>
<td>22.</td>
<td>DescFold</td>
<td>protein.cau.edu.cn/DescFold</td>
</tr>
<tr>
<td>23.</td>
<td>Inub</td>
<td><a href="http://inub.cse.buffalo.edu">http://inub.cse.buffalo.edu</a></td>
</tr>
<tr>
<td>24.</td>
<td>ModLink+</td>
<td><a href="http://sbi.imim.es/modlink">http://sbi.imim.es/modlink</a></td>
</tr>
<tr>
<td>25.</td>
<td>SAM-T08</td>
<td>compbio.soe.ucsc.edu/SAM_T08/T08-query.html</td>
</tr>
<tr>
<td>26.</td>
<td>MODELLER 9V8</td>
<td>salilab.org/modeler</td>
</tr>
<tr>
<td>27.</td>
<td>HOMCOS</td>
<td>strcomp.protein.osaka-u.ac.jp/homcos</td>
</tr>
</tbody>
</table>
Table 3.3 Description of various modules of Structural Analysis and Verification Server

<table>
<thead>
<tr>
<th>SAVS Modules</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROCHECK</td>
<td>Checks the stereochemical quality of a protein structure by analyzing residue-by-residue geometry and overall structure geometry.</td>
</tr>
<tr>
<td>WHAT_CHECK</td>
<td>Derived from a subset of protein verification tools from the WHATIF program (Vriend et al., 1990), this does extensive checking of many sterochemical parameters of the residues in the model.</td>
</tr>
<tr>
<td>ERRAT</td>
<td>Analyzes the statistics of non-bonded interactions between different atom types and plots the value of the error function versus position of a 9-residue sliding window, calculated by a comparison with statistics from highly refined structures.</td>
</tr>
<tr>
<td>VERIFY_3D</td>
<td>Determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigned a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to good structures.</td>
</tr>
<tr>
<td>PROVE</td>
<td>Calculates the volumes of atoms in macromolecules using an algorithm which treats the atoms like hard spheres and calculates a statistical Z-score deviation for the model from highly resolved (2.0 Å or better) and refined (R-factor of 0.2 or better) PDB-deposited structures.</td>
</tr>
</tbody>
</table>
3.2.1 Binding site prediction

The selected protein models were prepared for docking using the Protein preparation module of Schrodinger software. The binding pockets of the predicted models were identified using the Sitemap module of Schrodinger (Halgren et al., 2007, Halgren et al., 2009). Receptor grid was generated for the protein structures on all the identified binding sites. A grid is a 3D box that defines the region of the receptor to be docked. It also determines the position and size of the active site.

3.3 Molecular docking studies

The docking studies, after protein preparation and ligand generation, were carried out using the Glide module of the Schrodinger suite (Friesner et al., 2004, Halgren et al., 2004, Friesner et al., 2006) with either default options or modified settings. In general, rigid body docking has been performed in all the experiments, or else it has been mentioned. The work can be expanded in detail with respect to flexible docking studies also, involving molecular dynamics, for better precision.

3.3.1 Protein preparation

A typical downloaded 3D structure from PDB is not usable as it is for molecular modeling studies. It consists only of heavy atoms and may include a co-crystallized ligand, water molecules, metal ions, and cofactors. Some structures are multimeric, and may need to be reduced to a single unit. Because of the limited resolution of X-ray experiments, it can be difficult to distinguish between NH and O, and the placement of these groups must be checked. PDB structures may be missing information on connectivity, which must be assigned, along with bond orders and formal charges. Therefore, it becomes essential to prepare the downloaded crystal structure for further computational studies (as shown in Figure 3.4). This was done using the Protein preparation wizard of Schrodinger suite, wherein the bond orders were assigned to the protein, hydrogens added, metals treated and water molecules 5Å beyond hetero groups deleted (as it is expected that water molecules beyond this distance will not have any impact on protein-ligand interaction). Hydrogens were
then optimized using exhaustive sampling option and the protein was minimized to RMSD limit from the starting structure of 0.3Å using the Impref module of Impact with OPLS_2001 force field (Impact version 5.7, Schrodinger, LLC, New York, 2011).

![Figure 3.4 Steps in preparation of PDB protein for docking studies](image)

**Figure 3.4 Steps in preparation of PDB protein for docking studies**

### 3.3.2 Ligand generation

The 2D structure of all the ligands was first drawn using the Maestro interface of Schrodinger suite (Maestro version 9.2, Schrodinger, LLC, New York, NY, 2011). Maestro is the Graphical User Interface (GUI) for all Schrodinger computational modules and contains tools for building, displaying, and manipulating chemical structures. It also allows loading, organizing and storing these structures and associated data in order to set up, submit, monitor, and visualize the results of calculations on these structures.

In order to carry out the computational studies, it is important to generate the 3D structure of the ligands in consideration. This was carried out using the Ligprep module of Schrodinger suite (Chen et al., 2010). When used in the simplest way, LigPrep produces a single, low-energy 3D structure with correct chiralities for each successfully processed input structure. Further, LigPrep can also produce a number of structures from each input structure with various ionization states, tautomers, stereochemistries, and ring conformations, and eliminate molecules using various criteria including molecular weight or specified numbers and types of functional groups present. Further, the generation of probable bioactive conformations was carried out using the Confgen module of Schrodinger suite (Chen et al., 2010, Watts et al., 2010), which covers a broad range of conformational space. ConfGen carefully examines the structure of the ligand to understand where
to expect local minima as a function of rotations about rotatable bonds. It then systematically generates the conformations that arise from various combinations of these local minima. Thus it provides a broad and fairly uniform coverage of the available conformational space. This systematic approach used in ConfGen avoids the enormous amount of re-sampling of conformations that occurs in most conformational searching methods designed for exhaustive sampling. In addition to sampling rotatable bonds ConfGen also samples ring conformations, chiral nitrogen atom inversions and amide bond conformations. In many cases where the ligands in consideration were similar to the co-crystallized ligand, these were first aligned to the co-crystallized bioactive conformation and then used for further studies.

### 3.3.3 Ligand-receptor docking

The ligand receptor docking was carried out using MOE software ([http://www.chemcomp.com](http://www.chemcomp.com)) and Glide module of Schrodinger suite ([Friesner et al., 2004, Friesner et al., 2006, Halgren et al., 2004](http://www.chemcomp.com)). The complete docking process is summarized as a flowchart in **Figure 3.5**. For the available 3D crystal structure of the target protein with a cocrystallised ligand, the conformation of the co-crystallized ligand was taken as standard for establishing the docking protocol. In absence of any co-crystallized ligand, the docked pose of a known inhibitor was taken as standard. The grid was generated either using the co-crystallized ligand or based upon the information about the binding site residues of the target protein. Glide searches for favorable interactions between one or more ligand molecules and the receptor (target protein). Glide can be run in rigid or flexible docking modes; the latter automatically generates conformations for each input ligand. The combination of position and orientation of a ligand relative to the receptor, along with its conformation in flexible docking, is referred to as a ligand pose.
Figure 3.5. Ligand-receptor docking flowchart.

Protein Data Bank (PDB) → Homology Model

Target Protein

Prepare Protein for docking → Identify binding site

Establish docking protocol

Docking Ligands to the Target Protein

Use rigid or flexible docking as per the target protein

Minimisation of the Protein-Ligand complex

Scoring of the complex

Rank order the complex based on the consensus of various scoring

Evaluation and Optimization

Interpret the results with respect to the literature data available and try to draw some logical inference for further optimization of the ligands
The ligand poses that Glide generates pass through a series of hierarchical filters that evaluate the ligand’s interaction with the receptor. By default, Schrödinger’s proprietary GlideScore multi-ligand scoring function is used to score the poses. The process of rigid docking and subsequent flexible refinement is elaborated in Figure 3.6.

Figure 3.6. Rigid body docking and flexible refinement of the ligand-receptor complex.
The estimate of the relative binding affinities of the ligands with the proteins was determined using the Prime MMGBSA module of Schrodinger (Prime, version 3.0, Schrodinger, LLC, New York, NY, 2011). The ranking of the ligands based on the calculated binding energies (MMGBSA DG Bind) is expected to agree reasonably well with ranking based on experimental binding affinity, particularly in the case of congeneric series. The calculation of the “MMGBSA DG Bind” [dG(1)] property is calculated using the equation 

\[ dG(1) = E_{\text{complex(minimized)}} - (E_{\text{ligand(minimized)}} + E_{\text{receptor(minimized)}}) \]

Prime MMGBSA also accounts for the flexibility of the protein, which at times is very important for drawing any inference. In addition to this, Induced fit docking was also used for accounting the protein flexibility, wherever necessary.

3.4 Fragment based drug designing

Principle aim in structure-based fragment work is to find the energetically favourable binding mode(s), which can lead to valuable insights into the nature of the binding site and key interactions responsible for molecular recognition. Finding the most druggable pockets of target proteins, and identifying molecular fragments or functional groups that tend to bind there, are important steps in fragment based drug design (Congreve et al., 2008, Loving et al., 2009). The process flowchart is given in Figure 3.7.
Figure 3.7. Process flowchart of fragment-based drug design.

Select Target Protein

- Identify binding clefts within the active site
- Dock fragments on the clefts
- Identify good binders (as potential fragments)
- Link these potential fragments using a linker
- Dock the final molecule/s for optimization and further validation

Fragment Library (Rule of 3)
{Molecular weight ≤ 300
HB Donor ≤ 3
HB Acceptor ≤ 3
LogP ≤ 3
No. of rotatable bonds ≤ 3}
It has been observed that the small organic compounds cluster in the important pockets of the binding site, and this provides information on their druggability (Congreve et al., 2008, Loving et al., 2009). FTMAP server (Brenke et al., 2009, Cencic et al., 2011) was used to find the binding surface regions of the protein-ligand complex, which are often called ‘hot spots’, in order to identify the clefts/pockets within the binding site where small drug-like compounds will bind with high affinity. An illustrative example of the small fragment binding to the hot-spots is given in Figure 3.8.

Figure 3.8. Screening and linking of fragments after their optimization within the hot-spots.
The available online library of fragments from several vendors such as Chembridge (http://www.chembridge.com/screening_libraries/fragment_library), Maybridge (http://www.maybridge.com/portal/alias__Rainbow/lang_en/tabID__230/DesktopDefault), Schrodinger (http://www.schrodinger.com/productpage/14/5/73) etc. were downloaded and combined into a huge fragment library of about 1500 diverse fragments. This fragment library, in combination with the inputs generated from FTMAP server, was used for ligand optimization of the identified scaffolds for the target protein to arrive at a novel potent inhibitor of the protein. The regions for modification of the scaffolds were identified and therefore, a library of virtual molecules was generated with the attachment of desirable fragments on those regions. Virtual screening of this library was then carried out in order to arrive at a potent inhibitor of the identified protein target.

3.5 Quantitative Structure Activity Relationship

Quantitative structure-activity relationship (QSAR) equation is a mathematical equation that correlates the biological activity to a wide variety of physical or chemical parameters (Maloney et al., 1962, Fujita et al., 1964, Hansch et al., 1969, Livingstone et al., 2000, Hansch et al., 2001). QSAR is one of the most important methods in chemometrics, which give information that is useful for drug design and medicinal chemistry (Tuppurainen et al., 1999, Marder et al., 2001). The structures of the compounds taken in the data set for Quantitative Structure Activity Relationship (QSAR) studies were modeled using the 3D sketcher of the software cerius2 from Accelrys Inc (http://www.accelrys.com). All the molecules were energy minimized using the Cerius2.OFF module with default parameters of Smart Minimizer. The process flowchart is explained in Figure 3.9.

3.5.1 Descriptor calculation

E-state indices (Kier et al., 1990, Gregorio et al., 1998, Rose et al., 2002), electronic, information content (Melagraki et al., 2006), spatial, structural, thermodynamic and
topological descriptors (Estrada et al., 2001) were calculated using the Cerius2 4.10 software.

---

**Figure 3.9.** Process flowchart explain the Quantitative Structure Activity Relationship (QSAR) studies.
package. Topological indices are 2D descriptors based on graph theory concepts (Kier et al., 1976, Kier et al., 1986, Katritzky et al., 1993). These indices have been widely used in QSPR and in QSAR studies. They help to differentiate the molecules according mostly to their size, degree of branching, flexibility, and overall shape. Some of the descriptors included in the study are listed and described in Table 3.4.

Table 3.4. List of Descriptors used in the study.

<table>
<thead>
<tr>
<th>Type</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-state indices</td>
<td>Electrotopological-state indices</td>
</tr>
<tr>
<td>Electronic</td>
<td>Sum of partial charges, sum of formal charges, dipole moment, energy of the high occupied orbital, energy of the lowest unoccupied orbital, Superdelocalizability.</td>
</tr>
<tr>
<td>Information content</td>
<td>Information of atomic composition index, information indices based on the A-matrix, information indices based on the D-matrix, multigraph information content indices</td>
</tr>
<tr>
<td>Spatial</td>
<td>Radius of gyration, Jurs descriptors, shadow indices, area, density, PMI, $V_m$</td>
</tr>
<tr>
<td>Structural</td>
<td>Number of chiral centers, molecular weight, number of rotatable bonds, number of hydrogen-bond acceptors, number of hydrogen-bond donors</td>
</tr>
<tr>
<td>Thermodynamic</td>
<td>Log of the partition coefficient, log of the partition coefficient atom-type value, desolvation free energy of water, desolvation free energy of octanol, heat of formation, molar refractivity</td>
</tr>
<tr>
<td>Topological</td>
<td>Wiener index, Zagreb index, Hosoya index, Kier and Hall molecular connectivity index, Balaban indices</td>
</tr>
</tbody>
</table>

3.5.2 Regression analysis

The total number of descriptors calculated were about 200, some of the descriptors were rejected because they contain a value of zero for all the compounds. Further the inter-correlation of descriptors was taken in to account and highly correlated descriptors were grouped together and descriptor with highest correlation with biological activity was taken from the group (Shi et al., 2007). From descriptors thus remained, the selection of variables to obtain the QSAR models were carried out using genetic function approximation (GFA) method. GFA is genetics based method.
of variable selection, which combines Holland’s genetic algorithm (GA) with Friedman’s multivariate adaptive regression splines (MARS) (Rogers et al., 1994, Shi et al., 1998). The GFA method works in the following way: first of all a particular number of equations (set at 100 by default in the Cerius2 software) are generated randomly. Then pairs of “parent” equations are chosen randomly from these equations and “crossover” operations are performed at random. In the present study, the number of crossing over was set by default at 5000. The goodness of each progeny equation is assessed by Friedman’s lack of fit (LOF) score, which is given by following formula

\[ \text{LOF} = \frac{\text{LSE}}{(1 - \frac{c+dp}{m})^2} \]

Where LSE is the least-squares error, \( c \) is the number of basis functions in the model, \( d \) is smoothing parameter, \( p \) is the number of descriptors and \( m \) is the number of observations in the training set. The smoothing parameter, which controls the scoring bias between equations of different sizes, was set at default value of 1.0 and the new term was added with a probability of 50%. Only the linear equation terms were used for model building, which is set by default in the software. The best equation out of the 100 equations was taken based on the statistical parameters such as regression coefficient, adjusted regression coefficient, regression coefficient cross validation and F-test values.

### 3.5.3 Validation test

Further statistical significance of the relationship between activity and chemical structure descriptors was obtained by randomization procedure. The test was done by repeatedly permuting the activity values of the data set and using the permuted values to generate QSAR models and then comparing the resulting scores with the score of the original QSAR model generated from non-randomized activity values. If the original QSAR model is statistically significant, its score should be significantly better than those from permuted data (Deswal et al., 2006). The randomized test was performed at 90%, 95%, 98% and 99% confidence interval. The higher the confidence level, the more randomization tests are run. In this direction, nine trials
were run at 90% confidence level, 19 trials at 95%, 49 trials at 98% and 99 trials at 99% confidence level.

To further check the inter correlation of descriptors variance inflation factor (VIF) analysis was performed. VIF value is calculated from \(1/1 - r^2\), where \(r^2\) is the multiple correlation coefficient of one descriptor’s effect regressed on the remaining molecular descriptors. If VIF value is larger than 10, information of descriptors can be hidden by correlation of descriptors (Shapiro et al., 1998, Jaiswal et al., 2004).

It has been shown that a high value of statistical characteristic need not be the proof of a highly predictive model (Golbraikh et al., 2002, Roy et al., 2008). Hence, in order to evaluate the predictive ability of our QSAR model, we used the method described by Golbraikh et al (Golbraikh et al., 2002) and Roy et al (Roy et al., 2008). The values of correlation coefficient of predicted and actual activity and correlation coefficient for regressions through the origin (predicted versus observed activities and vice versa) were calculated using the Regression of analysis toolpak option of excel sheet and other parameters were calculated as reported by Golbraikh and Roy.

To arrive at the predictive \(R^2\) (\(R^2_{\text{pred}}\)) the following equation was used (Roy et al., 2008):

\[
R^2_{\text{pred}} = 1 - \frac{\sum (Y_{\text{pred(Test)}} - Y_{\text{Test}})^2}{\sum (Y_{\text{Test}} - \bar{Y}_{\text{training}})^2}
\]

where \(Y_{\text{pred(Test)}}\) and \(Y_{\text{Test}}\) are the predicted and observed activity values respectively, of the test set compounds and \(\bar{Y}_{\text{training}}\) is the mean activity values of the training set. Further evaluation of the predictive ability of the model was done by determining the value of \(rm^2\) by the equation (Roy et al., 2008):

\[
rm^2 = R^2 \left(1 - \sqrt{R^2 - R_0^2}\right)
\]

where \(R^2\) is the squared correlation coefficient between observed and predicted values and \(R_0^2\) is the squared correlation coefficient between observed and predicted values without intercept.
The value of $k$ and $k'$, slopes of the regression line of the predicted activity vs. actual activity and vice versa were calculated using the following equations (Roy et al., 2008):

\[
k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad \text{and} \quad k' = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2}
\]

where $\tilde{y}_i$ and $y_i$ are the predicted and actual activities, respectively of the test set.

Leave-25%-out has been shown to give an almost unbiased estimator of the generalisation properties of statistical models (Nargotra et al., 2009), and this cross-validation method was applied for the training set in order to get a sensible criterion for model selection and comparison.

### 3.6 Pharmacophore Modeling

Pharmacophore modeling is one of the major elements of drug design in the absence of structural data of the target receptor (Deng et al., 2006, Mustafa et al., 2009). The Pharmacophore modeling and 3D-QSAR studies were carried out using PHASE 3.3 implemented in the molecular modeling package from Schrodinger (Dixon et al., 2006, Dixon et al., 2006). Phase is a versatile product for pharmacophore perception, structure alignment, activity prediction, and 3D database searching. It provides support for lead discovery, SAR development, lead optimization, and lead expansion and hence is well suited to drug discovery projects for which no receptor structure is available. The algorithm within the software module utilizes fine-grained conformational sampling and a range of scoring techniques to identify common pharmacophore hypotheses, which convey characteristics of 3D chemical structures that are purported to be critical for binding. A given hypothesis may be combined with known activity data to create 3D-QSAR models that identify overall aspects of molecular structure that govern activity. These models may be used in conjunction with the hypothesis to mine a 3D database for molecules that are most likely to exhibit strong activity toward the target. The flowchart for pharmacophore modeling is explained in Figure 3.10.
3.6.1 Pharmacophore Hypothesis Generation.

PHASE can identify the spatial arrangements of functional groups that are common and essential for the biological activity of the ligands under investigation. It supplies a built-in set of six pharmacophore features i.e., Hydrogen bond acceptor (A), Hydrogen bond donor (D), Hydrophobic group (H), Negatively charged group (N), Positively charged group (P) and Aromatic ring (R). Hypotheses were generated by a systematic variation of number of sites and the number of matching active compounds. Common pharmacophore hypotheses (CPH) were considered. Further,
the best CPH was selected depending on the survival score, until at least one hypothesis was found and scored successfully. The hypotheses were scored using default parameters for site, vector, volume, selectivity, number of matches, and energy terms. The regression analysis was performed by constructing a series of models with an increasing number of PLS factors. Pharmacophore-based 3D-QSAR models were generated for the hypotheses using the training set with specified PLS factors and a grid spacing of 1 Å. The evaluation of generated CPHs was performed by correlating the observed and the estimated activity for the training. PLS analyses were performed in which a series of models were constructed with an increasing number of PLS factors. Score hypotheses step was employed to align the actives to the hypotheses and calculate the score for the actives. CPHs of significant statistical values were selected for molecular alignments.

3.6.2 Validation of Pharmacophore Model

For accurate and reliable predictions of biological activities of new compounds, the main target was to develop QSAR models, which were statistically robust both internally as well as externally. The data set was divided into a training set and a test set as external validation is considered to be a conclusive proof for judging predictability of a model. The training set was used to generate pharmacophore model and prediction of the activity of test set was used as a method to validate the proposed models. The robustness of the developed pharmacophore hypotheses was internally validated by statistical parameters i.e., squared correlation coefficient (R2) and variance ratio (F). Validation is a crucial aspect of pharmacophore design, particularly when the model is built for the purpose of predicting activities of molecules in external test series. The correlation between the experimental and predicted activities of the test set molecules was determined.

3.7 Wet lab studies

Wet lab validations for certain studies have been carried out in collaboration with Pharmacology and Clinical Microbiology Division, IIIM, Jammu