Development of Predictive Quantitative Structure-Activity Relationship Models of Epipodophyllotoxin Derivatives

PRADEEP KUMAR NAIK, ABHISHEK DUBEY, and RISHAY KUMAR

Epipodophyllotoxins are the most important anticancer drugs used in chemotherapy for various types of cancers. To further, improve their clinical efficacy a large number of epipodophyllotoxin derivatives have been synthesized and tested over the years. In this study, a quantitative structure-activity relationship (QSAR) model has been developed between percentage of cellular protein-DNA complex formation and structural properties by considering a data set of 130 epipodophyllotoxin analogues. A systematic stepwise 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. A statistically significant model ($r^2_{(train)} = 0.721$; $q^2_{cv} = 0.678$) was obtained with descriptors such as solvent-accessible surface area, heat of formation, Balaban 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 root mean square error between the experimental and predicted percentage of cellular protein–DNA complex formation (PCPDCF) was 0.194 and $r^2_{(test)} = 0.689$, revealing good predictability of the QSAR model. (*Journal of Biomolecular Screening* 2010:1194-1203)

Key words: epipodophyllotoxin, quantitative-structure activity relationship, genetic algorithm, variable selection, ADME model builder

INTRODUCTION

EPIPODOPHYLLOTOXINS ARE THE GLUCOSIDIC DERIVATIVES of podophyllotoxin, which has been used in chemotherapy of various types of cancer, including small cell lung cancer, testicular carcinoma, lymphoma, and Kaposi's sarcoma.[1-3] Etoposide (VP-16) and teniposide (VM-26) are the most successful widely prescribed chemotherapeutic agents. Efforts for further improving their clinical efficacy by overcoming drug resistance, myelosuppression, and poor bioavailability problems[4] associated with them have continued to be challenging. Over the years, a number of laboratories throughout the world have engaged in the synthesis and testing of epipodophyllotoxin derivatives[5-8] to prepare new more 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 further improve design of potentially bioactive compounds, there is a need for the development of predictive quantitative SAR (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 an inherent limitation in aligning 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 the unavailability of the X-ray structure. We should have some knowledge or hypothesis regarding active conformations of the molecules under study as a prerequisite for structural alignment. Nevertheless, especially for structurally diverse molecules, unambiguous 3D alignment makes 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, in this QSAR study, we have applied topological, electronic, geometrical, and energy-based descriptors calculated directly from the 2D and 3D structure of the molecules. This approach is simple, fast, and straightforward. It benefits in predicting the activities of a large set of molecules

Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India.

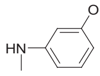
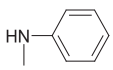
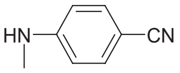
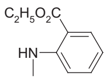
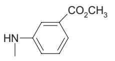
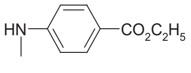
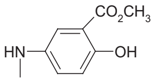
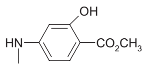
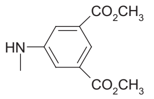
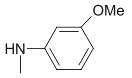
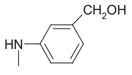
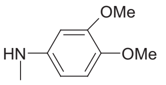
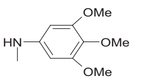
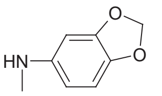
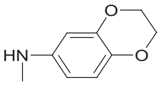
Received May 1, 2010, and in revised form Jun 4, 2010. Accepted for publication Jun 23, 2010.

Supplementary material for this article is available on the *Journal of Biomolecular Screening* Web site at http://jbx.sagepub.com/supplemental.

Journal of Biomolecular Screening 15(10); 2010
DOI: 10.1177/1087057110380743

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. In addition, these techniques are not computationally intensive and are practically automated. The behavior of the QSAR model is examined with a variety of statistical parameters,[19] and the contribution of various descriptors is analyzed.

Table 1. List of Epipodophyllotoxin Analogues and Their Experimental Activities

Compound No.	*R-Group*	*Scaffold Type* **(Fig. 1)**	*Log (PCPDCF)*
1	—OH	1	1.63
2	-NHCH$_2$CH$_2$OCH$_3$-	1	2.04
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		1	2.26
18		1	1.67
19		1	2.21
20		1	2.45

(continued)

MATERIALS AND METHODS

Data set

A total of 130 epipodophyllotoxin analogues (**Table 1**) were used in the study, and were synthesized and tested under the same conditions in the same laboratory.[5-8] To generate statistically robust and, most important, validated models, all compounds in the original data set were divided randomly into 100 molecules in the training set and 30 molecules in the 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 originally were 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.

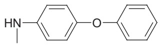
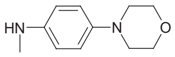
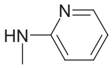
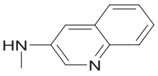
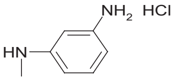
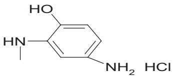
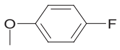
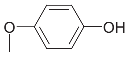
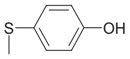
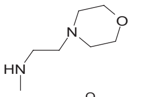
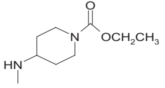
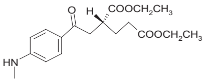
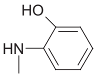
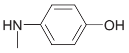
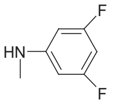
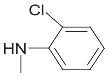
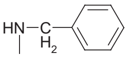
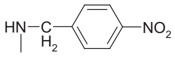
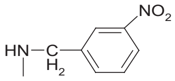
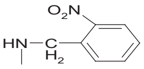
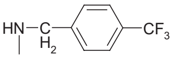
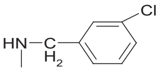
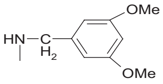
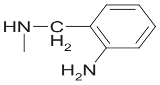
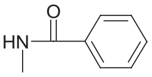
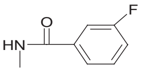
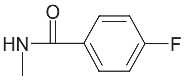
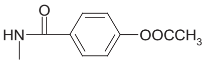
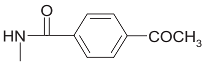
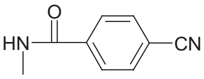
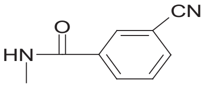
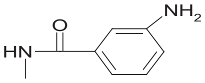
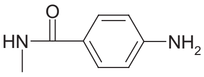
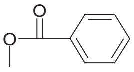
Building of molecular structures

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

Descriptor calculation

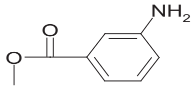
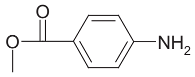
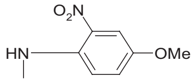
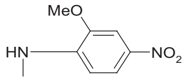
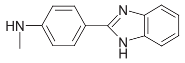
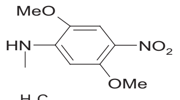
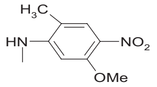
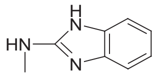
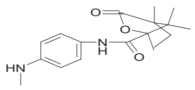
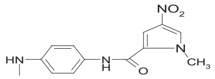
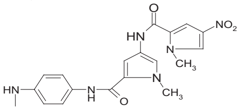
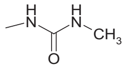
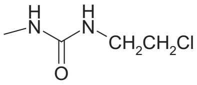
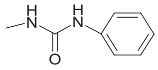
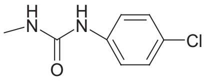
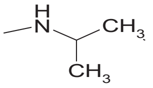
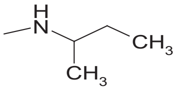
All the molecular descriptors such as E-state indices; log P; superpendentic index; structural, symmetrical, and topological

Table 1. (continued)

Compound No.	*R-Group*	*Scaffold Type* **(Fig. 1)**	*Log (PCPDCF)*
21		1	1.99
22		1	2.15
23		1	1.99
24		1	2.09
25		1	2.15
26		1	1.04
27		1	1.76
28		1	1.53
29		1	1.00
30		1	1.89
31		1	1.23
32		1	1.92
33		1	2.18
34		1	2.32
35		1	2.06
36		1	1.51
37		1	3.26
38		1	2.33
39		1	2.11
40		1	2.16
41		1	2.00
42		1	2.20
43		1	2.16
44		1	2.26
45		1	2.25
46		1	2.06
47		1	2.07
48		1	2.14
49		1	2.09
50		1	2.20
51		1	2.17
52		1	2.17
53		1	2.08
54		1	1.97

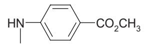
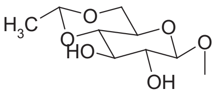
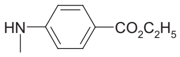
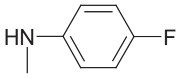
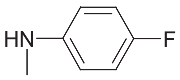
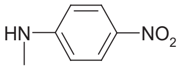
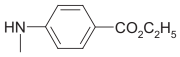
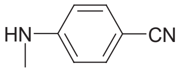
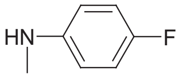
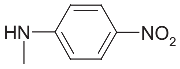
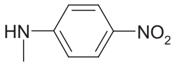
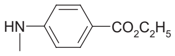
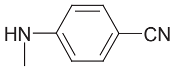
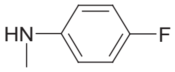
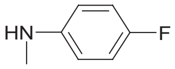
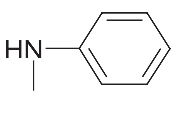
(continued)

Table 1. (continued)

Compound No.	*R-Group*	*Scaffold Type* **(Fig. 1)**	*Log (PCPDCF)*
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

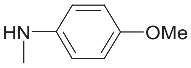
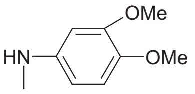
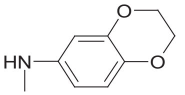
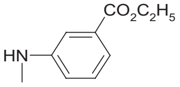
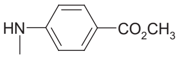
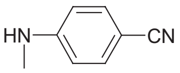
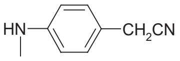
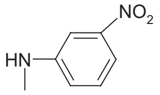
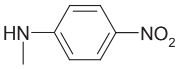
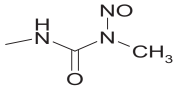
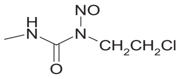
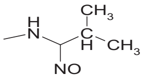
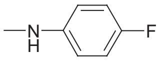
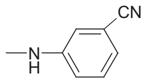
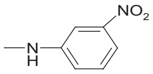
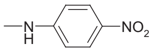
(continued)

Table 1. (continued)

Compound No.	*R-Group*	*Scaffold Type* **(Fig. 1)**	*Log (PCPDCF)*
72		1	2.32
73		2	0.79
74	**—OH**	2	1.19
75		3	1.34
76		3	1.04
77		4	1.72
78		5	1.88
79		5	2.10
80		5	2.10
81		5	2.03
82		3	1.36
83		6	0.90
84		6	0.95
85		6	1.08
86		6	0.90
87		7	2.07
88		7	2.02

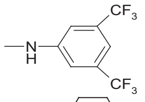
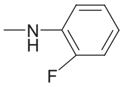
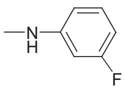
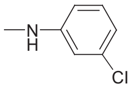
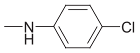
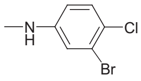
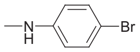
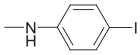
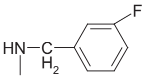
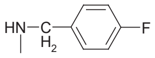
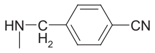
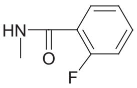
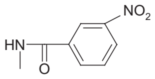
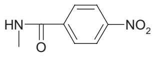
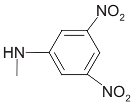
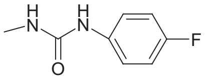
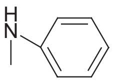
(continued)

Table 1. (continued)

Compound No.	*R-Group*	*Scaffold Type* **(Fig. 1)**	*Log (PCPDCF)*
89		7	1.98
90		7	1.84
91		7	2.08
92		7	1.97
93		7	2.24
94		7	2.16
95		7	2.04
96		7	1.88
97		7	2.30
98		8	1.61
99		8	0.85
100		9	0.00
101	—NHCH$_2$CH$_2$OH	1	2.08
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

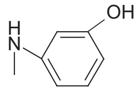
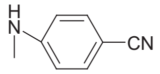
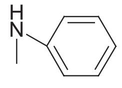
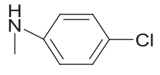
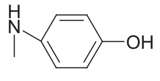
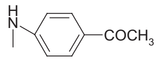
(continued)

Table 1. (continued)

Compound No.	*R-Group*	*Scaffold Type* **(Fig. 1)**	*Log (PCPDCF)*
108		1	1.32
109		1	2.08
110		1	2.20
111		1	1.71
112		1	2.00
113		1	1.79
114		1	2.25
115		1	1.81
116		1	2.33
117		1	2.23
118		1	2.45
119		1	2.11
120		1	1.93
121		1	2.20
122		1	1.30
123		1	2.07
124		3	0.95

(continued)

Table 1. (continued)

Compound No.	*R-Group*	*Scaffold Type* **(Fig. 1)**	*Log (PCPDCF)*
125		3	0.60
126		3	1.52
127		7	2.11
128		7	1.89
129		7	1.92
130		7	2.17

PCPDCF, percentage of cellular protein–DNA complex formation.

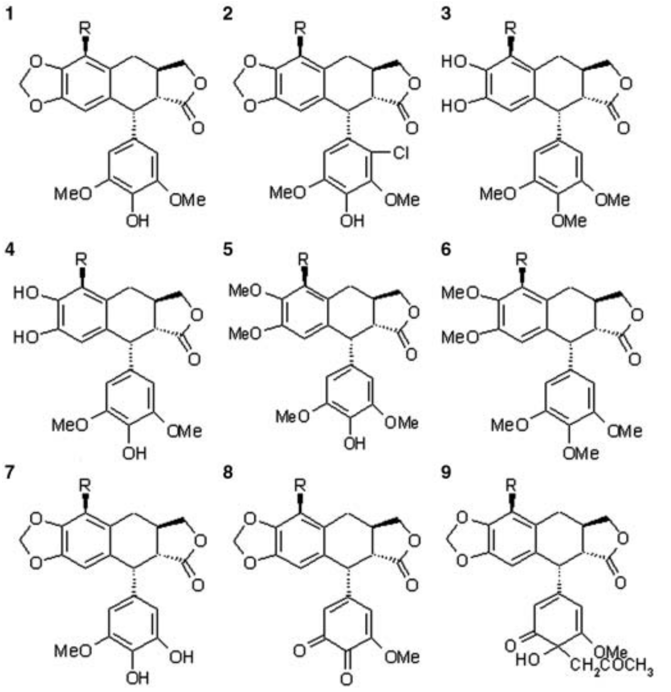


FIG. 1. The various scaffold structures used for building the epipodophyllotoxin analogues.

descriptors; lead likeness; electronic Wang-Ford atomic charge and extended Huckel partial charge functions; bulk; moments; orbital energies; molecular connectivity indexes; gravitational indexes; hydrophobicity; and steric and thermodynamic factors 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**.

Screening of descriptors and development of QSAR model

A set of 372 molecular descriptors was calculated using the ADME Model Builder software package (version 4.5). A systematic search in the order of missing value test, zero test, correlation coefficient, and genetic algorithm was performed to determine significant descriptors using the ADME Model Builder (version 4.5) software package (Fujitsu, Fukuoka, Japan). Any molecular descriptor that 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). 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 multicollinearity (high multiple correlations between subsets of the variables). From the descriptors thus remaining, the selection of variables to obtain the QSAR models was carried out using a genetic algorithm implemented in the ADME Model Builder (version 4.5) software package. The genetic algorithm (GA) works in the following way: first, a particular number of equations (set at 100 in this study) are generated randomly. Then pairs of "parent" equations are chosen randomly from this set of 100 equations, and progeny equations are generated performing "crossover" and "mutations" operations at random. The parameters set used for the GA included the following: mutation 0.1, crossover 0.9, number of generations 1000, r^2 floor limit 50% (lowest acceptable r^2), and objective function r^2/N_par. The goodness of each progeny equation is assessed by the objective function, which is a mathematical function used for ranking the progeny equation. It favors the equation that has the highest correlation (r^2) with the biological activity, while minimizing the number of molecular descriptors (*N_par*). Finally, the 2 best equations were selected (one with 5 descriptors and the other with 6 descriptors) based on objective function for comparison. The effect of each molecular descriptor on the statistical quality of the model developed was assessed by applying the brute-force approach. Initially, the QSAR equation was developed by taking a single descriptor that had the highest correlation with the biological activity. To this equation, the second best descriptor was added, and likewise we increased the number of molecular descriptors one by one and evaluated the effect of adding new terms on the statistical quality of the model. 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}).

Table 2. List of Descriptors Used in the Study

Type	*Descriptors*
E-state indices	Electrotopological 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, nonbonded strain energy, torsional strain energy, total strain energy of molecule
Lead likeness	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), Balaban indices

Validation of the QSAR model

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

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

where y_{pred}, y_{exp}, and $\bar{y}$ 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, 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 Microsoft 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]:

$$q^2_{test} = 1 - \frac{\sum(Y_{pred_{Test}} - Y_{Test})^2}{\sum(Y_{Test} - \bar{Y}_{Training})^2}$$

where $Y_{pred_{Test}}$ and Y_{Test} are the predicted value based on the 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.

We also carried out the leave-10%-out and leave-20%-out cross-validation for the training set to further validate the QSAR model internally.

To check the intercorrelation 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 PCPDCF were randomly divided into a training set of 100 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 is ideal as training and test sets, as both 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 a cutoff value of 0.6, and a multicollinearity test with a cutoff value of 0.9, we have discarded the most likely descriptors, resulting in 218 descriptors. Further additional descriptors were discarded by applying the GA, and finally 6 descriptors (2 equations were selected: one equation with 5 descriptors and the other with additional descriptors in comparison to the previous one) were selected for the development of the QSAR equation. The initial QSAR equation was developed by considering a single molecular descriptor that showed the highest correlation with the

Table 3. Statistical Assessment of Quantitative Structure-Activity Relationship (QSAR) Equations with Varying Number of Descriptors

No. of Descriptors	*QSAR Equation*	r^2	q^2	*PRESS*	F-*Value*
1	Log PCPDCF = 1.23 + 2.98 * SASA	0.26	0.22	15.17	34.72
2	Log PCPDCF = 0.589 + 3.00 * SASA + 0.148 * ES_Sum_aasC	0.43	0.35	12.67	36.03
3	Log PCPDCF = 1.59 + 2.71 * SASA + 0.148 * ES_Sum_aasC – 0.00277 * NATM	0.48	0.40	11.66	29.14
4	Log PCPDCF = 7.52 + 2.89 * SASA + 0.0906 * ES_Sum_aasC – 0.0818 * NATM – 2.51 * Balaban index	0.71	0.67	6.37	58.33
5	Log PCPDCF = 7.25 + 2.97 * SASA + 0.108 * ES_Sum_aasC – 0.0813 * NATM – 2.48 Balaban index – 0.000794 * HOF	0.72	0.68	6.24	48.71
6	Log PCPDCF = 7.33 + 2.87 * SASA + 0.106 * ES_Sum_aasC – 0.0761 * NATM – 2.39 * Balaban index – 0.000757 * HOF + 0.0182 * ESP_minimum	0.73	0.67	6.37	41.12

HOF, heat of formation; NATM, number of atomic classes; PCPDCF, percentage of cellular protein–DNA complex formation; SASA, solvent-accessible surface area.

Table 4. Regression Properties of Molecular Descriptors with Experimental Activities (Log PCPDCF)

Descriptor	*SASA*	*ES_Sum_aasC*	*NATM*	*Balaban Index*	*HOF*
R^2	0.261	0.162	0.107	0.039	0.01
Intercept (B)	1.228	1.318	3.281	3.074	2.05
Standard error of B	0.128	0.150	0.391	0.566	0.114

HOF, heat of formation; NATM, number of atomic classes; PCPDCF, percentage of cellular protein–DNA complex formation; SASA, solvent-accessible surface area.

biological activity (PCPDCF). To this equation, by applying a brute-force approach, we increased the number of parameters in the QSAR equation one by one and evaluated the effect of adding 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 number of descriptors was restricted to 5 in the final QSAR model. By graphing these molecular descriptors versus activity (log PCPDCF), all have poor correlation to experimental activity of the set of ligands (**Table 4**).

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

$$\text{Log PCPDCF} = 7.25 + 2.97 * \text{SASA} + 0.108 * \text{ES\_Sum\_aasC} - 0.0813 * \text{NATM} - 2.48 \text{ Balaban index} - 0.000794 * \text{HOF} \quad (1)$$
$$(n = 100;\ r^2_{train} = 0.721;\ s = 0.24;\ \text{PRESS} = 6.240;\ r^2_{adj} = 0.707;\ q^2_{cv} = 0.678;\ F\text{-test} = 48.71)$$

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 the adjusted correlation coefficient for degrees of freedom, F-test is the measure of variance that compares 2 models differing by 1 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 the 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 the training and test sets of molecules as reported in Supplementary Tables S1 and S2. 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. Leave-10%-out and leave-20%-out cross-validation for the training set was also performed. The value of r^2 and q^2_{cv} in all 10 cycles was greater than 0.5 (r^2 value 0.627-0.809 and q^2_{cv} value 0.605-0.78) using leave-10%-out cross-validation. Based on the leave-20%-out cross-validation technique, the value of r^2 and q^2_{cv} in all 5 cycles was greater than 0.5 (r^2 value 0.657-0.794 and q^2_{cv} value 0.601-0.752), hence further validating the model internally.

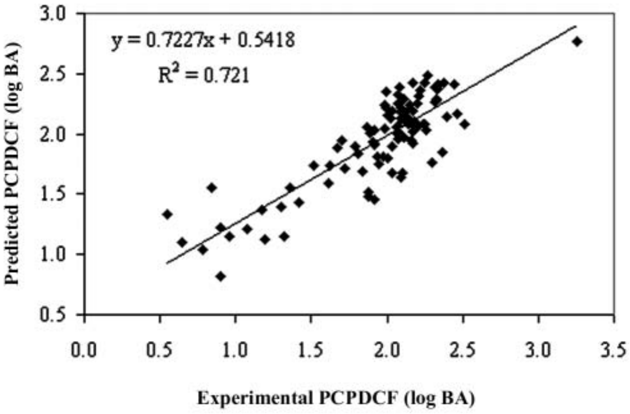


FIG. 2. Relationship between predicted and experimental percentage of cellular protein–DNA complex formation (PCPDCF) as per Equation (1) of the training set compounds.

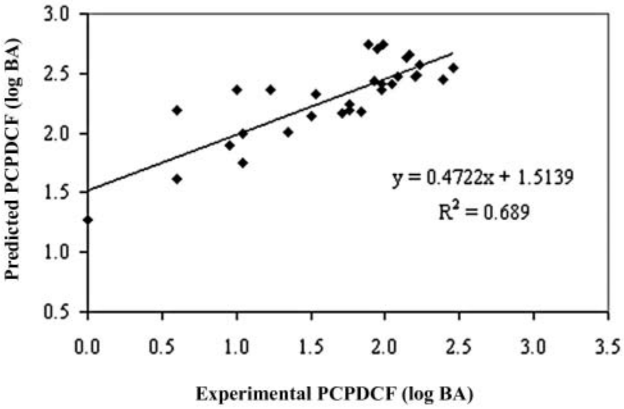


FIG. 3. Relationship between predicted and experimental percentage of cellular protein–DNA complex formation (PCPDCF) as per Equation (1) of the test set compounds.

The intercorrelation 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, VIF analysis was performed. In this model, the VIF values of these descriptors are 1.02 (solvent-accessible surface area [SASA]), 1.09 (ES_Sum_aasC), 1.27 (number of atomic classes [NATM]), 1.34 (Balaban index), and 1.02 (heat of formation [HOF]), which are less than the threshold value of 10.[22,23]

Satisfied with the robustness of the QSAR model (developed using the training set), we have applied the QSAR model to a test data set of epipodophyllotoxin analogues. 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. Supplementary Table S2 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 anticancer drugs based on epipodophyllotoxin derivatizations.

Descriptors interpretation

Based on the developed QSAR model, it is observed that the important molecular descriptor that contributes to the potentiating activity is the 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 the Balaban J-index[27] with a negative 32% value (since the coefficient is negative). The Balaban index is a type of topological descriptor and calculated based on the 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 that is the sum of the standard value for the atom type and the perturbation from the other 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 electrotopological state index of the aromatic carbon atoms linked by single bonds. This is well supported 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 a higher complex formation. The contributions of the other two descriptors, NATM and HOF, are very low with a contribution of 7% and 1.4%, respectively. The descriptors that have been used for constructing the QSAR model in the present work encoded electronic, geometrical, and topological aspects of molecules.

CONCLUSION

We have compiled a virtual library of epipodophyllotoxin analogues built through structural modification of the scaffold structure of natural podophyllotoxin. QSAR modeling was done 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 and electrotopological state indices, as well as 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 sets). The calculated PCPDCF value of a set of structural analogues demonstrates good 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 design (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.

ACKNOWLEDGMENT

We thank the authorities of Scube Scientific Software for supplying the ADME Model Builder for a trial period.

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 Cassady JM & Douros J (eds): *Anticancer Agents Based on Natural Products Models*. New York: Academic Press, 1980:319-335.
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, et al: 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, et al: 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 I, et al: 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 Kubinyi, H (ed): *3D QSAR in Drug Design: Theory Methods and Applications*. Leiden, The Netherlands: ESCOM, 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. Paper presented at Alfred Benzon Symposium No. 42, Munksgaard, Copenhagen, Denmark, July 1997.
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. de Gregorio C, 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, λ. *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 Deliv Rev* 2001;46:3-26.
17. Goldberg DE: *Genetic Algorithm in Search Optimization and Machine Learning*. Reading, MA: Addison-Wesley, 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;41:1339-1346.
20. Schrodinger LLC: http://www.schrodinger.com. Accessed April 24, 2007.
21. ADME Works Model Builder, version 4.5. Fukuoka, Japan: Fujitsu Kyushu System Engineering Ltd., 2007.
22. Golbraikh A, Tropsha A: Beware of q^2. *J Mol Graph Model* 2002;20:269-276.
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.
27. Balaban AT: Highly discriminating distance-based topological index. *Chem Phys Lett* 1982;80:399-404.

Address correspondence to:
Pradeep Kumar Naik, Ph.D.
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, Distt.- Solan, Himachal Pradesh, India (Pin-173215)

E-mail: pknaik73@rediffmail.com