COMPUTATIONAL SCEENING OF EPIPODOPHYLLOTOXIN ANALOGUES AS ANTICANCER DRUGS
(Docking, Pharmacophore and 3D QSAR approaches)

By

RACHIT MADAN
051524

MAY-2009

Submitted in partial fulfillment of the Degree of Bachelor of Technology

**DEPARTMENT OF
BIOINFORMATICS & BIOTECHNOLOGY
JAYPEE UNIVERSITY OF INFORMATION
TECHNOLOGY - WAKNAGHAT, SOLAN, HP, INDIA**

CERTIFICATE

This is to certify that the work entitled **"COMPUTATIONAL SCEENING OF EPIPODOPHYLLOTOXIN ANALOGUES AS ANTICANCER DRUGS (Docking, MMGBSA, Pharmacophore and 3D QSAR approaches)"** submitted by **RACHIT MADAN** in partial fulfillment for the award of degree of Bachelor of Technology in Bioinformatics of Jaypee University of Information Technology has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Dr. Pradeep Naik
(Project Coordinator)
Sr. Lecturer, Bioinformatics Department.
Jaypee University of Information Technology
Waknaghat, Solan (H.P.).

ACKNOWLEDGEMENT

It has been a great pleasure working under the able guidance of faculty & staff in the Department of Bioinformatics and Biotechnology at Jaypee University of Information Technology, during our study as a B.Tech student. Thanks a lot to Dr. Pradeep Kumar Naik, my project advisor and committee chair, for his financial and scientific support; otherwise, this research work would never have been possible.

Many people have contributed to this project in a variety of ways over the past few months. To the individuals who have helped me, I again express my appreciation. I also acknowledge the many helpful comments received from our teachers of the concerned department. I am indebted to all those who provided reviews & suggestions for improving the results and the topics covered in this project, and extend my apologies to anyone I may have failed to mention.

RACHIT MADAN
051524

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

- A : Angstron
- ADME : Absorption, Distribution, Metabolism,Excretion
- BuryP : penalty for buried polar groups
- CoMFA : Comparitive molecular field analysis
- CoMSIA : Comparitive molecular similarity indices analysis
- Coul : Coulomb energy
- eMrAcE : automated mechanism of Multi-Ligand Bimolecular Association with energies
- FEB : Free energy of Binding
- GB/SA : Gaussian smooth dielectric constant function
- HP : Highly Potent
- HBond : Hydrogen bonding term
- Htopo : Human Topoisomerase IIa
- Lipoi : Lipophillic contact term
- MD : Molecular Dynamics
- Metal : metal binding term
- MM : Macro Model
- N : Napthalene
- NTL : Non-Lactonic Tetralines
- pIC50 : Predicted Biological Activity
- QSAR : Quantitative Structure Activity Relationship
- SD : Standard Deviation
- vdW : van der Waal energy

Abstract

Epipodophyllotoxin and its structural derivatives, a class of Topoisomerase inhibitors, have been the objective of numerous studies to prepare better and safer anti-cancer drugs. A library of Epipodophyllotoxin analogues has been designed consisting of 154 analogues. Their molecular interactions and binding affinities with Human Topoisomerase II protein (2RGR) have been studied using the docking-molecular mechanics based generalized Born/surface area (MM-GB/SA) solvation model. Quantitative structure activity relationships were developed between the biological activity (IC_{50}) of these compounds and molecular descriptors like docking score and binding free energy. For all the cases the r^2 was in the range of 0.57 to 0.831 indicating good data fit and r^2_{cv} was in the range of 0.60 to 0.83 indicating that the predictive capabilities of the models were acceptable. In addition a linear correlation was observed between the calculated binding free energy and pIC_{50} for the inhibitors with correlation coefficient (r^2) in the range of 0.115 to 0.0938, suggesting that the docked structure orientation and the interaction energies are reasonable. Low levels of root mean square error for the majority of inhibitors establish the docking and prime MMGBSA based prediction model as an efficient tool for generating more potent and specific inhibitors of Topoisomerase by testing rationally designed lead compounds based on Epipodophyllotoxin derivatization. Ligand based approach was also used in which the r^2 was in the range of 0.83 to 0.22.

EPIPODOPHYLLOTOXIN

Epipodophyllotoxin are alkaloids naturally occurring in the root of American Mayapple plant (*Podophyllum peltatum*). Some epipodophyllotoxin derivatives are currently used in the treatment of cancer. These include etoposide and teniposide. They act as anti-cancer drugs by inhibiting topoisomerase II. Podoplyllotoxin exhibits high cytotoxic activity against various cancer cell lines, but its severe toxic side effects had prevented it from being directly used as a therapeutic agent and this has promoted the search for derivates whit a greater therapeutic window.

Unfortunately, several drawbacks such as myleosuppresion, anemia, metabolic inactivation, development of drug resistance, severe gastrointestinal side effects cytotoxicity towards normal cell and poor bioactivity, still exist during the administration of these drugs, so extensive structural modification of Epipodophyllotoxin at various positions have been undertaken to, discover and develop more potent and less toxic anticancer agents which are currently being tested in phase I or II clinical trials for treatment of various cancer.

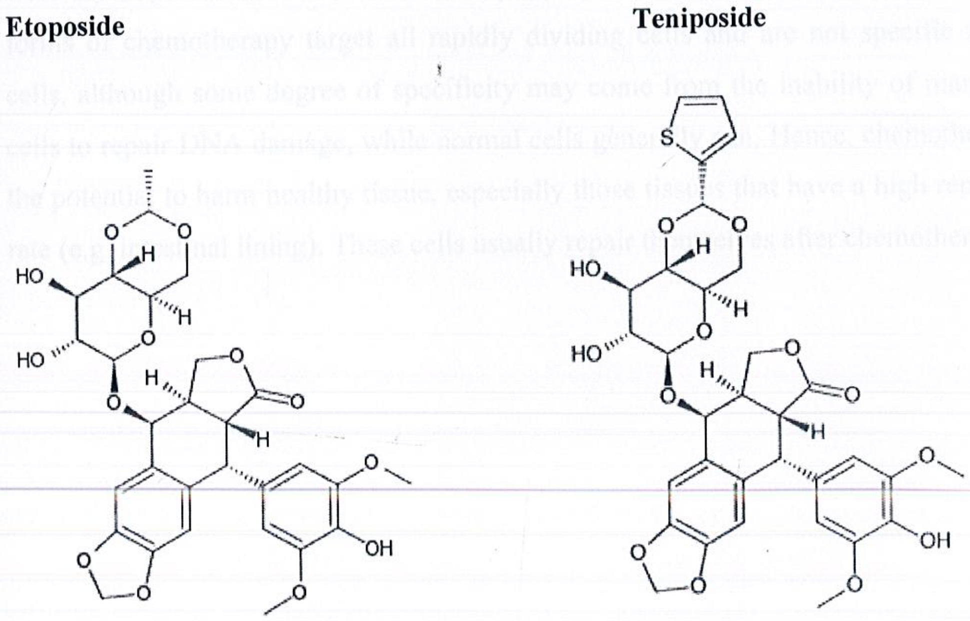
Drug resistance is the most important challenge in cancer treatment research. Clinically cells can't be totally killed by using a single drug for a long period of time, as they will become resistant to this drug and other drug with similar mechanism of action. In order to ensure efficacy the optimal administration schedule involves a combination of two or more drugs, especially drugs with different mechanism of action.

Human Topoisomerase II (topo II) is a cellular target for a number of widely used antitumor agents such as Etoposide.

Human Topoisomerase is highly expressed in highly proliferating cells, and plays an essential role in replication, transcription and chromosome organization. The depletion of topo II eventually results in cell death; hence this highly conserved nuclear enzyme is an important target for tumor chemotherapy. All the Topoisomerase II targeted anticancer drugs clinically used for their antitumor activities belong to topo II poisons. It is noteworthy that a wide range of topo II-targeted inhibitors are commonly classified as topo II poisons and catalytic inhibitors. Htopo II poisons have played an important role in chemotherapy against tumor for several decades.

Etoposide is a semi-synthetic podophyllotoxin derivative that has been used in cancer treatment since the early 1970s. Etoposide **(VP-16)** is an inhibitor of the enzyme topoisomerase II. It is used as a form of chemotherapy for cancer.

Teniposide (Vumon, VM-26) is a chemotherapeutic medication mainly used in the treatment of childhood acute lymphocytic leukemia. It is in a class of drugs known as podophyllotoxin derivatives and slows the growth of cancer cells in the body.



Etoposide

Teniposide

Figure 1: *Structure of Epipodophyllotoxin Conjugates*

Cancer

Cancer is a class of diseases in which a group of cells display *uncontrolled growth* (division beyond the normal limits), *invasion* (intrusion on and destruction of adjacent tissues), and sometimes *metastasis* (spread to other locations in the body via lymph or blood). These three malignant properties of cancers differentiate them from benign tumors, which are self-limited, and do not invade or metastasize. Most cancers form a tumor but some, like leukemia, do not.

Cancer is fundamentally a disease of regulation of tissue growth. In order for a normal cell to transform into a cancer cell, genes which regulate cell growth and differentiation must be altered. Genetic changes can occur at many levels, from gain or loss of entire chromosomes to a mutation affecting a single DNA nucleotide.

Chemotherapy is the treatment of cancer with drugs ("anticancer drugs") that can destroy cancer cells. In current usage, the term "chemotherapy" usually refers to *cytotoxic* drugs which affect rapidly dividing cells in general, in contrast with *targeted therapy* (see below). Chemotherapy drugs interfere with cell division in various possible ways, e.g. with the duplication of DNA or the separation of newly formed chromosomes. Most forms of chemotherapy target all rapidly dividing cells and are not specific to cancer cells, although some degree of specificity may come from the inability of many cancer cells to repair DNA damage, while normal cells generally can. Hence, chemotherapy has the potential to harm healthy tissue, especially those tissues that have a high replacement rate (e.g. intestinal lining). These cells usually repair themselves after chemotherapy.

BIOSYNTHESIS OF EPIPODOPHYLLOTOXIN

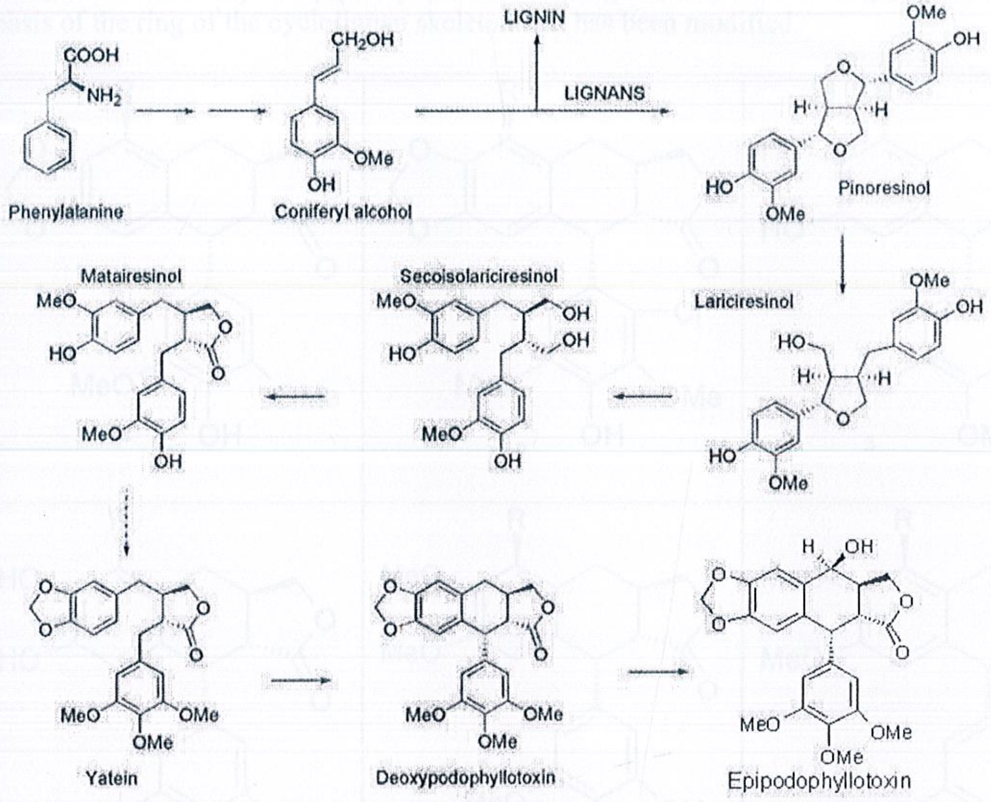
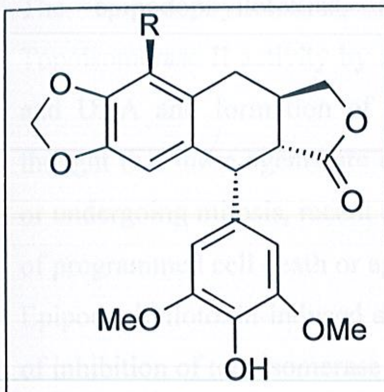


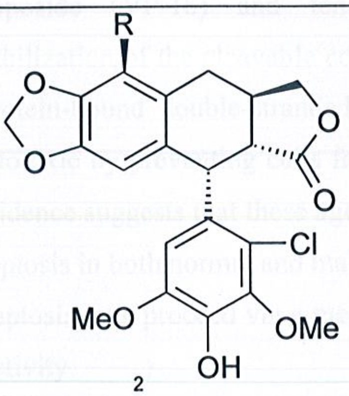
Figure 2. Biosynthetic pathway of synthesis of Epipodophyllotoxin.

Design of structural analogs of Epipodophyllotoxin

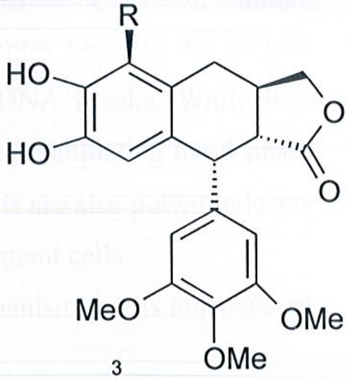
To facilitate the study of Epipodophyllotoxin analogues, they have been grouped on the basis of the ring of the cyclolignan skeleton that has been modified.



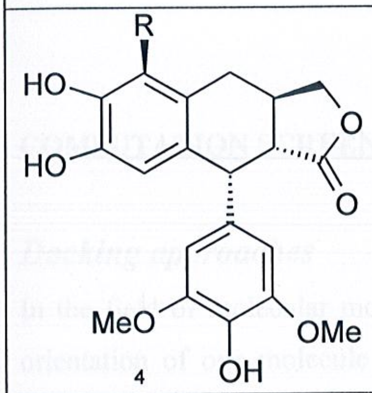
1



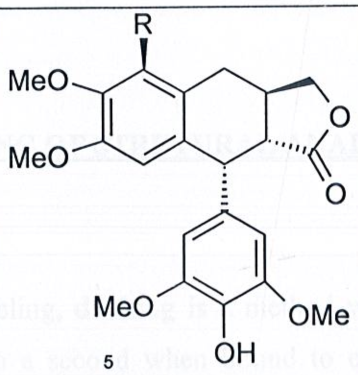
2



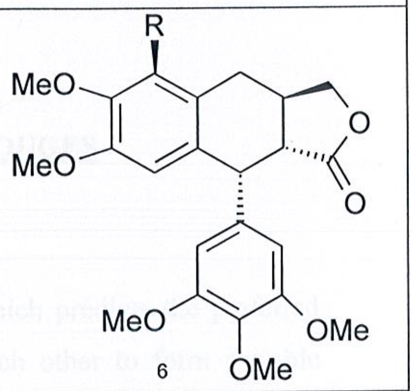
3



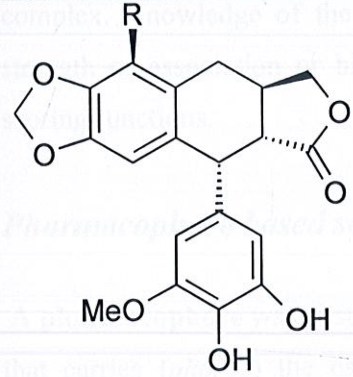
4



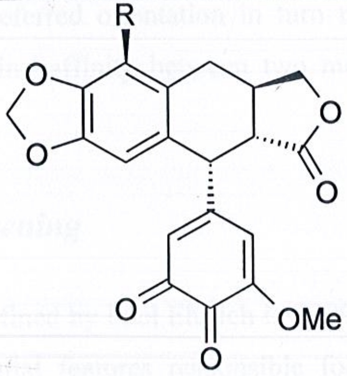
5



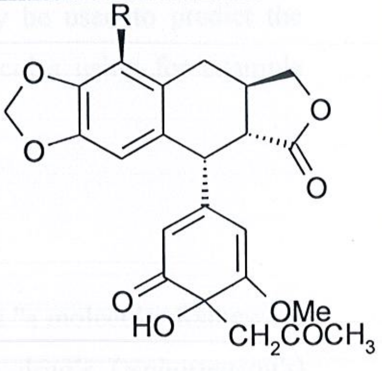
6



7



8



9

TABLE 1: STRUCTRUAL ANALOGUES OF EPIPODOPHYLLOTOXIN

THE MODE OF ACTION OF EPIPODOPHYLLOTOXIN

The epipodophyllotoxins, etoposide (VP-16) and teniposide (VM-26), inhibit Topoisomerase II activity by stabilization of the cleavable complex between the enzyme and DNA and formation of protein-bound double-stranded DNA breaks. While it is thought that these agents are cytotoxic by preventing cells from completing the S phase or undergoing mitosis, recent evidence suggests that these agents are also potent inducers of programmed cell death or apoptosis in both normal and malignant cells. Epipodophyllotoxin-induced apoptosis may proceed via a mechanism that is independent of inhibition of topoisomerase activity

COMPUTATION SCREENING OF STRUTURAL ANALOUGES

Docking approaches

In the field of molecular modeling, **docking** is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using for example scoring functions.

Pharmacophore based screening

A **pharmacophore** was first defined by Paul Ehrlich in 1909 as "a molecular framework that carries (*phoros*) the essential features responsible for a drug's (*=pharmacon's*) biological activity" (Ehrlich. *Dtsch. Chem. Ges.* 1909, 42: p.17). In 1977, this definition was updated by Peter Gund to "a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity" (Gund. *Prog. Mol. Subcell. Biol.* 1977, 5: pp 117–143). The IUPAC definition of a pharmacophore is "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response".

In modern computational chemistry, pharmacophores are used to define the essential features of one or more molecules with the same biological activity. A database of diverse chemical compounds can then be searched for more molecules which share the same features located a similar distance apart from each other.

Typical pharmacophore features are for where a molecule is hydrophobic, aromatic, a hydrogen bond acceptor, a hydrogen bond donor, a cation, or an anion. The features need to match different chemical groups with similar properties, in order to identify novel ligands. Ligands receptor interactions are typically "polar positive", "polar negative" or "hydrophobic". A well-defined pharmacophore model includes both hydrophobic volumes and hydrogen bond vectors.

QSAR screening

We have used QSAR in our study to develop a predictive model for newly developed Epipodophyllotoxin compounds. Traditional QSAR studies have been used since the early 1970s to predict activities of untested molecules. The pharmacophore based, quantum mechanics (QM) based and physicochemical based 3D-QSAR have been employed to build QSAR models for a wide range of applications and were shown good predictivity. With a wide range of molecular structures and their complementary activities, it has been assumed that the most important criteria for a systematic study of 3D-QSAR have been satisfied. Although comparative molecular field analysis (CoMFA) are statistically excellent and offer good predictive performance, they are inherently limited by the need to align the database molecules correctly within 3D space. The determination of the 'active' conformation that each compound will retain is a critical issue due to unavailability of X-ray structure. We should have some knowledge or hypothesis regarding active conformations of the molecules under study as a prerequisite for structural alignment. Hence, the developed models based on CoMFA may not suit to drug design, because of a false conformational hypothesis. However, we were motivated to explore possible alternatives that would use alignment free descriptors derived from 2D or 3D molecular topology and thus alleviate frequent ambiguity of structural alignment typical for 3D QSAR methods.

Quantitative structure-activity relationship (QSAR) is one of the most important methods in chemometrics, which give information that is useful for drug design and medicinal chemistry. A QSAR equation is a mathematical equation that correlates the biological activity to a wide variety of physical or chemical parameters. There are many examples available in literature in which QSAR models have been used successfully for the screening of compounds for biological activity.

Three-dimensional quantitative structure-activity relationships (*3DQSAR*) involve the analysis of the quantitative relationship between the biological activity of a set of compounds and their three-dimensional properties using statistical correlation methods.

The 3D QSAR method is categorized to the structure-based and ligand-based manners, respectively. The structure-based 3D QSAR is only available for the case where the 3D structures of a target protein or its homologue bound to the active compound has been experimentally solved using the X-ray crystal structure analysis.

When a homology-modeled target protein is used for this purpose, the numerous combinations of the side chain rotamers in the ligand-binding site must be considered in order to obtain a linear relationship between the *in vitro* activity of the compounds and the interaction energy (or score) of the complexes.

On the other hand, the ligand-based 3D QSAR is useful, when a fine 3D structure of the target protein is not known experimentally.

ADME is an acronym in pharmacokinetics and pharmacology for **a**bsorption, **d**istribution, **m**etabolism, and **e**xcretion, and describes the disposition of a pharmaceutical compound within an organism. The four criteria all influence the drug levels and kinetics of drug exposure to the tissues and hence influence the performance and pharmacological activity of the compound as a drug:

Absorption

Before a compound can exert a pharmacological effect in tissues, it has to be taken into the bloodstream — usually via mucous surfaces like the digestive tract (intestinal absorption). Uptake into the target organs or cells needs to be ensured, too. This can be a serious problem at some natural barriers like the blood-brain barrier. Factors such as poor compound solubility, chemical instability in the stomach, and inability to permeate the intestinal wall can all reduce the extent to which a drug is absorbed after oral administration. Absorption critically determines the compound's bioavailability. Drugs that absorb poorly when taken orally must be administered in some less desirable way, like intravenously or by inhalation (e.g. zanamivir).

Distribution

The compound needs to be carried to its effector site, most often via the bloodstream. From there, the compound may distribute into tissues and organs, usually to differing extents.

Metabolism

Compounds begin to be broken down as soon as they enter the body. The majority of small-molecule drug metabolism is carried out in the liver by redoxa enzymes, termed cytochrome P450 enzymes. As metabolism occurs, the initial (parent) compound is converted to new compounds called metabolites. When metabolites are pharmacologically inert, metabolism deactivates the administered dose of parent drug and this usually reduces the effects on the body. Metabolites may also be pharmacologically active, sometimes more so than the parent drug.

Excretion/Elimination

Compounds and their metabolites need to be removed from the body via excretion, usually through the kidneys (urine) or in the feces. Unless excretion is complete, accumulation of foreign substances can adversely affect normal metabolism.

There are three sites where drug excretion occurs. The kidney is the most important site and it is where products are excreted through urine. Biliary excretion or faecal excretion is the process that initiates in the liver and passes through to the gut until the products are finally excreted along with waste products or faeces. The last method of excretion is through the lungs e.g. anaesthetic gases.

Excretion of drugs by the kidney involves 3 main mechanisms:

- Glomerular filtration of unbound drug.
- Active secretion of (free & protein-bound) drug by transporters e.g. anions such as urate, penicillin, glucuronide, sulphate conjugates) or cations such as choline, histamine.
- Filtrate 100-fold concentrated in tubules for a favourable concentration gradient so that it may be reabsorbed by passive diffusion and passed out through the urine.

CHAPTER 2

Receptors for inhibition

Topoisomerase: Receptor for Epipodophyllotoxin

Topoisomerases are enzymes that unwind and wind DNA, in order for DNA to control the synthesis of proteins, and in order for DNA to reproduce. They cut the DNA, and at the end of the process connect it again. Topoisomerases are isomerase enzymes that act on the topology of DNA.

Once cut, the ends of the DNA are separated, and a second DNA duplex is passed through the break. Following passage, the cut DNA is religated. Type IIA topoisomerases are essential in the separation of daughter strands at the end of replication. This function is performed by topo II in eukaryotes and by topo IV in prokaryotes. Failure to separate these strands leads to cell death. Type IIA topoisomerases have the special ability to relax DNA to a state below that of thermodynamic equilibrium, a feature unlike type IA, IIA, and IIB topoisomerases. This ability, known as topology simplification.
Type IIA topoisomerases consist of several key motifs: an N-terminal GHKL ATPase domain (for Gyrase, Hsp, Kinase and MutL), a Toprim domain (sometimes called a Rossman fold). Eukaryotic type II topoisomerases are homodimers (A_2).
The central core of the protein contains a Toprim fold and a DNA binding core that contains a Winged Helix domain (WHD), often referred to as a CAP domain since it was first identified to resemble the WHD of Catabolite Activator Protein. The catalytic tyrosine lies on this WHD. The Toprim fold is a Rossman fold that contains three invariant acidic residues that coordinate a magnesium ion that is involved in DNA cleavage and DNA relegation. The topo II core was later solved in two new conformations, one by Fass et al. (Nature Structure Biology 1999, PDB ID = 1BJT) and one by Dong et al. (Nature 2007, PDB ID = 2RGR). The Fass structure shows that the Toprim domain is flexible and that this flexibility can allow the Toprim domain to coordinate with the WHD to form a competent cleavage complex. This was eventually substantiated by the Dong et al. structure that was solved in the presence of DNA.

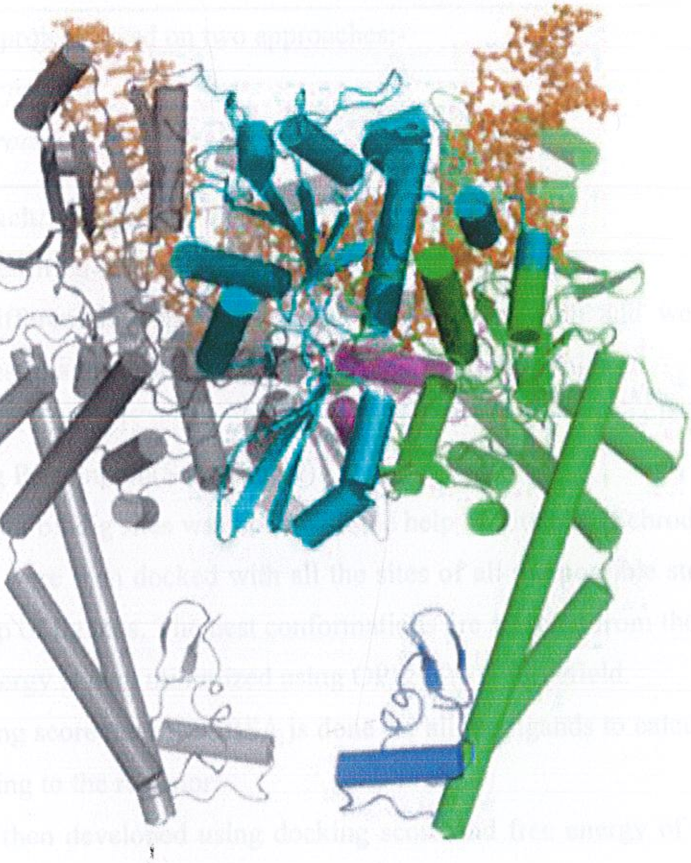


***Figure.3**: Structure of yeast topoisomerase II bound to a doubly-nicked 34-mer duplex DNA (PDB ID =2RGR). The Toprim fold is colored cyan, the DNA is colored orange, the HTH is colored magenta, and the C-gate is colored purple. Notice that the DNA is bent by ~160 degrees through an invariant isoleucine (Ile833 in yeast).*

CHAPTER 3

Procedure approached

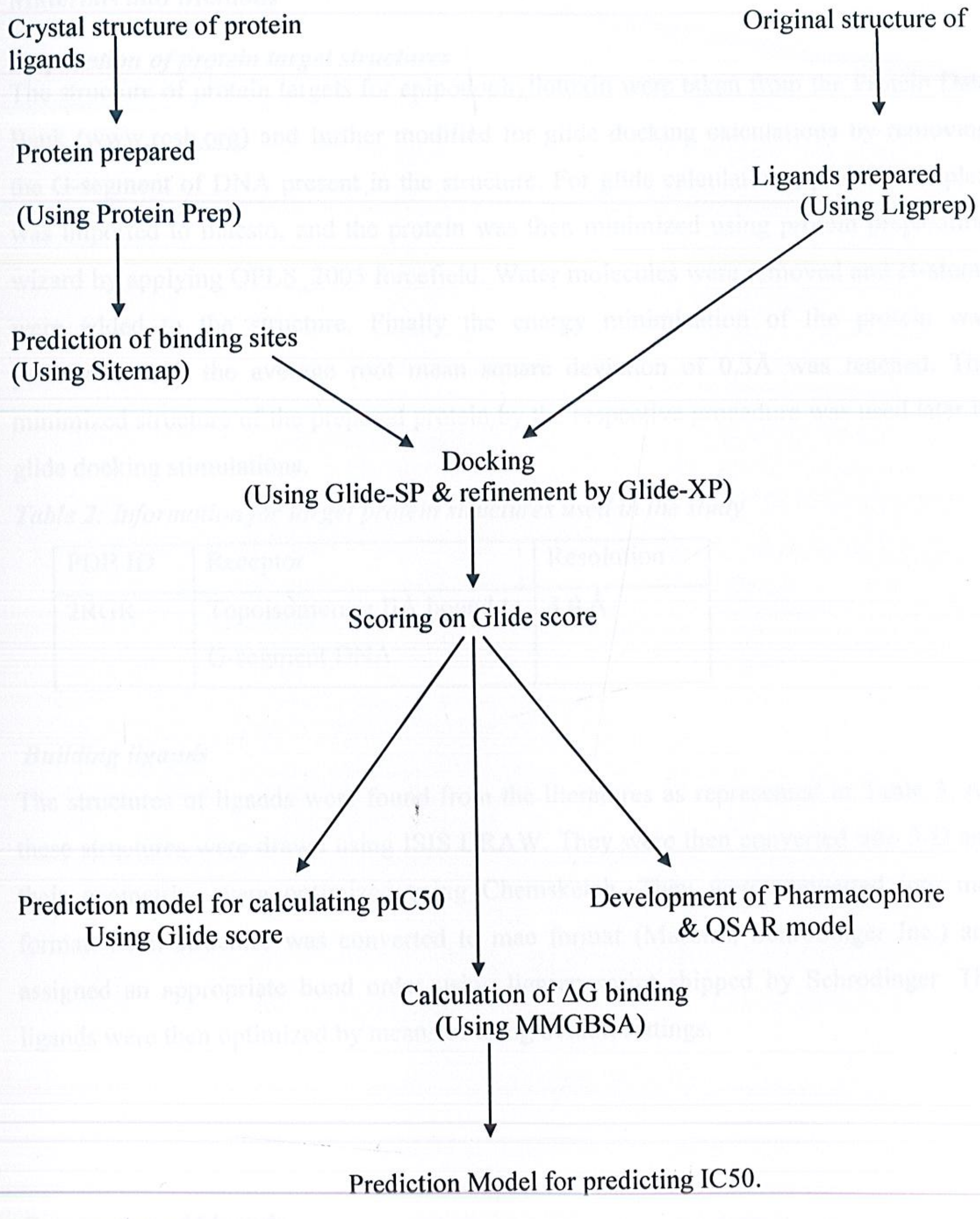
We are carrying out this project based on two approaches:-

1) ***"Structure based approach"***
2) ***"Ligand based approach"***

Structure based approach: ***Receptor structure known, Ligand known***

1) The ligands were identified from.
2) The structure of different ligands were drawn using Chemsketch and were then incorporated in Maestro and were prepared using Ligprep (Schrodinger).
3) Structure of the target receptors was downloaded from PDB (Protein Data Bank) and were prepared using Proteinprep(Schrodinger).
4) Prediction of different bindig sites was done with the help of Sitemap (Schrodinger).
5) The target proteins were then docked with all the sites of all the possible structures of a particular group of ligands. The best conformations are selected from the results obtained and the energy is then minimized using OPLS_2005 forcefield.
6) Based on the docking score the MMGBSA is done for all the ligands to calculate the free energy of binding to the receptor.
7) QSAR models are then developed using docking score and free energy of binding with the help of regression analysis.
8) Finally, the pharmacophore model was developed for Epipodophyllotoxin analogues.

Flow chart representation


Crystal structure of protein ligands
Original structure of
Protein prepared
(Using Protein Prep)
Ligands prepared
(Using Ligprep)
Prediction of binding sites
(Using Sitemap)
Docking
(Using Glide-SP & refinement by Glide-XP)
Scoring on Glide score
Prediction model for calculating pIC50
Using Glide score
Development of Pharmacophore
& QSAR model
Calculation of ΔG binding
(Using MMGBSA)
Prediction Model for predicting IC50.

Materials and Methods

Preparation of protein target structures

The structure of protein targets for epipodophyllotoxin were taken from the Protein Data Bank (www.rcsb.org) and further modified for glide docking calculations by removing the G-segment of DNA present in the structure. For glide calculations, protein complex was imported to maesto, and the protein was then minimized using protein preparation wizard by applying OPLS_2005 forcefield. Water molecules were removed and H-atoms were added to the structure. Finally the energy minimization of the protein was performed until the average root mean square deviation of 0.3Å was reached. The minimized structure of the prepared protein by the respective procedure was used later in glide docking stimulations.

Table 2: *Information for target protein structures used in the study*

PDB ID	Receptor	Resolution
2RGR	Topoisomerase IIA bound to G-segment DNA	3.0 Å

Building ligands

The structures of ligands were found from the literatures as represented in Table 3. All these structures were drawn using ISIS DRAW. They were then converted into 3-D and their geometries were optimized using Chemsketch. They were converted into mol format. Each structure was converted to mae format (Maestro, Schrodinger Inc.) and assigned an appropriate bond order using ligprep script shipped by Schrodinger. The ligands were then optimized by means of using default settings.

Preparation of Ligands

Ligprep is a robust collection of tools designed to prepare high quality, all-atom 3D structures for large numbers of drug-like molecules, starting with 2D or 3D structures in SD or Maestro format. The simplest use of Ligprep produces a single, low-energy, 3D structure with correct chiralities for each successfully processed input structure. Ligprep can also produce a number of structures from each input structure with various ionization states, tautomers, steriochemisteries, and ring conformations, and eliminate molecules using various criteria including molecular weight or specified numbers and types of functional groups present. The Ligprep process consists of a series of steps that perform conversions, apply corrections to the structures, generate variations on the structures, eliminate unwanted structures, and optimize the structures. Numbers of conformations for each ligand was changed from 32 to 6. All other values were kept at default settings.

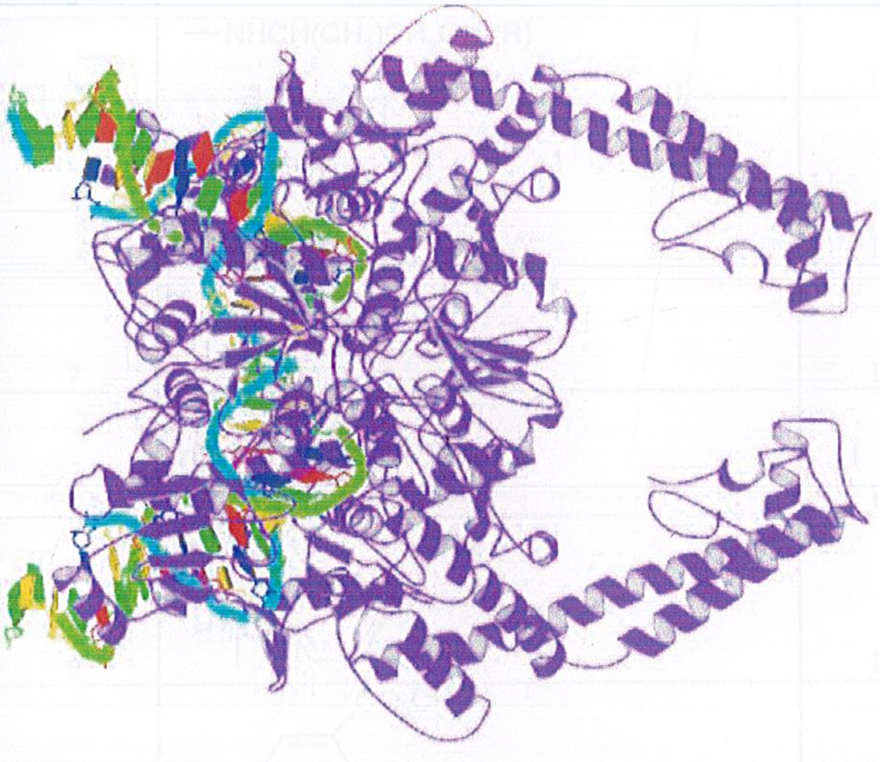
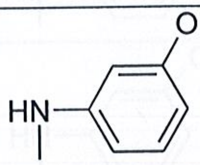
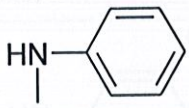
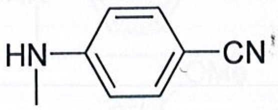
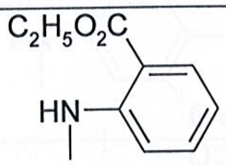
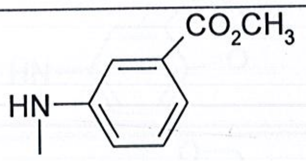
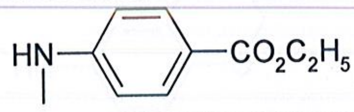
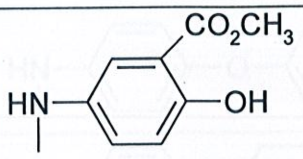
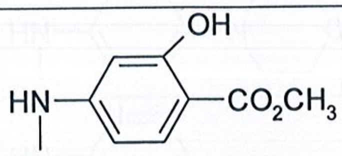
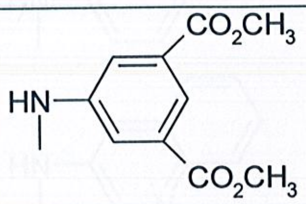
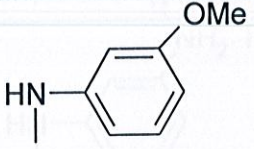
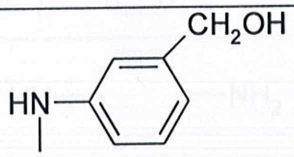
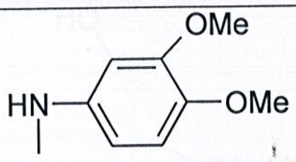
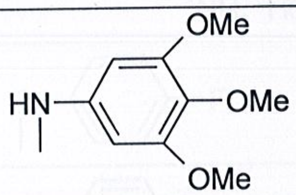
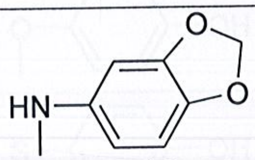
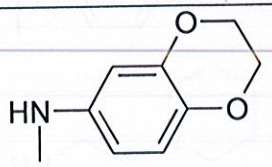
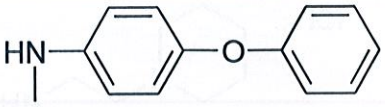
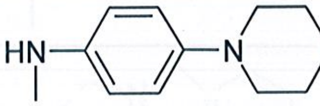
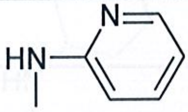
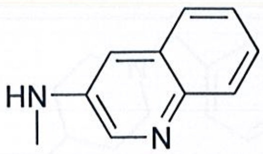
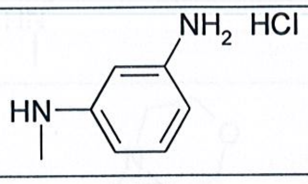
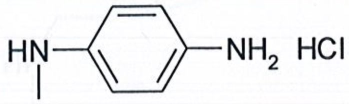
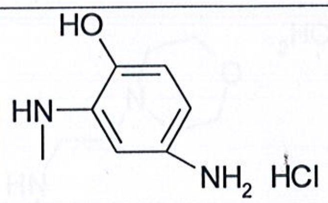
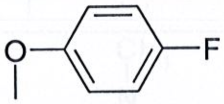
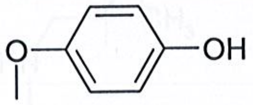
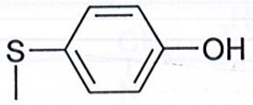
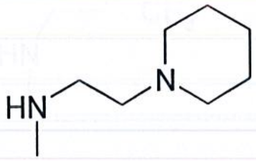


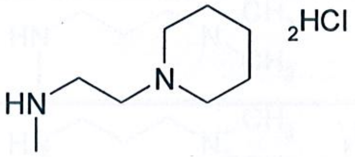
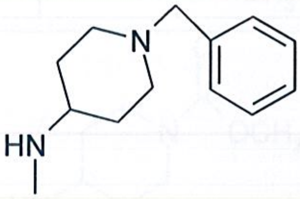
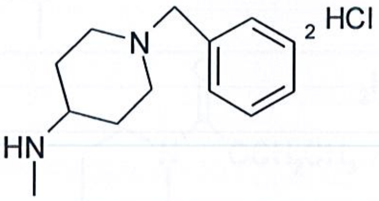
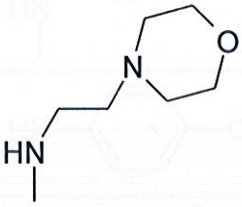
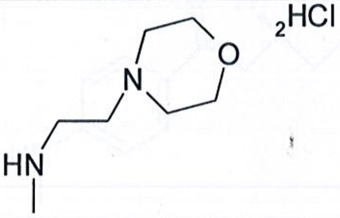
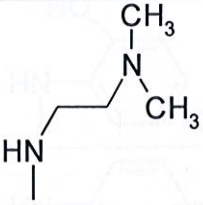
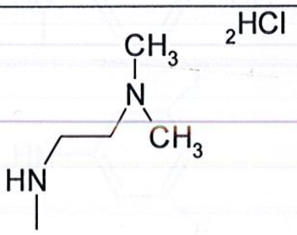
Figure 4: *Ribbon representation of Topoisomerase II (PDB ID: 2RGR)*

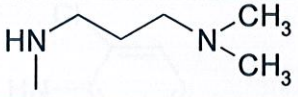
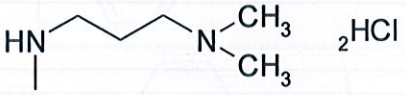
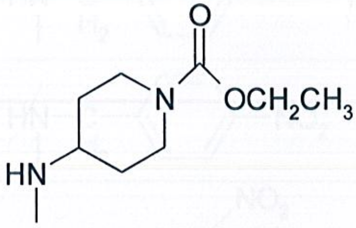
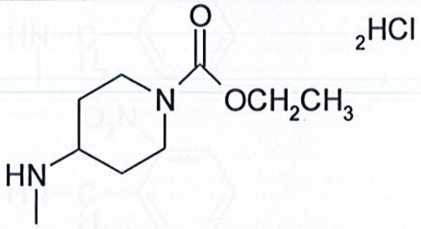
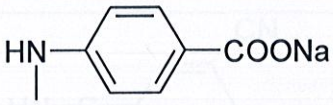
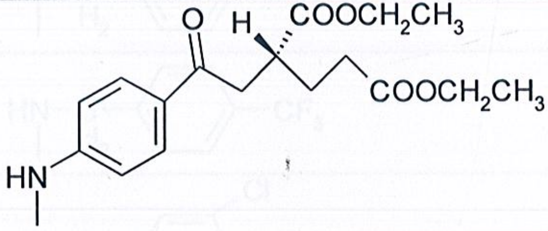
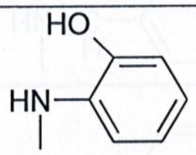
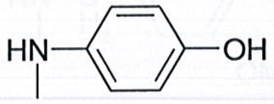
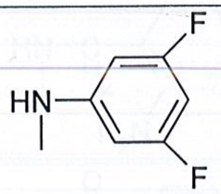
Table 3: *List of Epipodophyllotoxin analogs and their basic properties.*

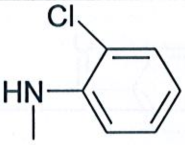
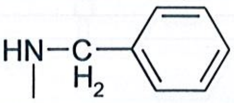
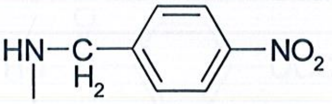
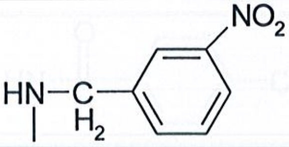
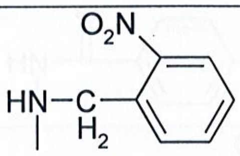
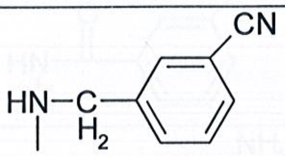
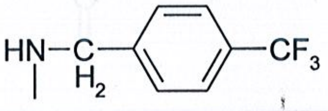
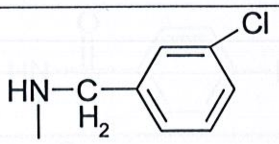
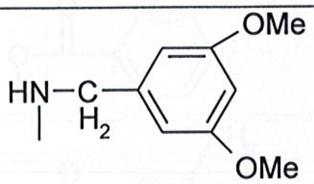
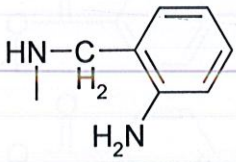
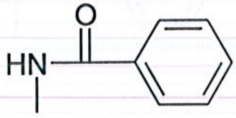
Compound	R	Structure Type	Cellular Protien-DNA Complex Formation
1	$-OH$	1	42.2
2	$-NHCH_2CH_{20}CH_3$	1	110.8
3	$-NHCH_2CH{=}CH_2$	1	84.1
4	$-NHCH_2CH(OH)CH_3$ (R)	1	167.2
5	$-NHCH(CH_3)CH_2OH$ (R)	1	161.7
6		1	290
7		1	243
8		1	211
9		1	4
10		1	249
11		1	207

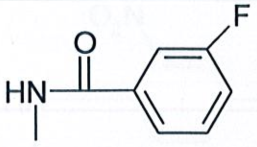
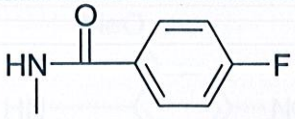
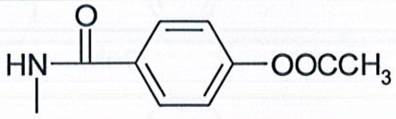
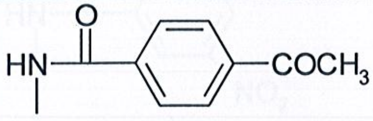
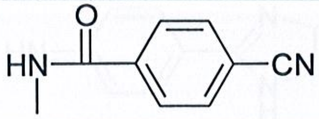
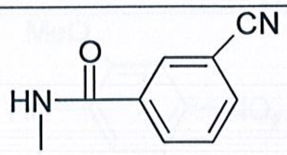
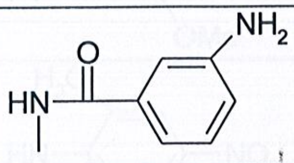
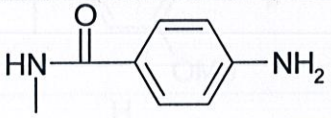
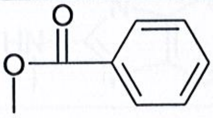
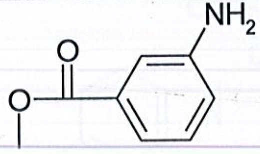
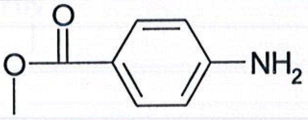
12		1	83
13		1	129
14		1	50
15		1	104
17		1	235
18		1	180
19		1	47
20		1	164
21		1	279

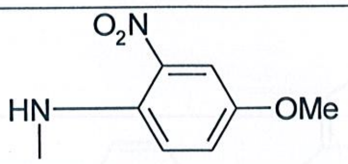
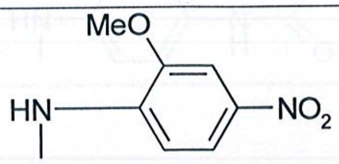
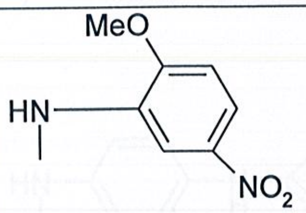
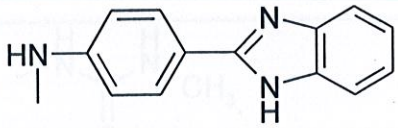
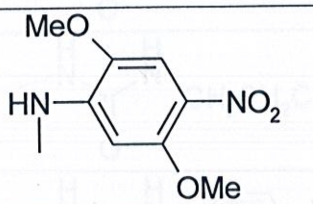
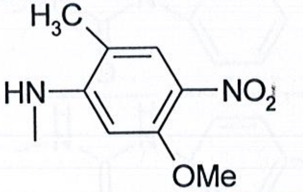
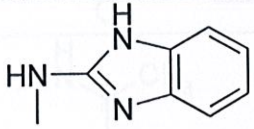
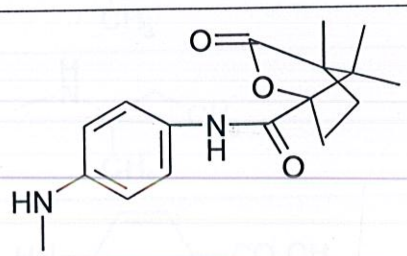
22		1	97
23		1	140
24		1	97
25		1	123
26		1	140
27		1	330
28		1	11
29		1	57
30		1	34
31		1	10
32		1	190

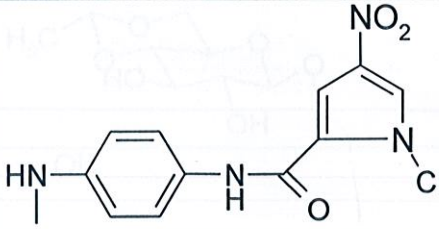
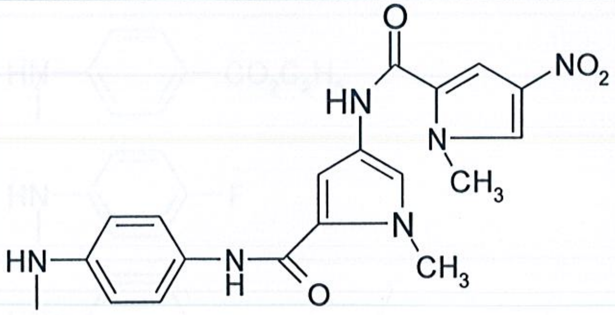
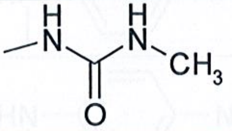
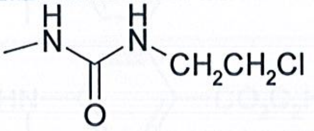
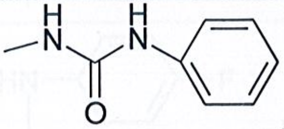
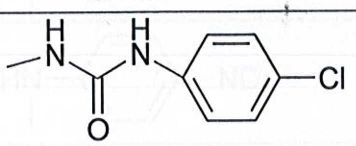
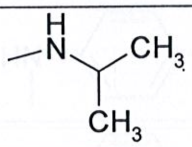
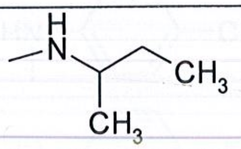
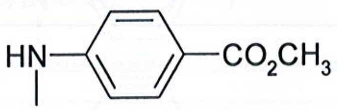
33		1	183
34		1	83
35		1	172
36		1	77
37		1	140
38		1	203
39		1	183

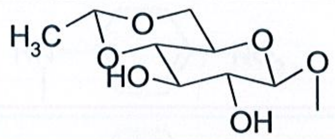
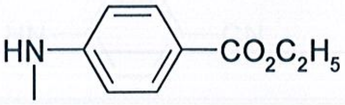
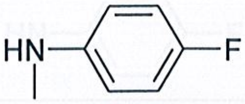
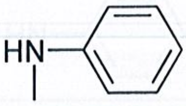
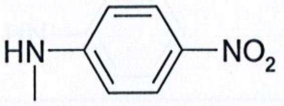
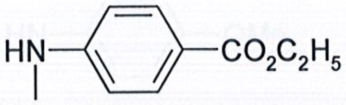
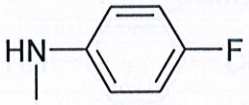
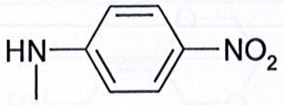
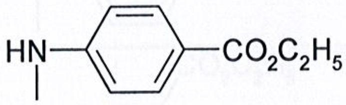
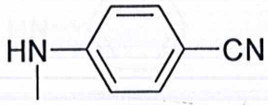
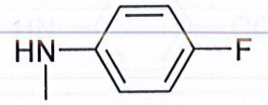
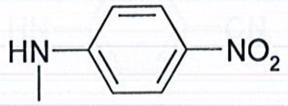
40		1	186
41		1	179
42		1	17
43		1	138
44		1	6.9
45		1	83
46		1	151
47		1	211
48		1	115

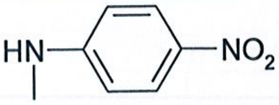
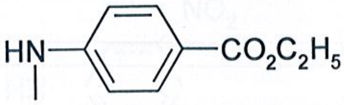
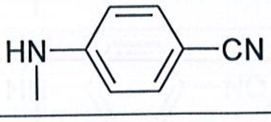
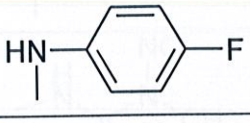
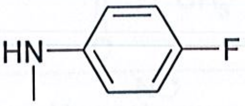
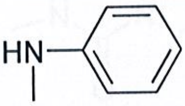
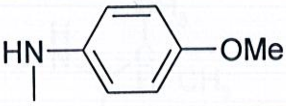
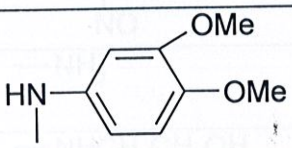
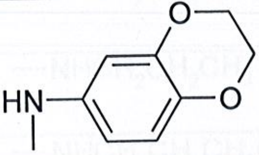
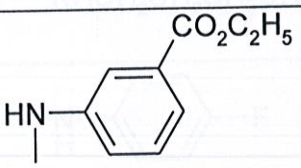
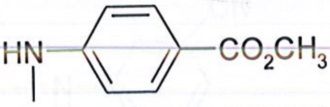
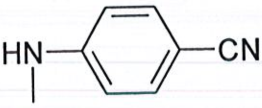
49		1	32
50		1	181
51		1	216
52		1	130
53		1	144
54		1	225
55		1	99
56		1	159
57		1	144
58		1	184
59		1	177

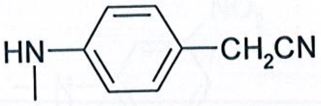
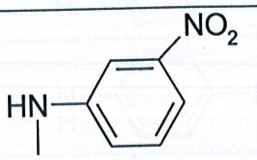
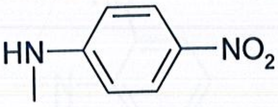
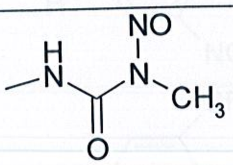
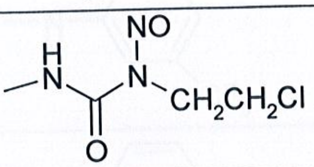
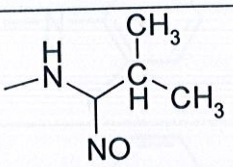
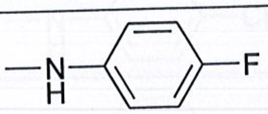
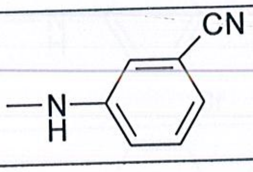
61		1	116
62		1	117
63		1	137
64		1	124
65		1	159
66		1	149
67		1	149
68		1	120
69		1	94
70		1	100
71		1	94

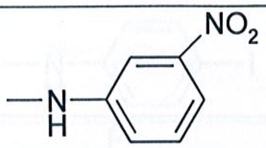
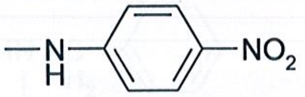
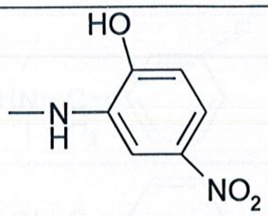
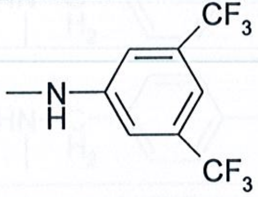
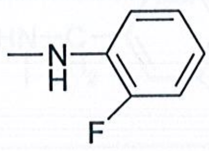
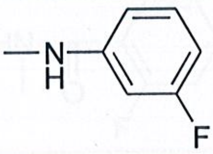
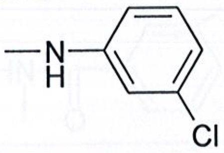
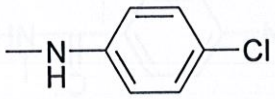
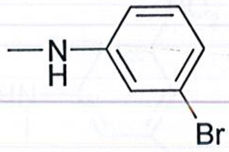
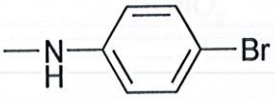
72		1	15
73		1	83
74		1	12
75		1	128
76		1	4.4
77		1	3.5
78		1	58
79		1	88

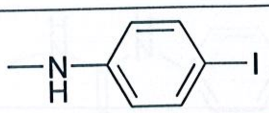
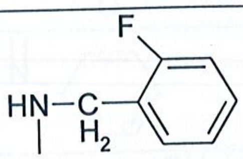
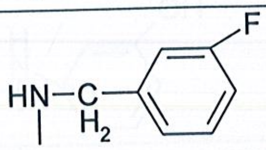
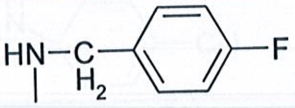
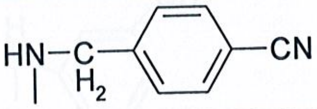
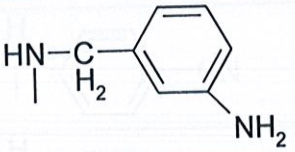
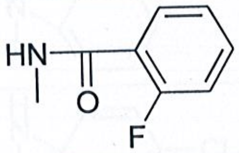
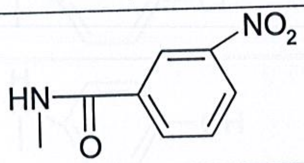
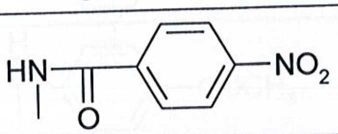
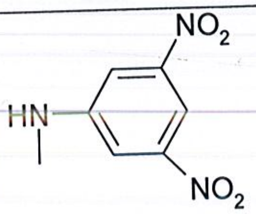
80		1	100
81		1	26
82		1	81
83		1	143
84		1	148
85		1	125
86		1	109
87		1	73
88		1	207

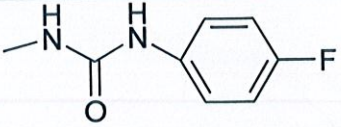
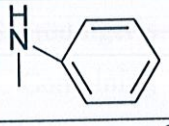
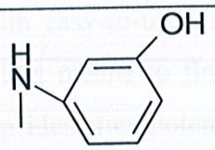
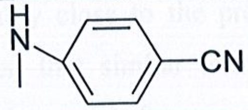
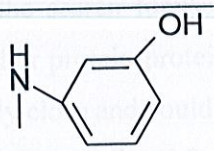
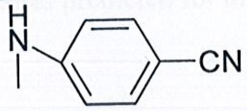
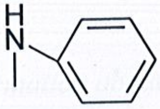
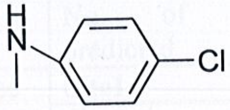
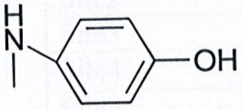
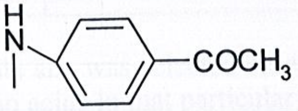
89		2	6.1
90	—OH	2	15.6
91		3	22
92		3	11
93		4	4
94		4	99
95		4	138
96		4	52
97		5	75
98		5	127
99		5	125
100		5	108
101		3	23

102		6	8
103		6	9
104		6	12
105		6	8
106		7	117
107		7	105
108		7	96
109		7	69
110		7	119
111		7	94
112		7	175
113		7	146

114		7	109
115		7	75
116		7	200
117		8	41
118		8	7
119		9	1
120	—NH_2	1	36.4
121	—$NHCH_2CH_2OH$	1	121.4
122	—$NHCH_2CH_2CH_3$	1	69.7
123	—$NHCH_2CH_2CH_2OH$	1	89.2
124		1	213
125		1	137

126		1	230
127		1	323
128		1	15
129		1	21
131		1	121
132		1	158
133		1	51
134		1	99
135		1	62
136		1	179

137		1	64
138		1	126
139		1	216
140		1	169
141		1	284
142		1	191
143		1	128
144		1	86
145		1	160
146		1	20

147		1	118
148		3	9
149		3	4
150		4	62
151		4	18
152		3	33
153		7	128
154		7	77
155		7	83
156		7	147

Site map

SiteMap is Schrödinger's program for identifying, evaluating, and visualizing ligand binding sites. Combining a novel algorithm for rapid binding site identification and evaluation with easy-to-use property visualization tools, SiteMap provides researchers with an efficient means to find and better exploit the characteristics of ligand binding sites. SiteMap identifies potential ligand binding sites by linking together "site points" that are suitably close to the protein surface and sufficiently well sheltered from the solvent. Given that similar terms dominate the site scoring function, this approach ensures that the search focuses on regions of the protein most likely to produce tight protein-ligand or protein-protein binding. Subsites are merged into larger sites when they are sufficiently close and could be bridged in solvent-exposed regions by ligand atoms. The different sites predicted for different receptors were represented in Table 4.

***Table 4**: Information about sites predicted for different receptors.*

Receptor	No. of sites predicted	Docking score	No. of Ligands Docked	Site selected
Topoisomerase II	**Site1**	**-9.24 to -4.37**	154	Site 1
	Site2	-8.03 to -0.7	150	
	Site3	-7.64 to -0.75	151	
	Site4	-6.82 to 0.4	148	
	Site5	-5.34 to 1.86	150	

The appropriate site was selected based on the docking scores obtained and the presence of active amino acids in that particular site.

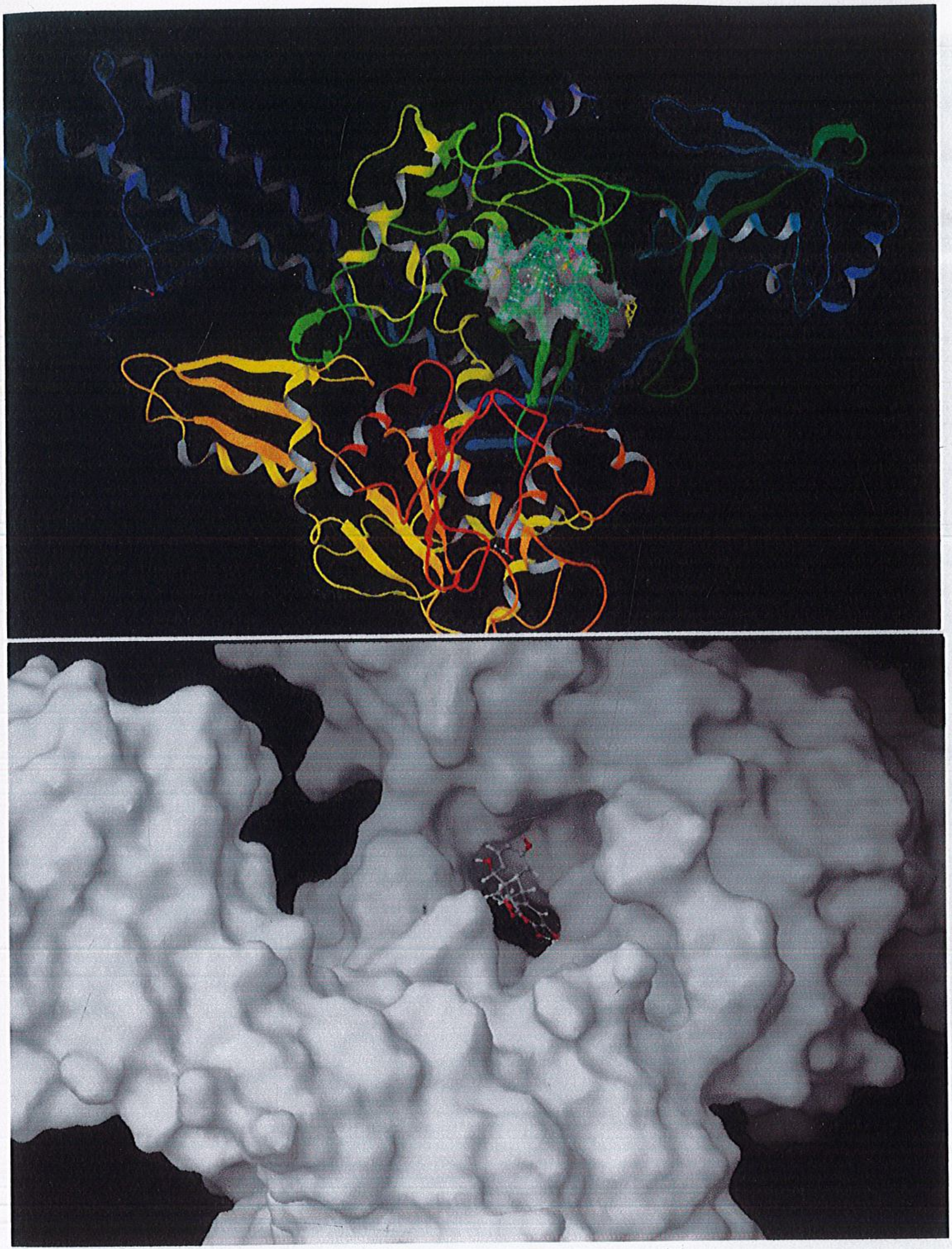


***Figure 5:** Ribbon and Surface view of Epipodophyllotoxin analogues docked to site 1 of Topoisomerase II*

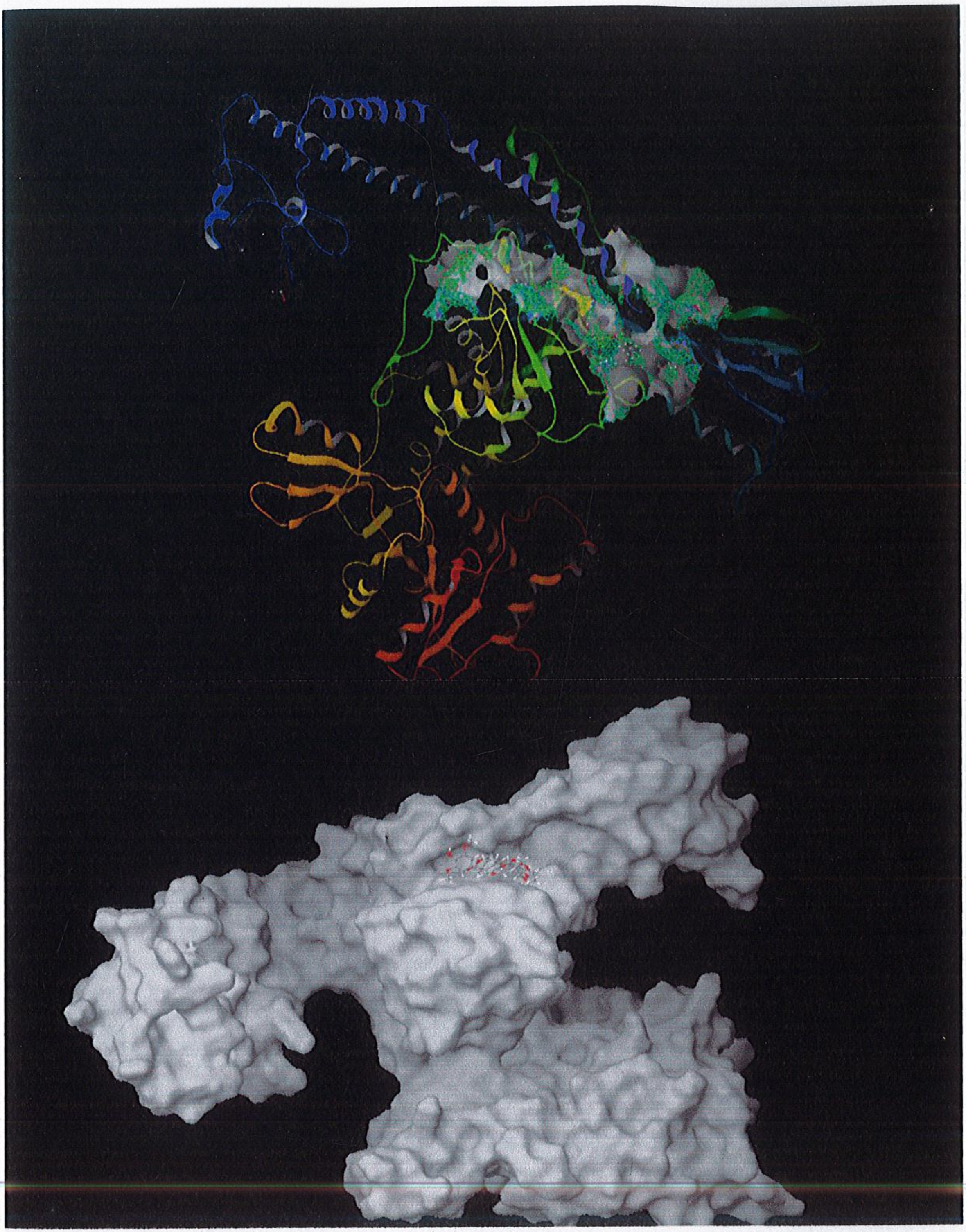


Figure 6: *Ribbon and Surface view of Epipodophyllotoxin analogues docked to site 2 of Topoisomerase II*

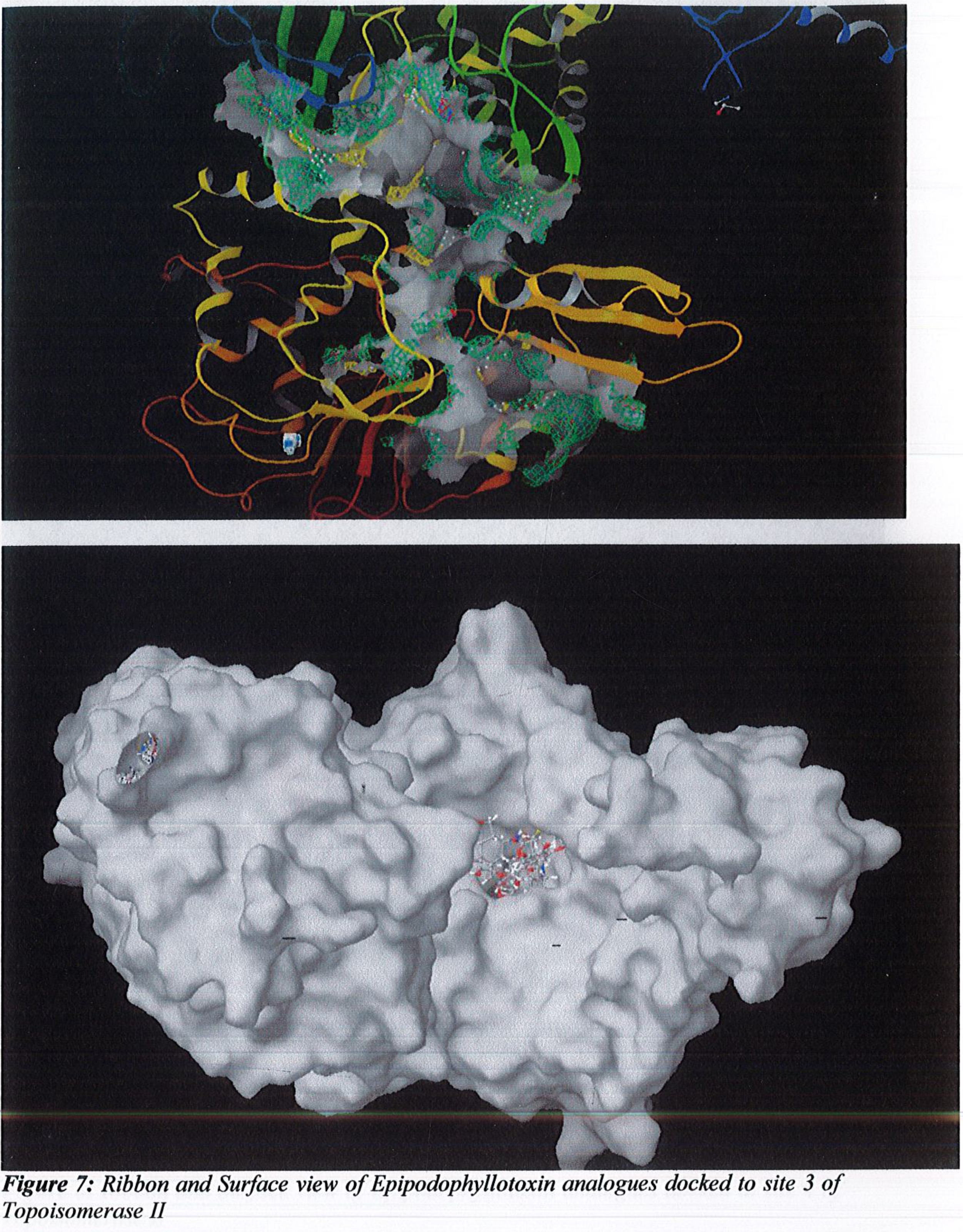


***Figure** 7: Ribbon and Surface view of Epipodophyllotoxin analogues docked to site 3 of Topoisomerase II*

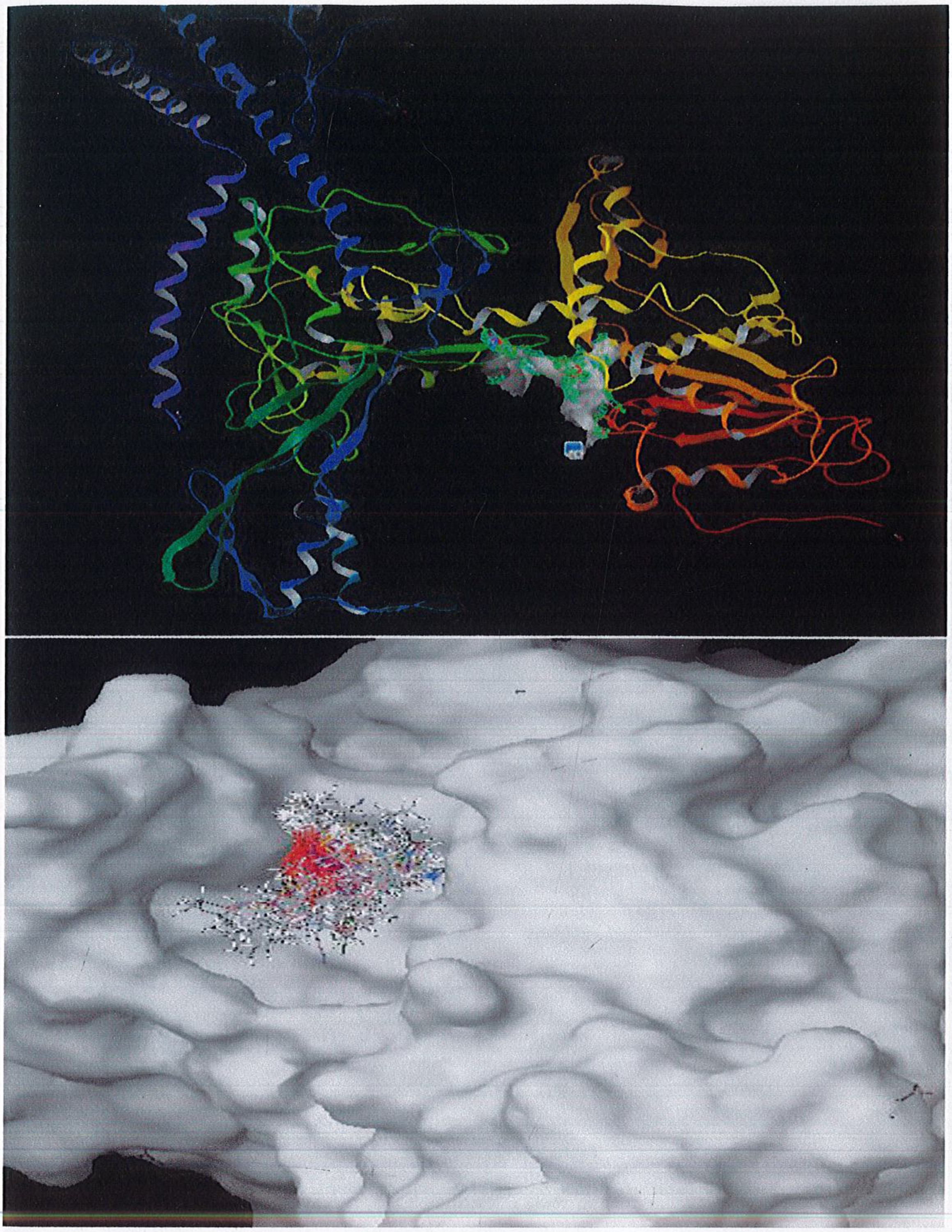


***Figure 8:** Ribbon and Surface view of Epipodophyllotoxin analogues docked to site 4 of Topoisomerase II*

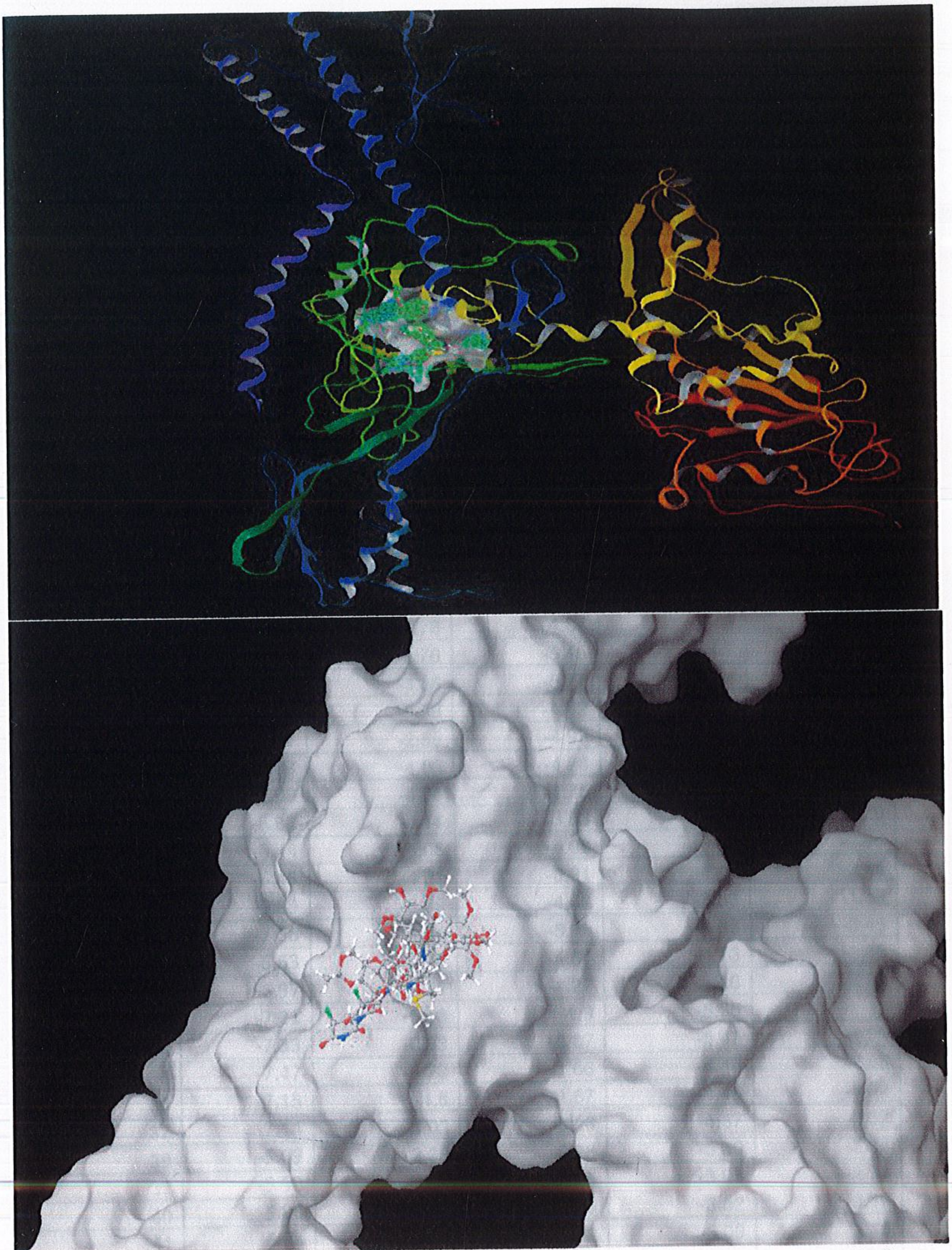


Figure 9: Ribbon and Surface view of Epipodophyllotoxin analogues docked to site 5 of Topoisomerase II

Table 5: Comparative Docking score for each site

Ligand name	Site 1	Site 2	Site 3	Site 4	Site 5
1	-6.17	-5.38	-4.78	-4.74	-1.7
2	-5.86	-3.72	-5.92	-5.05	-1.61
3	-5.63	-4.72	-2.36	-3.47	-2.14
4	-6.92	-3.85	-1.94	-5.56	-3.08
5	-6.86	-4.26	-4.27	-3.16	-2.98
6	-7.32	-4.9	-5.63	-4.14	-3.08
7	-5.76	-4.74	-1.96	-2.37	-2.09
8	-7.36	-3.95	-3.68	-4.24	-0.92
9	-7.37	-5.35	-4.31	-4.74	0.49
10	-6.69	-4.18	-4.97	-4.24	-1.92
11	-6.88	-5	-4.41	-3.72	-1.99
12	-7.13	-5.33	-6.07	-4.84	-1.87
13	-6.71	-4.88	-4.87	-5.39	-2.84
14	-7.64	-5.21	-5.1	-4.2	-1.57
15	-6.57	-3.96	-4.97	-3.46	-2.86
17	-7.39	-4.94	-4.64	-5.05	1.86
18	-7.82	-5.35	-4.86	-5.05	-3.81
19	-7.91	-4.78	-4.84	-4.43	-2.83
20	-6.67	-3.76	-5.16	-4.7	-2.17
21	-7.7	-3.88	-5.77	-5.14	-3.74
22	-7.48	-5.99	-4.42	-3.47	1.82
23	-4.57	-2.59	-4.93	-4.43	-2.62
24	-5.63	-4.06	-4.78	-3.68	-0.83
25	-6.99	-4.35	-1.75	-5.66	-3.02
26	-6.37	-4.11	-3.53	0.4	-2.26
27	-6.44	-5.08	-4.08	-3.88	-2.96
28	-6.45	-4.91	-0.91	-4.45	-2.85
29	-5.89	-3.39	-4.43	-3.33	-1.37
30	-7.58	-4.84	-5.29	-4.63	-3.45
31	-6.63	-3.78	-4.54	-4.73	-1.86
32	-6.97	-5.21	-1.27	-3.6	-2.91
33	-7.21	-5.18	-1.9	-5.36	-1
34	-7.33	-0.93	-3.37	-5.47	-1.98
35	-7.47	-6.22	-1.66	-4.46	-2.65
36	-7.15	-4.5	-1.57	-4.45	-3.02
37	-7.14	-4.02	-5.26	-4.87	-2.67
38	-5.09	-2.38	-6.28	-3.69	-2.38
39	-5.98	-7.38	-5.85	-3.03	-2.33
40	-5.96	-3.87	-3.15	-3.52	-3.18
41	-6.38	-3.76	-3.2	-5.09	-2.15
42	-6.8	-5.45	-4.63	-4.66	0.5
43	-7.21	-4.35	-5.37	-4.55	-2.17
44	-7.4	-5.11	-7.6	-4.34	-2.17
45	-7.83	-2.98	-5.47	-4.39	-3.16

46	-7.02	-4.42	-5.84	-5.27	-3.2
47	-7.47	-4.5	-6.35	-4.46	-2.62
48	-7.49	-4.18	-3.58	-4.54	-1.94
49	-6.75	-3.88	-3.98	-5.17	-2.99
50	-6.82	-0.74	-1.28	-5.22	-2.03
51	-6.79	-3.82	-3.87	-5.26	-1.44
52	-6.83	-3.41	-3.54	-4.85	-2.41
53	-7.14	-0.75	-1.99	-5.26	-1.74
54	-7.78	-3.08	-2.42	-5.17	-0.83
55	-8.05	-5.87	-2.27	-5.33	-2.02
56	-7.27	-3.45	-2.72	-4.77	-2.11
57	-7.19	-1.44	-2.62	-5.53	-1.47
58	-7.5	-4.35	-1.88	-5.34	-2.73
59	-7.99	-3.41	-5.7	-4.6	-2.44
61	-5.8	-3.79	-7.33	-3.66	0.76
62	-9.24	-3.98	-4.89	-2.73	-2.27
63	-7.7	-5.21	-4.7	-4.94	-3.49
64	-4.37	-2.34	-5.79	-4.1	-1.83
65	-7.01	-3.53	-4.26	-2.59	-2.1
66	-5.98	-5.15	-3.91	-2.49	-2.03
67	-8.6	-3.93	-3.43	-5	-2.65
68	-6.3	-2.3	-5.66	-5.97	-2.02
69	-7.43	-2.86	-5.24	-3.01	-2.35
70	-8.95	-3.83	-6.16	-5.75	-2.52
71	-8.79	-3.87	-4.36	-5.59	-2.4
72	-7.24	-6.19	-4.96	-4.08	-2.95
73	-7.31	-5.81	-5.43	-4.3	-0.59
74	-7.46	-3.56	-5.79	-4.82	1.58
75	-5.89	-5.25	-5.28	-4.59	-1.79
76	-7.52	-4.76	-4.79	-4.63	-3.19
77	-7.51	-3.81	-5.35	-4.85	-2.43
78	-6.21	-6.69	-5.81	-4.79	-2.07
79	-8.44	-5.44	-5.42	-2.49	-3.59
80	-8.81	-4.45	-5.1	-4.67	-2.22
81	-8.4	-4.92	-2.54	-6.35	-1.57
82	-6.82	-4.46	-3.2	-4.53	-2.61
83	-6.24	-4.33	-5.87	-4.89	-2.15
84	-7.04	-5.04	-4.55	-4.01	-2.23
85	-8.01	-4.8	-5.39	-4.81	-1.89
86	-5.17	-3.72	-0.75	-3.45	-2.37
87	-5.95	-3.86	-3.1	-2.31	-2.33
88	-7.91	-4.51	-4.13	-3.52	-1.85
89	-5.65	-4.07	-6.92	-6.82	-2.47
90	-6.37	-4.91	-4.6	-2.62	-1.83
91	-7.53	-6.01	-5.59	-5.48	-3.38
92	-7.82	-1.28	-6.64	-3.09	0.71
93	-7.84	-6.1	-6.89	-5.25	-2.97

94	-7.48	-8.03	-6.89	-0.45	-3.5
95	-7.16	-6.19	-5.16	-6.2	-3.65
96	-5.41	-7.69	-7.42	-5.81	-2.5
97	-6.88	-5.01	-4.98	-3.88	0.03
98	-7.91	-3.03	-4.26	-3.51	-2.78
99	-7.26	-5.51	-5.17	-3.82	-2.75
100	-6.52	-4.13	-6.18	-4.27	-2.93
101	-7.69	-4.11	-5.99	-5.57	-1.4
102	-6.75	-3.71	-4.42	-3.48	-3.08
103	-5.92	-5.18	-4.77	-1.71	-1.95
104	-6.92	-3.6	-4.64	-4.14	-2.6
105	-7.39	-3.29	-4.22	-4.49	-1.61
106	-7.72	-4.62	-5.3	-5.23	-2.37
107	-7.33	-4.62	-5.12	-5.85	-3.62
108	-6.66	-3.79	-6.99	-4.53	-3.66
109	-6.83	-4.5	-7.49	-5	-3.78
110	-5.94	-4.71	-7.24	-4.73	-4.12
111	-6.74	-4.48	-7.26	-5.37	-0.06
112	-6.03	-4.15	-6.27	-4.7	-4.17
113	-6.83	-4.3	-5.7	-6.1	-3.57
114	-6.75	-4.07	-6.47	-5.52	-0.02
115	-6.33	-3.34	-5.53	-6.41	-2.02
116	-7.36	-4.28	-6.05	-0.86	-2.68
117	-5.42	-3.24	-3.96	-4.24	-1.76
118	-5.53	-1.74	-2.74	-3.79	-1.32
119	-5.31	-5.11	-5.1	-4.9	-3.03
120	-8.95	-4.02	-1.45	-4.06	-2.7
121	-5.94	-5.33	-5.87	-3.24	-1.54
122	-5.28	-4.89	-4.97	-3.46	-3.1
123	-6.39	-3.26	-6.97	-4.71	-3.22
124	-5.78	-4.35	-4.19	-3.9	-1.9
125	-6.84	-4.21	-4.59	-5.1	-2.35
126	-6.95	-3.89	-5.78	-5.18	1.19
127	-7.34	-4.01	-3.21	-4.3	-3.66
128	-7.36	-4.22	-5.37	-5.39	-3.74
129	-7.94	-4.72	-4.12	-5.13	-2.13
131	-6.78	-4.79	-2.55	-3.25	-2.07
132	-6.01	-3.75	-4.09	-4.11	-1.97
133	-6.7	-3.83	-3.06	-3.57	-2.41
134	-7.38	-4.3	-4.33	-3.68	-2.11
135	-6.52	-3.84	-4.89	-2.97	-1.75
136	-7.39	-4.72	-3.61	-4.65	-2.78
137	-6.1	-4.81	-3.29	-4.08	-1.96
138	-7.27	-0.7	-5.18	-5.44	-2.29
139	-7.74	-4.16	-1.7	-5.13	-2.42
140	-7.35	-3.92	-1.71	-5.49	-1.83
141	-7.16	-0.87	-5.6	-5.5	-2.62

142	-6.68	-3.92	-4.78	-5.06	-2.91
143	-7.66	-4.67	-6.85	-3.76	-2.32
144	-7.3	-4.74	-4.66	-4.17	-2.45
145	-7.74	-4.14	-4.81	-3.56	-2.47
146	-6.7	-3.53	-5.81	-2.5	-2.5
147	-9.18	-4.34	-5.64	-4.86	-2.42
148	-7.9	-3.75	-7.17	-4.74	-4
149	-7.84	-3.94	-7	-5.73	-4.74
150	-7.14	-3.96	-6.96	-1.06	-4.08
151	-8.83	-4.84	-7.64	-4.21	-5.34
152	-7.92	-5.47	-3.44	-1.04	0.25
153	-7.59	-4.66	-6.13	-4.02	-3.13
154	-7.64	-4.33	-6.03	-5.73	-2.84
155	-7.84	-4.5	-6.22	-6.14	-2.79
156	-6.44	-4.28	-6.4	-5.15	-4.24

Glide Docking & Scoring function

The purpose of the dock application is to search for favourable binding configurations between one or smaller, flexible ligands and a rigid macromolecular target, which is usually a protein. For each ligand, a number of configurations called poses were generated and scored in an effort to determine favorable binding modes.

In this study we have used Glide (Schrodinger) for studying the molecular interaction of Epipodophyllotoxin with Topoisomearse IIa. Glide is a program which performs Glide-based Ligand Docking with Energetics and searches for favorable interactions between one or more typically small ligand molecules and a typically larger receptor molecule, usually a protein. Schrodinger recommends the performance of test calculations with different scaling factors for the receptor and ligand atom van der waal radii, because steric repulsive interactions might otherwise be overemphasized, leading to rejection of overall correct binding modes of active compounds. To soften the potential for nonpolar parts of the receptor, we scaled van der waal radii of receptor atoms by 1.00 with partial atomic charge 0.25. Grid boxes of size 3Å, 10Å were generated for each binding site by checking where ligands are binding appropriately. The receptor grid file generated was used for docking, with the option of flexible docking and SP (Standard Precision) mode and later on the docking result is refined using XP (Xtra Precision). Glide generates conformations internally and passes these through a series of filters. At first it places the ligands center at various grid positions of 1Å grid and rotates it through Euler angles. At this stage, crude score values and geometric filters weed out unlikely binding modes. The next filter stage involves a grid based forcefield evaluation and refinement of docking solutions including torsional and rigid movement of the ligands. The OPLS_2005 forcefield is used for this purpose. A small number of surviving docking solutions can then be subjected to a Monte Carlo procedure to try to minimize the score. The final energy evaluation is done with Glide score and a single best pose is generated as the output for a particular ligand.

Glide score (G-score) is given by:-

G-score = a*vdW + b*Coul + Lipo + H-bond + Metal + BuryP + RotB + Site

Where, vdW – van der Waal's energy; Coul – Coulomb energy; Lipo – Lipophilic contact term; H-bond – Hydrogen bonding term; Metal – Metal binding term; BuryP – Penalty for burried polar groups; RotB – Penalty for freezing rotatable bonds; Site – Polar interactions in the active site. And the coefficients of vdW and Coul are: a = 0.065, b = 0.130 for XP Glide.

Glide Methodology

In the docking application, ligands are docked to the full macromolecular target (the receptor); however, the search for binding modes is usually constrained to a specific, smaller region of the receptor called the *site*. The Dock application has been designed to dock drug-sized molecules in relatively small active sites; attempts to dock large ligands, like polypeptides, or dock ligands into very large sites will likely fail.

The Dock application is divided into a number of stages:

1. ***Conformational Analysis.*** Ligands are treated in a flexible manner by rotating rotatable bonds; ring conformations are not searched. Alternatively, ligand conformations may be supplied in a conformation database, in which case no further conformational analysis is conducted.
2. ***Placement.*** A collection of poses is generated from the pool of ligand conformations. The dock application provides a framework for the integration of multiple placement methodologies; each such placement methodology will have different properties.

3. ***Pharmacophore Filtering.*** Optionally, the generated poses are constrained to satisfy an arbitrary pharmacophore query. Such a query is used to bias the search towards known important interactions.
4. ***Scoring.*** Each pose generated by the placement methodology is subjected to scoring in an effort to identify the most favourable poses. The Dock application provides a framework for the integration of multiple scoring methodologies; each such scoring methodology will have different properties. Typically, scoring functions emphasize favourable hydrophobic, ionic and hydrogen bond contacts. The top scoring poses are output after scoring.

Calculated energies and predicted ΔG binding

The site that showed best docking results for each receptor was selected and the associated ligands were then subjected to MMGBSA (Schrodinger) to predict the free energy of binding. In this the calculation is performed first on the receptor, then on the ligand, and finally on the complex based on searches conducted on the receptor, each ligand, and each ligand-receptor complex. The energy difference was then calculated using the equation:

$$\Delta E = E_{complex} - E_{ligand} - E_{protein}$$

The full effects of relaxation and solvation are also included in this mode. To minimize the calculation period, there was a substructure (active site) defined in the receptor molecule, where actually the ligands are binding.

Results and Discussions

Docking results

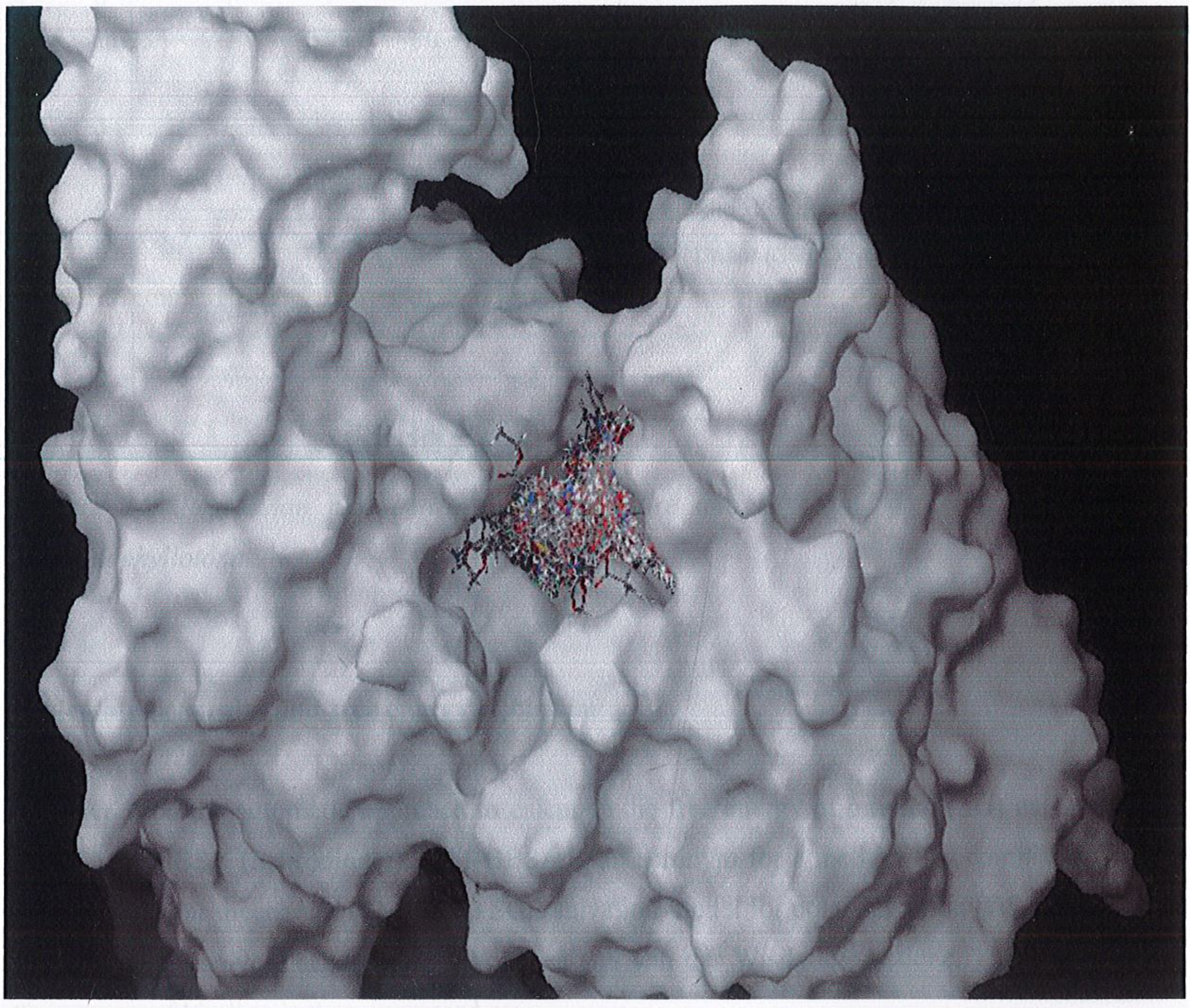
Epipodophyllotoxin analogues were docked to 5 different binding sites of Topoisomerase II predicted by Sitemap. Docking scores for each site was obtained and the site with the best docking score (site 1) was then selected for developing a prediction model using regression analysis between Glide-XP score and experimental logIC50.

Table 6: Docking result for Epipodophyllotoxin for site 1(the best scored site).

S.NO.	ANALOGUES	GLIDE XP-SCORE	GLIDE SP-SCORE	S.NO.	ANALOGUES	GLIDE SP-SCORE	GLIDE XP_SCORE
1.	1	-6.17	-5.47376	**83.**	85	-8.01	-5.13925
2.	2	-5.86	-5.3916	**84.**	86	-5.17	-5.33627
3.	3	-5.63	-5.31671	**85.**	87	-5.95	-5.43775
4.	4	-6.92	-6.09771	**86.**	88	-7.91	-4.61656
5.	5	-6.86	-6.47935	**87.**	89	-5.65	-6.17779
6.	6	-7.32	-5.01998	**88.**	90	-6.37	-5.29894
7.	7	-5.76	-4.89151	**89.**	91	-7.53	-6.08997
8.	8	-7.36	-4.99612	**90.**	92	-7.82	-6.24806
9.	9	-7.37	-4.83905	**91.**	93	-7.84	-5.75175
10.	10	-6.69	-5.69472	**92.**	94	-7.48	-5.69992
11.	11	-6.88	-4.92302	**93.**	95	-7.16	-5.85445
12.	12	-7.13	-5.89587	**94.**	96	-5.41	-5.88773
13.	13	-6.71	-5.29382	**95.**	97	-6.88	-5.20773
14.	14	-7.64	-6.08163	**96.**	98	-7.91	-5.25953
15.	15	-6.57	-5.30825	**97.**	99	-7.26	-5.06046
16.	17	-7.39	-4.92807	**98.**	100	-6.52	-4.89774
17.	18	-7.82	-5.10971	**99.**	101	-7.69	-5.62512
18.	19	-7.91	-5.63606	**100.**	102	-6.75	-5.05003
19.	20	-6.67	-4.94584	**101.**	103	-5.92	-5.56885
20.	21	-7.7	-5.12896	**102.**	104	-6.92	-5.06948
21.	22	-7.48	-5.50239	**103.**	105	-7.39	-5.14873
22.	23	-4.57	-6.10654	**104.**	106	-7.72	-4.96695
23.	24	-5.63	-5.60703	**105.**	107	-7.33	-4.90784
24.	25	-6.99	-5.42739	**106.**	108	-6.66	-5.37886
25.	26	-6.37	-5.15179	**107.**	109	-6.83	-5.3492
26.	27	-6.44	-5.69219	**108.**	110	-5.94	-5.55694
27.	28	-6.45	-5.24118	**109.**	111	-6.74	-5.86403
28.	29	-5.89	-5.03315	**110.**	112	-6.03	-5.38139
29.	30	-7.58	-4.48982	**111.**	113	-6.83	-5.28112
30.	31	-6.63	-5.02489	**112.**	114	-6.75	-5.43315
31.	32	-6.97	-5.91018	**113.**	115	-6.33	-5.47913
32.	33	-7.21	-6.1938	**114.**	116	-7.36	-5.08252
33.	34	-7.33	-6.46337	**115.**	117	-5.42	-4.0798
34.	35	-7.47	-6.0361	**116.**	118	-5.53	-4.49184
35.	36	-7.15	-6.22293	**117.**	119	-5.31	-5.33836
36.	37	-7.14	-6.03056	**118.**	120	-8.95	-6.4308
37.	38	-5.09	-5.58594	**119.**	121	-5.94	-6.21417
38.	39	-5.98	-5.68333	**120.**	122	-5.28	-5.38223
39.	40	-5.96	-5.93844	**121.**	123	-6.39	-6.35273

40.	41	-6.38	-5.98289	**122.**	124	-5.78	-4.89598
41.	42	-6.8	-6.17601	**123.**	125	-6.84	-4.66755
42.	43	-7.21	-5.90867	**124.**	126	-6.95	-5.39905
43.	44	-7.4	-5.37857	**125.**	127	-7.34	-4.75739
44.	45	-7.83	-6.26663	**126.**	128	-7.36	-5.85617
45.	46	-7.02	-5.38832	**127.**	129	-7.94	-5.85398
46.	47	-7.47	-5.44587	**128.**	131	-6.78	-4.34866
47.	48	-7.49	-5.06805	**129.**	132	-6.01	-5.22866
48.	49	-6.75	-4.8772	**130.**	133	-6.7	-5.09722
49.	50	-6.82	-5.84313	**131.**	134	-7.38	-4.73163
50.	51	-6.79	-6.23343	**132.**	135	-6.52	-5.23838
51.	52	-6.83	-6.20537	**133.**	136	-7.39	-4.95122
52.	53	-7.14	-6.33844	**134.**	137	-6.1	-4.91656
53.	54	-7.78	-6.13163	**135.**	138	-7.27	-5.8818
54.	55	-8.05	-6.09802	**136.**	139	-7.74	-6.34442
55.	56	-7.27	-5.8347	**137.**	140	-7.35	-6.20567
56.	57	-7.19	-6.28999	**138.**	141	-7.16	-5.77321
57.	58	-7.5	-6.08706	**139.**	142	-6.68	-5.89063
58.	59	-7.99	-4.93328	**140.**	143	-7.66	-5.18361
59.	61	-5.8	-5.20611	**141.**	144	-7.3	-5.18602
60.	62	-9.24	-5.69289	**142.**	145	-7.74	-5.08316
61.	63	-7.7	-6.35853	**143.**	146	-6.7	-5.47552
62.	64	-4.37	-5.19224	**144.**	147	-9.18	-5.83514
63.	65	-7.01	-5.34029	**145.**	148	-7.9	-5.40602
64.	66	-5.98	-5.11658	**146.**	149	-7.84	-6.14558
65.	67	-8.6	-5.71335	**147.**	150	-7.14	-5.44363
66.	68	-6.3	-4.84856	**148.**	151	-8.83	-6.53644
67.	69	-7.43	-4.44258	**149.**	152	-7.92	-5.94094
68.	70	-8.95	-5.24605	**150.**	153	-7.59	-4.88829
69.	71	-8.79	-5.10466	**151.**	154	-7.64	-5.01902
70.	72	-7.24	-5.18885	**152.**	155	-7.84	-5.06125
71.	73	-7.31	-5.89341	**153.**	156	-6.44	-5.57075
72.	74	-7.46	-5.67278				
73.	75	-5.89	-5.5491				
74.	76	-7.52	-5.35529				
75.	77	-7.51	-5.77947				
76.	78	-6.21	-5.50424				
77.	79	-8.44	-6.28917				
78.	80	-8.81	-6.29347				
79.	81	-8.4	-6.27098				
80.	82	-6.82	-5.30961				
81.	83	-6.24	-4.7524				
82.	84	-7.04	-5.22302				

***Figure 10:** Epipodophyllotoxin analogues docked to DNA Cleavage binding sites of Topoisomerase II.*



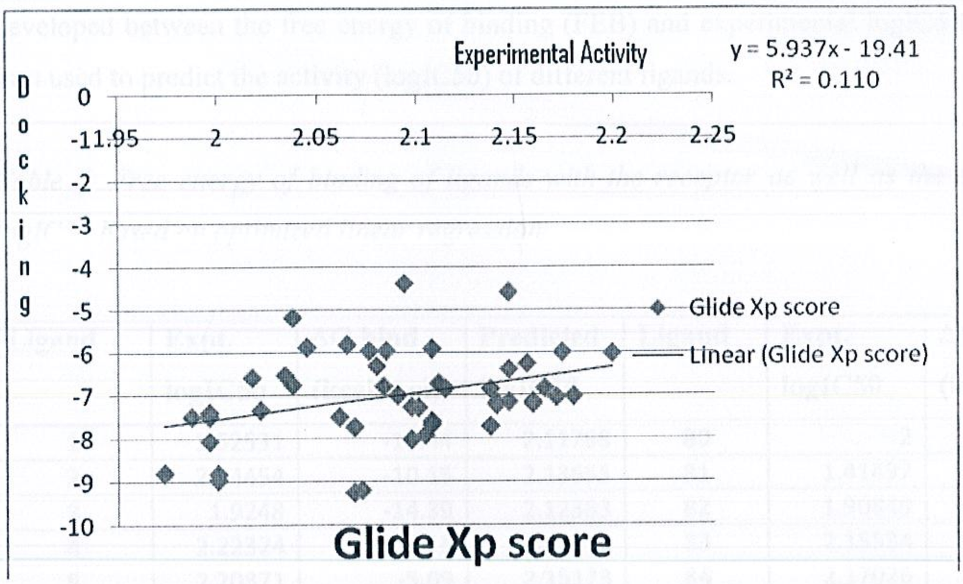


***Figure 11:** Regression plot between docking score and experimental logIC50 for Epipodophyllotoxin.*

Calculated energies and free energy of binding

We have used the automatic approach to calculate the free energy of binding (ΔG) using Prime MM-GBSA (Schrodinger). MM-GBSA is an acronym for a method that combines OPLS molecular mechanics (EMM), an SGB salvation model for polar salvation (GSGB), and a nonpolar solvent accessible surface area and van der Waals interactions. The total free energy of binding is then expressed as:

ΔGbind = Gcomplex – (Gprotein + Gligand)

Where, G = EMM + GSGB + GNP

ΔGbind = Ligand binding energy

Gcomplex = MM-GBSA-Ecomplex energy

Gprotein = MM-GBSA-Eprotein Energy of the receptor without the ligand,

Gligand = MM-GBSA-Eligand Energy of the unbound ligand

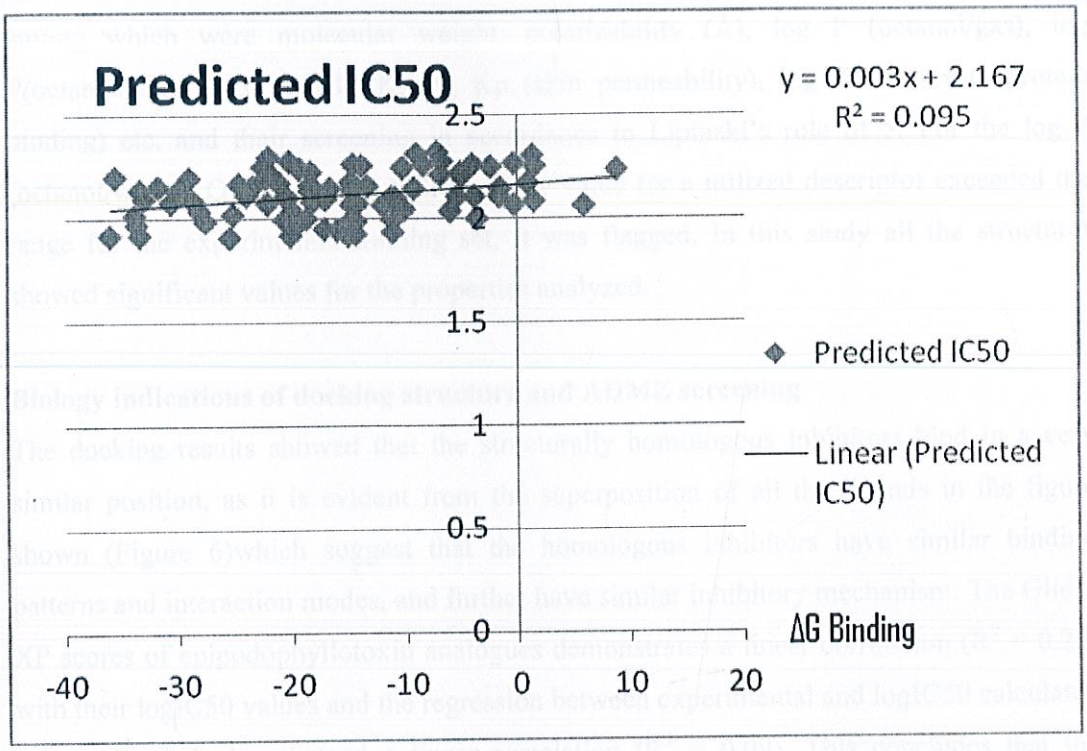
The ligand in the unbound state is minimized in SGB solvent but is not otherwise sampled. In the calculation of the complex, the ligand is minimized in the context of the receptor. The protein is held fixed in all calculations. A free energy relationship was developed between the free energy of binding (FEB) and experimental logIC50 which in turn used to predict the activity (logIC50) of different ligands.

***Table** 7: Free energy of binding of ligands with the receptor as well as the calculated logIC50 based on optimized linear regression.*

Ligand	Expt. logIC50	ΔG bind (kcal/mol)	Predicted logIC50	Ligand	Expt. logIC50	ΔG bind (kcal/mol)	Predicted logIC50
1	1.62531	-16.44	2.11768	**80**	2	-18.75	2.11075
2	2.04454	-10.15	2.13655	**81**	1.41497	-3.34	2.15698
3	1.9248	-14.39	2.12383	**82**	1.90849	-25.45	2.09065
4	2.22324	-5.15	2.15155	**83**	2.15534	-31.73	2.07181
5	2.20871	-5.09	2.15173	**84**	2.17026	-24.65	2.09305
6	2.4624	-13.09	2.12773	**85**	2.09691	-32.97	2.06809
7	2.38561	-25.17	2.09149	**86**	2.03743	-6.08	2.14876
8	2.32428	-7.23	2.14531	**87**	1.86332	-7.13	2.14561
9	0.60206	-20.03	2.10691	**88**	2.31597	-9.06	2.13982
10	2.3962	-23.2	2.0974	**89**	0.78533	-41.7	2.0419
11	2.31597	-28.34	2.08198	**90**	1.19312	-23.87	2.09539
12	1.91908	-35.99	2.05903	**91**	1.34242	-14.67	2.12299
13	2.11059	-20.52	2.10544	**92**	1.04139	-37.99	2.05303
14	1.69897	-18.14	2.11258	**93**	0.60206	1.87	2.17261
15	2.01703	-33.79	2.06563	**94**	1.99564	-11.15	2.13355
17	2.37107	-9.96	2.13712	**95**	2.13988	-33.35	2.06695
18	2.25527	1.42	2.17126	**96**	1.716	2.45	2.17435
19	1.6721	-22.04	2.10088	**97**	1.87506	-22.58	2.09926
20	2.21484	-6.93	2.14621	**98**	2.1038	-11.28	2.13316
21	2.4456	-29.66	2.07802	**99**	2.09691	-20.48	2.10556
22	1.98677	-14.37	2.12389	**100**	2.03342	-19.92	2.10724
23	2.14613	-16.66	2.11702	**101**	1.36173	-13.04	2.12788
24	1.98677	-27.56	2.08432	**102**	0.90309	-3.26	2.15722
25	2.08991	-28.7	2.0809	**103**	0.95424	-13.28	2.12716
26	2.14613	-9.3	2.1391	**104**	1.07918	-16.36	2.11792
27	2.51851	-11.73	2.13181	**105**	0.90309	-33.71	2.06587
28	1.04139	-10.22	2.13634	**106**	2.06819	-0.42	2.16574
29	1.75587	-25.6	2.0902	**107**	2.02119	-25.16	2.09152
30	1.53148	-12.31	2.13007	**108**	1.98227	-19.39	2.10883
31	1	-28.17	2.08249	**109**	1.83885	-24.7	2.0929
32	2.27875	-6.83	2.14651	**110**	2.07555	-22.16	2.10052
33	2.26245	-4.07	2.15479	**111**	1.97313	-33.63	2.06611
34	1.91908	-6.47	2.14759	**112**	2.24304	-19.36	2.10892
35	2.23553	8.73	2.19319	**113**	2.16435	-18.4	2.1118
36	1.88649	-13.7	2.1259	**114**	2.03743	-28.09	2.08273

37	2.14613	-1.65	2.16205	**115**	1.87506	-11.97	2.13109
38	2.3075	1.23	2.17069	**116**	2.30103	-22.06	2.10082
39	2.26245	-8.46	2.14162	**117**	1.61278	-11.52	2.13244
40	2.26951	-0.44	2.16568	**118**	0.8451	-11.12	2.13364
41	2.25285	-2.49	2.15953	**119**	0.004321	-11.09	2.13373
42	1.23045	-5.54	2.15038	**120**	1.5611	-4.71	2.15287
43	2.13988	-14.7	2.1229	**121**	2.08422	1.37	2.17111
44	0.83885	16.81	2.21743	**122**	1.84323	-14.6	2.1232
45	1.91908	-20.12	2.10664	**123**	1.95036	-17.14	2.11558
46	2.17898	-14.82	2.12254	**124**	2.32838	-29.07	2.07979
47	2.32428	-21.13	2.10361	**125**	2.13672	-18.4	2.1118
48	2.0607	5.79	2.18437	**126**	2.36173	-5.8	2.1496
49	1.50515	-18.51	2.11147	**127**	2.5092	7.76	2.19028
50	2.25768	-13.91	2.12527	**128**	1.17609	-21.67	2.10199
51	2.33445	-28.19	2.08243	**129**	1.32222	-19.76	2.10772
52	2.11394	-23.96	2.09512	**131**	2.08279	-18.2	2.1124
53	2.15836	-23.42	2.09674	**132**	2.19866	-28.99	2.08003
54	2.35218	-19.44	2.10868	**133**	1.70757	-28.1	2.0827
55	1.99564	-20.69	2.10493	**134**	1.99564	-14.74	2.12278
56	2.2014	-22.3	2.1001	**135**	1.79239	-33.94	2.06518
57	2.15836	-13.4	2.1268	**136**	2.25285	-22.54	2.09938
58	2.26482	-14.64	2.12308	**137**	1.80618	-2.11	2.16067
59	2.24797	-10.3	2.1361	**138**	2.10037	-10.91	2.13427
61	2.06446	-12.01	2.13097	**139**	2.33445	-7.27	2.14519
62	2.06819	-14.77	2.12269	**140**	2.22789	-18.37	2.11189
63	2.13672	-29.99	2.07703	**141**	2.45332	-23.43	2.09671
64	2.09342	-17.07	2.11579	**142**	2.28103	-20.24	2.10628
65	2.2014	-21.63	2.10211	**143**	2.10721	-5.44	2.15068
66	2.17319	-21.12	2.10364	**144**	1.9345	-11.15	2.13355
67	2.17319	0.8	2.1694	**145**	2.20412	-35.53	2.06041
68	2.07918	-30.6	2.0752	**146**	1.30103	-36.5	2.0575
69	1.97313	-19.68	2.10796	**147**	2.07188	-14.27	2.12419
70	2	-3.68	2.15596	**148**	0.95424	-14.27	2.12419
71	1.97313	-20.03	2.10691	**149**	0.60206	-17.65	2.11405
72	1.17609	-13.75	2.12575	**150**	1.79239	-15.88	2.11936
73	1.91908	-33.35	2.06695	**151**	1.25527	2.47	2.17441
74	1.07918	-23.67	2.09599	**152**	1.51851	-7.7	2.1439
75	2.10721	-1.72	2.16184	**153**	2.10721	-3.91	2.15527
76	0.64345	-22.85	2.09845	**154**	1.88649	-9.86	2.13742
77	0.54407	-21.59	2.10223	**155**	1.91908	-15.03	2.12191
78	1.76343	-30.74	2.07478	**156**	2.16732	-14.68	2.12296
79	1.94448	-15.98	2.11906				

***Figure 12**: Linear regression plot between predicted and experimental logIC50 calculated using (after deleting some outliers) binding for Epipodophyllotoxin.*



ADME Screening

We have analyzed 44 physically significant descriptors of Topoisoomerase II receptor, among which were molecular weight, polarizability (Å), log P (octanol/gas), log P(octanol/water), log p MDCK, log Kp (skin permeability), log Khsa (serum protein binding) etc. and their screening in accordance to Lipinski's rule of 5. For the log P (octanol/water), QP%, and log HERG, if the value for a utilized descriptor exceeded the range for the experimental training set, it was flagged. In this study all the structures showed significant values for the properties analyzed.

Biology indications of docking structure and ADME screening

The docking results showed that the structurally homologous inhibitors bind in a very similar position, as it is evident from the superposition of all the ligands in the figure shown (Figure 6)which suggest that the homologous inhibitors have similar binding patterns and interaction modes, and further have similar inhibitory mechanism. The Glide-XP scores of epipodophyllotoxin analogues demonstrates a linear correlation (R^2 = 0.21) with their logIC50 values and the regression between experimental and logIC50 calculated using ΔG bind also showed a linear correlation (R^2 = 0.09). This concludes that the structural analogues implemented in this study are significantly related to their activity. Also this proved the reasonability and reliability of the docking results. Further ADME screening provided a peered analysis for the final selection of the potential inhibitors from the Topoisomerase II binding site.

Table 8: Screening of ADME properties for Topoisomerase II using Qikprop simulations.

Ligand	QPlogP o/w	QPlog HERG	QPP Caco	QPP MDCK	Lipinski Rule of 5
1	1.839	-3.612	597.061	283.304	0
2	1.789	-4.659	333.598	167.067	0
3	2.543	-4.996	383.599	194.29	0
4	1.761	-4.803	162.681	76.874	0
5	1.696	-5.252	119.411	55.033	0
6	3.088	-4.879	468.276	217.872	0
7	3.813	-5.316	1225.712	616.435	0
8	3.247	-4.984	306.581	137.837	1
9	4.459	-5.558	603.27	286.49	1
10	3.496	-5.599	338.74	153.529	1
11	3.474	-5.276	217.804	95.253	1
12	3.149	-5.037	116.295	48.343	2
13	3.508	-5.199	153.049	65.05	2
14	3.193	-5.624	124.22	51.913	2
15	3.864	-5.156	1195.812	600.198	1
17	3.053	-4.861	409.886	188.662	1
18	4.183	-4.887	1357.337	688.287	1
19	4.214	-5.123	2016.936	1056.054	2
20	3.45	-4.479	1490.344	761.469	1
21	3.814	-4.989	1460.461	744.979	1
22	5.628	-6.515	1737.809	899.008	2
23	4.094	-5.256	1441.501	734.531	1
24	3.454	-5.22	795.736	386.45	0
25	4.068	-5.715	809.954	393.918	1
26	2.911	-5.039	335.933	152.155	0
27	2.945	-5.009	393.351	180.449	0
28	2.365	-5.147	167.392	71.664	1
29	4.084	-4.592	1271.55	1159.87	0
30	3.435	-4.641	690.822	331.683	0
31	3.798	-4.88	515.011	302.907	1
32	2.594	-6.547	85.66	42.517	1
33	2.579	-6.376	98.367	49.374	1
34	3.51	-6.963	126.429	64.761	1
35	3.748	-7.61	84.627	41.964	1
36	1.474	-5.883	78.996	38.954	1
37	1.583	-6.038	96.773	48.51	1
38	1.776	-5.923	71.813	35.14	0

39	1.692	-5.731	68.399	33.338	0
40	2.088	-5.945	75.498	37.093	0
41	2.022	-5.778	75.079	36.87	0
42	2.683	-5.532	74.451	33.026	2
43	2.745	-5.279	95.146	43.052	2
44	3.234	-3.63	35.286	16.94	1
45	5.167	-6.981	102.164	42.026	3
46	3.25	-5.159	740.452	357.513	0
47	2.955	-4.398	363.441	165.665	0
48	4.451	-4.874	1330.578	2192.647	1
49	4.338	-5.124	1415.69	1554.837	1
50	3.217	-4.911	280.291	138.408	0
51	2.924	-6.32	39.427	16.613	2
52	2.785	-6.114	29.597	12.185	2
53	3.019	-5.863	110.142	50.431	2
54	2.929	-6.429	93.686	42.338	1
55	4.484	-6.026	362.394	800.783	1
56	4.139	-6.009	460.287	582.78	1
57	3.794	-5.821	476.772	245.766	1
58	2.937	-6.197	170.361	80.804	1
59	3.512	-5.289	642.73	306.797	1
61	3.87	-5.178	1486.24	1371.064	1
62	3.811	-5.415	1432.943	1320.303	1
63	2.935	-5.462	362.541	165.222	2
64	2.976	-5.346	208.535	90.879	1
65	2.7	-5.241	223.687	98.037	1
66	2.796	-5.61	188.581	81.517	1
67	2.588	-5.43	182.696	78.771	1
68	2.56	-5.391	197.228	85.564	1
69	3.784	-5.173	707.068	340.123	1
70	2.668	-4.947	167.564	71.743	1
71	2.809	-5.072	161.628	69	1
72	3.502	-5.102	517.989	242.976	2
73	3.331	-5.277	194.442	84.259	2
74	4.199	-5.496	107.172	44.257	2
75	5.119	-7.087	987.952	488.27	2
76	3.385	-5.163	251.467	111.261	2
77	3.617	-5.178	232.97	102.441	2
78	3.6	-5.682	574.694	271.85	1
79	4.275	-5.573	414.057	190.738	2
80	3.973	-6.622	97.248	39.845	2
81	4.759	-6.065	65.956	26.188	2
82	1.696	-2.521	458.063	285.121	0
83	2.685	-3.227	518.752	808.734	1
84	3.067	-4.212	662.597	375.55	1
85	3.537	-4.408	399.505	655.586	1
86	2.562	-4.376	383.019	193.972	0

87	2.846	-4.812	413.312	210.607	0
88	3.459	-5.462	333.671	151.048	1
89	0.685	-4.058	435.203	264.47	2
90	1.852	-3.209	630.132	355.109	0
91	3.411	-5.612	112.616	46.692	1
92	3.58	-5.132	334.407	274.319	0
93	2.885	-5.451	170.774	73.23	0
94	2.201	-5.74	16.712	5.938	2
95	2.872	-5.447	60.157	23.708	1
96	3.196	-5.754	160.855	124.337	0
97	3.788	-5.316	192.423	83.314	2
98	3.818	-5.564	198.543	98.817	2
99	3.885	-5.806	268.506	119.431	1
100	5.114	-5.878	1546.68	1434.588	2
101	3.005	-5.457	52.051	20.275	2
102	3.907	-5.656	432.8	200.088	2
103	4.738	-6.279	1284.457	648.43	1
104	3.305	-5.362	861.988	421.342	1
105	4.746	-5.205	3542.925	3514.775	1
106	3.323	-4.878	355.976	292.98	0
107	3.076	-5.213	530	249.072	0
108	3.091	-4.859	514.882	241.402	0
109	3.217	-5.175	366.282	167.065	1
110	3.006	-4.783	550.803	259.655	1
111	2.261	-5.066	55.134	21.576	1
112	2.637	-5.207	124.818	52.184	1
113	2.318	-4.975	113.255	46.978	0
114	2.705	-5.589	90.999	37.084	1
115	2.337	-5.249	43.995	16.906	2
116	2.431	-5.237	72.714	29.1	2
117	-0.305	-2.771	38.992	18.978	1
118	0.592	-3.322	33.939	50.721	2
119	1.192	-5.497	20.8	8.322	2
120	1.565	-4.389	199.692	95.941	0
121	1.444	-4.789	147.695	69.249	0
122	2.738	-4.92	410.372	208.987	0
123	1.631	-4.844	99.227	45.051	0
124	4.026	-5.04	1259.448	1148.727	0
125	3.107	-5.049	278.963	124.466	1
126	3.16	-5.042	200.981	87.326	2
127	3.291	-5.16	219.087	95.86	2
128	2.48	-4.844	53.295	20.799	2
129	5.866	-5.253	2275.339	10000	2
131	3.783	-4.292	1494.803	950.537	0
132	4.064	-5.111	1231.09	1120.933	0
133	4.313	-5.227	1241.025	1542.388	1
134	4.35	-4.933	1269.224	1581.369	1

135	4.345	-5.207	1248.902	1613.689	1
136	4.355	-5.11	1573.918	2067.133	1
137	4.36	-5.131	1314.005	1718.588	1
138	3.488	-5.732	296.961	238.068	1
139	3.856	-6.052	338.762	306.996	1
140	3.816	-5.955	402.67	370.725	1
141	2.802	-6.236	89.691	40.39	1
142	2.631	-5.769	103.311	47.059	1
143	3.881	-5.243	851.013	653.221	1
144	2.783	-5.366	144.857	61.295	2
145	2.885	-5.396	83.371	33.736	2
146	2.413	-5.084	21.141	7.656	2
147	3.738	-4.249	537.405	573.366	1
148	3.536	-5.385	319.46	144.106	0
149	3.149	-6.113	98.836	40.548	0
150	2.301	-5.938	37.923	14.398	0
151	2.337	-5.61	55.735	21.83	0
152	2.788	-5.316	172.423	43.314	0
153	2.887	-4.981	456.355	211.883	0
154	3.591	-5.292	363.206	408.642	0
155	2.348	-4.856	148.542	62.982	0
156	2.467	-5.04	146.923	62.241	1

CHAPTER 4
Pharmacophore Based QSAR

A Pharmacophore was first defined by Paul Ehrlich in 1909 as "a molecular framework that carries (Phoros) the essential features responsible for a drug's (=pharmacon's) biological activity" .In 1977,this definition was updated by Peter Gund to "a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity. The IUPAC definition of a pharmacophore is "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response".

In modern computational chemistry, pharmacophores are used to define the essential features of one or more molecules with the same biological activity. A database of diverse chemical compounds can then be searched for more molecules which share the same features local a similar distance apart from each other.

Typical pharmacophore features are for where a molecule is hydrophobic, aromatic, a hydrogen bond acceptor, a hydrogen bond donor, a cation, or an anion. The features need to match different chemical groups with similar properties, in order to identify novel ligands. Ligands receptor interactions are typically "polar positive", "polar negative", or "hydrophobic". A well-defined pharmacophore model includes both hydrophobic volumes and hydrogen bond vectors.

A pharmacophore is a specific, three dimensional maps of biological properties common to all active conformations of a set of ligands which exhibit a particular activity. Conceptually, a pharmacophore is a distillation of the functional attributes of ligands which accomplish a specific task.

Phamacophores are conceptual templates for drug design. Once it is extracted from a set of ligands, a pharmacophore can be used as a model for the design of other molecules that can accomplish the same activity. The problem of pharmacophore identification is to generate the pharmacophore from structural data describing ligands and their interaction with the receptor. As many protein structures are described as sets of points, pharmacophore identification is commonly reduced directly to the problem of finding points common to all functional ligand conformations. This is a geometric problem called the Largest Common Pointest problem. However, since precisely congruent points are never actually part of the data, it is more accurate to classify this as the Largest Approximate Common Pointest problem. We begin by considering two manifestations of this problem, where only a few ligand conformations are considered, and very exacting methods are applied to determing the pharmacophore. We then describe pharmacophore identification when considering numerous ligand conformations.

Pharmacophore model is now recognized as integral components of lead discovery and lead optimization, and the continuing need for improved pharmacophore based tools has driven the development of PHASE by employing a novel, tree-based partitioning algorithm, PHASE exhaustively identifies spatial arrangements of functional groups that are common and essential to the biologic activity of a set of high affinity ligands. These pharmacophore hypotheses are validated in a number of ways, including their ability to:

1) rationalize the binding affinities of a training set of molecules of varying activity,

2) Successfully predict the affinities of a test set of molecules, and

3) Selectively retrieve known activities from a database of drug-like molecules. In addition, PHASE uniquely offers the ability to distinguish multiple binding modes through a bi-directional clustering approach applied to bit string representations of the ligand/hypothesis space.

Phase is a more recently developed pharmacophore modeling package. It follows a hypothesis generation step, with a grid-based 3D QSAR method, in which the grid positions of atoms in molecules overlaid to the hypotheses are correlated to their activities using a Partial-least-squares (PLS) fitting approach.

Computer-aided molecular design is frequently split into disciplines that focus on either structure-based or ligand-based techniques. When sufficient information is available or inferable about the structure of the biological target and its binding site, then it is possible to invoke a structure based approach, wherein specific ligand-receptor interactions are studied to help identify new molecules with activity towards the target. If, however, knowledge about the structure of the target is limited, but a sufficient number of actives have already been identified, then ligand-based methods provide alternative ways of leveraging the available information into models that can help identify new actives.

Figure.13: Summarizes the major tasks and workflows supported by PHASE.

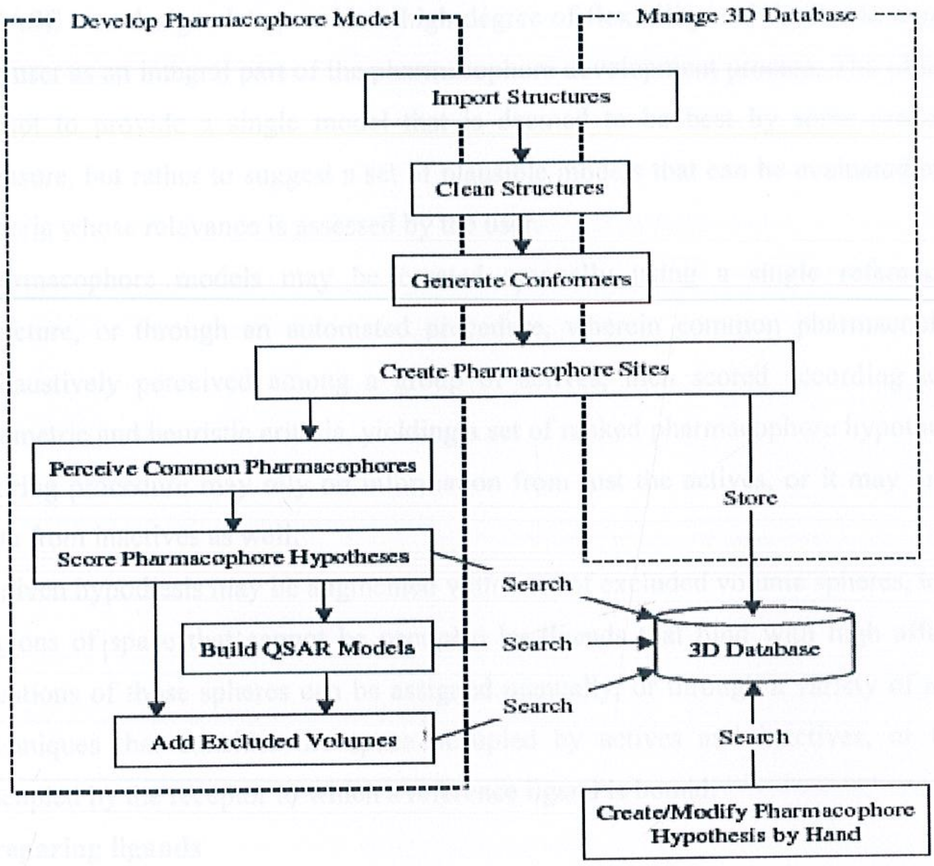

Develop Pharmacophore Model
Manage 3D Database
Import Structures
Clean Structures
Generate Conformers
Create Pharmacophore Sites
Perceive Common Pharmacophores
Score Pharmacophore Hypotheses
Build QSAR Models
Add Excluded Volumes
Store
Search
Search
Search
Search
3D Database
Create/Modify Pharmacophore Hypothesis by Hand

Fig. 1 PHASE Workflows

PHASE Methodology

We used Epipodophyllotoxin analogues and build the model. We only included the compounds/ligands whose activity values were known. The model for Epipodophyllotoxin was build using logIC50 values. Pharmacophore model was build using **Phase.**

PHASE was designed to provide a high degree of flexibility and feedback, emphasizing the user as an integral part of the pharmacophore development process. The ultimate goal is not to provide a single model that is deemed to be best by some predetermined measure, but rather to suggest a set of plausible models that can be evaluated by diverse criteria whose relevance is assessed by the user.

Pharmacophore models may be created manually using a single reference ligand structure, or through an automated procedure, wherein common pharmacophores are exhaustively perceived among a group of actives, then scored according to various geometric and heuristic criteria, yielding a set of ranked pharmacophore hypotheses. This scoring procedure may rely on information from just the actives, or it may incorporate data from inactives as well.

A given hypothesis may be augmented with a set of excluded volume spheres, to map out regions of space that cannot be occupied by ligands that bind with high affinity. The locations of those spheres can be assigned manually, or through a variety of automated techniques that consider the space occupied by actives and inactives, or the space occupied by the receptor to which a reference ligand is bound.

Preparing ligands

Before undertaking the tasks of pharmacophore model development and 3D database creation, low-energy, 3D structures must be available for each molecule of interest. Accordingly, PHASE incorporates a structure cleaning step utilizing LigPrep, which attaches hydrogens, converts 2D structures to 3D, generates sterioisomers, and, optionally, neutralizes charged structures or determines the most probable ionization state at a user defined pH. PHASE also allows for the importation of 3D structures prepared outside its own workflow. Because one does not generally know the structure that a given molecule will adopt if and when it binds to a target protein, it is customary to represent each molecule as a series of 3D structures that sample the thermally accessible conformational states. For purposes of pharmacophore model development, PHASE provides two built-in approaches, both of which employ the macromodel conformational search engine. The first approach involves a rapid torsion angle search followed by minimization of each generated structure using either the MMFF's or OPLS_2005 forcefield, with implicit GB/SA or distance dependent dielectric solvent model. The torsion search samples ring conformations, invertible pseudo-chiro nitrogens, and all rotatable bonds within a core region, which includes everything from the center of a molecule out to, but not including, the last rotatable bond along each path.

Structures with high energies are eliminated, as are structures with close non bonded contacts. The minimized structures that are ultimately obtained are filtered through a user-defined relative energy window, typically 5-10cal/mol, and a redundancy check, where any two structures within 1Kcal of each other are deemed to be equivalent if all corresponding pairs of heavy atoms in the two structures are within a user-defined distance, typically 1-2Å. By varying these parameters in conjuction with the maximum number of conformers initially sampled, any desired level of conformational coverage may be achieved.

Creating pharmacophore sites

For purposes of pharmacophore model development, each ligand structure is represented by a set of points in 3D space, which coincide with various chemical features that may facilitate non-covalent binding between the ligand and its target receptor. These pharmacophore sites are characterized by type, location, and if applicable, directionality. In accordance with the most cited explanations of ligand-receptor binding, PHASE provides six built-in types of pharmacophore features: hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobe (H), negative ionizable (N), positive ionizable (P), and aromatic ring(R). In addition, users may define up to three custom feature types(X, Y, Z) to account for characteristics that don't fit clearly into any of the six built-in categories.

Rings, isopropyl groups,t-butyl groups, various halogenated moieties , and chains as long as four carbons are each treated as a single hydrophobic site. Chains of five or more carbons are broken into smaller fragments containing between two and four carbons and each fragment is designated as separate hydrophobic site. The location rH of a given hydrophobic site is a weighted average of the positions of the non-hydrogen atoms in the associated fragment:

$$r_{H}=\frac{\sum_i s_i t_i r_i}{\sum_i s_i t_i}$$

Here s_i is the solvent-accessible surface area of atom I, computed using a probe radius of 1.4Å, and t_i is a hydrophobicity factor that ranges from 0 and 1. Polar atoms(O,N,S)are assigned a hydrophobicity of 0,whereas halogens and carbons at least three bonds from any polar atom receive a value of 1.negative and positive ionizable sites are modeled as a single point located on a formally charged atom, or at centroid of a group of atoms over which the ionic charge is shared. Finally, if the user so chooses, aromatic rings may be distinguished from other hydrophobic groups, and designated as a separate type of pharmacophore feature (i.e "R" rather than "H"). In that case, a single site is placed at the centroid of each aromatic ring, and a two headed vector normal to the plane of the ring is associated with the site. Unlike acceptors and donors, an aromatic feature cannot be represented as a pure projected point.

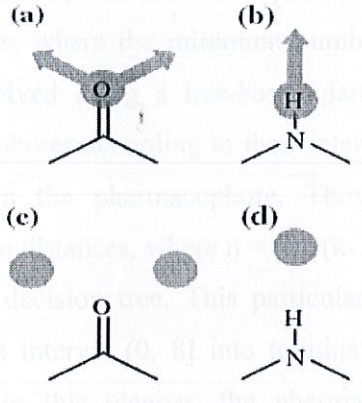


Fig. 2 Hydrogen bond acceptor and donor mappings based on the use of vector features (a, b) and pure projected points (c, d)

Figure 14: *Hydrogen bond acceptor and donor*

Figure 15: Hydrophobic features.

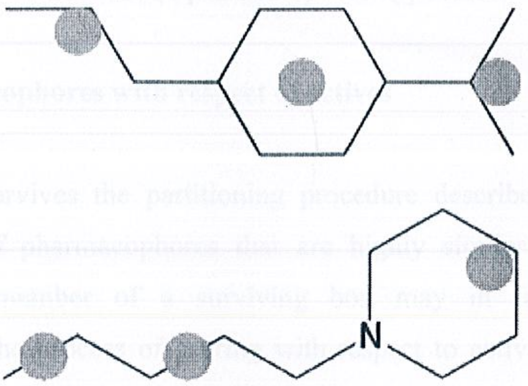


Fig. 3 Hydrophobic feature mappings

By default, PHASE will look for pharmacophores that are common to all actives, but this condition can be relaxed so that a common pharmacophore need match only a subset of the actives. The ligands that are matched may vary from one common pharmacophore to another, so the user does not have to choose any particular subset of ligands, just the number. Henceforth, we shall say that a common pharmacophore must match a minimum required number of actives, where the minimum number is set by the user. Common pharmacophores are perceived using a tree-based partitioning technique that groups together similar pharmacophores according to their intersite distances, i.e., the distances between pairs of sites in the pharmacophore. Thus a k-point pharmacophore is represented by a vector of n distances, where n = (k_ (k-1))/2. Each intersite distance d is filtered through a binary decision tree. This particular tree has a depth of four and partitions distances on the interval (0, 8] into terminal nodes that are 1 Å wide. By filtering all n distances in this manner, the pharmacophore is assigned to an n-dimensional box, whose sides are equal in length to the terminal node width, which we denote as e. all pharmacophores of a given variant (i.e., a particular combination of the feature types A, D, H, etc.) that are mapped into the same terminal box are considered to be similar enough to facilitate identifications of a common pharmacophore. So if each of a minimum required number of actives contributes at least one pharmacophore to a particular box, then that box represents a common intersite distance pharmacophore, each member of the box is a candidate pharmacophore hypothesis.

Scoring pharmacophores with respect to actives

Each box that survives the partitioning procedure described in the previous section contains a set of pharmacophores that are highly similar in the space of intersite distances. Any member of a surviving box may in fact constitute a common pharmacophore; the process of scoring with respect to actives is designed to filter out these sorts of inappropriate pharmacophores and to identify within each box a top ranked representative, henceforth referred to as the pharmacophore hypothesis for that box. Hypotheses are assigned a score comprised of geometric and heuristic factors that can be weighted according to the user's preference. At this point only information from the actives is used; a subsequent procedure may be invoked to consider information from inactives and adjust hypothesis scores accordingly. Each pharmacophore from a surviving box is treated temporarily as a reference in order to assign a score. Accordingly, all other non-reference pharmacophores from that box and its neighboring boxes are aligned, one-by-one, to the reference pharmacophore, and the quality the alignments are measured using two criteria: 1) the root-mean-squared deviation in the site point positions and 2) the average cosine of the angles formed by corresponding pairs of vector features (acceptors, donors and aromatic rings). These factors are combined with separate weights to yield a combined site + vector score for each non-reference pharmacophore I that's been aligned to the reference:

$$\text{Site\_Vector\_Score}_i = w_{site}\text{Site\_Score}_i + w_{vector}\text{Vector\_Score}_{i},$$

Where

$$\text{Site\_Score} = 1 - \text{RMSD}_i / \text{cutoff}_{RMSD}$$

$$\text{Vector\_Score} = \frac{1}{n_v}\sum_{j=1}^{n_v} \cos\theta_{ij}$$

The parameters w_{site}, w_{vector}, and cutoff_{RMSD}, are user-adjustable (with default values of 1.0, 1.0 and 1.2, respectively), n_v is the number of vector features in the hypothesis and h_{ij} is the angle between the jth vector feature in the non-reference pharmacophore and the corresponding vector feature in the reference pharmacophore. PHASE supports the use of a conformationally independent property, such as –logKi, to bias the selection of reference ligands to favor those with higher activities, or higher values of whichever property is incorporated. The weighted property term is combined with Site_Vector_Score to yield an overall Reference Score:

Reference Score = ¼ Site vector score + wpropPropref

Here, Propref is the value of the conformationally independent property for the ligand contributing the reference pharmacophore. After all the pharmacophores in a box have been treated as a reference, the one yielding the highest Reference Score is selected as the hypothesis to represent that box. The ligand that contributes the reference pharmacophore is referred to as the reference ligand for that hypothesis. Further refinement may then be done using volume scoring, selectivity scoring, reference ligand relative conformational energy, and the number of actives matched. Volume scoring measures how well each non-reference ligand overlays with the reference ligand, based on the van der Waals models of the structures and taking into account all heavy atoms:

$$\text{Volume\_Score}_i = \text{Volume}_{i,\ \text{common}}/\text{Volume}_{i,\ \text{total}}$$

$\text{Volume}_{i,\ \text{common}}$ is the common or overlapping volume between ligand I and the reference ligand, and $\text{Volume}_{i,\ \text{total}}$ is the total volume occupied by both ligands. The overall Volume Score for a hypothesis is the average obtained from applying the above formula to all non-reference ligands i. Volume Score is then added to Reference Score with its own adjustable weight (1.0 by default). The total active score for a given hypothesis is then:

$$\begin{aligned}\text{Active\_Score} = \text{Reference\_Score} &+ w_{\text{volume}}\ \text{Volume\_Score} \\ &+ \text{wselectivity Selectivity\_Score} - \text{wconf EConfref} \\ &+ w_{match}^{M-1}\end{aligned}$$

Scoring pharmacophores with respect to inactives

PHASE provides a means for penalizing hypothesis that fail to discriminate actives from the inactives, thus effectively elevating pharmacophore models composed only of features that are essential for high-affinity binding. A k-point hypothesis is scored with respect to inactives by searching the pharmacophore space of those inactives and finding all m-point matches to the hypothesis, where 3<= m <= k. The best match I provided by each inactive is determined by way of a fitness score:

Fitness I = ¼ w Site Score I + w Vector Score I + w Volume Score i

$$RMSD_i = [\frac{m}{k} RMSD_{i,m}^2 + \frac{k-m}{k} cutoff_{RMSD}^2]^{1/2}$$

For a valid hypothesis, all inactives should ideally exhibit relatively low fitness, so the overall score is reduced by the average fitness observed across a set of N inactives, multiplied by a user-adjustable weight winactive (1.0 by default)

$$Adjusted - Score = Active - score - w_{inactive} \frac{1}{N} \sum_{i=1}^{N} Fitness_i$$

Target pharmacophores

The protocol for this investigation was first described by Patel et al., who compared the programs Catalyst/ HipHop, Disco, and GASP for their ability to generate pharmacophores in accordance with ligand features that were overlaid in X-ray complexes from the PDB. For each of five proteins, a number of complexes were visually inspected to identify a target pharmacophore, which was used as a standard to establish the accuracy of all programs. More specifically, the ligands for a given protein were run through the automated pharmacophore perception workflow of each program, and the pharmacophore models produced that contained the correct features were aligned to the target pharmacophore to obtain an RMSD in the corresponding site point positions. For a given program, the pharmacophore with the lowest RMSD was judged to be the best, irrespective of how it was ranked by the program's own scoring function. Clean structures and generate conformations steps of PHASE were not performed as the ligands used were docked structures. Common pharmacophore models containing pharmacophore models were generated using default settings and relaxing the requirements, as necessary, that all actives match the pharmacophore. Scoring with respect to actives was also done using default parameter values, with incorporation of a binary property to promote the selection of reference ligands in accordance with the ligand used to define each target pharmacophore. Relative conformational energy was not incorporated into the scoring process, nor was there any scoring with respect to inactives. Each hypothesis that emerged from the flexible analysis and which contained the correct features was aligned to its associated target pharmacophore using a standard least-squares technique, and an RMSD in the matching site point positions was computed. When the pharmacophore contained more than one occurrence of a particular feature type, each possible mapping to the target pharmacophore was considered. After processing all hypotheses in this manner, the one yielding the lowest RMSD was selected, and its overall ranking according to the PHASE scoring function was recorded.

Build QSAR

The best scored hypothesis was then used to build QSAR model. For this, the whole data set was divided into training and test set randomly. Then the pharmacophore based QSAR model was build for the test set using the build QSAR module of PHASE. Here, the model is generated using PLS method and PLS factor 3 was used. The model thus obtained was further improved by removing some outliers to significant correlation between experimental and predicted activity.

Table 9: *Reasonably good correspondence in the structures. Root-mean-squared deviations (Å) from the target pharmacophore for some scoring hypothesis.*

Hypothesis No.	**Hypothesis Name**	**R2**	**RMSE**
1	AADR.370	0.4356	0.4236
2	AADR.351	0.4375	0.4093
3	AADR.350	0.4383	0.478
4	AADR.4294	0.4965	0.4738
5	**ADPR.2**	**0.574**	**0.453**
6	HHRR.207	0.4785	0.5117
7	AAHH.193	0.342	0.485
8	AAHH.23	0.434	0.502
9	DDRR.7	0.239	0.418
10	ADHR.1023	0.45	0.497

Table 10: *The experimental and predicted activities (predicted by PHASE) and fitness score of training set.*

Ligand	Expt. activity	Predicted activity	Fitness score	Ligand	Expt. Activity	Predicted activity	Fitness score
119	0.05	0.53	1.06	**114**	2.04	1.81	1.65
77	0.54	1.31	1.63	**48**	2.06	2	1.55
149	0.6	1.8	1.56	**62**	2.07	2.06	1.68
76	0.64	1.93	2.11	**106**	2.07	1.99	2.01
89	0.79	0.162	1.73	**68**	2.08	1.99	1.69
118	0.85	0.68	1.15	**110**	2.08	1.86	1.57
102	0.9	0.87	1.26	**121**	2.08	1.72	1.77
103	0.95	1.14	1.45	**64**	2.09	2	2.05
92	1.04	1.33	1.26	**98**	2.1	1.99	2.03
28	1.04	1.31	1.48	**138**	2.1	2.1	2.12
104	1.08	0.91	1.22	**143**	2.11	2	1.98
128	1.18	1.8	1.71	**13**	2.11	1.78	1.57
90	1.19	1.84	1.74	**75**	2.11	2.06	1.57
151	1.26	1.65	1.77	**43**	2.14	2.11	2.5
146	1.3	1.88	1.73	**95**	2.14	2	1.47
91	1.34	1.8	1.65	**37**	2.15	2.19	1.96
101	1.36	1.38	1.23	**26**	2.15	1.86	1.74
152	1.52	1.8	1.64	**57**	2.16	2.14	1.59
30	1.53	1.92	1.8	**53**	2.16	2.18	1.58
117	1.61	1.73	1.16	**83**	2.16	1.94	1.56
1	1.63	1.88	1.77	**67**	2.17	1.78	1.68
19	1.67	1.93	2.22	**66**	2.17	1.92	1.77
133	1.71	1.88	1.78	**46**	2.18	2.2	1.62
29	1.76	1.7	1.62	**56**	2.2	2.18	3
150	1.79	1.88	1.77	**65**	2.2	1.94	1.56
137	1.81	1.8	1.76	**20**	2.21	2.03	1.62
122	1.84	1.9	2.64	**4**	2.22	2.21	2.27
87	1.86	2.19	2.25	**35**	2.24	2.11	1.94
97	1.88	1.88	1.63	**112**	2.24	1.86	1.63
36	1.89	1.89	1.58	**136**	2.25	1.99	2.08
82	1.91	1.86	1.57	**41**	2.25	2.18	2.27

155	1.92	1.94	1.51	**33**	2.26	2.05	2.48
73	1.92	1.78	1.54	**50**	2.26	2.26	2.02
12	1.92	1.92	1.73	**39**	2.26	2.2	1.57
3	1.92	2.13	2.21	**116**	2.3	1.86	2.01
79	1.94	1.96	1.49	**38**	2.31	2.23	1.22
71	1.97	1.94	1.63	**47**	2.32	1.94	1.64
69	1.97	1.86	1.66	**139**	2.33	2.18	2.24
22	1.99	1.96	1.82	**51**	2.33	2.25	2.52
24	1.99	1.94	1.76	**54**	2.35	2.28	2.22
70	2	1.99	1.62	**126**	2.36	1.99	2.24
55	2	2.11	1.58	**7**	2.39	1.98	1.74
94	2	1.98	1.7	**2**	2.4	1.82	1.73
15	2.02	1.87	1.69	**21**	2.45	1.99	2.1
100	2.03	2.02	1.66	**141**	2.45	2.16	2.62

***Figure 16**: The regression plot between the experimental activity and the activity predicted by PHASE*

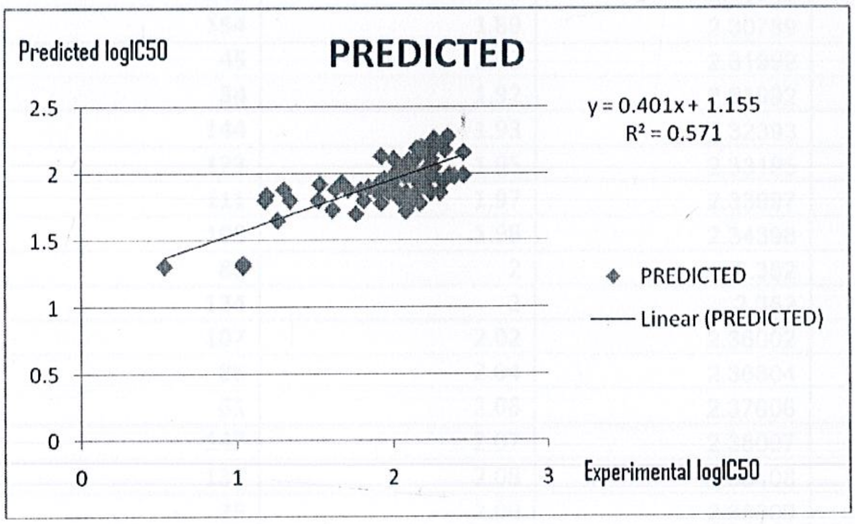


***Table 11**: The experimental and calculated activities (using model for training set) of test set.*

Ligand	Expt. Activity	Calculated Activity	Error
93	0.6	1.7906	1.1906
9	0.6	1.7906	1.28684
44	0.84	1.88684	1.0709
105	0.9	1.9109	1.03095
148	0.95	1.93095	1.001
31	1	1.951	0.98308
74	1.08	1.98308	0.94318
72	1.18	2.02318	0.86323
42	1.23	2.04323	0.84932
129	1.32	2.07932	0.79541
81	1.41	2.11541	0.74551
49	1.51	2.15551	0.66556
120	1.56	2.17556	0.6717
14	1.7	2.2317	0.53972
96	1.72	2.23972	0.53576
78	1.76	2.25576	0.50779
135	1.79	2.26779	0.49784
109	1.84	2.28784	0.46388
115	1.88	2.30388	0.42789
154	1.89	2.30789	0.42992
45	1.92	2.31992	0.39992
34	1.92	2.31992	0.40393
144	1.93	2.32393	0.40195
123	1.95	2.33195	0.38997
111	1.97	2.33997	0.37398
108	1.98	2.34398	0.372
80	2	2.352	0.352
134	2	2.352	0.36002
107	2.02	2.36002	0.34804
86	2.04	2.36804	0.33606
61	2.06	2.37606	0.32007
147	2.07	2.38007	0.31408
131	2.08	2.38408	0.30809
25	2.09	2.38809	0.3021
85	2.1	2.3921	0.2921
99	2.1	2.3921	0.29611
153	2.11	2.39611	0.28611
52	2.11	2.39611	0.29814
63	2.14	2.40814	0.26814
125	2.14	2.40814	0.27215

23	2.15	2.41215	0.26616
113	2.16	2.41616	0.26017
84	2.17	2.42017	0.25017
156	2.17	2.42017	0.2622
145	2.2	2.4322	0.2322
132	2.2	2.4322	0.23621
5	2.21	2.43621	0.23423
140	2.23	2.44423	0.22225
59	2.25	2.45225	0.20626
18	2.26	2.45626	0.19626
58	2.26	2.45626	0.20428
142	2.28	2.46428	0.20032
88	2.32	2.48032	0.16032
8	2.32	2.48032	0.16032
11	2.32	2.48032	0.16433
124	2.33	2.48433	0.17037
17	2.37	2.50037	0.1424
10	2.4	2.5124	0.13646
6	2.46	2.53646	0.07
27	2.52	1.7906	1.1906

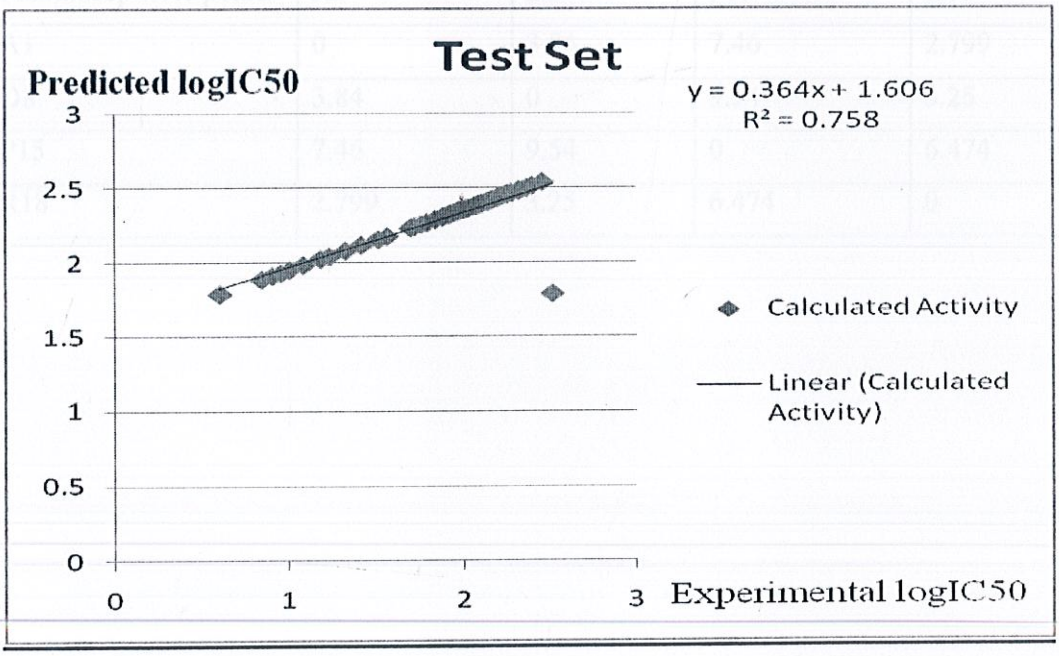


***Figure** 17: The regression plot between the experimental activity and the activity predicted by PHASE for Test data set*

The regression between experimental and predicted activity showed a linear correlation with significant R2 (0.571) which proves the reliability and accuracy of the results and also the results for test set also showed less deviation, therefore this model can be used for the prediction of biological activity of Epipodophyllotoxin compounds with unknown biological activity.

The common pharmacophore for the ligands contain the following four pharmacophoric points: one hydrogen bond acceptors: A5; one hydrogen donor D8; one positive ionizable P15 and one planar group: R7 (shown as a brown ring) in figure 14.

***Table 12:** The distance between the pharmacophoric points.*

Pharmacophoric points	A1	D8	P15	R18
A1	0	3.84	7.46	2.799
D8	3.84	0	9.54	3.25
P15	7.46	9.54	0	6.474
R18	2.799	3.25	6.474	0

***Figure18**: Top ranked four pharmacophore features and the distance between the pharmacophoric groups for training set of epipodophyllotoxin*

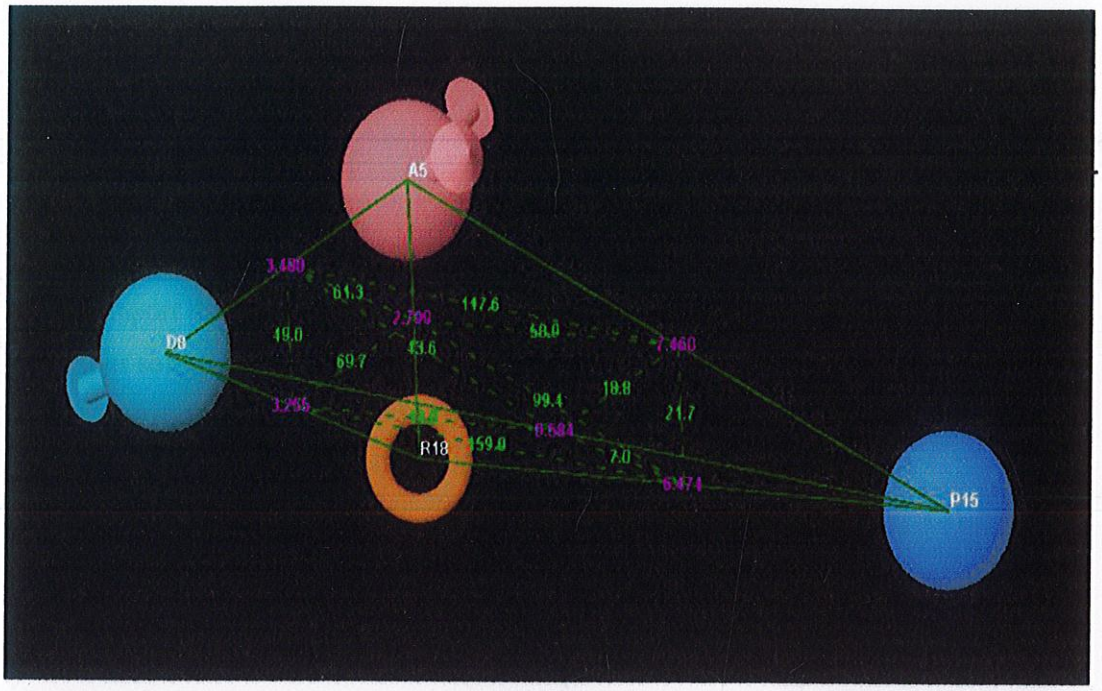
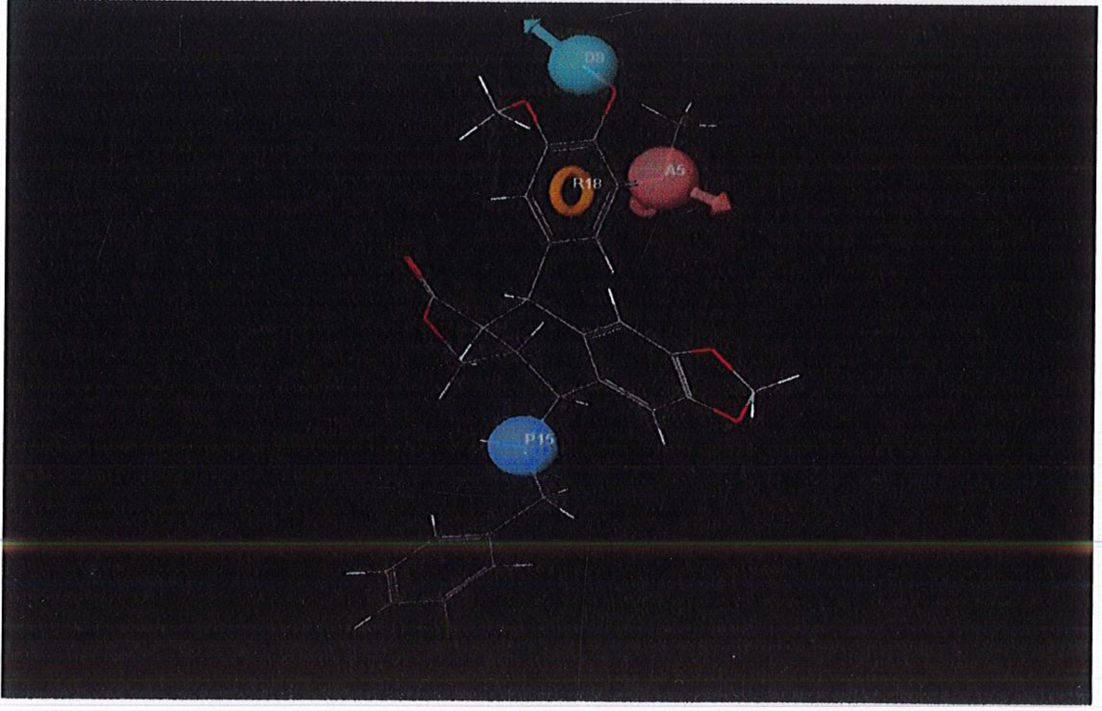


***Figure 19**: Top ranked four pharmacophore features displayed for training set of epipodophyllotoxin*



CHAPTER 5
LIGAND BASED QSAR

Procedure Approached

QSAR based on only Ligands, Receptor not considered

1) Import the mol files and sdf files of all ligands with known activity values in various software for calculation of descriptors.
2) Convert the data obtained into CSV files.
3) Apply different test to filter out the required descriptors :
   I. Zero test
   II. Correlation test
   III. Missing value test
4) Generate MLR and PLS model for the prediction of activity.
5) Develop QSAR model with appropriate set of descriptors.

Descriptor calculation

Descriptors like log P, structural, symmetry, topological, physiochemical electronic Wang-Ford atomic charge and extended Huckel partial charge, moments, orbital energies, molecular connectivity indexes, hydrophobicity, steric and thermodynamic factors and topological descriptors were calculated using PREADMET, CHEM OFFICE, MOPAC and ADME Model Builder software package. The Superpendentic index is computed from the pendent matrix. These descriptors help to differentiate the molecules mostly according to their size, degree of branching, flexibility and overall shape. Some of the descriptors included in the study are listed and described in Table 13.

Table 13: List of descriptors used in the study.

Electronic	Partial positive surface area (AM1), partial negative surface area (AM1), relative positive charge (AM1), relative negative charge (AM1), relative positive charged surface area (AM1), relative negative charged surface area (AM1), weighted positive charged partial SA (AM1), weighted negative charged partial SA (AM1), fractional negative charged partial SA (AM1), heat of formation, dipole moments, energy of the highest occupied orbital, energy of the lowest unoccupied orbital, electronegativity, hardness, mean partial charge on H atoms, most negative partial charge on H atom, most positive partial charge on H atom, most negative partial charge on C atom, mean partial charge on C atoms, most positive partial charge on O atom, mean partial charge on O atoms, most negative partial charge on heteroatom, mean partial charge on heteroatoms, most positive partial on heteroatom
Information content	Information of atomic composition index, superpendentivity index, superpendentivity index Carbon only
Structural	Topological symmetry, geometrical symmetry, combined symmetry, conformational flexibility indices, molecular distance edge descriptors, moment of inertia indices, geometric moment indices, number of single bonds, number of aromatic bonds
Topological	Wiener index, Kier and Hall molecular connectivity indices, path count and length descriptors, topological polar surface area (TPSA), Balban indices.
Constitutional	Atom count, Bond count, Ring count, Functional group count, Chemical feature count
Geometrical & Physicochemical	2D van der Waals chemical features surface area, H-bond donor surface area, Hydrophobic surface area, Polar surface area, Fraction of 2D-VSA polar, AlogP98 atomic types, Polarizability

Regression analysis

The total number of descriptors calculated was about 1300 including electrostatic, physiochemical, constitutional & others descriptors calculated. A systematic search was performed to determine significant descriptors. Some of the descriptors were rejected because they contain a value of zero for all the compounds. In order to minimize the effect of collinearity and to avoid redundancy correlation matrix developed with a cutoff value of 0.6 and the variables physically removed from the analysis which shows exact linear dependencies between subsets of the variables and multicollinearity (high multiple correlations between subsets of the variables). From descriptors thus remained, the set of descriptors that would give the statistically best QSAR models were selected from the large pool using a Genetic function approach. The genetic algorithm starts with the creation of a population of randomly generated parameter sets. The usage probability of a given parameter from active set is 0.5 in any of the initial population sets. The sets are then compared according to objective function. The form of objective function favors sets that have the R^2 as high as possible, while minimizing the number of parameters used as descriptors. The higher the score, higher is the probability that a given set will be used for the creation of the next generation of sets. Creation of a consecutive generation involves crossovers between set contents, as well as mutations. The parameters set used for genetic algorithm includes: mutation 0.1, crossover 0.9, population 300, number of generations 1000, R^2 floor limit 50% and objective function was R^2/N_par. The algorithm runs until the desired number of generations is reached. Equations were developed between the observed activity and the descriptors. The best model was selected based on the r^2, r^2_{adj}, F-ratio and q^2. r^2 is an indication of the model data fit.

Validation test

The predictive capability of the equation (q^2) is determined using leave-one-out cross validation method. The relation for q^2 is as shown below.

$$q^2 = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n}(y_{exp} - y_{pred})^2}{\sum_{i=1}^{n}(y_{exp} - \bar{y})^2}$$

Where y_{pred}, $y_{\exp}$ and y are the predicted, experimental and mean values of activity, respectively. A large F indicates that the model fit is not a chance occurrence. R^2 and R^2_{adj} above a value of 0.6 indicate good model fit while q^2 above 0.55 indicates good predictive capability for the model. Further statistical significance of the relationship between the cytotoxic activity and chemical structure descriptors was obtained by randomization process. The test set 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-randimized activity values. If the original QSAR model is statistically significant, its score should be significantly better than that from permuted data.

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 descriptor can be hidden by correlation of descriptors.

It has been shown that a high value of statistical characteristics need not be the proof of a highly predictive model. Hence, in order to evaluate the predictive ability of our QSAR model, we used the method described by Golbraikh et al. and Roy et al. The values correlation coefficient of predicted and actual activities and correlation coefficient for regressions through the origin (predicted vs. observed activities and vice versa) were calculated using the regression of analysis ToolpakA option of excel sheet and other parameters were calculated as reported by the above authors. To arrive at the predictive R^2 (R^2_{pred}) the following equation was used. [48].

$$r^2_{pred} = 1 - \frac{\sum_n (y_{pred_{test}} - y_{test})^2}{\sum_n (y_{test} - \bar{y}_{training})^2}$$

Where $Y_{pred_{Test}}$ and Y_{Test} are the predicted and observed activity values, respectively, of the Test set compounds and $\bar{Y}_{Training}$ is the mean activity value of the Training set. Further evaluation of the predictive ability of the model was done by determining the value of rm^2 by the following equation:

$$rm^2 = R^2(1 - \left|\sqrt{R^2 - R_o^2}\right|)$$

Where *R^2 is the square correlation coefficient between observed and predicted values and* R_0^2 is the squared correlation coefficient between observed and predicted values without intercept. The values 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:

$$k = \frac{\sum y_i y_i'}{\sum y_i'^2} \; and \; k' = \frac{\sum y_i y_i'}{\sum y_i^2}$$

Where y_i'and *yi* are the predicted and actual activities, respectively.

Results and Discussion

The 154 active compounds considered as potential Topoisomerase II inhibitors were segregated into 124 training and 30 test sets. The experimental IC_{50} values for these compounds set are available. With the wide range of difference between the IC_{50} values and the large diversity in the structures, the combined data set of 124 molecules and 30 molecules are ideal to be considered as training and test set, as both the sets does not suffer from bias, due to the similarity of the structures. The various molecular descriptors as described in Table 13 were calculated initially. By applying missing value test, zero test and correlation test with cutoff value of 0.6 on each category of descriptors separately we have discarded the most likely parameters. Further more if required the parameters were discarded by applying genetic algorithm for development of QSAR model. Taking a brute force approach, we increased the number of parameters in the QSAR equation one by one and evaluated the effect of addition of new term on the statistical quality of the model.

Results for Topological Descriptors

ic50 = 7.13 + 8.04 880a - 2.32 689a - 0.866 612a - 2.85 882a + 0.779 906a +0.000964 821a - 0.000717 823a - 0.0112 767a - 1.24 1017a + 0.388 650a - 0.274 891a -eq (1)

N = 117; r2=0.77; r2(adj) = 0.703; F-test = 28.12; q2 = 0.602

It was found that some compounds were outliers with prediction error in between 1.00 to 2.00. The quality of the above QSAR model has been improved further by removing these compounds.

ic50 = 10.3 + 8.10 880a - 2.22 689a - 0.835 612a - 2.79 882a + 0.782 906a + 0.00103 821a - 0.000775 823a - 0.0129 767a - 1.49 1017a + 0.463 650a - 0.183 891a -eq (2)

N = 100; r2= 0.823; r2(adj) = 0.809; F-test = 68.18; q2 = 0.717

Where N is the number of compounds in the training set, R^2 is the squared correlation coefficient, S is the estimated standard deviation about the regression line, R^2_{adj} is the square of adjusted correlation coefficient for degree of freedom, F is the measure of variance which compares two models differing by one or more variables to see if the more complex model is more reliable than the less complex one, the model is supposed to be good if the F-test is above a threshold value and q^2 is the square of the correlation coefficient of the cross-validation. The QSAR model developed in this study is statistically (r^2 = 0.823, q^2 = 0.717,) best fitted and consequently used for prediction of activity of training and test sets of molecules as reported in Table 14 and Table 15

***Table 14**: Statistical assessment of QSAR equations with varying number of descriptors for Topological set of Descriptors.*

No. of Descriptors	Equation	r^2	SD	Press	q^2
1	ic50 = 3.56 + 10.2 880a	53.80%	16.6922	8.28506	0.503657
2	ic50 = 3.42 + 10.1 880a - 1.79 689a	60.70%	16.6922	7.17617	0.555205
3	ic50 = 3.94 + 10.4 880a - 2.12 689a - 0.646 612a	64.30%	16.6922	7.20498	0.568362
4	ic50 = 4.60 + 9.72 880a - 2.08 689a - 0.592 612a - 2.40 882a	67.00%	16.6922	7.4246	0.570088
5	ic50 = 3.20 + 9.18 880a - 2.12 689a - 0.654 612a - 2.68 882a + 0.662 906a	71.20%	16.6922	171.572	0.578963
6	ic50 = 3.20 + 9.18 880a - 2.12 689a - 0.654 612a - 2.68 882a + 0.662 906a	74.60%	16.6922	41.7261	0.598245
7	ic50 = 3.29 + 8.96 880a - 2.44 689a - 0.819 612a - 2.88 882a + 0.703 906a + 0.00112 821a - 0.000568 823a	76.30%	16.6922	11.4139	0.623687
8	ic50 = 5.25 + 8.74 880a - 2.17 689a - 0.789 612a - 2.67 882a + 0.714 906a +0.000917 821a - 0.000573 823a - 0.0122 767a	78.60%	16.6922	11.6523	0.635478
9	ic50 = 6.97 + 8.52 880a - 2.36 689a - 0.846 612a - 2.66 882a + 0.718 906a +0.000940 821a - 0.000580 823a - 0.0111 767a - 1.12 1017	79.90%	16.6922	48.807	0.679543
10	ic50 = 7.13 + 8.04 880a - 2.32 689a - 0.866 612a - 2.85 882a + 0.779 906a +0.000964 821a - 0.000717 823a - 0.0112 767a - 1.24 1017a + 0.388 650a	81.10%	16.6922	88.8322	0.700365
11	ic50 = 10.3 + 8.10 880a - 2.22 689a - 0.835 612a - 2.79 882a + 0.782 906a + 0.00103 821a - 0.000775 823a - 0.0129 767a - 1.49 1017a + 0.463 650a - 0.183 891a	82.30%	16.6922	63.0589	0.716955

Table 15: The observed and predicted activity of training set.

Compound No.	Observed & Predicted toxicity [logIC50]			Compound No.	Observed & Predicted activity [logIC50]		
	Observed	Predicted	Error		Observed	Predicted	Error
3	1.9248	2.248042	0.323	**69**	1.97313	2.057214	0.084
4	2.22324	2.248042	0.025	**79**	1.41497	1.248802	0.166
5	2.20871	2.032321	0.176	**80**	1.90849	2.056592	0.148
6	2.4624	2.218443	0.244	**81**	2.15534	2.022624	0.133
7	2.38561	2.319412	0.066	**82**	2.17026	1.890894	0.279
8	2.32428	2.278314	0.046	**83**	2.09691	2.256106	0.159
11	2.31597	2.044973	0.271	**84**	2.03743	1.97437	0.063
12	1.91908	2.034268	0.115	**86**	2.31597	2.18178	0.134
13	2.11059	2.013532	0.097	**87**	0.78533	0.958474	0.173
14	1.69897	1.74904	0.05	**89**	1.34242	1.479321	0.137
15	2.01703	2.298842	0.282	**92**	1.99564	1.949048	0.047
16	2.37107	2.24628	0.125	**93**	2.13988	1.954779	0.185
17	2.25527	2.186643	0.069	**95**	1.87506	1.826752	0.048
19	2.21484	2.270043	0.055	**96**	2.1038	1.981361	0.122
20	2.4456	2.169514	0.276	**97**	2.09691	2.055911	0.041
21	1.98677	2.106063	0.119	**98**	2.03342	2.267318	0.234
22	2.14613	2.285743	0.14	**99**	1.36173	1.763344	0.402
23	1.98677	2.23056	0.244	**100**	0.90309	1.044391	0.141
24	2.08991	2.220385	0.13	**104**	2.06819	2.077252	0.009
25	2.14613	2.159339	0.013	**105**	2.02119	2.16686	0.146
26	2.51851	2.270218	0.248	**106**	1.98227	2.264597	0.282
28	1.75587	2.017343	0.261	**108**	2.07555	1.996179	0.079
30	1	0.964714	0.035	**109**	1.97313	2.087919	0.115
32	2.26245	2.302626	0.04	**110**	2.24304	2.031832	0.211
34	2.23553	2.22661	0.009	**111**	2.16435	2.068945	0.095
36	2.14613	2.140093	0.006	**113**	1.87506	1.993216	0.118
37	2.3075	2.331682	0.024	**114**	2.30103	2.034683	0.266
38	2.26245	2.331682	0.069	**116**	0.8451	0.603354	0.242
39	2.26951	2.321239	0.052	**117**	0.004321	0.305086	0.301
40	2.25285	2.321239	0.068	**118**	1.5611	1.655892	0.095
42	2.13988	2.00885	0.131	**121**	1.95036	1.88136	0.069
43	0.83885	0.83851	0	**124**	2.36173	2.056944	0.305
44	1.91908	1.854972	0.064	**127**	1.32222	1.54058	0.218
45	2.17898	2.265131	0.086	**128**	2.08279	1.768261	0.315
46	2.32428	2.346601	0.022	**129**	2.19866	2.348752	0.15
47	2.0607	2.165102	0.104	**131**	1.99564	2.173819	0.178
49	2.25768	2.092257	0.165	**134**	1.80618	1.945832	0.14
50	2.33445	2.022068	0.312	**135**	2.10037	2.043532	0.057
52	2.15836	2.215818	0.057	**136**	2.33445	2.194731	0.14
53	2.35218	2.267061	0.085	**137**	2.22789	1.986459	0.241
54	1.99564	2.055511	0.06	**138**	2.45332	2.200585	0.253
55	2.2014	2.031872	0.17	**139**	2.28103	2.23036	0.051

56	2.15836	2.055511	0.103	**140**	2.10721	2.268217	0.161
57	2.26482	2.161712	0.103	**143**	1.30103	1.56799	0.267
58	2.24797	2.121251	0.127	**150**	2.10721	2.331679	0.224
59	2.06446	2.075144	0.011	**151**	1.88649	2.204569	0.318
60	2.06819	2.052955	0.015	**152**	1.91908	2.184904	0.266
60	2.13672	2.020878	0.116	**153**	2.16732	2.068722	0.099
62	2.09342	2.15294	0.06				
63	2.2014	1.916547	0.285				
64	2.17319	1.962627	0.211				
66	2.07918	2.161112	0.082				
67	1.97313	2.242423	0.269				
68	2	1.967073	0.033				

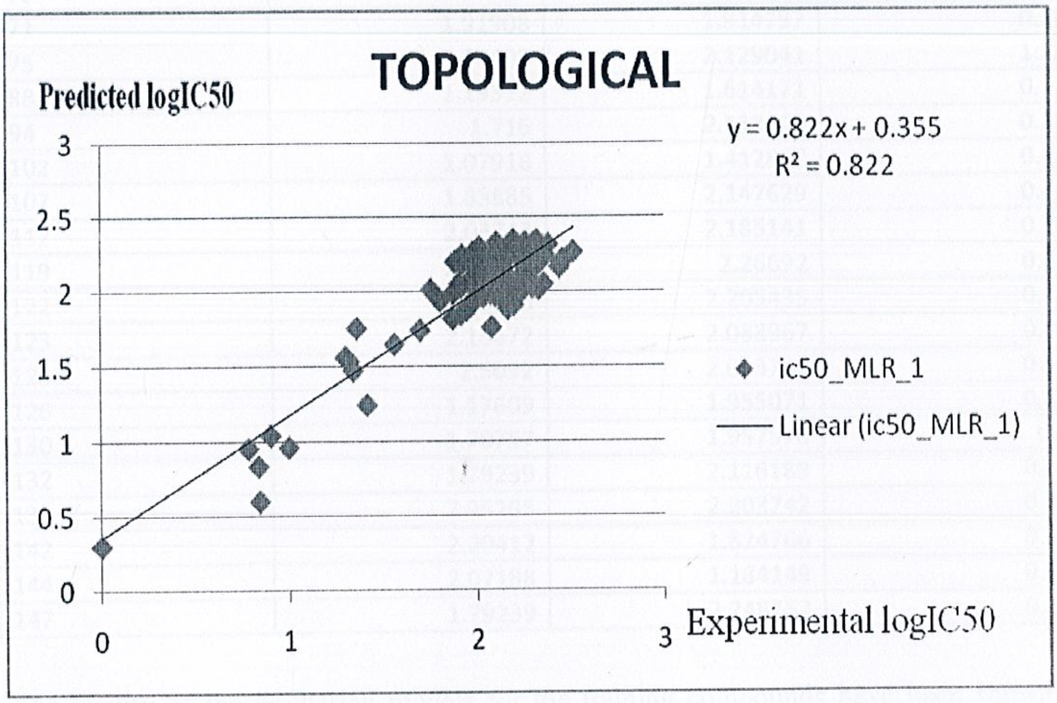


Figure 20*: The graph of predicted and actual activity for trainingset after the removal of 17 outliers*

***Table 16**: The observed and predicted activity of test set.*

Compound number	Actual Toxicity	Predicted Toxicity	Error
1	1.62531	2.36096	0.736
2	2.04454	2.083009	0.038
9	0.60206	2.001322	1.399
10	2.3962	2.090289	0.306
18	1.6721	2.152403	0.48
29	1.53148	2.170127	0.639
31	2.27875	2.302626	0.024
33	1.91908	2.22661	0.308
35	1.88649	2.140093	0.254
51	2.11394	2.100938	0.013
65	2.17319	2.16896	0.004
70	1.17609	1.953803	0.778
71	1.91908	1.814797	0.104
75	0.54407	2.129041	1.585
88	1.19312	1.614171	0.421
94	1.716	2.113349	0.397
102	1.07918	1.412827	0.334
107	1.83885	2.147629	0.309
112	2.03743	2.185141	0.148
119	2.08422	2.26692	0.183
122	2.32838	2.205435	0.123
123	2.13672	2.088967	0.048
125	2.5092	2.053713	0.455
126	1.17609	1.955071	0.779
130	1.70757	1.957576	0.25
132	1.79239	2.116189	0.324
133	2.25285	2.203742	0.049
142	2.20412	1.824766	0.379
144	2.07188	1.184149	0.888
147	1.79239	2.248752	0.456

The quality of the prediction models for the training compounds have been shown in Figure 17. The regression coefficient (r^2) and the cross-validation coefficient (q^2) of the QSAR model were 0.822 and 0.717, respectively revealed good predictive capabilities.

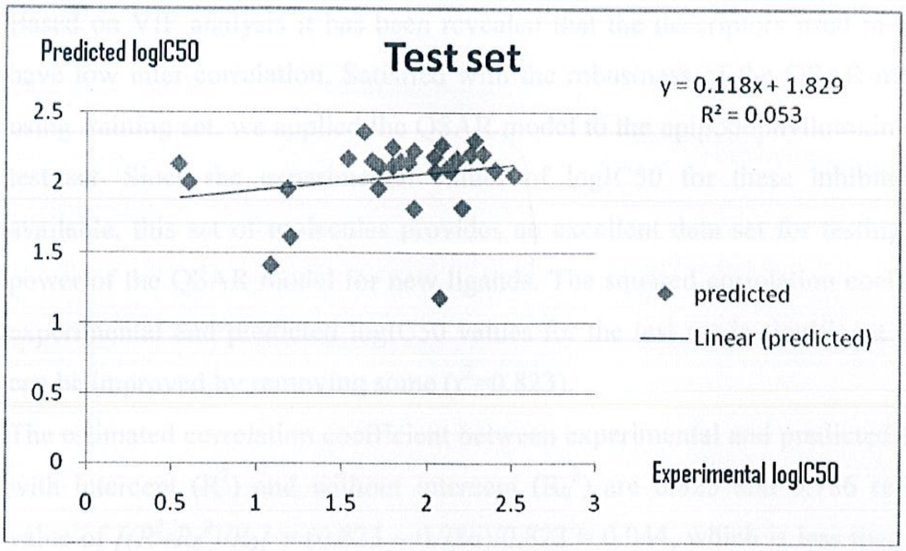


***Figure 21**: The graph of predicted and actual activity for Test data set.*

The inter-correlation of the descriptors used in the final model was very low which is in conformity to the study that for a statistically significant model, it is necessary that the descriptors evolved in the equation should not be inter-correlated with each other. The correlation matrix for the used descriptors is shown. To further check the inter-correlation of descriptors variance inflation factor (VIF) analysis was performed. In this model, the VIF values of these descriptors are:-

Descriptors	**VIF**
612a	1.18624
650a	1.210654
689a	1.262626
1017a	1.166861
767a	1.183432
821a	1.164144
823a	1.207729
880a	1.310616
882a	1.144165
891a	1.218027
906a	1.089325

Based on VIF analysis it has been revealed that the descriptors used in the final model have low inter-correlation. Satisfied with the robustness of the QSAR model developed using training set, we applied the QSAR model to the epipodophyllotoxin comprising the test set. Since the experimental values of logIC50 for these inhibitors are already available, this set of molecules provides an excellent data set for testing the prediction power of the QSAR model for new ligands. The squared correlation coefficient between experimental and predicted logIC50 values for the test set is significant (r^2 = 0.77), but can be improved by removing some (r^2=0.823).

The estimated correlation coefficient between experimental and predicted logIC50 values with intercept (R^2) and without intercept (R_0^2) are 0.823 and 0.786 respectively. The value of $[(R^2-R_0^2)/R_2]$ = (0.823 – 0.786)/0.823 = 0.044, which is less than 0.1. Being the value of q^2 = 0.717, the model corroborates with the criteria for a QSAR model to be highly predictive. Also the value of R^2_{pred} = 0.053 and rm^2 = 0.201 were found to be in the acceptable range, thereby indicating the good external predictability of the QSAR model

Table 17: List of descriptors used in the Topological equation.

Type	Descriptors
612a	2D Topological descriptors Autocorrelation descriptors (Geary, Mass) Order 3
650a	2D Topological descriptors Autocorrelation descriptors (Moran, AlogP98) Order 10
689a	2D Topological descriptors Autocorrelation descriptors (Moran, Mass) Order 5
767a	2D Topological descriptors Autocorrelation descriptors (Moreau-Bruto average, Mass) Order 6
821a	2D Topological descriptors Autocorrelation descriptors (Moreau-Bruto, E-state) Order 5
823a	2D Topological descriptors Autocorrelation descriptors (Moreau-Bruto, E-state) Order 7
880a	BCUT descriptors (AlogP98) Lowest eigenvalue 5 AlogP98
882a	BCUT descriptors (Charge) Highest eigenvalue 2 MPEOE charge
891a	BCUT descriptors (E-state) Highest eigenvalue 1 E-state
906a	BCUT descriptors (Electronegativity) Lowest eigenvalue 1 electronegativity
1017a	Information content related descriptors IC on the edge degree equality

Results for Final Set of Descriptors

Ic 50 = 5.01 + 6.40 880a - 1.81 689a - 0.785 232a + 0.740 650a - 0.840 612a - 0.00737 378a - 0.000730 823a +0.000077 Principal Moment - 3.62 20a -0.00342 384a **eq(1)**

N = 117; r2=0.712; r2(adj) = 0.709; F-test = 34.87; q2 = 0.564

It was found that some compounds were outliers with prediction error in between 1.00 to 2.00. The quality of the above QSAR model has been improved further by removing these compounds.

Ic 50 = 4.93 + 6.54 880a - 2.08 689a - 0.756 232a + 0.706 650a - 0.796 612a - 0.00737 378a - 0.000842 823a +0.000078 Principal Moment - 2.87 20a -0.000175 384a **-eq(2)**

N = 90; r2= 0.772; r2(adj) = 0.745; F-test = 56.65; q2 = 0.591

Where N is the number of compounds in the training set, R^2 is the squared correlation coefficient, S is the estimated standard deviation about the regression line, R^2_{adj} is the square of adjusted correlation coefficient for degree of freedom, F is the measure of variance which compares two models differing by one or more variables to see if the more complex model is more reliable than the less complex one, the model is supposed to be good if the F-test is above a threshold value and q^2 is the square of the correlation coefficient of the cross-validation.

The QSAR model developed in this study is statistically (r^2 = 0.772, q^2 = 0.591,) best fitted and consequently used for prediction of toxicity of training and test sets of molecules as reported in Table 19 and Table 20.

Table 18: Statistical assessment of QSAR equations with varying number of descriptors for Final set of Descriptors.

No. of Descriptors	Equation	r^2	SD	Press	q^2
1	LOG(PCPDF) = 3.40 + 9.01 880a	38.40%	11.6378	8.6393	0.257652
2	LOG(PCPDF) = 3.16 + 8.33 880a - 1.92 689a	48.10%	11.6378	7.32981	0.370172
3	LOG(PCPDF) = 3.33 + 7.58 880a - 1.81 689a - 0.531 232a	53.20%	11.6378	6.55733	0.339608
4	LOG(PCPDF) = 3.25 + 7.01 880a - 1.75 689a - 0.542 232a + 0.434 650a	55.90%	11.6378	6.7066	0.427403
5	LOG(PCPDF) = 3.88 + 7.52 880a - 1.83 689a - 0.592 232a + 0.464 650a - 0.694 612a	61.20%	11.6378	7.68551	0.436549
6	LOG(PCPDF) = 4.41 + 7.87 880a - 2.04 689a - 0.690 232a + 0.497 650a - 0.686 612a - 0.00832 378a	66.40%	11.6378	6.66377	0.423723
7	LOG(PCPDF) = 4.67 + 6.91 880a - 2.06 689a - 0.716 232a + 0.673 650a - 0.698 612a - 0.00728 378a -0.000769 823a	69.70%	11.6378	6.06161	0.479145
8	LOG(PCPDF) = 4.55 + 7.29 880a - 2.11 689a - 0.702 232a + 0.643 650a - 0.782 612a - 0.00720 378a -0.000800 823a +0.000066 Principal Moment	71.20%	11.6378	5.72788	0.507821
9	LOG(PCPDF) = 5.01 + 6.40 880a - 1.81 689a - 0.785 232a + 0.740 650a - 0.840 612a - 0.00737 378a -0.000730 823a +0.000077 Principal Moment - 3.62 20a	73.80%	11.6378	5.1034	0.561481
10	LOG(PCPDF) = 4.93 + 6.54 880a - 2.08 689a - 0.756 232a + 0.706 650a - 0.796 612a - 0.00737 378a -0.000842 823a +0.000078 Principal Moment - 2.87 20a -0.000175 384a	77.20%	11.6378	4.76481	0.590575

***Table 19**: The observed and predicted activities of training set.*

Compound No.	Observed & Predicted activities [logIC50]			Compound No.	Observed & Predicted activities [logIC50]		
	Observed	Predicted	Error		Observed	Predicted	Error
1	1.62531	2.078157	0.453	70	2.14613	2.005003	0.141
2	2.04454	1.973432	0.071	79	2.51851	2.338355	0.18
4	0.95424	1.362453	0.408	80	1.04139	1.8676	0.826
5	1.07918	1.980343	0.901	81	1.53148	2.080429	0.549
6	0.90309	1.790335	0.887	82	1	1.596404	0.596
7	2.06819	1.949503	0.119	83	2.22324	1.663078	0.56
8	2.02119	2.135781	0.115	84	2.27875	2.06655	0.212
9	1.98227	2.106407	0.124	86	1.91908	2.144172	0.225
12	1.97313	1.704257	0.269	87	2.23553	1.834469	0.401
13	2.24304	1.799938	0.443	89	1.88649	2.02403	0.138
14	1.91908	1.62431	0.295	92	2.14613	2.089685	0.056
15	2.16435	2.221688	0.057	93	2.3075	2.000607	0.307
17	1.87506	1.907912	0.033	95	2.26245	2.531742	0.269
18	2.30103	1.917034	0.384	96	2.20871	2.022979	0.186
19	1.61278	1.001979	0.611	97	1.23045	1.792367	0.562
20	0.8451	0.856836	0.012	98	2.13988	1.833428	0.306
21	0.004321	1.059311	1.055	99	0.83885	1.826212	0.987
23	2.08422	2.012407	0.072	100	2.17898	1.895666	0.283
24	1.84323	2.20042	0.357	104	2.32428	1.946503	0.378
26	1.95036	1.939144	0.011	105	2.0607	1.887996	0.173
27	2.32838	2.103588	0.225	106	1.50515	2.008413	0.503
28	2.13672	2.195909	0.059	108	2.25768	1.991031	0.267
29	2.36173	1.887528	0.474	109	2.33445	2.057418	0.277
30	2.5092	1.832557	0.677	110	2.4624	1.930036	0.532
32	1.32222	1.408144	0.086	111	2.11394	2.011417	0.103
33	2.08279	1.821785	0.261	113	2.15836	2.016583	0.142
37	1.99564	1.996933	0.001	114	2.35218	2.127241	0.225
38	1.79239	1.988323	0.196	116	1.99564	1.911199	0.084
42	2.33445	2.16982	0.165	117	2.2014	2.13062	0.071
43	2.22789	2.061241	0.167	118	2.26482	2.211905	0.053
44	2.45332	2.345713	0.108	121	2.24797	1.956957	0.291
45	2.28103	2.09515	0.186	124	2.06446	1.899561	0.165
46	2.10721	1.5997	0.508	127	2.06819	1.867079	0.201
47	2.01703	1.997833	0.019	128	2.13672	1.679764	0.457
48	1.9345	2.079498	0.145	129	2.2014	1.99336	0.208
49	2.20412	1.958235	0.246	131	2.17319	2.009565	0.164
50	1.30103	1.524762	0.224	134	2.17319	1.793601	0.38
51	2.07188	1.496708	0.575	135	1.97313	1.697703	0.275
52	0.95424	1.696834	0.743	136	2	1.471263	0.529
53	0.60206	1.798936	1.197	137	1.17609	1.725097	0.549
54	1.79239	1.663722	0.129	138	2.32428	2.1573	0.167
55	1.25527	2.163818	0.909	139	1.91908	1.646344	0.273

56	1.51851	1.840948	0.322	140	1.07918	1.640489	0.561
57	2.10721	1.916332	0.191	143	2.10721	1.537757	0.569
58	2.37107	1.978235	0.393	150	0.64345	1.239201	0.596
59	1.88649	1.691306	0.195	151	0.54407	1.600935	1.057
60	1.91908	1.665629	0.253	152	1.76343	1.473971	0.289
61	2.16732	1.458212	0.709	153	1.41497	1.401242	0.014
62	2.25527	1.859675	0.396				
63	1.6721	1.597594	0.075				
66	1.9248	1.789575	0.135				
67	1.98677	1.630881	0.356				
68	2.14613	2.077928	0.068				
69	1.98677	2.169111	0.182				

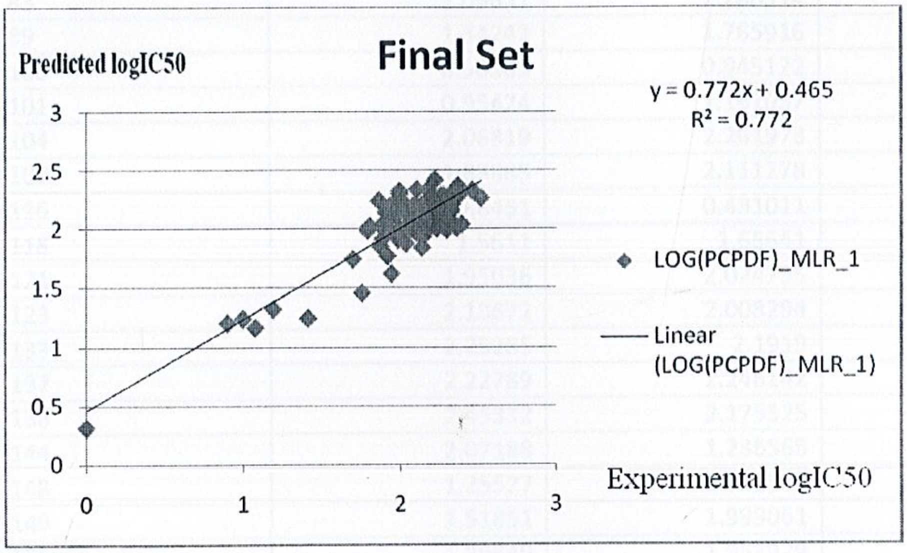


Figure 22: *The graph of predicted and actual activities for final set after the removal of 20 outliers*

***Table 20**: The observed and predicted activities of test set.*

Compound number	Actual Toxicity	Predicted Toxicity	Error
9	0.60206	2.102564	1.501
19	2.21484	2.165203	0.05
22	2.14613	2.198905	0.053
23	1.98677	2.184839	0.198
39	2.26951	2.213887	0.056
43	0.83885	1.807164	0.968
44	1.91908	2.030949	0.112
45	2.17898	2.569766	0.391
47	2.0607	2.220593	0.16
48	1.50515	2.172476	0.667
57	2.26482	2.228124	0.037
58	2.24797	2.264053	0.016
62	2.09342	2.203434	0.11
83	2.09691	2.280914	0.184
89	1.34242	1.765916	0.423
100	0.90309	0.945122	0.042
101	0.95424	1.141057	0.187
104	2.06819	2.261973	0.194
107	1.83885	2.111278	0.272
116	0.8451	0.491011	0.354
118	1.5611	1.65641	0.095
121	1.95036	2.024755	0.074
123	2.13672	2.008294	0.128
133	2.25285	2.1919	0.061
137	2.22789	2.248142	0.02
138	2.45332	2.175525	0.278
144	2.07188	1.236565	0.835
148	1.25527	1.901863	0.647
149	1.51851	1.999061	0.481
151	1.88649	1.953979	0.067

The quality of the prediction models for the training compounds have been shown in Figure 17. The regression coefficient (r^2) and the cross-validation coefficient (q^2) of the QSAR model were 0.772 and 0.591, respectively revealed good predictive capabilities.

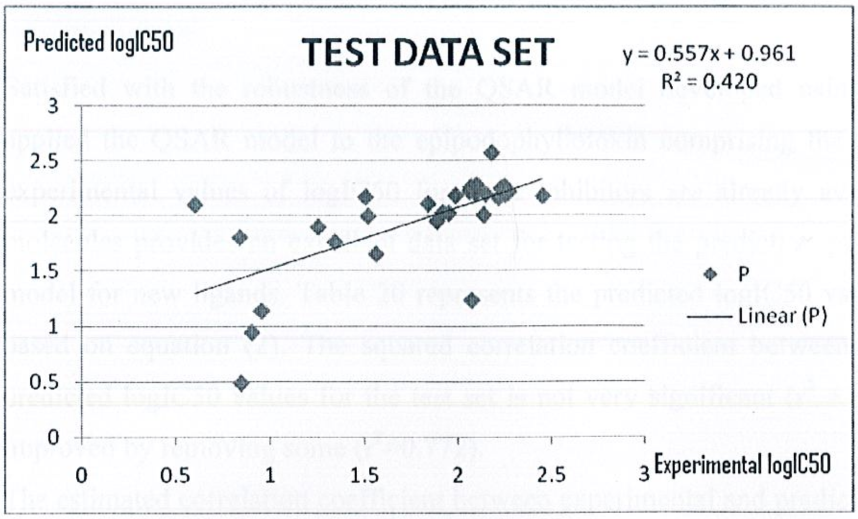


Figure 23: The graph of predicted and actual activities for Test data.

The inter-correlation of the descriptors used in the final model was very low which is in conformity to the study that for a statistically significant model, it is necessary that the descriptors evolved in the equation should not be inter-correlated with each other. The correlation matrix for the used descriptors is shown. To further check the inter-correlation of descriptors variance inflation factor (VIF) analysis was performed. In this model, the VIF values of these descriptors are:-

Descriptors	VIF
20a	1.426534
232a	1.169591
378a	1.128668
384a	1.092896
612a	1.131222
650a	1.278772
689a	1.215067
823a	1.270648
880a	1.485884
principal	1.123596

Satisfied with the robustness of the QSAR model developed using training set, we applied the QSAR model to the epipodophyllotoxin comprising the test set. Since the experimental values of logIC50 for these inhibitors are already available, this set of molecules provides an excellent data set for testing the prediction power of the QSAR model for new ligands. Table 20 represents the predicted logIC50 values of the test set based on equation (2). The squared correlation coefficient between experimental and predicted logIC50 values for the test set is not very significant ($r^2 = 0.751$), but can be improved by removing some (r^2=0.772).

The estimated correlation coefficient between experimental and predicted logIC50 values with intercept (R^2) and without intercept (R_0^2) are 0.772 and 0.76 respectively. The value of $[(R^2\text{-}R_0^2)/R_2]$ = (0.772 – 0.76)/0.772 = 0.0155, which is less than 0.1. Being the value of $q^2 = 0.591$, the model corroborates with the criteria for a QSAR model to be highly predictive. Also the value of $R^2_{pred} = 0.420$ and $rm^2 = 0.737$ were found to be in the acceptable range, thereby indicating the good external predictability of the QSAR model

Table 21: List of descriptors used in the Final equations.

Type	Descriptors
20a	Atom count Fraction of hetero atoms
232a	2D Electrostatic descriptors Partial charge of atom HRPCG (relative positive charge to H-bond donors atoms)
378a	2D Physicochemical descriptors Basic infomation Polarizability (MPEOE method)
384a	2D Physicochemical descriptors Basic information A Water solubility in buffer system (SK atomic types)
612a	2D Topological descriptors Autocorrelation descriptors (Geary, Mass) Order 3
650a	2D Topological descriptors Autocorrelation descriptors (Moran, AlogP98) Order 10
689a	2D Topological descriptors Autocorrelation descriptors (Moran, Mass) Order 5
823a	2D Topological descriptors Autocorrelation descriptors (Moreau-Bruto, E-state) Order 7
880a	BCUT descriptors (AlogP98) Lowest eigenvalue 5 AlogP98
Principal	Principal Moment

REFERENCES

1. Zhiyan xaio, Yun De Xaio, Jun Feng, alexander Golbraikh and Kuo-Hsiung Lee J.Med.Chem. **2001,** 22.
2. Sun, L.; McPhail, A. T.; Hamel, E.; Lin, C.M.; Hastie, S. B.; Chang, J.-J.; Lee, K-H. *J.Med. Chem.* **1993**, *36*, 544.
3. Yamashita, A.; Tawa, R.; Imakura, Y.; Shibuya, M.; Lee, K. H. *Mol. Pharmacol.,* **1994**, *47*, 1920.
4. *Schrodinger L. L. C. ,* http://www.schrodinger.com, (accessed: 24. 04. 2007).
5. Gordaliza, M.; Miguel del Corral, J.M.; Castro, M.A.; López-Vázquez, M.L.; San Feliciano, A.; García Grávalos, M.D.; Carpy, A. *Bioorg. Med. Che.* **1995**, *3*, 1203.
6. Yamashita, A.; Tawa, R.; Imakura, Y.; Shibuya, M.; Lee, K. H. *Mol. Pharmacol.,* **1994**, *47*, 1920.
7. 5. Weiss, S.G.; Tin-Wa, M.; Perdue, R.E. Jr.; Farnsworth, N.R. *J. Pharm. Sci.* **1975**, *64*, 95.
8. Warren Ross[2], Tom Rowe[3], Bonnie Glisson[4], Jack Yalowich and Leroy **Liu** 45, 4 (2007).
9. Kuo-Hsiung Lee, Yasuhiro Imakura, Mitsumasa Haruna, Scott A. Beers, Lee S. Thurston, Hua-Juan Dai, Chung-Hsiung Chen, Su-Ying Liu, Yuing-Chi Cheng *J. Nat. Prod.*, 1989, 52
10. Kamil Kucaa, Jiri Patockaab, Jirí Cabala. *Applied Biomedicine*. 1: 207-211 (2003).
11. Joseph L. Johnson, Bernadette Cusack, Thomas F. Hughes, Elizabeth H. McCullough, Abdul Fauq, Peteris Romanovskis, Arno F. Spatola, and Terrone L. Rosenberry. *Biological Chemistry*. 278, 40 (2003).
12. Fredrick M. Fishel. PI-51(2005).
13. Z. Siroka, J. Drastichova. *ACTA VET.73: 123-132 (2004).*
14. Lee, K H : Imakura, Y : Haruna, M : Beers, S A : Thurston, L S : Dai, H J : Chen, C H : Liu, S Y : Cheng, Y C 67 (2000).

15. Yvan Boublik, Pascale Saint-Aguet, Andree Lougarre, Muriel Arnaud, Francois Villate, Sandino Estrada-Mondaca and Didier Fournier. *Protein Engineering.* 15, 1 (2002).

16. Sinéad B. Walsch, Tracey A. Dolden, Graham D. Moores, Michael Kristensen, Terence Lewis, Alan L. Devonshire and Martin S. Williamson. *Biochem.* 359 (2001).

17. Martin Davies, Alan N. Bateson and Susan M. J. Dunn. Bioscience. 1, d214-233 (1996).

18. Alexander SPH, Mathie A, Peters JA. *Br J Pharmacol.* 153, 2 (2008).

19. www.wikipedia.com

20. Anyanwutaku, X Guo, HX Chen, Z Ji, KH Lee and YC Cheng 64: 65-73 (2008).