COMPUTATIONAL MODELLING OF EPIPODOPHYLLOTOXIN-MECHANISM, MODE OF INTERACTION AND PREDICTIVE ACTIVITY

BY:

Rishay Kumar (061520)

Abhishek Dubey (061501)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

MAY-2010

Submitted in partial fulfillment of the Degree of Bachelor of Technology

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

CERTIFICATE

This is to certify that the work entitled **"Computational modelling of epipodophyllotoxin-mechanism,mode of interaction and predictive activity"** submitted by Mr.Rishay Kumar (061520) and Mr.Abhishek Dubey (061501) 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 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 degree or diploma.

Dr. Pradeep Kumar Naik

(Project Coordinator)

Assistant professor

Dept. of Bioinformatics and Biotechnology

Jaypee Universityof Information Technology
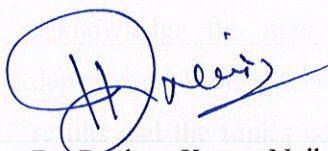
Waknaghat,Solan,Himachal Pradesh,India

2

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.I would like to thank Dr. Pradeep Kumar Naik,Senior Lecturer, Deptt. of Bioinformatics and Biotechnology, for giving us the opportunity to undertake our Btech Final project under his esteem guidance and guide us with the best of his knowledge and resources for the successful completion of this project and for his committed support in the forthcoming endeavors.

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.

Thanks & Regards:

Rishay Kumar & Abhishek Dubey

3

TABLE OF CONTENTS

LIST OF FIGURES

6

List of Abbreviations

- PCPDCF: Percentage of Cellular Protein-DNA Complex Formation

- QSAR: Quantitative Structure Activity Relationship

- TP-II: Topoisomerase II

- CoMFA: Comparitive molecular Field Analysis

- GFA: Genetic Function Approxiamtion

- MMFF: Molecular Mechanics Foce Field

- TNCG: Truncated Newton Conjugated Gradient

- HOF: Heat of Formation

- VIF: Variance Inflation Factor

- eMBrAcE: automated mechanism of Mul -Ligand Bimolecular Associa on with energies

- PdB: Protein Data Bank

- SASA: Solvent Accessible Surface Area

- FEB: Free Energy of Binding

7

Summary of the thesis

The present thesis describes two aspects of modern drug discovery in detail. Firstly, the challenges of **lead optimization** are addressed considering Topoisomerase II as drug target and epipodophyllotoxin structural derivatives as ligands. One important aspect of lead optimization is increasing the target-drug affinity as much as required. Using computer-aided drug design techniques, target-drug binding affinities are analyzed for TP-II which is one of the potent target for anticancer drug development. Secondlly, **structure activity prediction models** were developed based on structure centric and ligand based approaches for the virtual screening of potent inhibitors against TP-II.

Epipodophyllotoxin derivatives have important therapeutic value in the treatment of human cancers. These drugs kill cells by inhibiting the ability of topoisomerase II (TP II) to ligate nucleic acids that it cleaves during the double-stranded DNA passage reaction. The 3D structure of human TP IIα was modeled by following the homology modeling. A virtual library consisting of 143 epipodophyllotoxin derivatives has been developed. Their molecular interactions and binding affinities with modeled human TP IIα have been studied using the docking and Bimolecular Association with Energetics (eMBrAcE) developed by Schrödinger. Structure activity relationship models were developed between the experimental activity expressed in terms of percentage of intracellular covalent TP II-DNA complexes (ln PCPDCF) and molecular descriptors like docking score and free energy of binding. For both the cases the r^2 was in the range of 0.624-0.800 indicating good data fit and r^2_{cv} was in the range of 0.606-774 indicating that the predictive capabilities of the models were acceptable. Low levels of root mean square error for the majority of inhibitors establish the docking and eMBrAcE based prediction model as an efficient tool for generating more potent and specific inhibitors of human TP IIα by testing rationally designed lead compounds based on epipodophyllotoxin derivatization.

To improve their clinical efficacy and overcome the problems of drug resistance, myelosuppression and poor oral availability large number of epipodophyllotoxin derivatives have been synthesized and tested over the years. These data provides the base for the construction of an informative structure-activity relationship (SAR) model which can be utilized for the rapid prediction of anticancer activity of novel epipodophyllotoxin analogues and virtual

prescreening. A quantitative structure-activity relationship (QSAR) model has been developed between percentage of cellular protein-DNA complex formation (PCPDCF) and structural properties by considering a data set of 130 epipodophyllotoxin analogues based on 2D and 3D structural descriptors. A systematic step-wise searching approach of zero tests, missing value test, simple correlation test, multicollinearity test and genetic algorithm method of variable selection was used to generate the model. Statistically significant model ($r^2_{(train)} = 0.721$; $q^2_{cv} = 0.678$) was obtained with the descriptors like solvent accessible surface area (SASA), heat of formation (HOF), Balban index, number of atom classes and sum of E-state index of atoms. The robustness of the QSAR models was characterized by the values of the internal leave one out cross-validated regression coefficient (q^2_{cv}) for the training set and $r^2_{(test)}$ for the test set. The overall root mean square error (RMSE) between the experimental and predicted PCPDCF was 0.194 and $r^2_{(test)} = 0.689$; revealed good predictability of the QSAR model. QSAR model developed in this study shall aid in future designing of novel potent epipodophyllotoxin derivatives.

CHAPTER I

Introduction

Current Drug Discovery

Reducing time, resources and serendipity has been the major aim for the development of modern, rational drug discovery strategies.

Historically, most drugs have been discovered mainly by two of the following methods. The first is to modify a know starting molecule like a natural ligand, co-factor or simply an already existing drug. The first drug to market is rarely the best. COX-2 inhibitors, HIV protease inhibitors or the "patent-busting" around PDE5 inhibitors serve as quite productive examples for this approach [1-3]. For diseases where neither a drug nor natural template structure exists, the second route is applied, random screening. Being venerable this approach requires a certain serendipity to succeed. Cyclosporine and paclitaxel are prominent examples for drugs identified by screening corporate compound selections. The invention of combinatorial chemistry increased the size of the chemical collections and assay automation reduced time and resources necessary to screen large libraries. Still a certain degree of serendipity is required. Therefore, more rational routes for drug discovery have been sought [4-6].

Today, the whole drug discovery process from development to registration for the market is an impressively long, expensive, and risky challenge. On average, only one out of 10,000 originally synthesized or isolated compounds will clear all the hurdles on the way to becoming a commercially available drug [7] The process from first discovery to full development takes ~15 years to complete and costs approximately 800 million dollars. The process starts with finding the right molecular target to address a certain disease, i.e., a single gene or protein [8]. Interestingly, current drug therapy is based on less than 500 targets. Molecular biology efforts help understanding disease processes at the genetic level and to determine optimal targets [9]. The completion of the human genome together with new technologies like bioinformatics, genomics and proteomics allow to characterize more genetic and molecular processes in humans and other species, thus providing thousands of possible enzymes, receptors and ion channels as

10

potential new drug targets [10]. Unfortunately, having more targets does not necessarily lead to more drugs on the market.

IDEA

Discovery	0	Target selection
	2	HTS / VS
	5	Lead discovery / optimization
Development	6-7	Exploratory development "Clinic I"
	8-9	Safety / feasibility "Clinic II"
	10-12	Tolerability / efficacy "Clinic III"
	13-14	Registration
	15-16	Market introduction

years
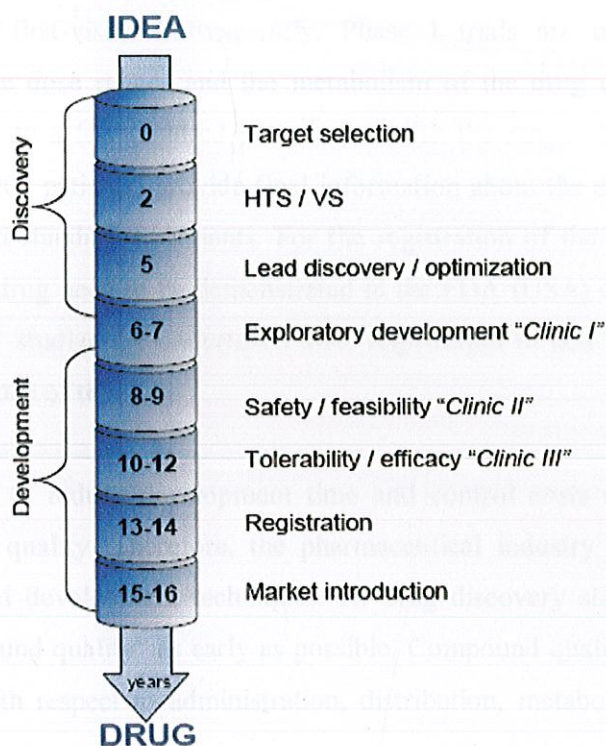
DRUG

Figure 1.1 Scheme of the pharmaceutical company's drug design and development strategy.

Once a target is validated, different screening strategies are applied to identify the so called lead, a prototypical structure that demonstrates an adequate activity. Considering a representative target portfolio, high-throughput screening (HTS) is presently the most widely applicable technology delivering chemistry entry points for drug discovery programmes [11]. Identified lead compounds are subsequently "optimized," i.e., altered in ways that both increase their potential efficacy as well as minimize any potential side effects [12]. The tools of modern chemistry (parallel and automated synthesis) and modern biology merge in these stages to give the best candidates for the drugs of the future.

To ensure the safety of administering a new drug candidate to healthy volunteers and patients, extensive toxicological and safety pharmacological profiles are done in both, in vitro

11

tests and animals studies. Once the initial safety profile is established, the drug candidate is given to a small group of patients or healthy volunteers to verify tolerance and possible effects to the human disease (first-man-first-visit). Subsequently, Phase I trials are undertaken to gain information about tolerable dose ranges and the metabolism of the drug candidate. Phase II clinical trials involving up to a few hundred patients establish the range of efficacious doses. Phase III trials (up to 10,000 patients) provide final information about the drug's effectiveness and provide comparison to standard treatments. For the registration of the new drug, quality, efficacy and safety of the drug need to be demonstrated to the FDA (USA) or EMEA (EU). At times, additional Phase IV studies are undertaken after registration to add new indications or improve existing formulations of the drug.

Future challenges are to reduce development time and control costs effectively without compromising safety and quality. Therefore, the pharmaceutical industry tries to constantly integrate new research and development techniques. At drug discovery state (Figure 1), this means to consider "compound quality" as early as possible. Compound quality comprises drug-likeness of a molecule with respect to administration, distribution, metabolism and excretion (ADME).

Another requirement is high affinity of the drug to its target to keep the necessary dosage as low as possible. This facilitates formulation and decreases the risk to experience possible side-effects. It is well recognized that even when compounds are identified from HTS they are not always suitable to embark on further chemistry exploration [12]. One promising and now frequently applied approach is to integrate computational models to predict the desired property at each step of the discovery process. In the last 20 years, those models and computational approaches to assist drug discovery have gained in importance and proven to be a resource-saving techniques to identify and optimize novel chemotypes in biologically active molecules.

12

Computational drug design

Computer-aided drug design is an essential part of the modern medicinal chemistry, and has led to the acceleration of many projects, and even to drugs on the market [13]. Its application ranges from early target validation to toxicity models at late discovery/early development transition state.

Assessing the druggability of a target is supported by the analysis of theoretical signaling pathway networks [13], comparing the putative binding pocket to existing ones [14,15], or generating homology models [16-19] if neither crystal nor NMR structures are available. Computational strategies for the target validation process have been reviewed by Wiemann, Blundell, Hajduk, and Keller [20-23]. Most extensively, computer-aided strategies are applied in the lead discovery and lead optimization stage. In addition to high throughput screening (HTS), the main lead discovery technology employed by most pharmaceutical companies today is virtual screening [24,25]. It involves the rapid assessment of large libraries of chemical structures and can be categorized as being either ligand-based or receptor-based. The compounds are compared to previously as active identified molecules (similarity searches) or matched to a pharmacophore derived from the target binding site (pharmacophore search). When the structure of the target protein is known, receptor-based docking can be employed. This approach aims to predict correctly the structure of the intermolecular complex formed between the target receptor and the ligand. To correctly dock a molecule, two technical challenges imply, (1) the pose generation (docking) of the ligand in the active site and (2) the evaluation of the different poses (scoring). Scoring requires estimation of the binding energy between protein and ligand and produces a relative rank-ordering between different ligand, docked to the same target. Examples of virtual screening have been published as referenced in the appendix. Different screening protocols and their successful application are reviewed by Jalaie, Pirard, Hou, Green, Langer and Schneider [26-31]. An alternative to screen for compounds is designing molecules with desired properties from scratch. Integration of so called de-novo design to drug discovery is reviewed by Schneider and Honma [32,33]. At lead optimization state, docking is applied to assess the modifications of the lead compound. As a series of molecules with known activity arises, three dimensional quantitative structure-activity relationships (3D QSAR) methods gain in importance. Using 3D

13

QSAR statistical correlations are established between the potencies of a series of structurally related compounds and one or more quantitative structural parameters, such as lipophilicity, polarity, and molecular size, by using multi linear regression analysis [34,35].

For lead compound series with convincing activities computational methods are also applied to estimate pharmacokinetic properties. Again docking, homology modeling and statistical correlations methodologies are used in order to evaluate potential Cyp450 inhibition [36,37], plasma protein binding [38], HERG channel blockage [39,40], bioavailability and toxicity [41,42].

Computer-aided drug design techniques are nowadays established and have emerged as key strategy to help assessing compounds [43]. Predicting chemical and biological properties with computational models identifies compounds that are likely to fail in primary, secondary and further downstream screen at significantly lower costs. Integration of computational approaches into the drug discovery process is nicely reviewed by Chin and Oprea [44, 6].

Podophyllotoxin and analogues design

The aryltetralin lactone podophyllotoxin occupies a unique position among lignan natural products since its glucopyranoside derivative was recognized as a potent antitumor factor [45]. This discovery entails a particularly fascinating account involving multitude investigations conducted over a period of more than a century [46]. The studies culminated in the structure elucidation of podophyllotoxin (1), the assessment of its biological activity and the discovery of its mode of action. Initial expectations regarding the clinical utility of podophyllotoxin were tempered largely due to its unacceptable gastrointestinal toxicity. This led chemists in the pharmaceutical research department of Sandoz, to investigate the possibility that the *Podophyllum* lignans might occur naturally as glycosides [46]. Using special procedures to inhibit enzymatic degradation, these researchers indeed obtained the podophyllotoxin-β-D-glucopyranoside (5) as the main component and its 4'-demethyl derivative (6) from the Indian *Podophyllum* species. Both of these glucosides and the glucosides (8) and (10) of α- and β-peltatin (7) and (9) (Figure 1) were also isolated from the American *P. peltatum*. Being less hydrophobic, the glucosides displayed lower toxicity than the aglucones but their cytostatic

14

activity was reduced to the same degree. The research efforts were then focused on a program to chemically modify both the glucosides and aglucones of a wide range of podophyllotoxin derivatives. Nearly 600 derivatives were prepared and tested over a period of about 20 years [46]. Somehow serendipitously, a radical change in the mechanism of action and a quantum step in therapeutic utility were effected by acetalization of the 4- and 6-hydroxy groups of the glucopyranose moiety using aldehydes eventually leading to the discovery of the clinically important anticancer drugs etoposide (11), etopophos (12) [47] and teniposide (13) (Figure 1). However, the clinical application of podophyllotoxin and its analogues in the treatment of cancer has been limited by severe toxic side effects during administration of the drugs [48,49].
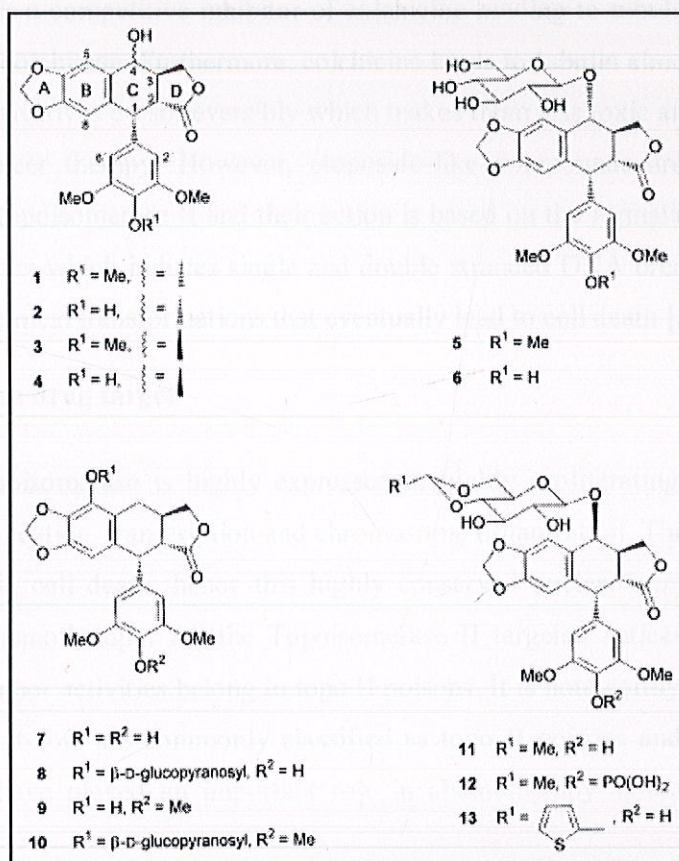

Figure 2. Structures of podophyllotoxin and its congeners and derivatives.

The main problem of most antineoplastic agents which substantially reduces their therapeutic usefulness lies in their scant selectivity because these substances affect cancer and normal cells alike and lead to the appearance of adverse side effects. Like most antineoplastic agents,

15

podophyllotoxin, etoposide and teniposide do not show specificity against tumour cells; rather they affect all cells, especially those in the active phase of division. This accounts for the adverse effects of these drugs: anaemia, catharsis and alopecia among others [50-52]. Thus, the studies aimed at improving antineoplastic agents mainly focused on the search for more selective drugs.

Mechanisms of action of podophyllotoxin

Lignans inhibit the polymerization of tubulin and DNA topoisomerase II [53, 50].Studies on Structure-Activity Relationships (SAR) have shown that podophyllotoxin like compounds preferentially inhibit tubulin polymerization which leads to arrest of the cell cycle in the metaphase [54]. It is a competitive inhibitor of colchicine binding to tubulin [55]. Its affinity is double than that of colchicine. Furthermore, colchicine binds to tubulin almost irreversibly while podophyllotoxin derivatives do so reversibly which makes them less toxic and hence more useful in the field of cancer therapy. However, etoposide-like compounds are potent irreversible inhibitors of DNA topoisomerase II and their action is based on the formation of a nucleic acid-drug-enzyme complex which induces single and double stranded DNA breaks as the initial step in a series of biochemical transformations that eventually lead to cell death [54].

Topoisomerase II as drug target

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.

Type IIA topoisomerases are essential in the separation of daughter strands during 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

16

IA, IB, and IIB topoisomerases. This ability, known as topology simplification, was first identified by Rybenkov et al. Science 1997. The hydrolysis of ATP drives this simplification, but a clear molecular mechanism for this simplification is still lacking. Several models to explain this phenomenon have been proposed, including two models that rely on the ability of type IIA topoisomerases to recognize bent DNA duplexes. Biochemistry, electron microscopy, and the recent structure of topo II bound to DNA reveals that type IIA topoisomerases bind at the apices of DNA, supporting this model. Type IIA topoisomerase operates through a "two-gate" mechanism, a mechanism supported by biochemistry as well as by structural work. A strand of DNA, called the polymorphic-delta segment, or PDS-segment is bound by a central DNA-binding gate (DNA-gate). A second strand of DNA, called the Transporter exchanger (sometimes called the transfer), or T-segment is captured by the oxidation of the N-terminal ATPase domain (the ATPase-gate) as two molecules of ATP are bound. Hydrolysis of ATP and release of a inorganic phosphate leads to the cleavage of the G-segment, as the catalytic tyrosines form a covalent phosphotyrosine bond with the 5' end of the DNA. This creates a four-base overhang and a double stranded break in the G-segment. As the DNA-binding gate separates, the T-segment is transferred through the G-segment. The G-segment is sealed, leading to the C-terminal gate (or C-gate) to open, allowing for the release of the T-segment. Release of product ADP leads to a reset of the system, and allows a second T-segment to be captured.

Computational strategy for virtual screening of potent lead molecules

It is widely appreciated that advancement in the biological component of drug development has catalyzed a shift in the strategies and tactics that underlie the drug discovery process. New information has evolved to describe disease states at the molecular rather than organism level which in turn presents those involved in drug development with a large array of well-defined targets. Additionally, economic factors are driving the need for a shorter lead-to-drug development time. A number of methodologies have evolved to integrate the higher degree of molecular information, number of new targets and need for efficiency. This integration has been most widely implemented in the coupling of high throughput screening (HTS) with high-output chemical synthesis. HTS relies on the development of efficient and reliable assays to permit the evaluation of a large number of compounds against a target in a rapid and often automated manner. The large volume of HTS data is modeled in order to assess structure-activity

17

relationships but problems arise when these models suffer from distortion by false positives. Combinatorial Chemistry, the synthesis of a very large number of compounds using a single scaffold and a diverse array of reactants, has also made an attempt to address the need for a large number of new drug leads. This methodology is severely hindered by the labor-intensive and cost effective measures required for its preparation and purification in such a large number of compounds. Virtual screening, using a computational approach to assess the interaction of an *in silico* library of small molecules and the structure of a target macromolecule, has arisen as an alternative method for the rapid identification of new drug leads. A great deal of effort has been made to create reliable and efficient software that evaluates the highly complex enthalpic and entropic nature of the interaction between small molecules and their macromolecular receptors. A typical virtual library screening (VLS) approach involves several stages (Figure 1.4) including parallel efforts that involve small molecule and macromolecule preparation. Various stages of the VLS methodology described in Figure 1.4 have been previously reviewed [56-58] and provide an excellent resource for detailed analyses of many of the components of this process.

High throughput screening has been routinely used for the identification of small molecule drugs of a specific target. One of the most widely used virtual screening approaches in rational drug design is Quantitative Structure Activity Relationship (QSAR) [59]. The QSAR is based on statistical analysis of the relationship between certain biological activities of a chemical against quantitative attributes of the structure of the chemical. The resultant statistical model may be used to predict the activity of an unknown chemical by its quantitative attributes calculated from its structure. The molecular modeling techniques: molecular docking and rescoring using Prime MMGB/SA were widely used for virtual screening of inhibitors for energetically favorable interaction with receptor and the energy score has been used to build models for prediction of the inhibitory activity (pIC_{50}) [60]. Further, the newly developed structure-based linear interaction energy method implementing a surface generalized Born (SGB-LIE) [61] continuum model for solvation could be used to build a binding affinity model for estimating the free energy of binding a diverse set of inhibitors. The LIE method has been applied on a number of protein-ligand systems with promising results, [62-64] producing small errors in the order of 1 kcal/mol for free energy prediction.

18

By generating QSAR, Docking, MM-GB/SA and SGB-LIE models of a sufficient pool of potential analogues of epipodophyllotoxin, these virtual screening methods can be extended to facilitate the search for the potential drugs with low toxicity and better biological activity against cancer.
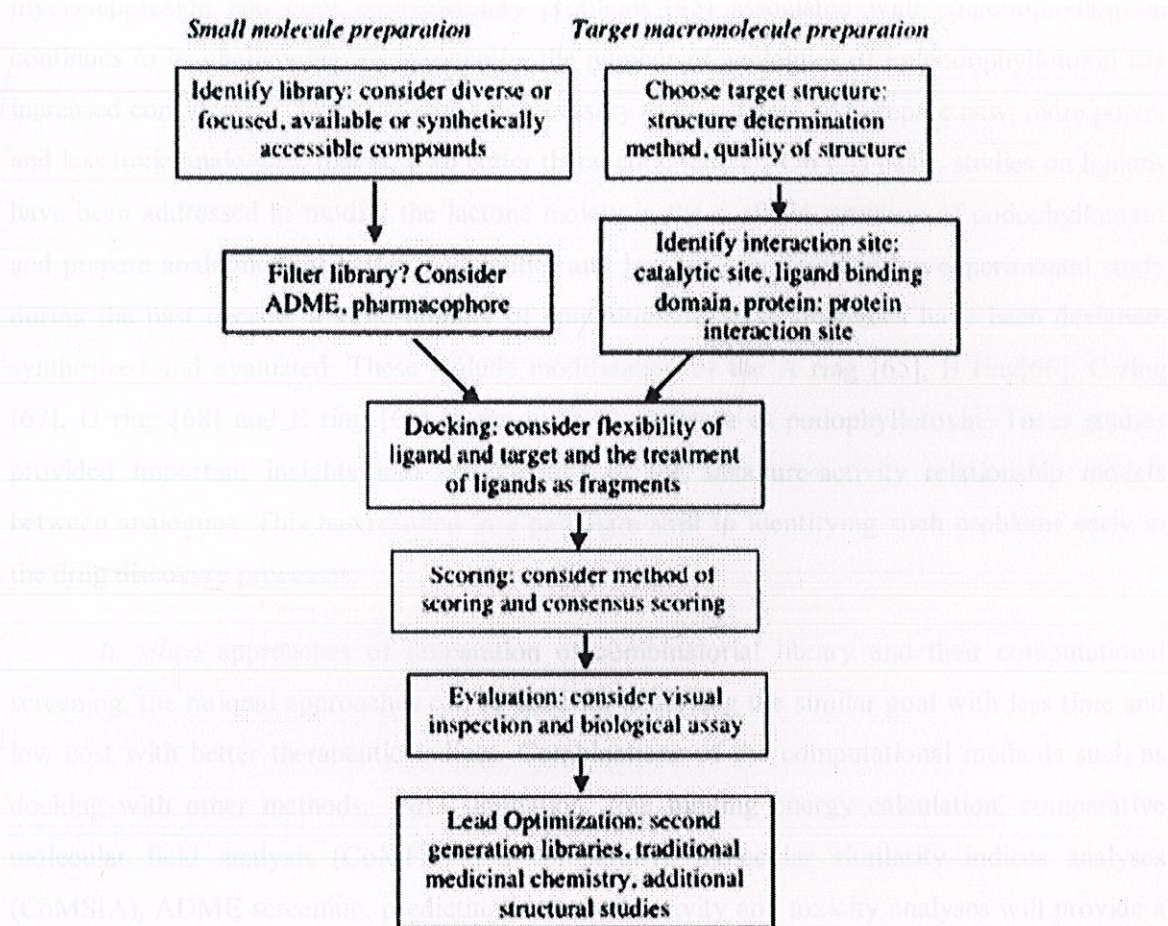
Small molecule preparation   **Target macromolecule preparation**

```
┌──────────────────────────────┐   ┌──────────────────────────────┐
│ Identify library: consider   │   │ Choose target structure:     │
│ diverse or focused,          │   │ structure determination      │
│ available or synthetically   │   │ method, quality of structure │
│ accessible compounds         │   │                              │
└──────────────────────────────┘   └──────────────────────────────┘
              │                                  │
              ▼                                  ▼
┌──────────────────────────────┐   ┌──────────────────────────────┐
│ Filter library? Consider     │   │ Identify interaction site:   │
│ ADME, pharmacophore          │   │ catalytic site, ligand       │
│                              │   │ binding domain, protein:     │
│                              │   │ protein interaction site     │
└──────────────────────────────┘   └──────────────────────────────┘
              │                                  │
              └──────────────┐   ┌───────────────┘
                             ▼   ▼
              ┌──────────────────────────────┐
              │ Docking: consider flexibility│
              │ of ligand and target and the │
              │ treatment of ligands as      │
              │ fragments                    │
              └──────────────────────────────┘
                             │
                             ▼
              ┌──────────────────────────────┐
              │ Scoring: consider method of  │
              │ scoring and consensus scoring│
              └──────────────────────────────┘
                             │
                             ▼
              ┌──────────────────────────────┐
              │ Evaluation: consider visual  │
              │ inspection and biological    │
              │ assay                        │
              └──────────────────────────────┘
                             │
                             ▼
              ┌──────────────────────────────┐
              │ Lead Optimization: second    │
              │ generation libraries,        │
              │ traditional medicinal        │
              │ chemistry, additional        │
              │ structural studies           │
              └──────────────────────────────┘
```

Figure 3. A typical virtual library scheme (VLS). Stages include both small molecule library preparation (choice of library, considerations for filtering) and target preparation (choice of structure of target and identification of binding site). In VLS, the library is docked into the target, scored and evaluated. Any possible leads are optimized in later stages.

Predicted models of structure activity relationship of epipodophyllotoxin

Efforts at improving their clinical efficacy further by overcoming the drug resistance, myelosuppresion and poor bioavailability problems [45] associated with epipodophyllotoxin continues to be challenging. Consequently, the number of analogues of epipodophyllotoxin has increased considerably. In this regard it is necessary to investigate and prepare new, more potent and less toxic analogues, that is, with better therapeutic indices. On this basis, studies on lignans have been addressed to modify the lactone moiety in the scaffold structure of podophyllotoxin and prepare analogues for better therapeutics and low toxicity. In many an experimental study during the past decade, a large number of epipodophyllotoxin analogues have been designed, synthesized and evaluated. These include modification of the A ring [65], B ring[66], C ring [67], D ring [68] and E ring [69] in the scaffold structure of podophyllotoxin. These studies provided important insights into development of the structure-activity relationship models between analogues. This has resulted in a paradigm shift in identifying such problems early in the drug discovery processes.

In silico approaches of preparation of combinatorial library and their computational screening, the rational approaches can be used for achieving the similar goal with less time and low cost with better therapeutic indices. Combinations of the computational methods such as docking with other methods, MD simulation, free binding energy calculation, comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analyses (CoMSIA), ADME screening, predicting biological activity and toxicity analyses will provide a lot of insights into the biological system in rational drug designing. These computational techniques can be used for the study of the interaction of enzymes with active agents and screening of the lead drug candidates. This knowledge provides researchers a unique chance to learn the functioning of these enzymes in biological conditions and finally helps them in designing a suitable agent to affect the function of an enzyme. This will decrease the amount of experimental work required to synthesize and test a large number of possible candidates. With the help of a variety of computational methods, traditional drug development has greatly benefited from the computational research. Expenses and the development period for a new drug

have been reduced considerably as computational methods can dramatically decrease the number of candidates of a drug which need to be synthesized and tested.

Docking, modeling, molecular simulation, QSAR, virtual screening, free energy calculations and data mining methods have been used directly in rational drug discovery projects to speed up development and help find good agents. These methods produce a lot of information on a variety of drug related research. They benefit basic scientific activities as well as industrious efforts. But most of these computational tools have their own limitations and they need further development on some basic, methodological, and application problems. A lot of applications have demonstrated that if a proper approach is chosen on a specific research, good results can be produced to solve targeted problems [70].

In this work we used computational methods to explore the binding structures, binding affinity and inhibition mechanism of active ligands in their corresponding receptors. Topoisomerase II/epipodophyllotoxin bio-system was used in the work. By the application of computational techniques, we tried to learn more about the bio-system, especially the TP-II and its mechanism of inhibition that produces information for a researcher to understand their biological functions affected by inhibitors. Further, we tried to assess how a ligand affects its receptor and what type of ligand will better inhibit TP-II. Also we tried to develop approaches to calculate the activity of a set of ligands by applying the ways of free energy of binding (FEB) and quantitative structure-activity relationship (QSAR). These are convenient approaches which can be used to normal set of compounds to benefit ligand activity evaluations in a rational drug design.

References

1. Wermuth, C. G. (2004) "Selective optimization of side activities: another way for drug discovery. "*J Med Chem,* **47** (6): 1303 - 1314.

2. Wermuth, C. G. (2006a) "Selective optimization of side activities: the SOSA approach. "*Drug Discov Today,* **11** (3-4): 160 - 164.

3. Wermuth, C. G. (2006b) "Similarity in drugs: reflections on analogue design. "*Drug Discov Today,* **11** (7-8): 348 - 354.

4. Lundqvist, T. (2005) "The devil is still in the details--driving early drug discovery forward with biophysical experimental methods. "*Curr Opin Drug Discov Devel,* **8** (4): 513 - 519.

21

5. Klebe, G. (2000) "Recent developments in structure-based drug design. "*J Mol Med,* **78** (5): 269 - 281.

6. Oprea, T. I. and Matter, H. (2004) "Integrating virtual screening in lead discovery. "*Curr Opin Chem Biol,* **8** (4): 349 - 358.

7. Gad, S. (2005). Drug Discovery in the 21st Century. In S. Gad (Ed.), *Drug Discovery Handbook* (pp. 1 - 10). Hoboken:Wiley.

8. Kubinyi, H. (2003) "Drug research: myths, hype and reality. "*Nat Rev Drug Discov,* **2** (8): 665 - 668.

9. Drews, J. (2000) "Drug discovery: a historical perspective. "*Science,* **287** (5460): 1960 – 1964.

10. Austin, C. P. (2004) "The impact of the completed human genome sequence on the development of novel therapeutics for human disease. "*Annu Rev Med,* **55** 1 - 13.

11. Baxter, C. A., Murray, C. W., Waszkowycz, B., Li, J., Sykes, R. A., Bone, R. G., Perkins, T. D. and Wylie, W. (2000) "New approach to molecular docking and its application to virtual screening of chemical databases. "*J Chem Inf Comput Sci,* **40** (2): 254 - 262.

12. Keseru, G. M. and Makara, G. M. (2006) "Hit discovery and hit-to-lead approaches. "*Drug Discov Today,* **11** (15-16): 741 – 748.

13. Coupez, B. and Lewis, R. A. (2006) "Docking and scoring--theoretically easy, practically impossible? "*Curr Med Chem,* **13** (25): 2995 – 3003.

14. Nettles, J. H., Jenkins, J. L., Bender, A., Deng, Z., Davies, J. W. and Glick, M. (2006) "Bridging chemical and biological space: 'target fishing' using 2D and 3D molecular descriptors. "*J Med Chem,* **49** (23): 6802 – 6810

15. Weber, A., Casini, A., Heine, A., Kuhn, D., Supuran, C. T., Scozzafava, A. and Klebe, G. (2004) "Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. "*J Med Chem,* **47** (3): 550-7.

16. Evers, A. and Klebe, G. (2004) "Ligand-supported homology modeling of g-protein-coupled receptor sites: models sufficient for successful virtual screening. "*Angew Chem Int Ed Engl,* **43** (2): 248 - 251.

17. Hillisch, A., Pineda, L. F. and Hilgenfeld, R. (2004) "Utility of homology models in the drug discovery process. "*Drug Discov Today,* **9** (15): 659 - 669.

18. Kairys, V., Fernandes, M. X. and Gilson, M. K. (2006) "Screening drug-like compounds by docking to homology models: a systematic study. "*J Chem Inf Model,* **46** (1): 365 - 379.

22

19. Rong, S. B., Enyedy, I. J., Qiao, L., Zhao, L., Ma, D., Pearce, L. L., Lorenzo, P. S., Stone, J. C., Blumberg, P. M., Wang, S. and Kozikowski, A. P. (2002) "Structural basis of RasGRP binding to high-affinity PKC ligands. "*J Med Chem,* **45** (4): 853 - 860.

20. Rong, S. B., Enyedy, I. J., Qiao, L., Zhao, L., Ma, D., Pearce, L. L., Lorenzo, P. S., Stone, J. C., Blumberg, P. M., Wang, S. and Kozikowski, A. P. (2002) "Structural basis of RasGRP binding to high-affinity PKC ligands. "*J Med Chem,* **45** (4): 853 - 860.

21. Hajduk, P. J., Huth, J. R. and Fesik, S. W. (2005a) "Druggability indices for protein targets derived from NMR-based screening data. "*J Med Chem,* **48** (7): 2518 - 2525.

22. Hajduk, P. J., Huth, J. R. and Tse, C. (2005b) "Predicting protein druggability. "*Drug Discov Today,* **10** (23-24): 1675 - 1682.

23. Wieman, H., Tondel, K., Anderssen, E. and Drablos, F. (2004) "Homology-based modelling of targets for rational drug design. "*Mini Rev Med Chem,* **4** (7): 793 - 804.

24. Walters, W., Stahl, M. and Murcko, M. (1998) "Virtual Screening - an overview. "*Drug Discovery Today,* **3** (4): 160 - 178.

25. Klebe, G. (2006) "Virtual ligand screening: strategies, perspectives and limitations. "*Drug Discov Today,* **11** (13-14): 580-94.

26. Green, D. V. (2003) "Virtual screening of virtual libraries. "*Prog Med Chem,* **41** 61 - 97.

27. Hou, T. and Xu, X. (2004) "Recent development and application of virtual screening in drug discovery: an overview. "*Curr Pharm Des,* **10** (9): 1011 - 1033.

28. Jalaie, M. and Shanmugasundaram, V. (2006) "Virtual screening: are we there yet? "*Mini Rev Med Chem,* **6** (10): 1159 – 1167.

29. Langer, T. and Krovat, E. M. (2003) "Chemical feature-based pharmacophores and virtual library screening for discovery of new leads. "*Curr Opin Drug Discov Devel,* **6** (3): 370 – 376.

30. Pirard, B. (2005) "Knowledge-driven lead discovery. "*Mini Rev Med Chem,* **5** (11): 1045 - 1052.

31. Schneider, G. and Bohm, H. J. (2002) "Virtual screening and fast automated docking methods. "*Drug Discov Today,* **7** (1): 64 - 70.

32. Honma, T. (2003) "Recent advances in de novo design strategy for practical lead identification. "*Med Res Rev,* **23** (5): 606 - 632.

33. Schneider, G. and Fechner, U. (2005) "Computer-based de novo design of drug-like molecules. "*Nat Rev Drug Discov,* **4** (8): 649 - 663.

34. Gohlke, H., Schwarz, S., Gundisch, D., Tilotta, M. C., Weber, A., Wegge, T. and Seitz, G. (2003) "3D QSAR analyses-guided rational design of novel ligands for the (alpha4)2(beta2)3 nicotinic acetylcholine receptor. "*J Med Chem,* **46** (11): 2031 - 2048.

35. Kuo, C. L., Assefa, H., Kamath, S., Brzozowski, Z., Slawinski, J., Saczewski, F., Buolamwini, J. K. and Neamati, N. (2004) "Application of CoMFA and CoMSIA 3D-QSAR and docking studies in optimization of mercaptobenzenesulfonamides as HIV-1 integrase inhibitors. "*J Med Chem,* **47** (2): 385 - 399.

36. Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T. and Vianello, R. (2005) "MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. "*J Med Chem,* **48** (22): 6970 - 6979.

37. Zamora, I., Afzelius, L. and Cruciani, G. (2003) "Predicting drug metabolism: a site of metabolism prediction tool applied to the cytochrome P450 2C9. "*J Med Chem,* **46** (12): 2313 - 2324.

38. Gleeson, M. P. (2007) "Plasma protein binding affinity and its relationship to molecular structure: an in-silico analysis. "*J Med Chem,* **50** (1): 101 - 112.

39. Farid, R., Day, T., Friesner, R. A. and Pearlstein, R. A. (2006) "New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. "*Bioorg Med Chem,* **14** (9): 3160 – 3173.

40. Pearlstein, R. A., Vaz, R. J., Kang, J., Chen, X. L., Preobrazhenskaya, M., Shchekotikhin, A. E., Korolev, A. M., Lysenkova, L. N., Miroshnikova, O. V., Hendrix, J. and Rampe, D. (2003) "Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. "*Bioorg Med Chem Lett,* **13** (10): 1829 - 1835.

41. Egan, W. J. and Lauri, G. (2002) "Prediction of intestinal permeability. "*Adv Drug Deliv Rev,* **54** (3): 273 – 289.

42. Lipinski, C. A., Lombardo, F., Dominy, B. W. and Feeney, P. J. (2001) "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. "*Adv Drug Deliv Rev,* **46** (1-3): 3 - 26.

43. Good, A. C., Krystek, S. R. and Mason, J. S. (2000) "High-throughput and virtual screening: core lead discovery technologies move towards integration. "*Drug Discov Today,* **5** (12 Suppl 1): 61 - 69.

44. Chin, D. N., Chuaqui, C. E. and Singh, J. (2004) "Integration of virtual screening into the drug discovery process. "*Mini Rev Med Chem,* **4** (10): 1053 – 1065. Clark, D. E. and Pickett, S. D. (2000) "Computational methods for the prediction of 'drug-likeness'. "*Drug Discov Today,* **5** (2): 49-58.

45. Jardine I (1980). In *Anticancer Agents based on Natural Product Models*, Cassady JM, Douras JD, Ed., Academic Press: New York, Charpter 9.

46. Stahelin HF & Wartburg AV (1991). The chemical and biological route form podophyllotoxin glucoside to etoposide: Ninth Cain Memorail Award Lecture. *Cancer. Res.* 51: 5-15.

47. Schacter L (1996). Etoposide phosphate: what, why, where, and how? *Seminars in Oncology.* 23: 1-7.

48. Keller-Juslen C, Kuhn M, von Wartburg A & Stahelin H (1971). Synthesis and antimitotic activity of glycosidic lignan derivatives related to podophyllotoxin. *J. Med. Chem.* 14: 936.

49. Weiss SG, Tin-Wa M, Perdue RE Jr, & Farnsworth NR (1975). Potential anticancer agents II: antitumor and cytotoxic lignans from Linum album (Linaceae). *J. Pharm. Sci.* 64: 95.

50. Buss AD & Waigh RD (1995). In *Natural Products as Leads for new Pharmaceuticals.* Wolff, M.E., Ed.; John Wiley & Sons, Inc: New York, Charper 24.

51. Sengupta SK (1995). *Cancer Chemotherapeutic Agents,* ed. by Foye WO De, American Chemical Society: Washington DC, Charper 5. p. 205-217.

52. Garth P & Milles PH (1991). *The toxicity of anticancer drugs*, Pergamon Press: New York.

53. Imbert TF (1998). Discovery of podophyllotoxins. *Biochimie* 80: 207-22.

54. Margolis RL & Wilson L (1978). Opposite end assembly and disassembly of microtubules at steady state in vitro. *Cell* 13: 1.

55. Cortese F, Bhattacharyya B & Wolff P (1977). Podophyllotoxin as a probe for the colchicine binding site of tubulin. *J. Biol. Chem.* 252: 1134.

56. Lyne PD (2002). Structure-based virtual screening: An overview. *Drug Disc. Today* 7: 1047-1055.

57. Oprea T & Matter H (2004). Integrating virtual screening in lead discovery. *Curr. Op. Chem. Biol.* 8:349-358.

58. Shoichet B, McGovern S, Wei B & Irwin J (2002). Lead discovery using molecular docking. *Curr. Op. Chem. Biol.* 6: 439-446.

59. Selassie CD, Mekapati SB & Verma RP (2002). QSAR: Then and Now *Curr. Top. Med. Chem.* 23: 1357-1379

60. Podlipnika C & Bernardib A (2007). Design of a focused virtual library to explore cholera toxin B-site. *Acta Chim. Slov.* 07 (54): 425–436.

61. Zhou RH, Friesner RA, Ghosh A, Rizzo RC, Jorgensen WL & Levy RM (2001). New linear interaction method for binding affinity calculations using a continuum solvent model. *J. Phys. Chem. B.* 105: 10388-10397.

62. Tominaga Y & Jorgensen WL (2004). General model for estimation of the inhibition of protein kinases using Monte Carlo simulations. *J. Med. Chem.* 47: 2534-2549.

63. Leiros HKS, Brandsdal BO, Andersen OA Os V, Leiros I, Helland R, Otlewski J, Willassen NP, & Smalas AO (2004). Trypsin specificity as elucidated by LIE calculations, X-ray structures, and association constant measurements. *Protein Sci.* 13: 1056-1070.

64. Ostrovsky D, Udier-Blagovic M & Jorgensen WL (2003). Analyses of activity for factor Xa inhibitors based on Monte Carlo simulations. *J. Med. Chem.* 46: 5691-5699.

65. Cho SJ, Kashiwada Y, Bastow KF, Cheng YC & Lee KH (1996). Antitumor agents 163 three-dimensional quantitative structure-activity relationship study of 4'-O-Demethylepipodophyllotoxin Analogs Using the Modified CoMFA/q^2-GRS Approach. *J. Med. Chem.* 39: 1396-1405.

66. Thurston LS, Irie H, Tani S, Han FS, Liu ZC, Cheng YC & Lee KH (1986). Antitumour agents. 78. Inhibition of human DNA topoisomerase II by podophyllotoxin and alpha-peltatin analogs. *J. Med. Chem.,* 29: 1547- 1550.

67. Beers SA, ImaKura Y, Dai HJ, Li DH, Cheng YC & Lee KH (1988). Anti-AIDS agents. 29[1]. Anti-HIV activity of modified podophyllotoxin derivatives *J. Natl. Prod.* 51: 901.

68. Wang JZ, Tian X & Tsumura H (1993). Antitumour activity of a new low immunosuppressive derivative of podophyllotoxin (GP-11) and its mechanisms. *Anti-Cancer Drug Des.* 8:193-202.

69. Gordaliza M, Castro MA, Miguel del Corral JM & San Feliciano A (2000). Antitumor properties of podophyllotoxin and related compound. *Curr. Pharm. Des.* 6:1811-1839.

70. Titmuss SJ, Keller PA & Griffith R (1999). Docking experiments in the flexible nonnucleoside inhibitor binding pocket of HIV-1 resease transcriptase. *Bioorg. Med. Chem.* 7: 1163-1170.

CHAPTER II

The Binding Modes and Binding Affinities of Epipodophyllotoxin Derivatives with Human Topoisomerase IIα

Introduction

Human DNA topoisomerase II (TP II) is a ubiquitous nuclear enzyme involved in the control of DNA topology [1-4]. During the catalytic cycle, the enzyme transiently cleaves dsDNA, passess an intact double helix through the break and reseals it. Due to the requirement of such a DNA strand passage activity in a number of critical nuclear processes, including replication, recombination and chromosome segregation, TP II is essential for the survival of proliferating eukaryotic cells [5]. Vertebrates contain two isoforms of the enzyme, TP IIα and β [1].Topoisomerase IIα (TP IIα) level increases during cell proliferation and this enzyme appears to be the isoform involved in mitosis [2]. To maintain DNA integrity during the strand passage event, the enzyme, a homodimer, forms a covalent phosphotyrosyl adduct between the catalytic Tyr^{804} of each monomer and a strand of the duplex. This covalent enzyme-cleaved DNA complex is referred to as the cleavage complex [1–4]. Because the covalent TP II-cleaved DNA complex (referred to as the *cleavage complex*) is normally a short-lived intermediate in the catalytic cycle of the enzyme, it is tolerated by the cell. However, when present in high concentrations, cleavage complexes become potentially toxic, promoting frameshift mutations, permanent double-stranded DNA breaks, illegitimate recombination, and apoptosis [6,7]. The cytotoxic potential of TP II has been exploited clinically by the development of anticancer drugs that generate high levels of covalent enzyme-DNA cleavage complexes [7]. A number of drugs, such as etoposide (VP16), teniposide (VM26), etc. increase TP II-mediated DNA breakage primarily by inhibiting the ability of the enzyme to religate cleaved nucleic acid molecules [8-12]. As a result, they dramatically increase levels of TP II-DNA cleavage complexes (DNA/TP II/drug ternary complex). The resulting DNA strand breaks initiate multiple recombination/repair pathways and trigger cell death pathways [13, 14].

Despite wide use of TP II-targeted drugs as antitumor agents, several limitations hamper their benefits. Efforts to improve their clinical efficacy further by overcoming the drug resistance, myelosuppresion and poor bioavailability problems [15] associated with them, continued to be challenging. Over the years a number of laboratories throughout the world have

been engaged in the synthesis and testing of epipodophyllotoxin derivatives [16-27] to prepare new but more potent and less toxic analogues, that is, with better therapeutic indices. The mechanism of action of any drug is very important in drug development. Generally, the drug compound binds with a specific target, a receptor, to mediate its effects. Therefore, suitable drug–receptor interactions are required for high activity. Understanding the nature of these interactions is very significant. Theoretical calculations, particularly the molecular docking method, seems to be a proper tool for gaining such understanding. The docking results obtained will give information on how the chemical structure of the drug should be modified to achieve suitable interactions and for the rapid prediction and virtual prescreening of anti-tumour activity.

The amino acid sequence of human TP IIα has been known but the three-dimensional structure remains unknown. In this study, therefore we have constructed the 3D structure of TP IIα by homology modeling and epipodophyllotoxin analogues and TP Iiα have been taken for interaction study between them. Of utmost importance in a structure-based drug design is the reliable filtering of putative hits in terms of their predicted binding affinity (scoring problem) which is based on the *in silico*-generated near native protein-ligand configurations (docking problem). Most of scoring functions used in docking programs are designed to predict binding affinity by evaluating the interaction between a compound and a receptor. However, it should be noted that ligand receptor recognition process is determined not only by enthalpic effects but also by entropic effects. Moreover, the scoring functions have a simplified form for the energy function to facilitate high throughput evaluation of a large number of compounds in a single docking run. These functions may be problematic when used with contemporary docking programs, and can result in a decrease of virtual screening accuracy. To overcome this problem, more precise but time consuming computational methodologies are necessary. Here, we have used and evaluated both docking and molecular mechanics based energy minimization of docking complex for computational modeling of epipodophyllotoxin and its derivatives as potent inhibitor of TP IIα.

Computational methods

Sequence analysis

The protein sequence of human TP IIα was obtained from the protein NCBI database. This protein comprises 1531 amino acids. Sequence similarity search with BLAST in Protein Data Bank (pdb) database gives only one similar protein (33.6% identical), Topoisomerase II (pdb ID: 1BJT) from yeast. This structure is determined at 2.50 Å resolution [28]. We performed the pairwise alignment of human TP IIα with IBJT as template using the homology module of PRIME [29]. We removed the mismatched sequence part (1-420) from the whole sequence of human TP IIα and then constructed the three-dimensional structure of human TP IIα. The sequence alignment after removing the part of mismatched sequence is shown in Figure 1. This fragment (430-1214) of human TP IIα consists of drug binding site which is proved to be the binding site for epipodophyllotoxin analogues [30,31].

Figure 1. Alignment of human IIα sequence with template (pdb ID: 1BJT).

Homology model construction

The homology model of human TP IIα was built using Prime [29] accessible through the Maestro interface (Schrodinger Inc.). All water molecules were removed. During the homology model building, Prime keeps the backbone rigid for the cases in which the backbone does not need to be remodeled due to gaps in the alignment. The model was screened for collision free side chain conformation model using a rotamer library database of Prime. Implicit solvation treatment was accompanied and protein model was subjected to energy minimization

```
1BJT          5   KKSDGTPKSRITNYPKLEDANKAGTKEGYKCTLVLTEGDSALSLAVAGLA    54
                  ||....:.:||...:|||:|||.:|.:....:|||:|||||||.:|||:||.
hTP-IIalpha   425 KKCSAVKENRIKGIPKLDDANDAGGRNSTECTLILTEGDSAKTLAVSGLG    474

1BJT          55  VVGRDYYGCYPLRGKKLNVREASADQILKNAEIQAIKKINGLQHREKYED   104
                  |||||.:||.:|||||:|||||||.:||::||||..|.:|:||||::|.|||
hTP-IIalpha   475 VVGRDKYGVFPLRGKILNVREASHKQINENAEINNIIKIVGLQYKKNYED   524

1BJT          105 ---TKSLRYGHLNIKTDQDHDGSHIKGLIINFLESSFLGLLDICQFLLEF  151
                  .|:|||:.:||||||||..|||||||:|||:...::..||...||.|||.|
hTP-IIalpha   525 EDSLKTLRYGKINIKTDQDQDGSHIKGLLINFIHHNMPSLLR-HPFLEEF  573

1BJT          152 ITPIIKVSITKPTKNTIAFYNNPDYEKUREEESHKFTWKQKYYKGLGTSL  201
                  ||||:||||   ..|..:|||::|:::|:|:....:...||.|||||||||.
hTP-IIalpha   574 ITPIVKVS---KNKQEMAFYSLPEFEEUKSSTPNHKKVKVKYYKGLGTST  620

1BJT          202 AQEVREYFSNLDRHLKIFHSLQGNDKDYIDLAFSEKKADDRREULRQYE-  250
                  ::|.:|||::|.||....|......|.:.|.||||||:.||||||...
hTP-IIalpha   621 SKEAKEYFADNKRHRIQFKYSGPEDDAAISLAFSKKQIDDRKEVLTNFME  670

1BJT          251 ----------PGTVL-DPTLKEIPISDFINKELILFSLADNIRSIPNVLD  289
                  |...|  ..|...:.:|||||||||||.:||.||||||:::|
hTP-IIalpha   671 DRRQRKLLGLPEDYLYGQTTTYLTYNDFINKEL ILFSNSDNERSIPSNVD   720

1BJT          290 GFKPGQRKVLYGCFKKNLKSELKVAQLAPYVSECTAYHKGEQSLAQTIIG  339
                  |.|||||||||.|||||:|.|:|||||..:||.:||||||..|||..:||.
hTP-IIalpha   721 GLKPGQRKVLFTCFKPNDKREVKVAQLAGSVAEHSSYHKGEHSLNNTIIN  770

1BJT          340 LAQNFVGSNNIYLLLPNGAFGTRATGGKDAAAARYIYTELNKLTRKIFHP  389
                  ||||||||||:||.:|.|.||||.:..|||.::..|||||:.|.|.:|.|
hTP-IIalpha   771 LAQNFVGSNNLNLLQPIGQFGTRLHGGKDSASPRYIFTKLSSLARLLFPP  820

1BJT          390 ADDPLYKYIQEDEKTVEPEUYLPILPHILVNGAEGIGTGVSTYIPPYNPL  439
                  .||...::|:.:|.:.:||||||:||:||:|||||||||||||||||||.||..
hTP-IIalpha   821 KDDHTLKFLYDDNQRVEPEUYIPIIPHVLINGAEGIGTGNSCKIPNFDVR  870

1BJT          440 EIIKNIRHLNNDEELEQNHPVFRGUTGTIEEIEPLRYRMYGRIEQIGDNV  489
                  ||:.|||.||:.||.:.|.|.||:|||:|||:|.||:.|.:.|.:.:::..
hTP-IIalpha   871 EIVNNIRPLNDGEEPLPNLPSYKNFKGTIEELAPNQYVISGEVAILNSTT  920

1BJT          490 LEITELPARTUTSTIKEYLLLG-LSGNDKIKPVIKOMEEQH-DDNIKFII  537
                  :||:|||.||||.|.|||.::..  |:|.:|.:|.:|.:|.|  |..:|||::
hTP-IIalpha   921 IEISELPVRTUTQTYKEQVLEPMLNGTEKTPPLITDYREYHTDTTVKPYVV  970

1BJT          538 TLSPEEKAKTRKIGFYERFKLISPISLMDNVAFDPHGKIKKYNSVNEILS  587
                  .::.:||:|..:.:|.:|||.:.:.::.||.|.|.::|||||||.:||.
hTP-IIalpha   971 KNTEEKLAEAERVGLHKVFKLQTSLTCNSNVLFDHVGCLKKYDTVLDILR  1020

1BJT          588 KFYYVRLEYYQRKKDHNSERLQUEVEKYSFQVKFIKMIIEKELTVTNKPR  637
                  :|:.:|||.||:.:.:::..|...|::|..||:||||:||..:|...:.:|||
hTP-IIalpha   1021 DFFELRLKYYGLRKEVLLGKLGAESAKLNNQARFILEKIDGKIIIENKPK  1070

1BJT          638 NAIIQELENLGFPPRFNKEGKPYYGSPNDEIAEQINDVKGATSDEEDEESS  687
                  ..:|:|:.|..:           .:|.:...........::||:|||.
hTP-IIalpha   1071 KELIKVLIQRGY-------------DSDPVKAVKEAQQKVPDEENEESD  1107

1BJT          688 HE-DTEN----VINGPEELYGTYEYLLGNRIVSLTKERYQKLLKQXOEXE  732
                  :| :||.   ..:||     |:.|||.|.:|.:||||:.|||.|.:.:.:|||
hTP-IIalpha   1108 NEKETEKSDSVTDSGP-----TFNYLLDNPLWYLTKEKKDELCPLRNEKE  1152

1BJT          733 TELENLLKLSARDIWNTDLKAF--EVGYQEFLQRDAEARGGNVPNKGSKT  780
                  .||:|:::|:..|.:|..|..|  |:::::|.:::.|.....  :|.||.|.
hTP-IIalpha   1153 QELDTLKRKSPSDLUKEDLATFIEELEAVEAKEKQDEQVG--LPGKGGKA  1200

1BJT          781 XGK    783
                  |||
hTP-IIalpha   1201 XGK    1203
```

Figure 1. Alignment of Human TP-IIα sequence with template (pdb ID: 1BJT).

Homology model construction

The homology model of the human TP IIα was built using Prime [29] accessible through the Maestro interface (Schrodinger, Inc.). All water molecules were removed. During the homology model building, Prime keeps the backbone rigid for the cases in which the backbone does not need to be reconstructed due to gaps in the alignment. The model was screened for unfavorable steric contacts and remodeled using a rotamer library database of Prime. Explicit hydrogens were added to the protein and the protein model was subjected to energy minimization

31

using the Macromodel (Prime version 1.5) force-field OPLS-2005. Energy minimization and relaxation of the loop regions were performed using 300 iterations in a simple minimization method. The steepest descent energy minimization was carried out until the energy showed stability in the sequential repetition. Model evaluation was performed in PROCHECK v3.4.4 [32] producing plots that were analyzed for the overall and residue-by-residue geometry. Ramachandran plot [33] provided by the program PROCHECK assured strong confidence for the predicted protein. There were only 0.2 % residues in the disallowed region and 0.8 % residues in generously allowed regions. Nevertheless, PROCHECK assured the reliability of the structure and the protein was subjected to VERIFY3D [34], made available by NIH MBI Laboratory Servers.

Ligand binding site prediction

Several works [35-37] revealed that the epipodophyllotoxin drug, etoposide target the catalytic core domain of human TP IIα at the DNA cleavage-ligation site (drug binding site). Further in vitro drug binding assay revealed two binding sites for etoposide on human TP IIα. One is the lower affinity site in the ATPase domain (266 amino acid; N-terminal fragment), while the second one (430 – 1214 amino acids) binds with higher affinity [30]. Out of the amino acid residues in this catalytic site of TP IIα site, directed mutagenesis study has revealed that Tyr^{804} is the critical residue involved in the binding of etoposide with TP IIα [31]. *In silico* prediction of binding site was done for the modeled structure of human TP IIα using SiteMap (Schrodinger package). SiteMap treat entire proteins to locate binding sites whose size, functionality, and extent of solvent exposure meet user specifications. SiteScore, the scoring function used to assess a site's propensity for ligand binding, accurately ranks possible binding sites to eliminate those not likely to be pharmaceutically relevant. It 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. As the 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. SiteMap evaluates sites using a series of properties. The binding site with highest site score was taken for docking of the epipodophytllotoxin analogues. The algorithm proceeds as

follows: the protein is projected onto a 3D grid with a step size of 1.0 Å; grid points are labeled as protein surface, or solvent using certain rules. A grid point is marked as protein in case of presence of at least one atom within 1.6 Å. After the solvent excluded surface is calculated the surface vertices' coordinates are stored. A sequence of grid points, which starts and ends with surface grid points and which has solvent grid points in between, is called a surface–solvent–surface event. If the number of surface–solvent–surface events of a solvent grid exceeds a minimal threshold of 6, then this grid is marked as pocket. Finally, all pocket grid points are clustered according to their spatial proximity. The clusters are ranked by the number of grid points in the cluster. The top three clusters are retained and their centers of mass are used to represent the predicted pocket sites. The binding pocket obtained by *in silico* studies on human TP IIα was consistent with the site directed mutagenesis studies.

Preparation of the ligands

A total number of 143 epipodophyllotoxin analogues (Table 1 and 2) collected from different published articles [16-27] were used in the study. These compounds were tested for their ability to form intracellular covalent topoisomerase II-DNA complexes. The assay procedures have been described earlier [16]. The activity data were originally expressed as the percentage of cellular protein-DNA complex formed (PCPDCF) and were transformed by taking the natural logarithm of PCPDCF i.e., ln (PCPDCF). These transformed activities were used in the development of prediction model. To generate statistically robust and most importantly, validated models, all compounds in the original data set were divided randomly into 110 molecules in training set and 33 molecules in the test set.

All these epipodophyllotoxin analogues were built from the various scaffold structure (Figure 2) and substitution of functional groups as mentioned in Table 1 and Table 2. We used Maestro-molecular builder for building the scaffold and structural derivatives. LigPrep [40] was used for final preparation of ligands. LigPrep is a utility of Schrödinger software suit that combines tools for generating 3D structures from 1D (Smiles) and 2D (SDF) representation. Ligprep searches for tatomers and steric isomers and performs a geometry minimization of ligands. Then The ligands were energy minimized using Macromodel module of Schrodinger with default parameters and applying molecular mechanics force fields (MMFFs). Truncated

Newton Conjugate Gradient (TNCG) minimization method was used with 500 iterations and convergence threshold of 0.05 kJ/mol.
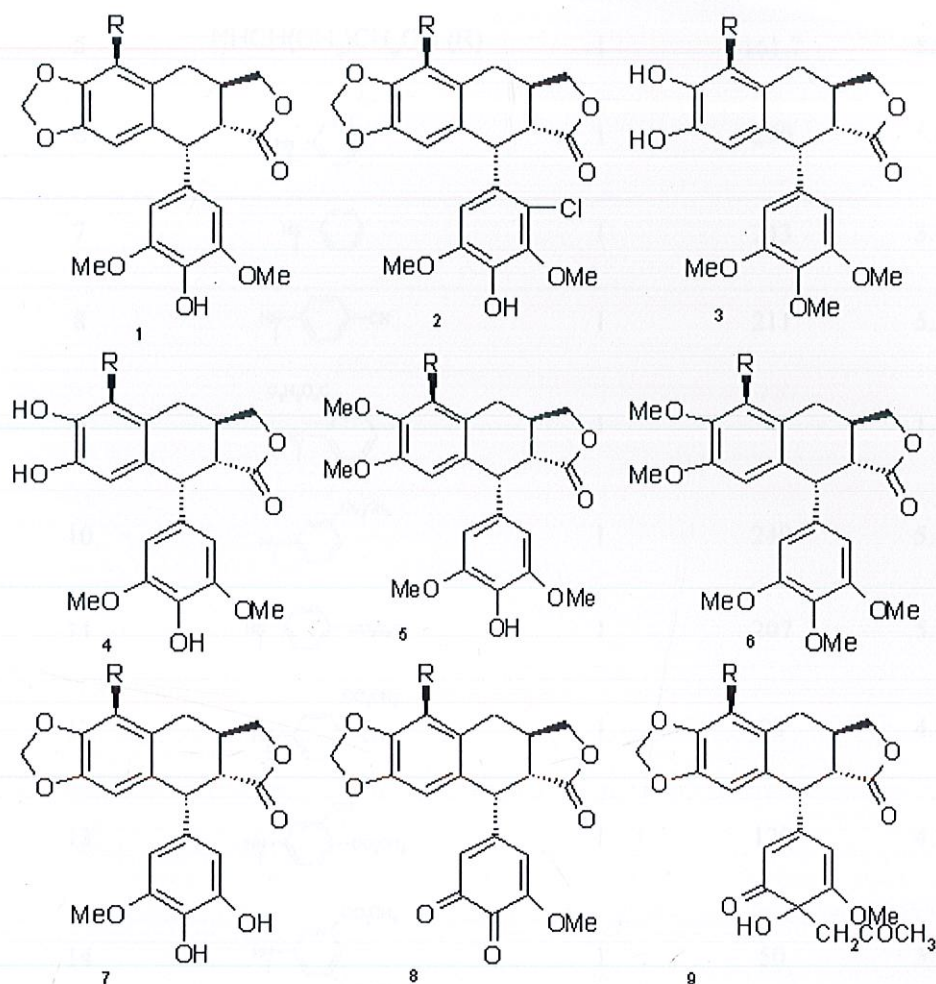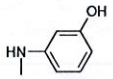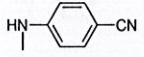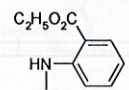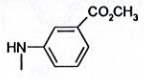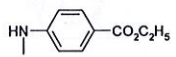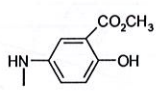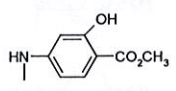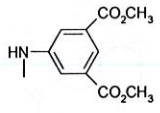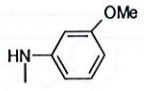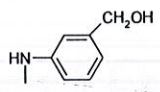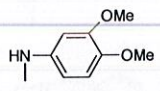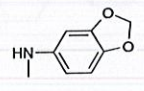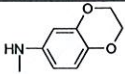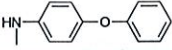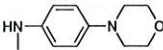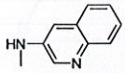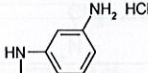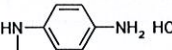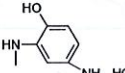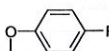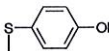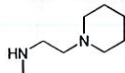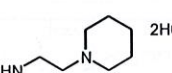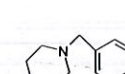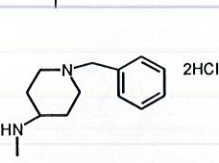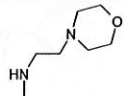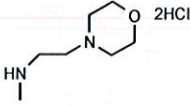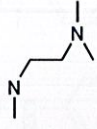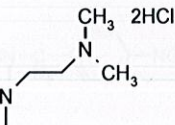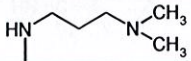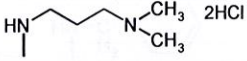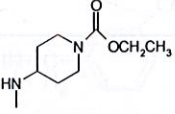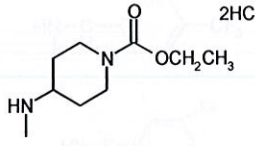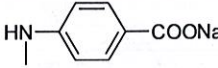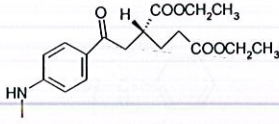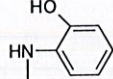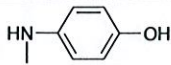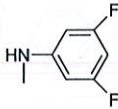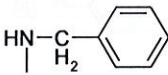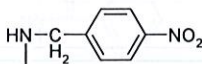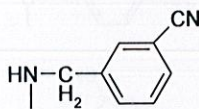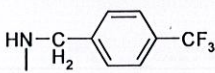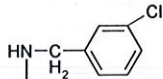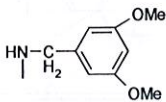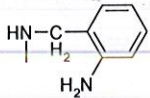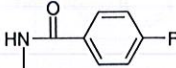

Figure 2. The various scaffold structures used for building the epipodophyllotoxin analogues

Table 1. Epipodophyllotoxin analogues (training set) with binding affinity (PCPDCF) against the human Topoisomerase IIα.
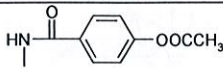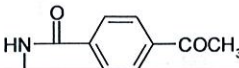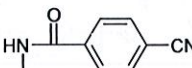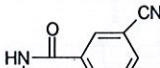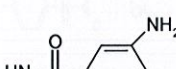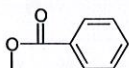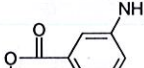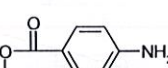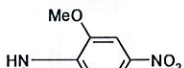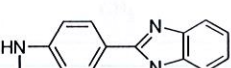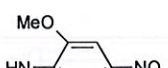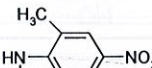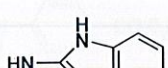
Ligand	R	Structure type	PCPDCF	Ln (PCPDCF)
1	—OH	1	42.2	3.742
2	—NHCH$_2$CH$_{20}$CH$_3$	1	110.8	4.708

34
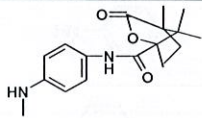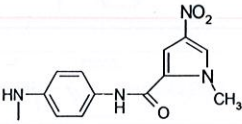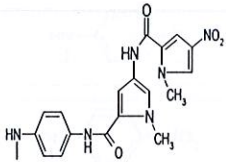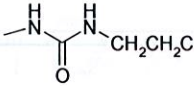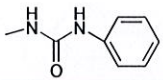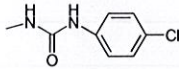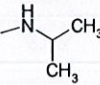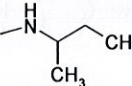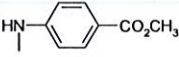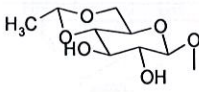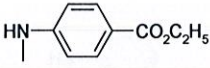
#	Structure		Value	Value
3	—NHCH$_2$CH=CH$_2$	1	84.1	4.432
4	—NHCH$_2$CH(OH)CH$_3$ (R)	1	167.2	5.119
5	—NHCH(CH$_3$)CH$_2$OH (R)	1	161.7	5.086
6		1	290	5.670
7		1	243	5.493
8		1	211	5.352
9		1	4	1.386
10		1	249	5.517
11		1	207	5.333
12		1	83	4.419
13		1	129	4.860
14		1	50	3.912
15		1	104	4.644
16		1	235	5.460
17		1	180	5.193
18		1	164	5.100
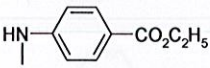
No.	Structure			
19		1	279	5.631
20		1	97	4.575
21		1	140	4.942
22		1	97	4.575
23		1	123	4.812
24		1	140	4.942
25		1	330	5.799
26		1	11	2.398
27		1	57	4.043
28		1	34	3.526
29		1	10	2.303
30		1	190	5.247
31		1	183	5.209
32		1	83	4.419
33		1	172	5.147

No.	Structure			
34	morpholine, HN–CH₃	1	77	4.344
35	morpholine, 2HCl	1	140	4.942
36	N, N	1	203	5.313
37	CH₃, N, CH₃, 2HCl	1	183	5.209
38	HN, N(CH₃)(CH₃)	1	186	5.226
39	HN, N(CH₃)(CH₃), 2HCl	1	179	5.187
40	OCH₂CH₃, HN	1	17	2.833
41	OCH₂CH₃, 2HCl, HN	1	138	4.927
42	HN–COONa	1	6.9	1.932
43	COOCH₂CH₃, COOCH₂CH₃, HN	1	83	4.419
44	HO, HN	1	151	5.017

No.	Structure		Value1	Value2
45	HN—⟨ ⟩—OH	1	211	5.352
46	HN—⟨ ⟩(F,F)	1	115	4.745
47	HN—⟨ ⟩(Cl)	1	32	3.466
48	HN—CH₂—⟨ ⟩	1	181	5.198
49	HN—CH₂—⟨ ⟩—NO₂	1	216	5.375
50	HN—CH₂—⟨ ⟩(NO₂)	1	130	4.868
51	HN—CH₂—⟨ ⟩(O₂N)	1	144	4.970
52	HN—CH₂—⟨ ⟩(CN)	1	225	5.416
53	HN—CH₂—⟨ ⟩—CF₃	1	99	4.595
54	HN—CH₂—⟨ ⟩(Cl)	1	159	5.069
55	HN—CH₂—⟨ ⟩(OMe,OMe)	1	144	4.970
56	HN—CH₂—⟨ ⟩(H₂N)	1	184	5.215
57	HN—C(=O)—⟨ ⟩—F	1	117	4.762

38

No.	Structure			
58		1	137	4.920
59		1	124	4.820
60		1	159	5.069
61		1	149	5.004
62		1	149	5.004
63		1	94	4.543
64		1	100	4.605
65		1	94	4.543
66		1	83	4.419
67		1	128	4.852
68		1	4.4	1.482
69		1	3.5	1.253
70		1	58	4.060

No.	Structure			
71	4-(methylamino)phenyl amide of pivalic-type ester (tert-butyl, O=)	1	88	4.477
72	HN(CH₃)—C₆H₄—NH—C(=O)— N-methylpyrrole with NO₂	1	100	4.605
73	bis-pyrrole amide with NO₂, N—CH₃	1	26	3.258
74	CH₃—NH—C(=O)—NH—CH₂CH₂Cl	1	143	4.963
75	CH₃—NH—C(=O)—NH—C₆H₅	1	148	4.997
76	CH₃—NH—C(=O)—NH—C₆H₄—Cl	1	125	4.828
77	CH₃—NH—CH(CH₃)CH₃	1	109	4.691
78	CH₃—NH—CH(CH₃)CH₂CH₃	1	73	4.290
79	HN(CH₃)—C₆H₄—CO₂CH₃	1	207	5.333
80	methyl glycoside (H₃C, HO, OH)	2	6.1	1.808
81	—OH	2	15.6	2.747
82	HN(CH₃)—C₆H₄—CO₂C₂H₅	3	22	3.091

No.	Structure			
83	HN(CH₃)–C₆H₅	4	4	1.386
84	HN(CH₃)–C₆H₄–NO_2	4	99	4.595
85	HN(CH₃)–C₆H₄–$CO_2C_2H_5$	4	138	4.927
86	HN(CH₃)–C₆H₄–F	4	52	3.951
87	HN(CH₃)–C₆H₄–NO_2	5	75	4.317
88	HN(CH₃)–C₆H₄–$CO_2C_2H_5$	5	127	4.844
89	HN(CH₃)–C₆H₄–CN	5	125	4.828
90	HN(CH₃)–C₆H₄–F	5	108	4.682
91	HN(CH₃)–C₆H₄–NO_2	3	23	3.135
92	HN(CH₃)–C₆H₄–NO_2	6	8	2.079
93	HN(CH₃)–C₆H₄–$CO_2C_2H_5$	6	9	2.197
94	HN(CH₃)–C₆H₄–CN	6	12	2.485
95	HN(CH₃)–C₆H₄–F	6	8	2.079
96	HN(CH₃)–C₆H₄–F	7	117	4.762
97	HN(CH₃)–C₆H₅	7	105	4.654

41
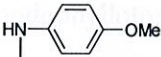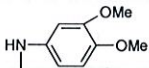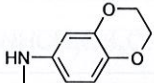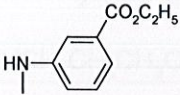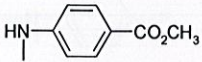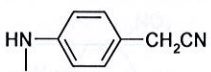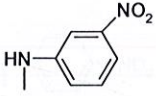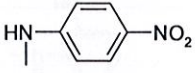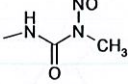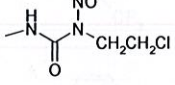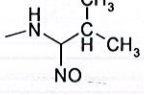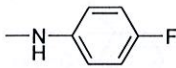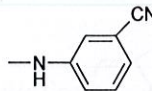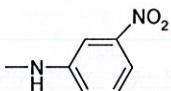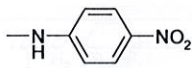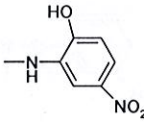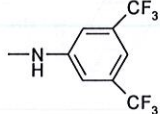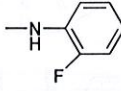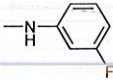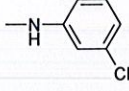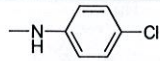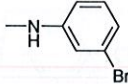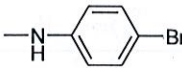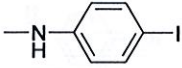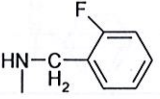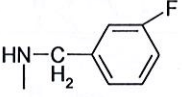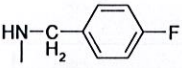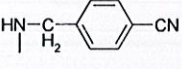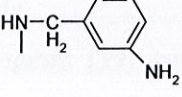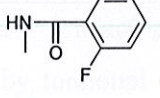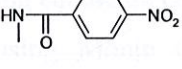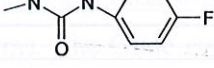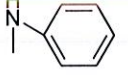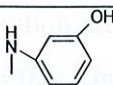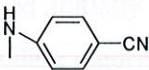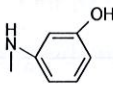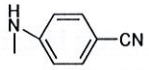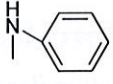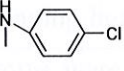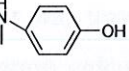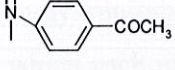
	R			
98	HN—⬡—OMe	7	96	4.564
99	HN—⬡(OMe)(OMe)	7	69	4.234
100	HN—benzodioxane	7	119	4.779
101	HN—⬡—CO$_2$C$_2$H$_5$	7	94	4.543
102	HN—⬡—CO$_2$CH$_3$	7	175	5.165
103	HN—⬡—CN	7	146	4.984
104	HN—⬡—CH$_2$CN	7	109	4.691
105	HN—⬡—NO$_2$	7	75	4.317
106	HN—⬡—NO$_2$	7	200	5.298
107	—N(H)—C(O)—N(NO)—CH$_3$	8	41	3.714
108	—N(H)—C(O)—N(NO)—CH$_2$CH$_2$Cl	8	7	1.946
109	—N(H)—CH(CH$_3$)—CH(NO)(CH$_3$)	9	1	0.000
110	—NH$_2$	1	36.4	3.595

Table 2. Substituted epipodophyllotoxin derivatives (Test set) with anti-tumour activities against the human Topoisomerase IIα used in this work.

Ligands	R	Structure type	PCPDCF	Ln(PCPDCF)
1	—NHCH$_2$CH$_2$OH	1	121.4	4.799
2	—NHCH$_2$CH$_2$CH$_3$	1	69.7	4.244
3	—NHCH$_2$CH$_2$CH$_2$OH	1	89.2	4.491
4		1	213	5.361
5		1	137	4.920
6		1	230	5.438
7		1	323	5.778
8		1	15	2.708
9		1	21	3.045
10		1	121	4.796
11		1	158	5.063
12		1	51	3.932

No.	Structure			
13	(4-Cl phenyl)-NH-CH structure	1	99	4.595
14	(3-Br phenyl)-NH-CH structure	1	62	4.127
15	(4-Br phenyl)-NH-CH structure	1	179	5.187
16	(4-I phenyl)-NH-CH structure	1	64	4.159
17	(2-F benzyl)-HN-CH₂ structure	1	126	4.836
18	(3-F benzyl)-HN-CH₂ structure	1	216	5.375
19	(4-F benzyl)-HN-CH₂ structure	1	169	5.130
20	(4-CN benzyl)-HN-CH₂ structure	1	284	5.649
21	(3-NH₂ benzyl)-HN-CH₂ structure	1	191	5.252
22	(2-F benzamide)-HN structure	1	128	4.852
23	(4-NO₂ benzamide)-HN structure	1	160	5.075
24	(4-F phenyl urea) structure	1	118	4.771
25	(phenyl)-N structure	3	9	2.197

No.	Structure			
26	(N-H, phenol OH structure)	3	4	1.386
27	(N-H, –CN structure)	4	62	4.127
28	(N-H, phenol OH structure)	4	18	2.890
29	(N-H, –CN structure)	3	33	3.497
30	(N-H, phenyl structure)	7	128	4.852
31	(N-H, –Cl structure)	7	77	4.344
32	(N-H, –OH structure)	7	83	4.419
33	(N-H, –COCH$_3$ structure)	7	147	4.990

Docking of the ligands

The Glide program [39] was used for docking study. The Glide docking algorithm performs a series of hierarchical searches for locations of possible ligand affinity within the binding site of a receptor. A rough positioning and scoring algorithm is applied during the initial search step, followed by torsional energy optimization on an OPLS-AA non-bonded potential energy grid for enduring candidate poses. The pose conformations of the very best candidates are further refined by using Monte Carlo sampling. Selection of the final docked pose is accomplished using a Glide score, which is a model energy function that combines empirical and force field based terms. The Glide score is a modified and extended version of the ChemScore function [40].

All the ligands were docked to the human TP IIα receptor using Glide 4.0. After ensuring that protein and ligands were in correct form for docking, the receptor-grid files were generated

45

using grid-receptor generation program by selecting the drug binding site, using van der Waals scaling of the receptor at 0.4. The default size was used for the bounding and enclosing boxes. The ligands were docked initially using the "standard precision" method and further refined using "xtra precision" Glide algorithm. For the ligand docking stage, van der Waals scaling of the ligand was set at 0.5. Out of the 50,000 poses that were sampled, 4,000 were taken through minimization (conjugate gradients 1,000) and the 30 structures having the lowest energy conformations were further evaluated for the favorable Glide docking score. A single best conformation for each ligand was considered for further analysis.

Molecular Mechanics and Free Energies of Binding

After obtaining preferable binding structure from docking simulation, the complex was partially minimized by relaxing ligand and atoms of side chains that are within 7Å away from the ligand while all other atoms were fixed. Bimolecular Association with Energetics (eMBrAcE) developed by Schrödinger was used for physics based rescoring procedure [41]. The eMBrAcE (*MacroModel v9.1*) program calculates binding energies between ligands and receptors using molecular mechanics energy minimization for docked conformations. eMBrAcE applies multiple minimizations, during which each of the specified pre-positioned ligand is minimized with the receptor. For each ligand, the protein-ligand complex ($E_{\text{lig-prot}}$), the free protein (E_{prot}), and the free ligand (E_{lig}) were all subjected to energy minimization in implicit solvent (generalized Born) [42,43]. It uses traditional molecular mechanics (MM) methods to calculate ligand-receptor interaction energies (G_{ele}, G_{vdW}, G_{solv}), with a Gaussian smooth dielectric constant function method [44] for electrostatic part of solvation energy and solvent-accessible surface for the nonpolar part of solvation energy. A conjugate gradient minimization protocol was used in all minimization. The nonpolar solvent-accessible surface area (SASA) of solvation energy was calculated using Qikprop program. The total free energy of binding, ΔG_{cald} is calculated using linear optimized multiple regression as follows:

$$\Delta G_{\text{cald}} = C + \alpha(\Delta G_{\text{vdW}}) + \beta (\Delta G_{\text{ele}}) + \gamma(\Delta G_{\text{solv}}) + \delta(\text{SASA})$$

where α, β, γ and δ are the coefficients for van der Walls, electrostatic, solvation energy terms and SASA, respectively; C is a constant. The approach is simple, fast and straightforward. It benefits the calculation of relative binding affinity needed to evaluate the activity of large set of molecules in rational drug design.

46

The eMBrAcE calculation was performed using the Ligand & Structure-Based Descriptors (LSBD) application of the Schrödinger software package. This calculation was applied to the ligand-receptor complex structures obtained from Glide docking.

The predictive capabilities of the proposed models were determined using leave-one-out cross validation method. The cross validation regression coefficient (q^2_{cv}) was calculated by the following equation.

$$q^2_{cv} = 1 - \frac{PRESS}{TOTAL} = 1 - \frac{\sum_{i-1}^{n}\left(y_{exp} - y_{pred}\right)^2}{\sum_{i-1}^{n}\left(y_{exp} - \overline{y}\right)^2}$$

where, y_{pred}, y_{exp} and $\overline{y}$ are the predicted, experimental and mean values of experimental activity, respectively. Also the accuracy of the prediction of the developed models were validated by F-value, r^2 and r^2_{adj}. A large F indicates that the model is fit and not a chance occurrence.

Results and Discussion

The atomic coordinates of human TP IIα was not available in Protein Data Bank, which necessitated the developing of a protein model. The final model, which we took for further analysis, consisted of 789 amino acid residues. We used both PROCHECK and the VERIFY3D softwares to check the quality of the modeled protein. Ramachandran plot obtained from the program PROCHECK, which checks the stereochemical quality of a protein structures, producing a number of postscript plots, analyzing its overall and residue-by-residue geometry, assured the reliability of the modeled protein with 91.3% residues in most allowed region and 7.8% in additional allowed region. There were only 0.2% residues in disallowed region and 0.8% in generously allowed region. The assessment with VERIFY3D, which derives a ''3D-1D'' profile based on the local environment of each residue, described the statistical preferences for: the area of the residue that is buried, the fraction of side-chain area that is covered by polar atoms (oxygen and nitrogen), and the local secondary structure. It also substantiated the reliability of the three dimensional structure. The residues that deviated from the standard conformational angles of Ramachandran plot were the members of N terminal domain of the

47

protein. This was an ignorable condition since the N-terminal end was not critical in our study. The distance of these residues to the active site residues also were found to be more than 10 Å, which suggested that those residues would interfere little with the binding of ligands in the drug binding site region of TP IIα. The structural comparison of template protein and human TP IIα model showed significant similarity in overall structure and binding site residues (Figure 3). Active site was identified considering the amino acids which are essential for binding of etoposide with TP IIα from experimental study. The output from the Sitemap program (Figure 4) showed coherent active sites for the target protein as reported in site directed mutagenesis study [24].
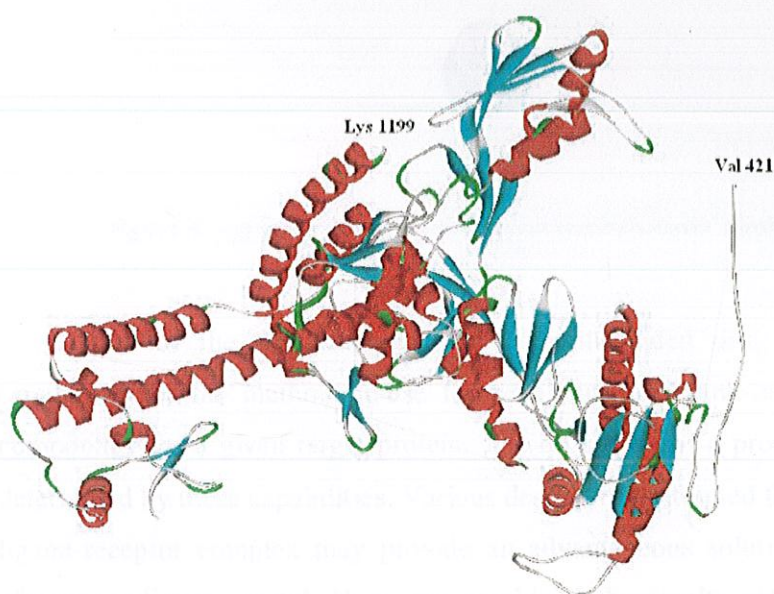
Figure 3. The structural comparison of template (pdb ID:1BJT) and modeled structure of human TP-IIα.
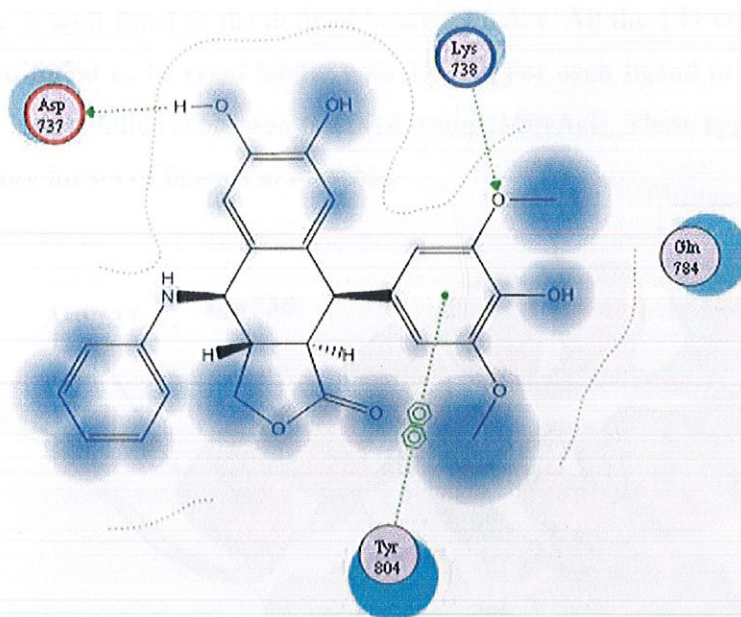
Figure 4: Ligplot of human TP-IIα-epipodophyllotoxin binding site.

One of the key challenges in computer-aided drug discovery is to maximize the capabilities of the method in use for predicting and rank-ordering the binding affinities of compounds for a given target protein. The efficiency of a prediction method is predominantly determined by these capabilities. Various descriptors extracted from the structural information on ligand-receptor complex may provide an advantageous solution to create a reliable binding-affinity-prediction model. Here, we combined the results obtained from a standard docking protocol with data from structure based calculations of free energy of binding (eMBrAcE) and then investigated the utility of both the methods on the virtual screening efficiency for epipodophyllotoxin derivatives.

Docking simulation of epipodophyllotoxin derivatives to the homology modeled TP IIα was performed using the Glide program (Schrodinger package). All the 143 epipodophyllotoxin ligands with known binding affinity expressed in terms of cellular protein-DNA complex formations (PCPDCF) were docked into the defined binding site. The binding mode of a cognate ligand within the binding site is represented in Figure 5. In this figure we can observe that the

49

molecule is well fitted to the defined binding pocket. All the 143 epipodophyllotoxin analogues were also found to be good binder with TP IIα. For each ligand in the virtual library, the pose with the lowest Glide score was rescored using eMBrAcE. These approaches predict the binding free energy for set of ligands to receptor.
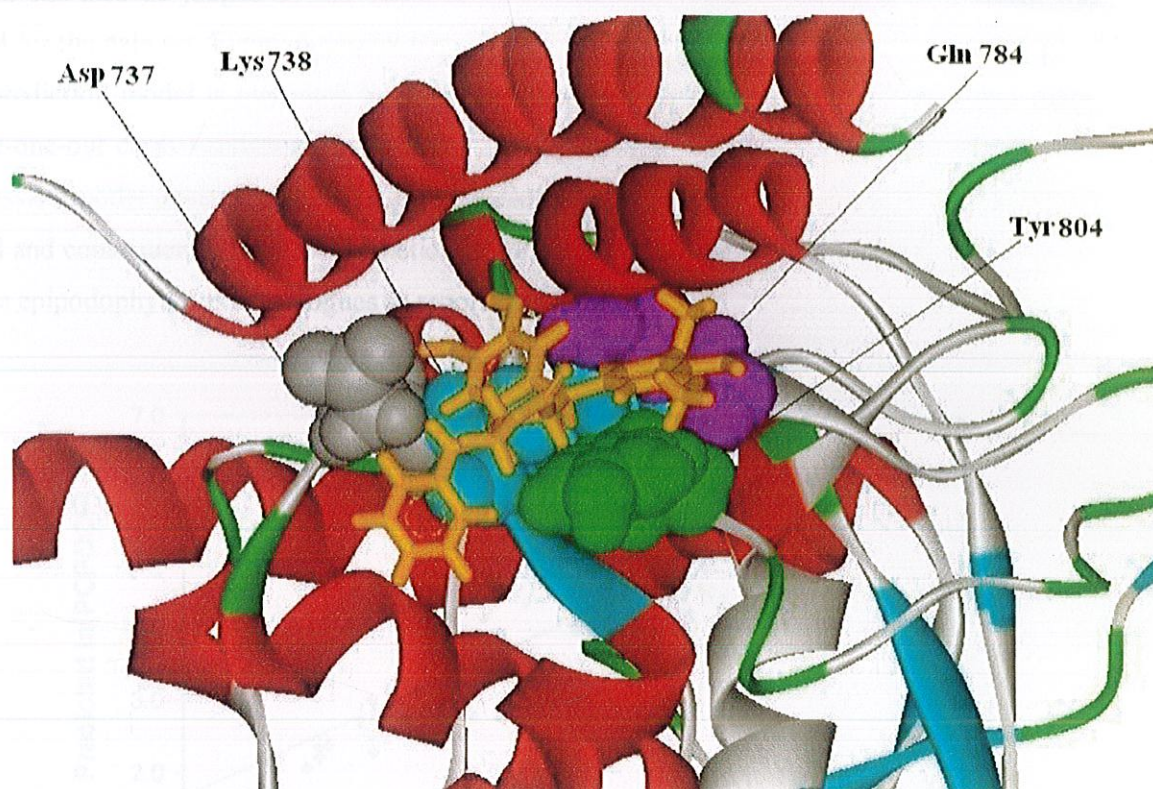
Figure: 5 Binding mode of epipodophyllotoxin derivative(5) within the binding site of Human TopoII.

Building models for prediction of binding affinity using Glide score

Prediction model for prediction of binding of epipodophyllotoxin with TP IIα were built by considering the Glide score (GScore) as a descriptor. The equation (1) of the model and the corresponding statistics are shown below:

$$Ln\,(PCPDCF) = 0.421\,(\pm\,0.302) - 1.479\,(\pm\,0.111) \times Gscore$$

$$(N = 110;\ r^2 = 0.624;\ s = 0.739;\ F = 179.02;\ r^2_{cv} = 0.606;\ PRESS = 61.76)$$

50

The root mean square error (RMSE) between the experimental PCPDCF and the predicted PCPDCF obtained by the regression model was 0.624, which is an indicator of the robustness of the fit and suggested that the calculated PCPDCF based on Glide score is reliable. The quality of the fit can also be judged by the value of the squared correlation coefficient (r^2), which was 0.624 for the data set. Figure 6 graphically shows the quality of fit. The statistical significance of the prediction model is evaluated by the correlation coefficient r^2, standard error, F-test value, leave-one-out cross-validation coefficient r^2_{cv} and predictive error sum of squares PRESS. The regression model developed in this study is statistically ($r^2_{cv} = 0.606$, $r^2 = 0.624$, F = 179.02) best fitted and consequently used for prediction of formation of complexes with TP IIα (ln PCPDCF) of the epipodophyllotoxin analogues as reported in Table 3.

Figure:6 Models for predic ng binding a nity (ln PCPDCF) of the epipodophyllotoxin deriva ves based on Glide score for the training set .

Table 3. Predicted ln (PCPDCF) of epipodophyllotoxin analogues using Glide score (XP) as a descriptor for Training set compounds.

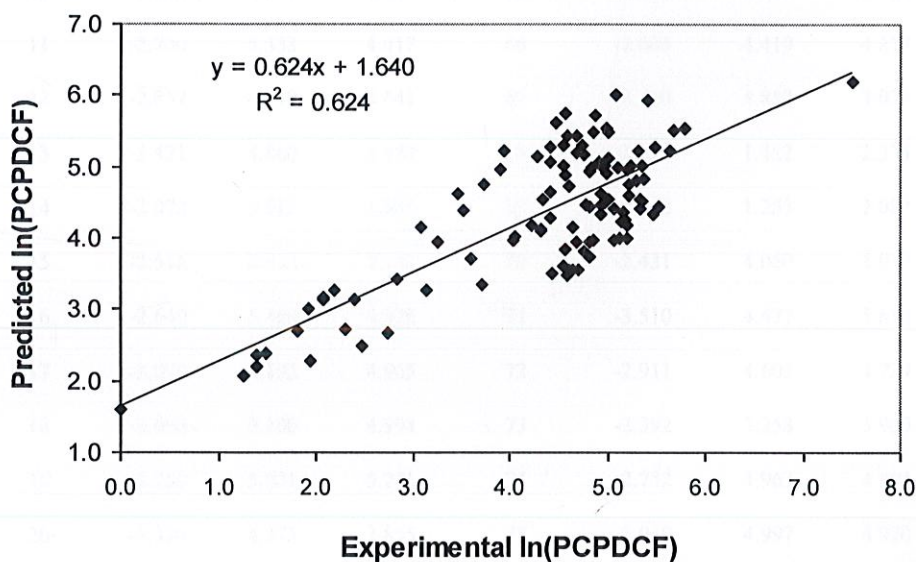| Ligand	GScore	Expt.	[a]Pred.	Ligand	GScore	Expt.	[a]Pred.
1	-2.927	3.742	4.753	56	-2.819	5.215	4.593
2	-2.124	4.708	3.565	57	-2.300	4.762	3.825
3	-2.087	4.432	3.509	58	-3.140	4.920	5.068
4	-2.413	5.119	3.992	59	-2.390	4.820	3.958
5	-3.771	5.086	6.003	60	-2.700	5.069	4.417
6	-3.432	5.670	5.500	61	-2.780	5.004	4.535
7	-3.290	5.493	5.290	62	-3.430	5.004	5.497
8	-3.110	5.352	5.024	63	-3.100	4.543	5.009
9	-1.310	1.386	2.360	64	-2.120	4.605	3.559
10	-2.720	5.517	4.447	65	-3.300	4.543	5.305
11	-2.700	5.333	4.417	66	-2.605	4.419	4.277
12	-2.852	4.419	4.641	67	-2.400	4.852	3.973
13	-3.421	4.860	5.484	68	-1.320	1.482	2.375
14	-3.070	3.912	4.965	69	-1.122	1.253	2.082
15	-2.516	4.644	4.145	70	-2.431	4.060	4.019
16	-2.640	5.460	4.328	71	-3.510	4.477	5.616
17	-3.070	5.193	4.965	72	-2.911	4.605	4.729
18	-3.090	5.100	4.994	73	-2.392	3.258	3.960
19	-3.250	5.631	5.231	74	-2.752	4.963	4.494
20	-2.320	4.575	3.855	75	-3.040	4.997	4.920
21	-2.650	4.942	4.343	76	-2.720	4.828	4.447
22	-3.604	4.575	5.755	77	-3.381	4.691	5.425
23	-2.248	4.812	3.749	78	-3.200	4.290	5.157
24	-3.114	4.942	5.030	79	-2.783	5.333	4.540
25	-3.450	5.799	5.527	80	-1.540	1.808	2.700
26	-1.830	2.398	3.129	81	-1.520	2.747	2.671
27	-2.394	4.043	3.964	82	-2.531	3.091	4.166
28	-2.681	3.526	4.389	83	-1.200	1.386	2.197

29	-1.560	2.303	2.730	84	-3.396	4.595	5.447
30	-3.100	5.247	5.009	85	-3.120	4.927	5.039
31	-2.410	5.209	3.988	86	-3.251	3.951	5.232
32	-3.138	4.419	5.065	87	-2.502	4.317	4.123
33	-2.585	5.147	4.247	88	-3.093	4.844	4.998
34	-2.793	4.344	4.555	89	-3.055	4.828	4.942
35	-2.737	4.942	4.472	90	-2.389	4.682	3.957
36	-3.240	5.313	5.216	91	-1.930	3.135	3.277
37	-2.549	5.209	4.194	92	-1.861	2.079	3.175
38	-2.930	5.226	4.757	93	-1.929	2.197	3.276
39	-2.587	5.187	4.250	94	-1.401	2.485	2.494
40	-2.030	2.833	3.425	95	-1.839	2.079	3.143
41	-2.843	4.927	4.629	96	-3.217	4.762	5.182
42	-1.741	1.932	2.998	97	-2.119	4.654	3.558
43	-3.274	4.419	5.267	98	-3.010	4.564	4.876
44	-3.182	5.017	5.130	99	-2.541	4.234	4.182
45	-2.770	5.352	4.521	100	-2.742	4.779	4.479
46	-3.303	4.745	5.310	101	-2.155	4.543	3.611
47	-2.850	3.466	4.639	102	-2.662	5.165	4.361
48	-3.900	7.502	6.193	103	-3.460	4.984	5.543
49	-2.987	5.375	4.841	104	-3.224	4.691	5.192
50	-3.577	4.868	5.715	105	-2.481	4.317	4.094
51	-3.140	4.970	5.068	106	-2.958	5.298	4.799
52	-3.730	5.416	5.941	107	-1.977	3.714	3.347
53	-2.070	4.595	3.484	108	-1.256	1.946	2.280
54	-2.401	5.069	3.975	109	-0.800	0.000	1.605
55	-3.110	4.970	5.024	110	-2.224	3.595	3.713

Linear optimization of energy parameters vs binding affinity

One docking structure from each molecule docking result was picked up as final docked structure in TP IIα and was imported into eMBrAcE for further calculations. As the Glide treats a receptor rigidly during docking simulation, an energy minimization was performed to the docked complex. A vdW energy and electrostatic energy between ligand and receptor as well as solvation energy were calculated for each minimized complex. Also solvent accessible surface area (SASA) change was calculated using Qikprop. All these energies are listed in Table 4. A scheme similar to Linear Response was used to develop a free energy of binding (FEB) relationship based on these energies, which can express the activity of these analogues. The final calculated FEB and activity (ln PCPDCF) of these analogues are listed in Table 4. The calculated activity has good correlation to the actual activity. The calculated FEB represents the actual activity well. Several papers have been reported, in which a reasonable correlation has been shown between calculated FEB and biological activity for a small set of ligands. Although these energy components are added directly together in most of these applications, it is still a challenge to apply these methods into large set of ligands. Normally, these different energy components (vdW, electrostatic, solvation) were calculated using more than one method. To the same set of structure, different force field or different methods will produce different values of energy. This suggests that these energy components need to be scaled before an equation is obtained to get a better expression for these energy components. A set of weights can be used to scale these energies to get free energy expression by linearly combining these energies. Some scoring functions [24] used this strategy, which were optimized using a test set of molecules. In the work, a linear combination strategy was used to express FEB by four energy components calculated from different methods. An expression of free energy, whose weight coefficients were optimized by a multiple regression, was obtained and successfully predicted the activity of a large set of ligands.

Table 4. Calculated Energies and Estimated Binding Free Energy (ΔG_{cald}) of epipodophyllotoxin analogues in the Training set .

Ligand	Expt.	ΔG_{vdW}	ΔG_{ele}	ΔG_{solv}	SASA	ΔG_{cald}[1]	Pred.[2]

54

	ln(PCPDCF)	(kcal/mol)	(kcal/mol)	(kcal/mol)		(kcal/mol)	ln(PCPDCF)
1	3.742	-120.4	-191.1	301.4	561.7	-5.498	4.032
2	4.708	-91.1	445.7	-444.1	639.8	-6.682	4.900
3	4.432	-80.4	300.7	-210.3	643.9	-6.099	4.472
4	5.119	-51.7	324.9	-297.8	236.4	-7.832	5.743
5	5.086	-68.3	275.0	-244.9	340.0	-7.366	5.401
6	5.670	-108.1	-26.4	139.7	184.3	-7.082	5.194
7	5.493	-143.1	45.6	91.5	400.7	-6.323	4.637
8	5.352	-129.0	-18.1	162.7	396.3	-6.216	4.559
9	1.386	-192.8	249.9	796.3	732.6	-1.934	1.418
10	5.517	-151.1	-48.1	153.0	275.7	-6.731	4.936
11	5.333	-134.2	-142.6	275.5	199.2	-6.751	4.950
12	4.419	-84.0	-40.9	106.1	728.3	-5.351	3.924
13	4.860	-125.4	57.8	36.5	349.5	-6.700	4.913
14	3.912	-106.5	-65.9	150.3	853.5	-4.783	3.507
15	4.644	-62.7	-257.1	280.7	317.0	-6.624	4.857
16	5.460	-143.9	65.8	55.3	416.0	-6.363	4.666
17	5.193	-41.2	-528.6	583.6	65.5	-6.965	5.108
18	5.100	-107.8	56.4	95.1	281.8	-6.725	4.931
19	5.631	-154.8	86.3	71.0	235.5	-6.882	5.047
20	4.575	-172.1	189.5	-5.3	106.2	-7.385	5.416
21	4.942	-116.1	243.9	-602.7	774.6	-7.264	5.327
22	4.575	-36.3	336.2	-314.0	668.1	-6.363	4.666
23	4.812	-139.9	641.2	-531.5	737.0	-6.204	4.549
24	4.942	-57.5	-388.0	469.8	291.4	-6.282	4.607
25	5.799	-82.8	957.8	-841.1	392.7	-7.929	5.814
26	2.398	-138.3	298.2	577.5	692.0	-2.849	2.089
27	4.043	-102.6	190.5	68.5	347.4	-6.296	4.617
28	3.526	-104.3	-167.0	272.1	665.0	-5.204	3.816

55

29	2.303	-76.0	-55.5	738.7	715.8	-2.985	2.189
30	5.247	-53.2	949.4	-867.3	724.7	-6.903	5.062
31	5.209	-131.4	1381.9	-1186.5	767.6	-6.955	5.100
32	4.419	-87.8	733.1	-708.1	722.1	-6.765	4.961
33	5.147	-93.1	966.1	-837.8	830.4	-6.354	4.659
34	4.344	-104.4	612.6	-484.1	659.7	-6.379	4.678
35	4.942	-67.2	344.5	-260.1	698.1	-6.011	4.408
36	5.313	-57.4	-15.3	76.1	651.0	-5.698	4.178
37	5.209	-100.7	-245.1	270.0	366.1	-6.441	4.724
38	5.226	-73.9	951.0	-934.1	694.1	-7.252	5.318
39	5.187	-128.5	1359.6	-1200.9	707.7	-7.273	5.334
40	2.833	-64.5	460.7	374.1	771.1	-3.038	2.228
41	4.927	-102.5	578.8	-461.2	798.9	-5.880	4.312
42	1.932	-68.6	-1185.2	949.2	722.6	-4.716	3.459
43	4.419	-82.1	-346.4	419.7	315.2	-6.282	4.607
44	5.017	-144.7	-303.5	423.6	116.8	-6.826	5.005
45	5.352	-35.2	-162.8	211.5	312.2	-6.712	4.922
46	4.745	-137.7	13.4	149.0	287.7	-6.574	4.821
47	3.466	-126.5	153.5	-4.9	638.0	-5.629	4.128
48	7.502	-69.6	136.2	-797.9	203.0	-10.301	7.554
49	5.375	-92.3	436.7	-317.1	322.7	-7.323	5.370
50	4.868	-76.3	186.4	-159.0	306.0	-7.349	5.389
51	4.970	-86.7	601.9	-403.1	241.1	-7.571	5.552
52	5.416	-94.2	507.7	-411.7	231.2	-7.848	5.755
53	4.595	-75.3	516.6	-416.9	765.3	-5.986	4.389
54	5.069	-108.2	499.7	-341.8	330.5	-7.238	5.308
55	4.970	-119.9	323.0	-243.1	309.4	-7.324	5.371
56	5.215	-85.3	536.1	-424.7	473.2	-6.990	5.126
57	4.762	-89.4	185.7	-17.4	334.7	-6.693	4.908

58	4.920	-182.8	-58.7	241.7	335.2	-6.183	4.534
59	4.820	-110.5	-104.7	173.9	393.9	-6.390	4.686
60	5.069	-77.5	-57.9	120.2	404.9	-6.475	4.748
61	5.004	-108.0	-113.8	251.2	258.5	-6.589	4.832
62	5.004	-145.3	74.1	105.5	231.2	-6.798	4.985
63	4.543	-56.2	-174.0	224.3	274.1	-6.808	4.992
64	4.605	-102.5	-272.1	359.0	321.3	-6.314	4.630
65	4.543	-88.7	-212.1	287.1	233.3	-6.774	4.967
66	4.419	-49.8	-62.8	119.7	241.3	-7.080	5.192
67	4.852	-133.3	263.5	-110.9	219.4	-7.254	5.320
68	1.482	-216.2	238.6	884.6	595.9	-2.083	1.527
69	1.253	-251.8	-326.9	906.4	761.0	-2.680	1.965
70	4.060	-105.1	-173.1	243.0	726.2	-5.115	3.751
71	4.477	-113.2	-14.7	116.9	447.2	-6.217	4.559
72	4.605	-181.5	43.9	79.2	318.1	-6.639	4.869
73	3.258	-164.4	-21.8	139.5	886.2	-4.572	3.353
74	4.963	-77.0	-131.3	246.0	300.4	-6.522	4.783
75	4.997	-109.4	46.3	57.2	246.5	-7.017	5.146
76	4.828	-87.3	-114.1	224.0	244.6	-6.757	4.955
77	4.691	-62.7	242.1	-214.6	319.5	-7.399	5.426
78	4.290	-84.0	354.4	-309.8	592.9	-6.539	4.795
79	5.333	-91.8	-78.4	195.1	257.4	-6.740	4.942
80	1.808	-124.9	-59.5	690.8	937.7	-2.370	1.738
81	2.747	-79.7	-285.4	270.9	576.4	-5.805	4.257
82	3.091	-113.3	89.2	-1.2	871.3	-4.950	3.630
83	1.386	-236.8	285.2	1278.0	270.6	-1.583	1.161
84	4.595	-104.7	-98.9	157.4	738.6	-5.235	3.839
85	4.927	-56.2	-167.1	207.5	477.8	-6.143	4.505
86	3.951	-83.5	3.9	69.7	712.2	-5.447	3.994

87	4.317	-86.0	6.5	91.5	795.7	-5.063	3.712
88	4.844	-74.8	-103.0	165.7	449.8	-6.245	4.580
89	4.828	-131.5	46.5	99.2	782.7	-4.958	3.636
90	4.682	-147.8	-49.3	208.9	407.4	-6.058	4.442
91	3.135	-85.7	-318.2	303.0	752.1	-5.134	3.765
92	2.079	-183.6	-205.8	362.5	996.8	-3.726	2.732
93	2.197	-78.2	3.0	72.8	970.3	-4.535	3.326
94	2.485	-98.5	-91.3	180.6	894.9	-4.583	3.361
95	2.079	-149.6	-19.3	170.0	980.8	-4.126	3.026
96	4.762	-36.3	-184.5	209.8	597.0	-5.768	4.230
97	4.654	-72.9	-19.4	77.4	666.0	-5.640	4.136
98	4.564	-64.1	-316.0	317.3	691.2	-5.302	3.888
99	4.234	-155.2	5.7	104.7	723.4	-5.222	3.829
100	4.779	-131.7	51.8	65.7	684.2	-5.421	3.975
101	4.543	-105.3	-148.8	231.3	744.6	-5.040	3.696
102	5.165	-91.5	-25.3	103.3	230.9	-7.068	5.183
103	4.984	-124.1	-71.3	160.8	693.9	-5.303	3.889
104	4.691	-105.2	-376.6	349.0	318.4	-6.598	4.839
105	4.317	-67.7	-352.3	386.9	681.8	-5.145	3.773
106	5.298	-191.6	-142.0	345.6	25.1	-7.052	5.172
107	3.714	-122.8	-7.3	103.8	666.2	-5.477	4.016
108	1.946	-50.3	-329.5	928.8	668.3	-3.055	2.241
109	0.000	-149.9	1062.9	372.4	728.6	-1.771	0.299
110	3.595	-143.3	1045.0	-138.7	572.7	-4.341	3.184

The equation of the model for calculation of FEB and the corresponding statistics are shown below:

$$\Delta G_{cald} = -8.28 - 0.00065\ \Delta G_{vdW} + 0.00227\ \Delta G_{ele} + 0.00387\ \Delta G_{solv} + 0.00351\ SASA \qquad (2)$$
$$(N = 110, r^2 = 0.800, s = 0.745, F = 105.1, r^2_{cv} = 0.774, PRESS = 65.81)$$

58

Predicted PCPDCF is estimated from ΔG_{cald} using the following relationship:

$$\Delta G_{binding} = -RT\ln K_{dissociated} \approx -RT \ln PCPDCF \qquad (3)$$

where 298 K is used in the work for temperature T, R is the gas constant in kcal/mol.

The statistical significance of the prediction model is evaluated by the correlation coefficient r^2, standard error s, F-test value, leave-one-out cross-validation coefficient r^2_{cv} and predictive error sum of squares PRESS. The regression model developed based on ΔG_{cald} in this study is statistically ($r^2_{cv} = 0.774$; $r^2 = 0.800$ and F = 105.1) best fitted and consequently used for prediction of PCPDCF of the epipodophyllotoxin analogues as reported in Table 4. The root mean square error (RMSE) between the experimental PCPDCF values and the predicted PCPDCF values obtained by the regression model was also very less (0.411), which is an indicator of the robustness of the fit and suggested that the calculated PCPDCF based on above structure based approach is reliable. Figure 7 graphically shows the quality of fit.

y = 0.828x + 0.736
$R^2 = 0.811$
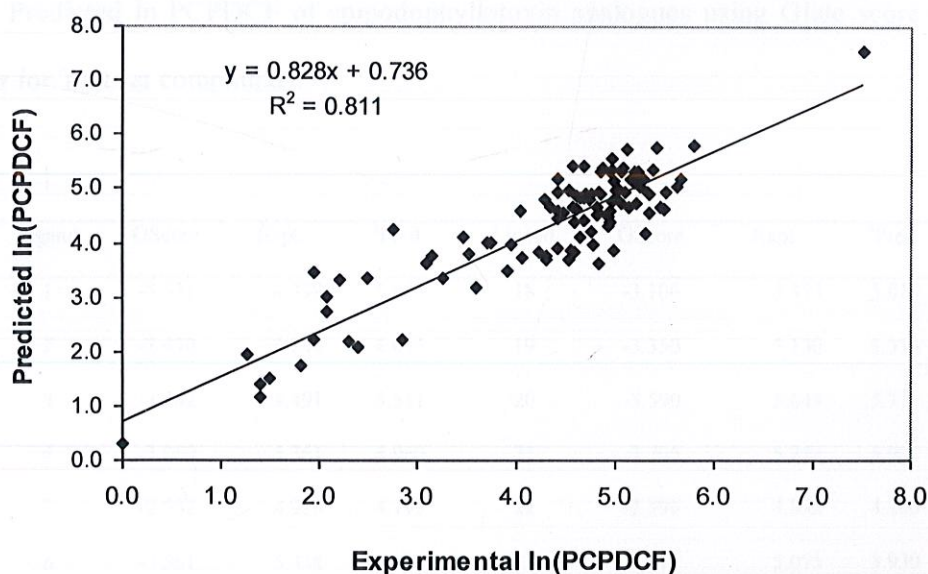
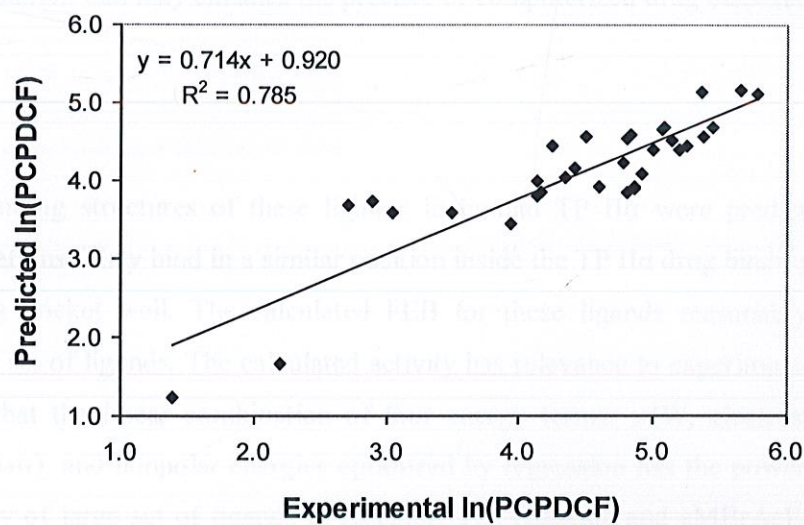**Figure: 7** Models for predic ng binding a nity (ln PCPDCF) of the epipodophyllotoxin deriva ves based on eMBrAcE(ΔG_{cald}) for the training set.

To judge the accuracy of the developed prediction models, we have taken a separate data set called as test set consisting of 33 compounds (Table 2). Their potencies and chemical structures were obtained from literature [23]. Experimentally determined biological activity of the drugs based on *in vitro* study is also provided in order to evaluate the accuracy of predictions. For all compounds, both the prediction models (equations 1 and 2) produce exactly the same trend for relative potencies, even though the exact magnitudes of these values do not match very well (Table 5 and Table 6). The overall RMSE between the experimental and predicted PCPDCF value was in the range of 0.517-0.527, which means that both the structure based modeling were able to predict the activity of 33 epipodophyllotoxin analogues more reliably. Figure 8 and 9 graphically shows the quality of fit for the test set. These results indicate that methodologies with a better prediction precision in binding affinities, though more time-consuming, can provide a significant advantage in prioritizing candidate compounds with high biological activity (low micromolar or nanomolar activity).

Table 5. Predicted ln PCPDCF of epipodophyllotoxin analogues using Glide score (XP) as a descriptor for Test set compounds.

Ligand	GScore	Expt.	[a]Pred.	Ligand	GScore	Expt.	[a]Pred.
1	-3.531	4.799	5.643	18	-3.106	5.375	5.014
2	-2.470	4.244	4.074	19	-3.350	5.130	5.376
3	-3.442	4.491	5.511	20	-3.590	5.649	5.731
4	-3.069	5.361	4.960	21	-3.705	5.252	5.901
5	-2.552	4.920	4.195	22	-2.596	4.852	4.260
6	-3.561	5.438	5.688	23	-2.372	5.075	3.930
7	-3.750	5.778	5.967	24	-3.050	4.771	4.932
8	-1.600	2.708	2.787	25	-1.955	2.197	3.312
9	-2.360	3.045	3.911	26	-1.350	1.386	2.418
10	-2.653	4.796	4.345	27	-2.501	4.127	4.120
11	-2.830	5.063	4.607	28	-2.330	2.890	3.867

12	-2.255	3.932	3.757	29	-1.947	3.497	3.301
13	-2.594	4.595	4.258	30	-2.288	4.852	3.805
14	-2.660	4.127	4.355	31	-2.517	4.344	4.143
15	-2.541	5.187	4.179	32	-2.197	4.419	3.671
16	-2.826	4.159	4.601	33	-2.377	4.990	3.936
17	-2.902	4.836	4.713				

Table 6. Calculated Energies and Estimated Binding Free Energy (ΔG_{cald}) of epipodophyllotoxin analogues in the Test set .

Ligand	Expt. ln(PCPDCF)	ΔG_{vdW} (kcal/mol)	ΔG_{ele} (kcal/mol)	ΔG_{solv} (kcal/mol)	SASA	ΔG_{cald}[1] (kcal/mol)	Pred.[2] ln(PCPDCF)
1	4.799	-31.1	257.1	-167.3	605.8	-6.197	4.545
2	4.244	-96.6	215.4	-165.6	652.6	-6.078	4.457
3	4.491	-125.8	457.4	-325.5	624.3	-6.228	4.567
4	5.361	-54.8	-99.1	134.6	270.3	-7.000	5.133
5	4.920	-28.4	-50.2	102.5	682.7	-5.583	4.094
6	5.438	-115.7	-241.4	311.8	326.7	-6.399	4.693
7	5.778	-82.1	-177.4	213.2	236.2	-6.975	5.115
8	2.708	-57.8	17.3	158.8	725.0	-5.044	3.699
9	3.045	-103.8	-56.8	187.4	773.0	-4.903	3.595
10	4.796	-164.9	-58.0	162.7	691.6	-5.247	3.848
11	5.063	-122.0	-110.3	216.4	357.9	-6.357	4.662
12	3.932	-154.7	-159.9	344.3	709.2	-4.721	3.462
13	4.595	-129.1	152.2	14.6	692.8	-5.363	3.932
14	4.127	-121.0	121.6	5.1	694.1	-5.470	4.011

61

15	5.187	-96.2	70.9	-6.1	592.1	-6.002	4.401
16	4.159	-83.4	-246.2	264.1	712.7	-5.261	3.858
17	4.836	-127.6	621.7	-492.8	695.9	-6.250	4.583
18	5.375	-121.4	402.5	-276.7	603.1	-6.241	4.577
19	5.130	-21.7	192.3	-221.2	719.2	-6.161	4.518
20	5.649	-138.3	530.4	-404.6	430.0	-7.043	5.164
21	5.252	-125.7	431.8	-335.2	695.0	-6.076	4.455
22	4.852	-117.0	104.2	24.1	716.2	-5.360	3.931
23	5.075	-109.6	-73.6	198.7	345.6	-6.394	4.689
24	4.771	-158.3	-27.4	171.3	511.1	-5.783	4.240
25	2.197	-15.4	-116.9	960.4	727.2	-2.266	1.662
26	1.386	-194.7	973.6	440.5	729.3	-1.679	1.231
27	4.127	-129.9	-11.4	112.4	737.6	-5.197	3.811
28	2.890	-151.8	-127.0	214.7	722.7	-5.102	3.741
29	3.497	-109.1	-176.7	258.4	774.9	-4.890	3.586
30	4.852	-85.7	-194.2	268.3	660.3	-5.309	3.893
31	4.344	-117.2	36.6	67.6	667.2	-5.517	4.046
32	4.419	-90.9	-40.1	77.1	667.8	-5.670	4.158
33	4.990	-113.4	21.0	108.0	492.2	-6.013	4.410

1. Calculated free energy of binding, ΔG_{cald} is calculated using equation 4 which has been developed using training set.
2. Predicted pIC_{50} is estimated from ΔG_{cald} using the following relationship: $\Delta G_{binding} = RT\ln Kdissociated \approx RT\ln IC_{50} = -RTpIC_{50}$, where 298K is used in the work for temperature T.
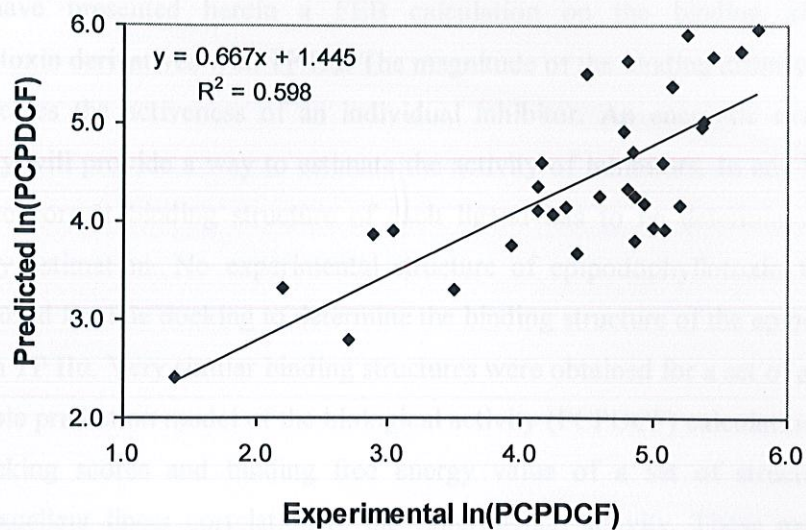
Figure:8 Models for predic ng binding a nity (ln PCPDCF) of the epipodophyllotoxin deriva ves based on Glide score for the test set .

Figure: 9 Models for predic ng binding a nity (ln PCPDCF) of the epipodophyllotoxin deriva ves based on eMBrAcE(ΔG_{cald}) for the test set.

We have presented herein a FEB calculation on the binding affinity of 143 epipodophyllotoxin derivatives with TP IIα. The magnitude of the binding affinity can be the key factor that decides the activeness of an individual inhibitor. An energetic evaluation of the binding affinity will provide a way to estimate the activity of inhibitors. In any binding energy calculation, the correct binding structure of each ligand has to be determined first prior to binding energy estimation. No experimental structure of epipodophyllotoxin with TP IIα is available. We used flexible docking to determine the binding structure of the epipodophyllotoxin analogues with TP IIα. Very similar binding structures were obtained for a set of analogues. This makes a credible prediction model of the biological activity (PCPDCF) calculation possible. The calculated docking scores and binding free energy value of a set of structural analogues demonstrate excellent linear correlation to the experimental activity. These models could be useful to predict the range of activities for new epipodophyllotoxin analogues. The information that we have expressed in this study may lead to the designing (synthesis) of more potent epipodophyllotoxin derivatives for inhibition of TP IIα. Although the current study does not involve a large number of receptors and test sets of compounds, our evaluation data should add valuable information that may enhance the practice of computerized drug discovery.

Conclusion

The binding structures of these ligands in human TP IIα were predicted by flexible docking simulations. They bind in a similar position inside the TP IIα drug binding site and try to fit the binding pocket well. The calculated FEB for these ligands reasonably predicted the activity of this set of ligands. The calculated activity has relevance to experimental activity. The result shows that the linear combination of four energy terms: vdW, electrostatic, solvation (electrostatic part), and nonpolar energies optimized by regression has the power to express the binding affinity of large set of ligands in receptor. The Docking and eMBrAcE demonstrate a good ability on the binding structure prediction and binding energy determination to produce reasonable energies. This work suggests that in the relative FEB calculation, which is essential in drug designing, the contribution of different energy terms can be scaled by a set of weight factors to arrive at relevant results. In practice, it is admitted that same energy term plays different role in different type of systems. This is one of the reasons that a reasonable activity model can be

obtained based on some energy terms. The calculation of solvation effect upon a ligand binding in a protein is a challenging work. This work and many other works have shown that solvation effect is an important driving force on ligand binding and a key factor in expression of activity of a set of ligands. In the work, GB and SASA methods were used to estimate the electrostatic and the nonpolar parts of solvation that produced satisfactory results in terms of useful experimental activities.

References:

[1] J.J. Champoux, DNA topoisomerases: structure, function, and mechanism, Annu. Rev. Biochem. 70 (2001) 369–413.

[2] J.C. Wang, Cellular roles of DNA topoisomerases: a molecular perspective. Nature Rev. Mol. Cell. Biol. 3(2002) 430–440.

[3] J.C. Wang, DNA topoisomerases, Annu. Rev. Biochem. 65 (1996) 635–692.

[4] D.A. Burden and N. Osheroff, Mechanism of action of eukaryotic topoisomerase II and drugs targeted to the enzyme, Biochim. Biophys. Acta. 1400 (1998) 139–154.

[5] T. Goto, C Holm, J.C. Wang , and D. Botstein, DNA topoisomerase II is required at the time of mitosis in yeast, Cell 41 (1985) 553-563.

[6] A.H. Corbett and N. Osheroff , When good enzymes go bad: conversion of topoisomerase II to a cellular toxin by antineoplastic drugs, Chem. Res. Toxicol. 6 (1993) 585-597.

[7] S.J. Froelich-Ammon and N. Osheroff, Topoisomerase Poisons: Harnessing the Dark Side of Enzyme Mechanism, J. Biol. Chem. 270 (1995) 21429-21432.

[8] G. Capranico and M. Binaschi, DNA sequence selectivity of topoisomerases and topoisomerase poisons, Biochim. Biophys. Acta 1400 (1998) 185–194.

[9] G. Capranico, G. Zagotto, and M. Palumbo, Development of DNA topoisomerase-related therapeutics: a short perspective of new challenges, Curr. Med. Chem. Anti-Canc. Agents, 4 (2004) 335–345.

[10] K.R. Hande, Etoposide: four decades of development of a topoisomerase II inhibitor, Eur. J. Cancer. 34 (1998) 1514–1521.

[11] K.R. Hande, Topoisomerase II inhibitors, Cancer Chemother. Biol. Response Modif 21 (2003) 103–125.

[12] E.L. Baldwin, and N. Osheroff, Etoposide, topoisomerase II and cancer, Curr. Med. Chem. Anti-Canc. Agents 5 (2005) 363–372.

[13] S.H. Kaufmann, Cell death induced by topoisomerase-targeted drugs: more questions than answers, Biochim. Biophys. Acta 1400 (1998) 195–211.

[14] O. Sordet, Q.A. Khan, K.W Kohn, and Y. Pommier, Apoptosis induced by topoisomerase inhibitors, Curr. Med. Chem. Anti-Canc. Agents 3 (2003) 271–290.

[15] J. Aisner, C.P. Belani, and L.A. Doyle, Etoposide: current status and future perspectives in the management of malignant neoplasms, Cancer Chemother. Pharmacol. 34 (1994) S118.

[16] K.H. Lee, Y. Imakura, M. Haruna, S.A. Beers, L.S. Thurston, H.J. Dai, and C.H. Chen, Antitumor Agents.107. New cytotoxic 4-alkylamino analogues of 4'-demethylepipodophyllotoxin as inhibitors of Human DNA Toposiomerase II, J. Nat. Prod. 52 (1989) 606-613.

[17] Z.Q. Wang, Y.H. Kuo, D. Schnur, J.P. Bowen, S.Y. Liu, F.S. Han , Y.C. Cheng, and K.H. Lee, Antitumor Agents 113. New 4-arylamino derivatives of 4'-O-demethyl-epipodophyllotoxin and related compounds as potent inhibitors of Human DNA toposiomerase II, J. Med. Chem. 33 (1990) 2660-2666.

[18] K.H. Lee, S.A. Beers, M. Mori, Z.Q. Wang, Y.H. Kuo, Li l, S.Y. Liu, J.Y. Cheng, F.S. Han, and Y.C. Cheng, Antitumor Agents.111. New 4-hydroxylated and 4-halogenated anilino derivatives of 4'-demethylepipodophyllotoxin as potent inhibitors of Human DNA toposiomerase II, J. Med Chem. 33 (1990) 1364-1368.

[19] X.M. Zhou, Z.Q. Wang, J.Y. Cheng , H.X. Chen, Y.C. Cheng, and K.H. Lee, Antitumor Agents. 120. New 4-substituted benzylamine and benzyl ether derivatives of 4'-O- demethyl-epipodophyllotoxin as potent inhibitors of Human DNA toposiomerase II, J. Med. Chem. 33 (1991) 3346-3350.

[20] H. Hu, Z.Q. Wang, S.Y. Liu, Y.C. Cheng, and K.H. Lee, Antitumor Agents. 123. Synthesis and Human DNA topoisomerase II inhibitory activity of 2'-chloro derivatives of Etoposide and 4β-(arylamino)-4'-O-demethylpodophyllotoxins, J. Med. Chem. 35 (1992) 866-871.

[21] Z.Q. Wang, H. Hu, H.X. Chen, Y.C. Cheng, and K.H. Lee, Antitumor Agents. 124. New 4β-substituted aniline derivatives of 6,7-O,O- demethylpodophyllotoxin as potent inhibitors of Human DNA toposiomerase II, J. Med. Chem. 35 (1992) 871-877.

[22] X.M. Zhou, Z.Q. Wang, H.X. Chen, Y.C. Cheng, and K.H. Lee, Antitumor Agents. 125. New 4-benzoylamino derivatives of 4'-O-demethyl-4-desoxypodophyllotoxin and 4-benzoyl derivatives of 4'-O- demethylpodophyllotoxin as potent inhibitors of Human DNA toposiomerase II, Pharm. Res. 10 (1993) 214-219.

[23] Z.Q. Wang, Y.C. Cheng, H.X. Chen, J.Y. Cheng, X. Guo, Y.C. Cheng, and K.H. Lee, Antitumor Agents. 126. Novel 4β-substituted anilino derivatives of 3,4'-O,O-didemethylpodophyllotoxin as potent inhibitors of Human DNA toposiomerase II, Pharm. Res. 10 (1993) 343-350.

[24] M. Miyahara, Y. Kashiwada, X. Guo, H.X. Chen, Y.C. Cheng, and K.H. Lee, Nitrosourea derivatives of 3',4'-dioxo-4-deoxypodophyllotoxin and urea derivatives of 4'-O-demethylpodophyllotoxin as potent inhibitors of Human DNA toposiomerase II, Heterocycles 39 (1994) 361-369.

[25] Y.L. Zhang, X. Guo, Y.C. Cheng, and K.H. Lee, Antitumor Agents. 148.Synthesis and biological evaluation of novel 4-amino derivatives of Etoposide with better pharmacological profiles, J. Med. Chem. 37 (1994) 446-452.

[26] Z. Ji, H.K. Wang, K.F. Bastow, X.K. Zhu, S.J. Cho, Y.C. Cheng, and K.H. Lee, Antitumor Agents. 177. Design, synthesis and biological evaluation of novel etoposide analogues bearing pyrrolecarboxamidino group as DNA topoisomerase II inhibitors, Bioorg. Med. Chem. Lett. 7 (1997) 607-612.

[27] X.K. Zhu, J. Guan, Y. Tachibana, K.F. Bastow, S.J. Cho, H.H. Cheng, Y.C. Cheng, M. Gurwith, and K.H. Lee, Antitumor Agents.194. Synthesis and biological evaluation of 4-β-Mono-,-Di-,and −trisubstituted aniline-4'-demethyl-podophyllotoxin and related compounds with improved pharmacological profiles, J. Med. Chem. 42 (1999) 2441-2446.

[28] D. Fass, C.E. Bogden , and J.M. Berger, Quaternary changes in topoisomerase II may direct orthogonal movement of two DNA strands, Nat. Struct. Biol. 6 (1999) 322-326.

[29] Prime version 1.5, Macromodel version 9.1, Schrodinger, LLC, New York, NY, 2005

[30] D. Leroy, A.V. Kajava, C. Frei, and S.M. Gasser, Analysis of etoposide binding to subdomains of human DNA topoisomerase II in the absence of DNA, Biochemistry 40 (2001) 1624–1634.

[31] J.C. Wang, Cellular roles of DNA topoisomerases: a molecular perspective, Nature Rev. Mol. Cell. Biol. 3 (2002) 430–440.

[32] R.A. Laskowski, M.W. MacArthur, D.S. Moss, and J.M. Thornton, PROCHECK: a program to check the stereochemical quality of protein structures, J. Appl. Crystallogr. 26 (1993) 283-291

[33] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, Stereochemistry of polypeptide chain configurations, J. Mol. Biol. 7 (1963) 95-99.

[34] D. Eisenberg, R. Luthy, and J.U. Bowie, VERIFY3D: Assessment of protein models with three-dimensional profiles, Methods. Enzymol. 277 (1997) 396-404.

[35] M.J. Robinson, A.H. Corbett, and N. Osheroff, Effects of topoisomerase II targeted drugs on enzyme-mediated DNA cleavage and ATP hydrolysis: evidence for distinct drug interaction domains on topoisomerase II, Biochemistry 32 (1993) 3638–3643.

[36] D.A. Burden and N. Osheroff, Mechanism of action of eukaryotic topoisomerase II and drugs targeted to the enzyme, Biochim Biophys Acta 1400 (1998) 139–154.

[37] G. Capranico, G. Zagotto, and M. Palumbo, Development of DNA topoisomeraserelated therapeutics: a short perspective of new challenges, Curr. Top. Med. Chem. 4 (2004) 335–345.

[38] Schrodinger LLC. , http://www.schrodinger.com, (accessed: 24. 04.2007).Schrodinger. 2000; Inc.: Portland, OR.

[39] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, and P.S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring 1 Method and assessment of docking accuracy, J. Med. Chem. 47 (2004) 1739-1749.

[40] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, and R.P. Mee, Empirical scoring functions: The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes, J. Comput-Aided Mol .Desig. 11 (1997) 425-445.

[41] O. Guvench, J. Weiser, P.S. Shenkin, I. Kolossváry, and W.C. Still, Application of the frozen atom approximation to the GB/SA continuum model for solvation free energy, J. Comput. Chem. 23 (2002) 214-221.

[42] X. Wu, J.L.S. Milne, M.J. Borgnia, A.V. Rostapshov, S. Subramaniam, and B.R. Brooks, A core-weighted fitting method for docking atomic structures into low-resolution maps: application to cryo-electron microscopy, J. Struct. Bio. 141 (2003) 63-76.

[43] N.P. Todorov, R.L. Mancera, and P.H. Monthoux, A new quantum stochastic tunneling optimisation method for protein-ligand docking, Chemical Physics Letters 369 (2003) 257-263.

[44] C.H. Reynolds, Estimating liphophilicity using GB/SA continuum solvation Model: a direct method for computing partition coefficients, J. Chem. Inf. Comput. Sci. 35 (1995) 738-742.

CHAPTER III

Development of Predictive Quantitative Structure Activity Relationship Models of Epipodophyllotoxin Derivatives

Introduction

Epipodophyllotoxins are the glucosidic derivatives of podophyllotoxin which have been used in the chemotherapy of various types of cancer, including small cell lung cancer, testicular carcinoma, lymohoma, and Kaposi's sarcoma [1-3]. Etoposide (VP-16) and teniposide (VM-26) are the most successful candidates widely prescribed chemotherapeutic agent. Efforts for improving their clinical efficacy further by overcoming the drug resistance, myelosuppresion and poor bioavailability problems [4] associated with them, were continued to be challenging. Over the years a number of laboratories throughout the world engaged in the synthesis and testing of epipodophyllotoxin derivatives [5-8] to prepare more new potent and less toxic analogues, that is, with better therapeutic indices. The proposed mechanism of epipodophyllotoxins' antitopoisomerase II activity is to inhibit the catalytic activity of the target enzyme by stabilizing the covalent topoisomerase II (TP-II) -DNA cleavable complex [9].

To construct an informative structure-activity relationship (SAR) model and improve further designing of potentially bioactive compounds there is a need for development of predictive QSAR models for the rapid prediction of inhibition of human TP-IIα of novel epipodophyllotoxin analogues and virtual prescreening. Comparative molecular field analysis (CoMFA) is one of the most popular methods for QSAR and is characterized by reasonable simplicity and a clear physiochemical sense of steric and electrostatic descriptors [10]. However, despite statistically excellent predictive performance, CoMFA has inherent limitation to align with the database molecules correctly within 3D space [11,12]. 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. Neverthless, especially for structurally diverse molecules, unambiguous 3D alignment still finds it difficult to initiate the CoMFA process.

We, as well as other researchers [13], 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 of 3D QSAR methods. Accordingly,

70

in this QSAR study, we have applied topological, electronic, geometrical and energy based descriptors calculated directly form the 2D and 3D structure of the molecules. The approach is simple, fast and straightforward. It benefits in predicting the activities of a large set of molecules in rational drug design. Furthermore, we have implemented the concept of variable selection, a process that has been investigated recently by a number of researchers [14-16] using genetic function approximation (GFA) [17,18] optimization algorithms. Variable selection techniques choose the most informative variables and eliminate irrelevant variables to improve the signal-to-noise ratio in the resulting models. Additionally, these techniques are not computationally intensive and are practically automated. The behavior of QSAR model is examined with a variety of statistical parameters [19] and the contribution of various descriptors are analyzed.

Materials and methods
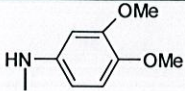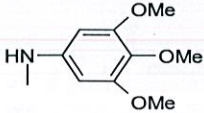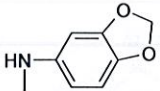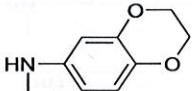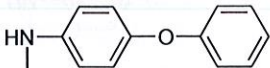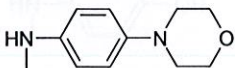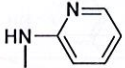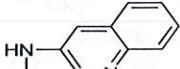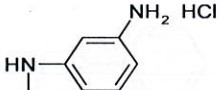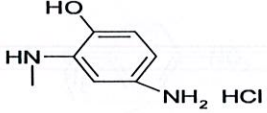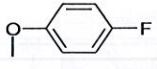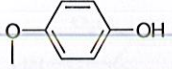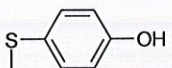
Data set
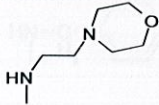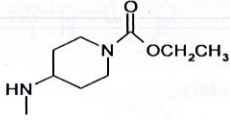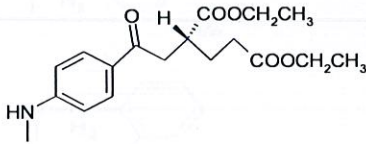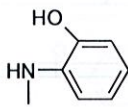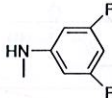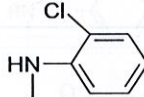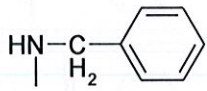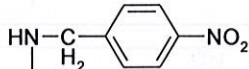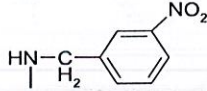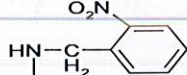
A total number of 130 epipodophyllotoxin analogues (Table 1) were used in the study which were synthesized and tested under the same conditions in the same laboratory [5-8]. To generate statistically robust and most importantly, validated models, all compounds in the original data set were divided randomly into 100 molecules in training set and 30 molecules in test set. All compounds in this study were evaluated for their ability to form intracellular covalent topoisomerase II-DNA complexes using human TP-IIα. at similar laboratory conditions and experimental setup. The assay system has been described previously by Lee et al [5]. The activity data are originally expressed as the percentage of cellular protein-DNA complex formation (PCPDF) and were transformed by taking the logarithm of PCPDF, i.e. $\log_{10}$ (PCPDF) and were used in subsequent variable selection as well as QSAR model development.
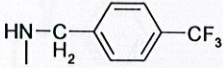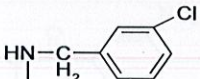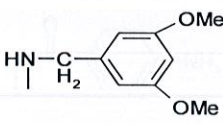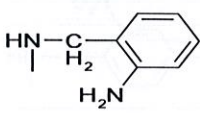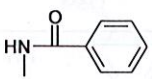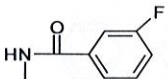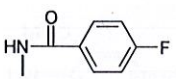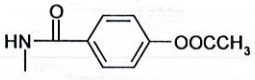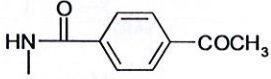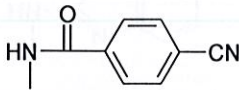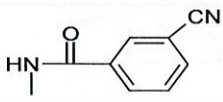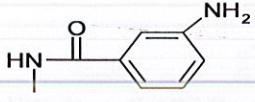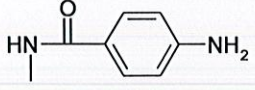
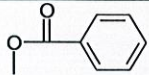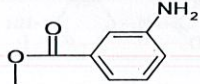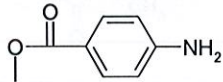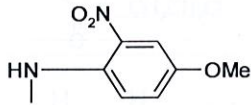**Table 1:** List of epipodophyllotoxin analogues and their experimental activities.
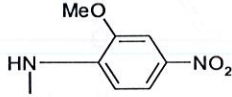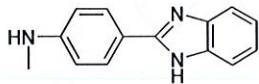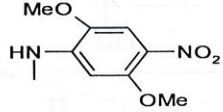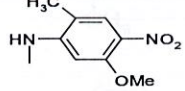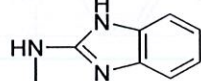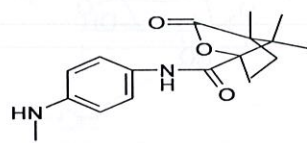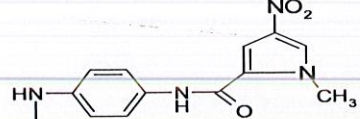
Compound No.	R-group	Scaffold (Figure 1)	type	Log (PCPDCF)
1	—OH	1		1.63
2	-NHCH$_2$CH$_2$OCH$_3$-	1		2.04

71
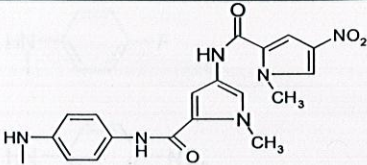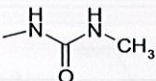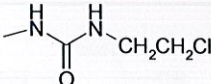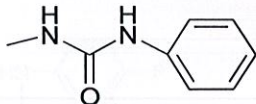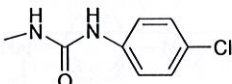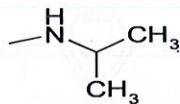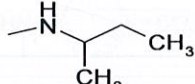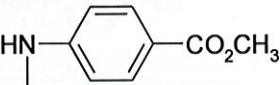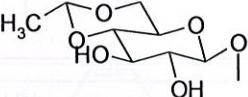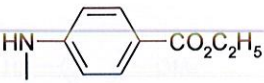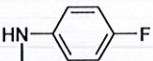
3	—NHCH$_2$CH=CH$_2$	1	1.92
4	—NHCH$_2$CH(OH)CH$_3$	1	2.22
5	—NHCH(CH$_3$)CH$_2$OH	1	2.21
6		1	2.46
7		1	2.39
8		1	2.32
9		1	0.60
10		1	2.40
11		1	2.32
12		1	1.92
13		1	2.11
14		1	1.70
15		1	2.02
16		1	2.37

17	(structure: HN-phenyl with OMe, OMe)	1	2.26
18	(structure: HN-phenyl with OMe, OMe, OMe)	1	1.67
19	(structure: HN-benzodioxole)	1	2.21
20	(structure: HN-benzodioxane)	1	2.45
21	(structure: HN-phenyl-O-phenyl)	1	1.99
22	(structure: HN-phenyl-morpholine)	1	2.15
23	(structure: HN-pyridine)	1	1.99
24	(structure: HN-quinoline)	1	2.09
25	(structure: HN-phenyl-NH₂·HCl)	1	2.15
26	(structure: HN-phenyl, HO, NH₂·HCl)	1	1.04
27	(structure: O-phenyl-F)	1	1.76
28	(structure: O-phenyl-OH)	1	1.53
29	(structure: S-phenyl-OH)	1	1.00

No.	Structure		
30	morpholine N-CH2CH2-N(CH3)H	1	1.89
31	ethyl piperidine-1-carboxylate, 4-N(CH3)H, OCH2CH3	1	1.23
32	HN(CH3)-C6H4-C(=O)CH2-CH(COOCH2CH3)-CH2CH2-COOCH2CH3	1	1.92
33	HN(CH3)-C6H4-OH (ortho), HO	1	2.18
34	HN(CH3)-C6H4-OH	1	2.32
35	HN(CH3)-C6H3(F)(F)	1	2.06
36	HN(CH3)-C6H4-Cl	1	1.51
37	HN(CH3)-CH2-C6H5	1	3.26
38	HN(CH3)-CH2-C6H4-NO2	1	2.33
39	HN(CH3)-CH2-C6H4-NO2	1	2.11
40	HN(CH3)-CH2-C6H4-NO2 (O2N)	1	2.16

No.	Structure		
41	HN–CH₂–C₆H₄–CF₃	1	2.00
42	HN–CH₂–C₆H₄(3-Cl)	1	2.20
43	HN–CH₂–C₆H₃(3,5-diOMe)	1	2.16
44	HN–CH₂–C₆H₄(2-NH₂)	1	2.26
45	HN–C(=O)–C₆H₅	1	2.25
46	HN–C(=O)–C₆H₄(3-F)	1	2.06
47	HN–C(=O)–C₆H₄(4-F)	1	2.07
48	HN–C(=O)–C₆H₄(4-OOCCH₃)	1	2.14
49	HN–C(=O)–C₆H₄(4-COCH₃)	1	2.09
50	HN–C(=O)–C₆H₄(4-CN)	1	2.20
51	HN–C(=O)–C₆H₄(3-CN)	1	2.17
52	HN–C(=O)–C₆H₄(3-NH₂)	1	2.17
53	HN–C(=O)–C₆H₄(4-NH₂)	1	2.08

54		1	1.97
55		1	2.00
56		1	1.97
57		1	1.18
58		1	1.92
59		1	2.11
60		1	0.64
61		1	0.54
62		1	1.76
63		1	1.94
64		1	2.00
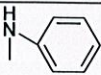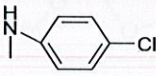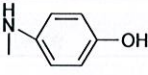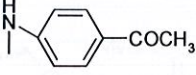
65		1	1.41
66		1	1.91
67		1	2.16
68		1	2.17
69		1	2.10
70		1	2.04
71		1	1.86
72		1	2.32
73		2	0.79
74		2	1.19
75		3	1.34
76		3	1.04

No.	Structure		
77	HN—⟨C6H4⟩—F	4	1.72
78	HN—⟨C6H4⟩—NO$_2$	5	1.88
79	HN—⟨C6H4⟩—CO$_2$C$_2$H$_5$	5	2.10
80	HN—⟨C6H4⟩—CN	5	2.10
81	HN—⟨C6H4⟩—F	5	2.03
82	HN—⟨C6H4⟩—NO$_2$	3	1.36
83	HN—⟨C6H4⟩—NO$_2$	6	0.90
84	HN—⟨C6H4⟩—CO$_2$C$_2$H$_5$	6	0.95
85	HN—⟨C6H4⟩—CN	6	1.08
86	HN—⟨C6H4⟩—F	6	0.90
87	HN—⟨C6H4⟩—F	7	2.07
88	HN—⟨C6H5⟩	7	2.02
89	HN—⟨C6H4⟩—OMe	7	1.98

No.	Structure		
90	HN(CH₃)– phenyl–3,4-(OMe)₂	7	1.84
91	HN(CH₃)– 2,3-dihydro-1,4-benzodioxine	7	2.08
92	HN(CH₃)– phenyl–$CO_2C_2H_5$	7	1.97
93	HN(CH₃)– phenyl–CO_2CH_3	7	2.24
94	HN(CH₃)– phenyl–CN	7	2.16
95	HN(CH₃)– phenyl–CH_2CN	7	2.04
96	HN(CH₃)– phenyl–3-NO_2	7	1.88
97	HN(CH₃)– phenyl–4-NO_2	7	2.30
98	CH₃–NH–C(=O)–N(N=O)–CH₃	8	1.61
99	CH₃–NH–C(=O)–N(N=O)–CH_2CH_2Cl		0.85
		8	
100	CH₃–NH–CH(N=O)–CH(CH₃)₂	9	0.00
101	–$NHCH_2CH_2OH$	1	2.08

79

102	—NHCH$_2$CH$_2$CH$_3$	1	1.84
103	—NHCH$_2$CH$_2$CH$_2$OH	1	1.95
104		1	2.33
105		1	2.14
106		1	2.36
107		1	2.51
108		1	1.32
109		1	2.08
110		1	2.20
111		1	1.71
112		1	2.00
113		1	1.79

No.	Structure		
114	—NH—C6H4—Br	1	2.25
115	—NH—C6H4—I	1	1.81
116	HN—CH2—C6H4—F (3-F)	1	2.33
117	HN—CH2—C6H4—F (4-F)	1	2.23
118	HN—CH2—C6H4—CN	1	2.45
119	HN—CO—C6H4—F (2-F)	1	2.11
120	HN—CO—C6H4—NO2 (3-NO2)	1	1.93
121	HN—CO—C6H4—NO2 (4-NO2)	1	2.20
122	HN—C6H3(NO2)2	1	1.30
123	HN—CO—NH—C6H4—F	1	2.07
124	HN—C6H5	3	0.95
125	HN—C6H4—OH	3	0.60
126	HN—C6H4—CN	3	1.52

127	H N— phenyl	7	2.11
128	H N— C₆H₄—Cl	7	1.89
129	H N— C₆H₄—OH	7	1.92
130	H N— C₆H₄—COCH₃	7	2.17

Building of molecular structures

All these epipodophyllotoxin analogues were built from the various scaffold structure (Figure 1) and the substitution of functional groups was carried out as mentioned in Table 1. We used Maestro-molecular builder for building the scaffold and structural derivatives. LigPrep [20] was used for final preparation of ligands. LigPrep is a utility of Schrödinger software suit that combines tools for generating 3D structures from 1D (Smiles) and 2D (SDF) representation and searching for tatutomers and steric isomers and performing a geometry minimization of ligands. The ligands were energy minimized using Macromodel module of Schrodinger with default parameters and applying molecular mechanics force fields (MMFFs). Truncated Newton Conjugate Gradient (TNCG) minimization method was used with 500 iterations and convergence threshold of 0.05 kJ/mol.

82
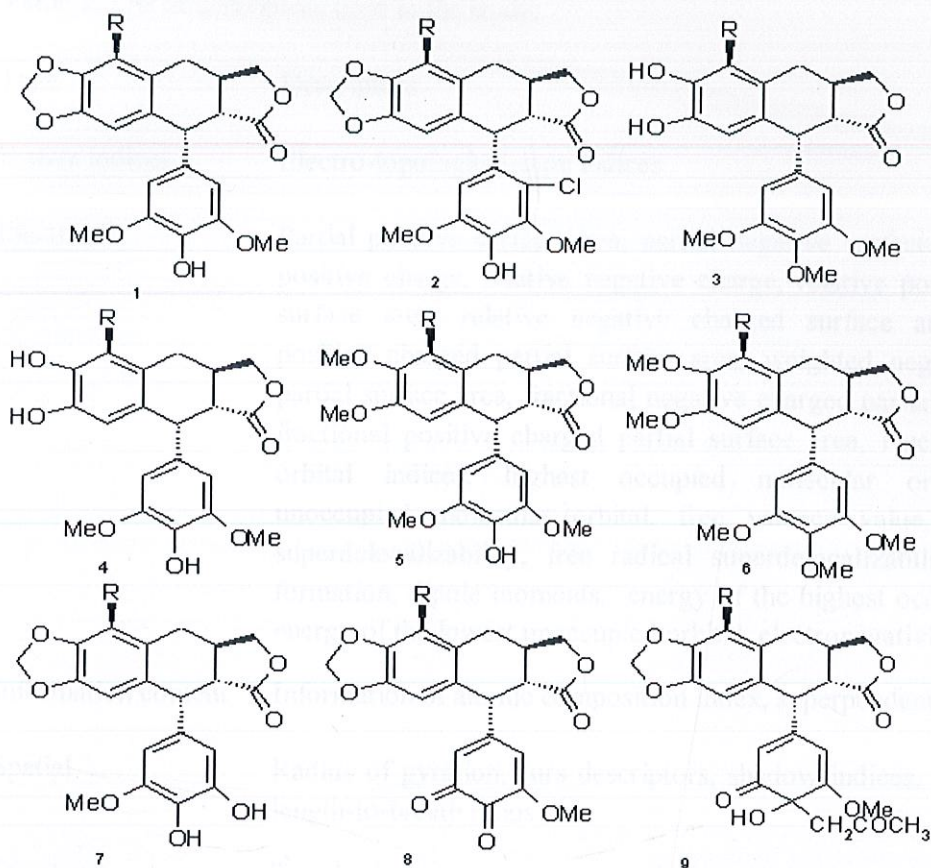
Figure 1. The various scaffold structures used for building the epipodophyllotoxin analogues.

Descriptor calculation

All the molecular descriptors such as E-state indices, log P, superpendentic index, structural, symmetrical, topological, lead likeness, electronic Wang-Ford atomic charge and extended Huckel partial charge functions, bulk, moments, orbital energies, molecular connectivity indexes , gravitational indexes, hydrophobicity, steric and thermodynamic factors and topological descriptors were calculated using ADME Model Builder software package (version 4.5) [21]. These descriptors help 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 2.

83

Table 2. List of descriptors used in the study.

Type	Descriptors
E-state indices	Electro-topological-state indices
Electronic	Partial positive surface area, partial negative surface area, relative positive charge, relative negative charge, relative positive charged surface area, relative negative charged surface area, weighted positive charged partial surface area, weighted negative charged partial surface area, fractional negative charged partial surface area, fractional positive charged partial surface area, Huckel molecular orbital indices, highest occupied molecular orbital, lowest unoccupied molecular orbital, free valence value, nucleophilic superdelocalizability, free radical superdelocalizability, heat of formation, dipole moments, energy of the highest occupied orbital, energy of the lowest unoccupied orbital, electronegativity, hardness
Information content	Information of atomic composition index, superpendentivity index
Spatial	Radius of gyration, Jurs descriptors, shadow indices, area, density, length-to-breath ratios
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.
Thermodynamic	Average energy, bond strain energy, angle strain energy, non-bonded strain energy, torsional strain energy, total strain energy of molecule.
Leadlikeness	LogP (Meylan, Howard), LogS, LogP(Moriguchi, Hirono).
Topological	Wiener index, Kier and Hall molecular connectivity indices, path count and length descriptors, topological polar surface area (TPSA), Balban indices.

Screening of descriptors and development of QSAR model

A set of 372 molecular descriptors were calculated using ADME Model Builder software package (version 4.5). A systematic search in the order of missing value test, zero test, correlation coefficient, multi-colinearity and genetic algorithm was performed to determine significant descriptors using ADME model Builder (version 4.5) software package (Fujitsu Inc.). Any parameter which was not calculated (missing value) for any number of the compounds in the data set was rejected in the first step. Some of the descriptors were rejected because they contained a zero value for all the compounds (zero tests). In order to minimize the effect of collinearity and to avoid redundancy, a correlation matrix was developed with a cutoff value of 0.6 and the variables were physically removed from the analysis which showed exact linear dependencies between subsets of the variables and multi-collinearity (high multiple correlations between subsets of the variables). From the descriptors, the set of descriptors that would give the statistically best QSAR models was selected by using a genetic function approach implemented in the ADME model Builder (version 4.5) software package. The genetic algorithm (GA) starts with the creation of a population of randomly generated parameter sets. The usage probability of a given parameter from the active set is 0.5 in any of the initial population sets. The sets are then compared according to their objective functions. The parameters set used for the GA includes: mutation 0.1, crossover 0.9, population 300, number of generations 1000, r^2 floor limit 50% and the objective function was r^2/N_par. The form of the objective function favors sets that have r^2 as high as possible, while minimizing the number of parameters used as descriptors. The higher the score, the higher the probability of a given set to 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 algorithm runs until the desired number of generations are reached. Equations were developed between the observed activity and the descriptors. The best equation was taken on the basis of statistical parameters such as squared regression coefficient (r^2) and leave one out cross-validated regression coefficient (q^2_{cv}).

Validation of QSAR model

The predictive capability of the QSAR equation was determined using leave-one-out cross validation method. The cross validation regression coefficient (q^2_{cv}) was calculated by the following equation.

85

$$q^2_{cv} = 1 - \frac{PRESS}{TOTAL} = 1 - \frac{\sum\limits_{i-1}^{n}\left(y_{exp} - y_{pred}\right)^2}{\sum\limits_{i-1}^{n}\left(y_{exp} - \overline{y}\right)^2}$$

where, y_{pred}, y_{exp} and $\overline{y}$ are the predicted, experimental and mean values of experimental activity, respectively. Also the accuracy of the prediction of the QSAR equation was validated by F-value, r^2 and r^2_{adj}. A large F indicates that the model fit is not a chance occurrence. It has been shown that a high value of statistical characteristics is not necessary as the proof of a highly predictive model [22,23]. Hence, in order to evaluate the predictive ability of our QSAR model, we used the method described by Golbraikh and Tropsha [22] and Roy and Roy [23] .The values of the correlation coefficient of predicted and actual activities and the correlation coefficient for regressions through the origin (predicted vs. actual activities and vice versa) were calculated using the regression of analysis Tool-pak option of Excel and other parameters were calculated as reported by the above authors [22,23]. The determination coefficient in prediction, q^2_{test} was calculated using the following equation [23]:

where $Y_{pred_{Test}}$ and Y_{Test} are $\qquad q^2_{test} = 1 - \dfrac{\sum\left(Y_{pred_{Test}} - Y_{Test}\right)^2}{\sum\left(Y_{Test} - \overline{Y}_{Training}\right)^2} \qquad$ the predicted value based on QSAR equation (model response) and experimental activity values, respectively, of the external test set compounds. $Y_{Training}$ is the mean activity value of the training set compounds.

To check the inter-correlation of descriptors, variance inflation factor (VIF) analysis was performed. The 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 the VIF value is larger than 10, information of descriptors can be hidden by correlation of descriptors [24,25] .

Results and Discussion

The 130 active compounds with their biological activity in terms of percentage of cellular protein DNA complex form (PCPDCF) were randomly divided into a training set of 100

86

compounds and a test set of 30 compounds. With the wide range of difference between PCPDCF values and the large diversity in the structures, the combined data set of 100 molecules and 30 molecules are ideal to be considered as training and test sets, as both the sets do not suffer from bias, due to the similarity of the structures. The various molecular descriptors (372 in total) as described in Table 2 were calculated initially. By applying a missing value test, a zero test, a correlation test with cutoff value of 0.6 and a multicollinearity test with cutoff value of 0.9 we have discarded the most likely parameters that resulted in 218 parameters. Further additional parameters were discarded by applying the GA and finally 5 parameters were selected for the development of the QSAR equation. 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 a new term on the statistical quality of the model. As the squared correlation coefficient, r^2, can be easily increased by the number of terms in the QSAR equation, we took the cross-validation correlation coefficient, q^2_{cv}, as the limiting factor for a number of descriptors to be used in the final model. It was observed that the q^2_{cv} value increased until the number of descriptors in the equation reached 5, as shown in Table 3. With further addition of parameters to the equation with 5 descriptors, there was a decrease in the q^2_{cv} value of the model. So, the numbers of descriptors were restricted to 5 in the final QSAR model.

The best significant relationship for the cytotoxic activity has been deduced to be

$$\text{Log PCPDCF} = 7.25 + 2.97*\text{SASA} + 0.108*\text{ES\_Sum\_aasC} - 0.0813*\text{NATM} - 2.48 \text{ Balban index} - 0.000794*\text{HOF} \qquad (1)$$

$(n = 100;\ r^2_{train} = 0.721;\ s = 0.24;\ \text{PRESS} = 6.240;\ r^2_{adj} = 0.707;\ q^2_{cv} = 0.678;\ F\text{-test} = 48.71)$

where n is the number of compounds in the training set, r^2_{train} 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-test 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_{cv} is the square of the correlation coefficient of the cross-validation using leave-one-out cross-validation technique. The QSAR model developed in this study is statistically ($r^2_{train} = 0.721$, $q^2_{cv} = 0.678$, F-test = 48.71) best fitted and consequently was used for prediction of PCPDCF of training and test sets of molecules as reported in Tables 4 and 5.

Table 4. Observed and predicted PCPDCF of Training set of epipodophyllotoxin derivatives.

| S.No. | PCPDCF (Log BA) | | | S.No. | PCPDCF (Log BA) | | |
	Observed	Predicted	Residual		Observed	Predicted	Residual
1	1.63	1.73	0.11	65	1.41	1.44	0.02
4	2.22	2.36	0.14	66	1.91	1.93	0.02
5	2.21	2.31	0.11	67	2.16	2.24	0.09
6	2.46	2.16	0.30	68	2.17	2.43	0.26
8	2.32	2.28	0.04	69	2.10	2.13	0.03
10	2.40	2.15	0.25	70	2.04	1.89	0.14
11	2.32	2.26	0.05	71	1.86	2.05	0.19
12	1.92	1.93	0.01	72	2.32	2.38	0.07
13	2.11	1.97	0.14	73	0.79	1.04	0.25
14	1.70	1.95	0.25	74	1.19	1.12	0.07
15	2.02	2.17	0.15	77	1.72	1.72	0.00
16	2.37	2.42	0.05	78	1.88	1.48	0.40
17	2.26	2.04	0.22	80	2.10	1.65	0.45
18	1.67	1.89	0.22	81	2.03	1.67	0.36
20	2.45	2.41	0.04	82	1.36	1.55	0.19
23	1.99	2.23	0.25	83	0.90	0.82	0.08
24	2.09	2.17	0.08	84	0.95	1.15	0.19
25	2.15	2.22	0.07	85	1.08	1.21	0.13
33	2.18	2.11	0.07	86	0.90	1.23	0.32
34	2.32	2.38	0.05	87	2.07	1.96	0.11
35	2.06	2.01	0.05	88	2.02	2.13	0.11
37	3.26	2.76	0.49	89	1.98	2.04	0.06
38	2.33	2.41	0.08	90	1.84	1.69	0.15
39	2.11	2.18	0.06	91	2.08	2.07	0.00
40	2.16	2.08	0.07	92	1.97	1.82	0.16
41	2.00	2.22	0.23	93	2.24	2.08	0.17

44	2.26	2.48	0.22	94	2.16	1.95	0.22
45	2.25	2.42	0.18	95	2.04	2.19	0.15
46	2.06	2.05	0.01	98	1.61	1.59	0.02
47	2.07	2.26	0.19	99	0.85	1.55	0.71
48	2.14	2.13	0.00	104	2.33	2.29	0.04
49	2.09	2.21	0.12	105	2.14	2.08	0.06
50	2.20	2.26	0.06	106	2.36	1.85	0.52
51	2.17	2.02	0.15	107	2.51	2.08	0.43
52	2.17	2.19	0.02	108	1.32	1.15	0.17
53	2.08	2.39	0.31	109	2.08	2.00	0.08
55	2.00	2.16	0.16	110	2.20	2.07	0.13
57	1.18	1.37	0.19	112	2.00	2.35	0.35
59	2.11	2.29	0.18	113	1.79	1.90	0.11
60	0.64	1.10	0.45	114	2.25	2.08	0.17
61	0.54	1.33	0.79	115	1.81	1.83	0.03
64	2.00	1.80	0.20	116	2.33	2.38	0.04
119	2.11	1.97	0.14	129	1.92	2.03	0.12
120	1.93	1.81	0.12	130	2.17	1.92	0.25
121	2.20	2.06	0.15	58	1.92	1.46	0.46
122	1.30	1.39	0.09	96	1.88	1.51	0.36
123	2.07	2.33	0.26	63	1.94	1.75	0.20
126	1.52	1.74	0.22	79	2.10	1.68	0.42
127	2.11	2.12	0.01	97	2.30	1.76	0.54
128	1.89	2.00	0.12	32	1.92	1.91	0.01

Table 5. Observed and predicted PCPDCF of Test set of epipodophyllotoxin derivatives.

| | PCPDCF (Log BA) | | | | PCPDCF (Log BA) | | |
S.No.	Observed	Predicted	Residual	S.No.	Observed	Predicted	Residual
2	2.04	2.41	0.37	100	0.00	1.27	1.27
3	1.92	2.44	0.52	101	2.08	2.47	0.38
9	0.60	2.19	1.59	102	1.84	2.18	0.33
21	1.99	2.74	0.75	103	1.95	2.71	0.76
22	2.15	2.63	0.49	111	1.71	2.16	0.46
26	1.04	2.00	0.95	117	2.23	2.57	0.35
27	1.76	2.24	0.49	118	2.45	2.54	0.09
28	1.53	2.33	0.80	124	0.95	1.89	0.94
29	1.00	2.37	1.37	125	0.60	1.61	1.01
30	1.89	2.75	0.86	62	1.76	2.19	0.43
31	1.23	2.36	1.13	19	2.21	2.49	0.27
36	1.51	2.14	0.63	7	2.39	2.45	0.07
42	2.20	2.48	0.28	54	1.97	2.41	0.44
43	2.16	2.66	0.50	56	1.97	2.37	0.39
76	1.04	1.75	0.70	75	1.34	2.01	0.67

The relationships between predicted (both training and test) activities and the corresponding experimental activities are shown in Figures 2 and 3. The r^2_{train} and q^2_{cv} have values of 0.721 and 0.678, respectively, which corroborate with the criteria for a QSAR model to be highly predictive [22]. The standard error of estimate for the model was 0.24, which is an indicator of the robustness of the fit and suggested that the predicted PCPDCF based on equation (1) is reliable.
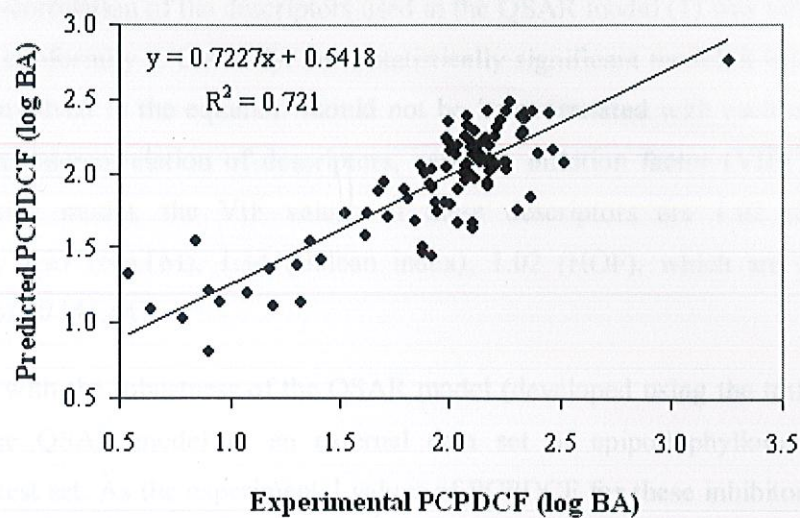
90

Figure 2. Relationship between predicted and experimental PCPDCF as per equation (1) of the Training set compounds.
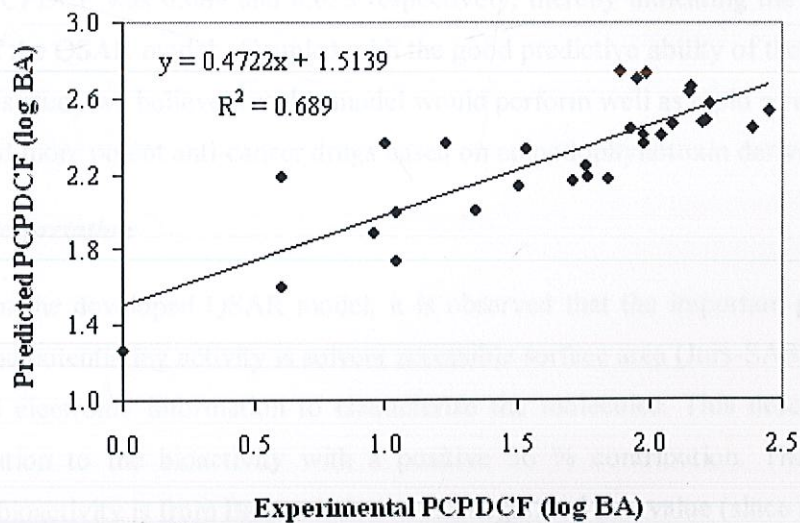

Figure 3. Relationship between predicted and experimental PCPDCF as per equation (1) of the Test set compounds.

91

The inter-correlation of the descriptors used in the QSAR model (1) was very low (below 0.6), which is in conformity to the study. For a statistically significant model, it is necessary that the descriptors involved in the equation should not be intercorrelated with each other [26]. To further check the intercorrelation of descriptors, variance inflation factor (VIF) analysis was performed. In this model, the VIF values of these descriptors are 1.02 (SASA), 1.09 (ES_Sum_aasC), 1.27 (NATM), 1.34 (Balban index), 1.02 (HOF), which are less than the threshold value of 10 [43,44].

Satisfied with the robustness of the QSAR model (developed using the training set), we have applied the QSAR model to an external data set of epipodophyllotoxin analogues constituting the test set. As the experimental values of PCPDCF 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 5 represents the predicted PCPDCF of the test set based on Equation (1). The overall root mean square error (RMSE) between the experimental and predicted PCPDCF was 0.194, which revealed good predictability. The estimated correlation coefficients ($r^2_{(test)}$) and the cross validated correlation coefficient ($q^2_{cv(test)}$) between experimental and predicted PCPDCF was 0.689 and 0.623 respectively, thereby indicating the good external predictability of the QSAR model. Coupled with the good predictive ability of the QSAR model developed in this study we believe that this model would perform well as rapid screening tools to uncover new and more potent anti-cancer drugs based on epipodophyllotoxin derivatizations.

Descriptors interpretation

Based on the developed QSAR model, it is observed that the important parameter that contributes to the potentiating activity is solvent accessible surface area (Jurs-SASA). It includes both shape and electronic information to characterize the molecules. This descriptor has the largest contribution to the bioactivity with a positive 36 % contribution. The next largest contribution to bioactivity is from Balban index with a negative 32 % value (since the coefficient is negative). Balban index is a type of topological descriptor and calculated based on 2D structure of the molecule. It is inversely proportional to the electronegativities and covalent radii of the atoms in the molecules. The third largest contribution to the biological activity comes from the descriptor ES_Sum_aasC with a positive 22.2 %. It is the E-state index of an atom type which is the sum of the standard value for the atom type and the perturbation from the other

92

atoms in the molecule. In this descriptor 'a' represents an aromatic bond, 's' is the single bond and 'C' is the carbon atom. Hence it describes the electro-topological-state index of the aromatic carbon atoms linked by single bonds. This is supported well if we compare the molecules consisting of aromatic rings and the molecules consisting of no aromatic rings. In general substitution of aromatic groups in the scaffold structure (R group) has higher complex formation. The contributions of other two descriptors such as number of atomic classes (NATM) and heat of formation (HOF) are very low with a contribution of 7% and 1.4%, respectively. The descriptors which have been used for constructing QSAR model in the present work encoded electronic, geometrical and topological aspects of molecules and revealed inhibition of human TP-IIα in cancer cells.

Conclusion

We have compiled a virtual library of epipodophyllotoxin analogues built through structural modification of scaffold structure of natural podophyllotoxin. QSAR modeling was done in the work to get insights into ligand:TP-IIα interactions and corresponding PCPDCF of epipodophyllotoxin analogues. We have demonstrated that the QSAR model developed in this study can be applied to estimate the PCPDCF with a high level of accuracy for a diverse set of epipodophyllotoxin analogues. Using a combination of topological, electro-topological-state indices, electronic and thermodynamic descriptors of chemical structures, we have built several robust QSAR models with high values of q^2_{cv} (for training sets) and predictive r^2_{test} (for test set). The calculated PCPDCF value of a set of structural analogues demonstrates excellent linear correlation to the experimental PCPDCF value. This model could be useful to predict the range of activities for new epipodophyllotoxin analogues. The information we have expressed in this study may lead to the designing (synthesis) of more potent epipodophyllotoxin derivatives for inhibition of human TP-IIα (anticancer activity) and facilitate the search for related structures with similar biological activity from a large number of databases.

93

References

[1] Beck WT, Chen M, Danks MK, Kim R, Wolverton JS: Drug resistance associated with altered DNA topoisomerase II. Adv Enzyme Regul. 1993; 33: 113-127.

[2] Jardine I, Podophyllotoxins In Anticancer Agents Based on Natural Products Models. Cassady JM Douros J Eds.Academic Press: New York **1980** 319-35.

[3] Issell BF: The Podophyllotoxin Derivatives VP-16-213 AND VM-26. Cancer Chemother Pharmacol 1982; 7: 73-80.

[4] Aisner J, Belani CP, Doyle LA. Etoposide: Current status and future perspectives in the management of malignant neoplasms. Cancer Chemother Pharmacol 1994 ; 34: 118-123.

[5] Lee KH, Imakura Y, Haruna M, Beers SA, Thurston LS, Dai HJ, Chen CH: Antitumor Agents.107: New cytotoxic 4-alkylamino analogues of 4'-demethylepipodophyllotoxin as inhibitors of Human DNA Toposiomerase II. J Nat Prod 1989; 52: 606-613.

[6] Wang ZQ, Kuo YH, Schnur D, Bowen JP, Liu SY, Han FS, Cheng YC, Lee KH: Antitumor Agents 113:New 4-arylamino derivatives of 4'-O-demethyl-epipodophyllotoxin and related compounds as potent inhibitors of Human DNA toposiomerase II. J Med Chem 1990 ; 33: 2660-2666.

[7] Lee KH, Beers SA, Mori M, Wang ZQ, Kuo YH, Li l, Liu SY, Cheng JY, Han FS, Cheng YC: Antitumor Agents.111: New 4-hydroxylated and 4-halogenated anilino derivatives of 4'-demethylepipodophyllotoxin as potent inhibitors of Human DNA toposiomerase II. J. Med Chem 1990 ; 33: 1364-1368.

[8] Zhou XM, Wang ZQ, Cheng JY, Chen HX, Cheng YC, Lee KH :Antitumor Agents. 120:New 4-substituted benzylamine and benzyl ether derivatives of 4'-O- demethyl-epipodophyllotoxin as potent inhibitors of Human DNA toposiomerase II. J Med Chem 1991 ; 33: 3346-3350.

[9] Osheroff N, Zechiedrich EL, Gale KC: Catalytic function of DNA Topoisomerase II. BioEssays 1991 ; 13: 269-275.

[10] Cramer RD, Depriest SA, Patterson DE, Hecht P: The Developing Practice of Comparative Molecular Field Analysis. In 3D QSAR in Drug Design: Theory methods and applications Kubinyi H Ed. ESCOM: Leiden 1993 443-485.

[11] Cho SJ, Tropsha A, Suffness M, Cheng YC, Lee KH:Antitumor Agents. 163:Three-dimensional quantitative structure-activity relationship study of 4'-O-Demethylepipodophyllotoxin analogues using the modified CoMFA/q2-GRS approach. J Med Chem 1996 ; 39: 1383-1395.

[12] Kim KH, Brusniak MYK, Perlman RS, UniSur-CoMFA: for stable and consistent 3D-QSAR.Alfred Benzon Symp 42 (Rational Molecular Design in Drug Research): 1998 67-86.

[13] Rogers D, Hopfinger AJ:Application of genetic function approximation to quantitative structure-activity relationship and quantitative structure-property relationships. J Chem Inf Comput Sci 1994 ; 34: 854-866.

[14] Gregorio deC, Kier LB, Hall LH:QSAR modeling with the electrotopological state indices: corticosteroids. J Comput Aid Mol Des 1998 ; 12: 557–561.

[15] Liu SS, Cao CZ, Li ZL: Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance edge (MDE) vector l. J Chem Inf Comput Sci 1998 ; 38: 387–394.

[16] Lipinski A, Lombardo F, Dominy B, Feeney P: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Del Rev 2001 ; 46: 3–26.

[17] Goldberg DE: Genetic algorithm in search optimization and machine learning , Addion-Wesley: Reading MA 1989.

[18] Forrest S: Genetic algorithms :Principles of natural selection applied to computation. Science 1993 ; 261: 872-878.

[19] Deswal S, Roy N: Quantitative structure activity relationship studies of aryl heterocycle-based thrombin inhibitors. J Med Chem 2006 ; 41(11): 1339–1346.

[20] Schrodinger LLC. http://www.schrodinger.com (accessed: 24. 04.2007). Schrodinger. 2000 Inc.: Portland OR.

[21] ADME Works ModelBuilder version 4.5 Fujitsu Kyushu System Engineering Ltd. 2007.

[22] Golbraikh A, Tropsha A:Beware of q^2. J. Mol Graph Model 2002 ; 20(4): 269–276.

[23] Roy PP, Roy K:On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci 2008 ; 27: 302–313.

[24] Jaiswal M, Khadikar PV, Scozzafava A, Supuran CT: Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. Bioorg Med Chem Lett 2004 ; 14: 3283–3290.

[25] Shapiro S, Guggenheim B:Inhibition of oral bacteria by phenolic compounds. Part 1 QSAR analysis using molecular connectivity. Quant Struct Act Relat 1998 ; 17: 327–337.

[26] Deswal S, Roy N: Quantitative structure activity relationship studies of aryl heterocycle-based thrombin inhibitors. Eur J Med Chem 2006 ; 41: 1339-1346.

LIST OF PUBLICATIONS

1. "Development of predictive structure activity relationship models of Epipododphyllotoxin".By:- Dr. Pradeep Kumar Naik, Abhishek Dubey and Rishay Kumar (2010). Communicated to Journal of Biomolecular screening.

2. "The Binding Modes and Binding Affinities of Epipodophyllotoxin Derivatives with Human Topoisomerase IIα".By:- Dr. Pradeep Kumar Naik, Abhishek Dubey and Rishay Kumar (2010). Communicated to Journal of Molecular Graphics and modelling-Elsevier.