

Antitumor Agents. 213.[†] Modeling of Epipodophyllotoxin Derivatives Using Variable Selection *k* Nearest Neighbor QSAR Method

Zhiyan Xiao, Yun-De Xiao, Jun Feng, Alexander Golbraikh, Alexander Tropsha,* and Kuo-Hsiung Lee*

Natural Products Laboratory, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

Received November 27, 2001

We have applied a variable selection *k* nearest neighbor quantitative structure–activity relationship (*k*NN QSAR) method to develop predictive QSAR models for 157 epipodophyllotoxins synthesized previously in our ongoing effort to develop potential anticancer agents. QSAR models were generated using multiple topological descriptors of chemical structures, including molecular connectivity indices (MCI) and molecular operating environment descriptors. The 157 compounds were separated into several training and test sets. The robustness of QSAR models was characterized by the values of the internal leave one out cross-validated R^2 (q^2) for the training set and external predictive R^2 for the test set. The significance of the training set models was confirmed by statistically higher values of q^2 for the original data set as compared to q^2 values for the same data set with randomly shuffled activities. *k*NN QSAR models were compared with those obtained with the comparative molecular field analysis method; the *k*NN QSAR approach afforded models with higher values of both q^2 and predictive R^2 . One of the best models obtained from *k*NN analysis using MCI as descriptors provided q^2 and predictive R^2 values of 0.60 and 0.62, respectively. QSAR models developed in these studies shall aid in future design of novel potent epipodophyllotoxin derivatives.

Introduction

Etoposide (VP-16, Figure 1a) and teniposide (VM-26, Figure 1b) are two semisynthetic glucosidic cyclic acetals of podophyllotoxin (Figure 1c). They are currently used in the chemotherapy for various types of cancer, including small cell lung cancer, testicular carcinoma, lymphoma, and Kaposi's sarcoma.^{2,3} To improve their clinical efficacy and overcome the problems of drug resistance, myelosuppression, and poor oral availability,^{4–6} we have been engaged for years in the synthesis and testing of epipodophyllotoxin derivatives.^{7–18} Interestingly, although podophyllotoxin is known as an antimicrotubule agent, the epipodophyllotoxins, its 4 β -congeners, are potent inhibitors of DNA topoisomerase II. The proposed mechanism of epipodophyllotoxins' antitopoisomerase II activity is to inhibit the catalytic activity of the target enzyme by stabilizing the covalent topoisomerase II–DNA cleavable complex.^{19,20}

An early structure–activity relationship (SAR) study²¹ suggested that the structural features essential for the antitopoisomerase activity include 4'-demethylation, 4-epimerization, and 4-substitution. On the basis of this assumption, most of our research efforts have been focused on exploring different 4-substituted 4'-*O*-demethylepipodophyllotoxin (DMEP) derivatives. Some of these derivatives (e.g., NPF and GL-331)^{16,22} have displayed better pharmacological profiles than those of etoposide. GL-331 has been successfully pushed into phase II clinical trials against gastric carcinoma, colon cancer, nonsmall cell carcinoma, and etoposide-resistant malignancies.²³

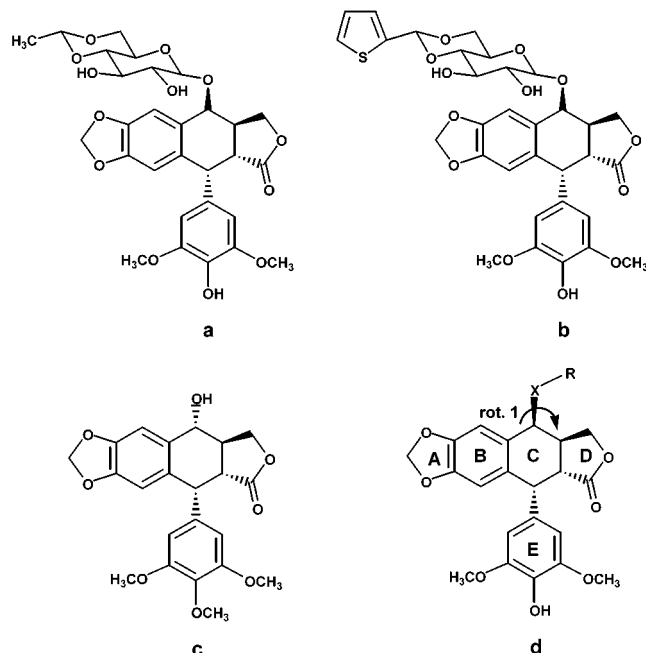


Figure 1. Representative epipodophyllotoxin derivatives. (a) Etoposide; (b) teniposide; (c) podophyllotoxin; (d) generalized structure of epipodophyllotoxins.

To construct an informative SAR model and improve further design of potentially bioactive compounds, a previous molecular modeling study from these groups²⁴ applied comparative molecular field analysis (CoMFA)²⁵ and CoMFA/ q^2 -guided region selection (GRS)²⁶ techniques to build three-dimensional (3D) quantitative SAR (QSAR) models for 102 DMEP derivatives. The contour plots of the final model matched well with the composite pharmacophore model proposed by MacDonald²⁷ and

* To whom correspondence should be addressed. Tel: 1-919-962-0066. Fax: 1-919-966-3893. E-mail: A.T., Alex_tropsha@unc.edu. K.-H.L., khlee@unc.edu.

[†] For part 212, see ref 1.

agreed with the hypothesis^{24,28} that the variously substituted region at C₄ is responsible for drug interaction with the DNA minor groove. On the basis of these models, several novel analogues were designed and synthesized.^{17,18,29} Some of them showed comparable or superior biological activities in terms of cellular protein–DNA complex formation,¹⁸ as compared to the etoposide prototype.

CoMFA is one of the most popular methods for QSAR and is characterized by reasonable simplicity and a clear physicochemical sense of steric and electrostatic descriptors.³⁰ However, despite many successful applications, several problems have persisted with this method. As we have shown previously,²⁴ the results of conventional CoMFA may often be nonreproducible due to a sometimes strong dependence of the cross-validated correlation coefficient, q^2 , on the orientation of rigidly aligned molecules on the users' terminal. Other groups have demonstrated that q^2 is dependent on the lateral shift of molecules along the CoMFA grid.³¹ We have provided a solution to this problem^{24,26} by developing a q^2 -GRS method, which was based on the rational selection of only the most significant regions in the steric and electrostatic fields of aligned molecules. Nevertheless, especially for structurally diverse molecules, unambiguous 3D alignment to initiate the CoMFA process is still a difficult task. In some reported cases, CoMFA has been effectively applied only by having the knowledge of 3D receptor structures.^{32,33}

We, as well as other researchers,³⁴ were motivated to explore possible alternatives that would use alignment-free descriptors derived from two-dimensional (2D) molecular topology and, thus, alleviate frequent ambiguity of structural alignment typical for 3D QSAR methods. Accordingly, in this QSAR study, we have applied topological indices developed on the basis of chemical graph theory.^{35,36} Furthermore, we have implemented the concept of variable selection, a process that has been investigated recently by a number of researchers^{34,37,38} using such optimization methods as evolutionary,^{39,40} genetic,^{41,42} or simulated annealing (SA) algorithms.^{43,44} From these considerations, we have developed the variable selection approach, k nearest neighbor (k NN) QSAR,⁴⁵ which employs topological descriptors of chemical structures. 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. They have produced predictive models that were comparable or superior to those obtained with conventional CoMFA.

In this paper, we have applied the k NN QSAR method⁴⁵ to a data set of 157 epipodophyllotoxin derivatives, which were synthesized and tested in one of our laboratories previously. The k NN QSAR models were generated using molecular connectivity indices (MCI)^{46,47} and molecular operating environment (MOE)⁴⁸ descriptors. MCI descriptors are derived from 2D molecular topology, while MOE descriptors incorporate a large variety of numerical descriptions, derived from both 2D molecular topology and 3D molecular topography. The latter descriptors, which combine both topological and topographical information, are perhaps more informa-

tive than topological indices; yet, they are still insensitive to 3D alignment.

To generate statistically robust and, most importantly, validated models, all compounds in the original data set were separated into several training and test sets, using specially designed diversity sampling methods. Variable selection QSAR models were generated for the training set and applied to predict biological activities of the test set. The robustness of the models was evaluated from the values of the internal leave one out cross-validated R^2 (q^2) for the training set and external R^2 (predictive R^2) for the test set. We have also attempted to apply the CoMFA²⁵ 3D QSAR method to the same data set but without much success. The k NN QSAR models afforded higher q^2 and predictive R^2 as compared to CoMFA. The results of our studies demonstrate the effectiveness of the k NN QSAR approach and provide rationale for further design and synthesis of novel potent epipodophyllotoxin derivatives.

Biological Activity Data

All compounds in this study were tested for their ability to form intracellular covalent topoisomerase II–DNA complexes. The assay procedures have been described previously.⁷ The activity data are originally expressed as the percentage of cellular protein–DNA complex formed (PCPDCF) and were transformed by taking the natural logarithm of PCPDCF, i.e., $\ln(\text{PCPDCF})$. These transformed activities were used in the subsequent CoMFA and variable selection studies.

Initially, multiple training and test sets were generated to investigate the relationship between q^2 and n_{var} . Then, a molecular diversity sampling tool, SAGE (stimulated annealing guided evaluation, see Computational Materials and Methods),⁴⁹ was applied to split the 157 epipodophyllotoxin derivatives into training (Table 1) and test (Table 2) sets on the basis of their chemical diversity in multidimensional descriptor space.

Computational Materials and Methods

CoMFA Method. Structures were generated, and CoMFA was performed with Sybyl molecular modeling software.⁵⁰ The default Sybyl settings were used except as otherwise noted. All calculations were performed on a Silicon Graphics Octane workstation.

Structure Alignment. The aligned structures of etoposide, **1–72**, and **120–146** were directly adapted from the previous study.²⁴ The structures of **72–119** and **147–156** were generated by modifying the X-ray crystal structure of **124**. Structure optimization, field-fit minimization, charge calculations, and structural alignment of all compounds **72–119** and **147–156** were performed as previously described.²⁴ The lowest-energy conformers were obtained using the genetic algorithm search method as implemented in Sybyl (population of 100 and generations of 5000 with randomized seed turned on). The torsion angle around the rotatable bond linking the C ring and R group (rot. 1 in Figure 1d) was modified manually to fit that of **124**.

Conventional CoMFA. CoMFA was performed with the QSAR option of Sybyl as described previously.²⁴ The steric and electrostatic field energies were calculated using sp³ carbon probe atoms with a +1 charge.

q^2 -GRS CoMFA. We executed q^2 -GRS CoMFA following the routine described earlier.²⁶ In contrast to standard CoMFA, this modified method leads to reproducible q^2 values that are independent of the orientation of aligned molecules on the user terminal. Additional details of this approach were discussed in the original publication.²⁶

Table 1. Structure and Activity of Training Set Molecules

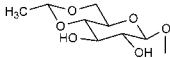
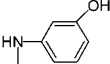


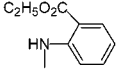
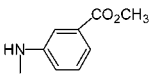
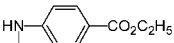
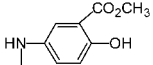
Compound	R	Structure	Cellular Protein-DNA	Ref.
		Type	Complex Formation (%)	
Etoposide		1	100	7
1	-OH	1	42.2	7
2	-NHCH ₂ CH ₂ OCH ₃	1	110.8	7
3	-NHCH ₂ CH=CH ₂	1	84.1	7
4	-NHCH ₂ CH(OH)CH ₃ (R)	1	167.2	7
5	-NHCH(CH ₃)CH ₂ OH (R)	1	161.7	7
6		1	290	8
7		1	243	8
8		1	211	8
9		1	4	8
10		1	249	8
11		1	207	8
12		1	83	8

Table 1. (Continued)

Compound	R	Cellular Protein-DNA		
		Structure Type	Complex Formation (%)	Ref.
13		1	129	8
14		1	50	8
15		1	104	8
17		1	235	8
18		1	180	8
19		1	47	8
20		1	164	8
21		1	279	8
22		1	97	8
23		1	140	8
24		1	97	8
25		1	123	8
26		1	140	8
27		1	330	8
28		1	11	8
29		1	57	8
30		1	34	8
31		1	10	8

Table 1. (Continued)

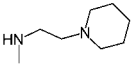
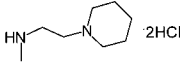
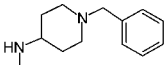
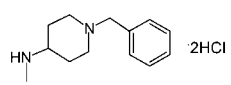
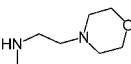
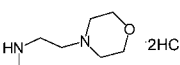
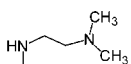
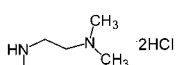
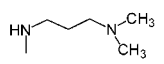
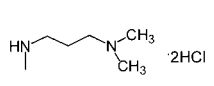
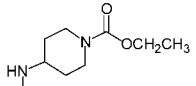
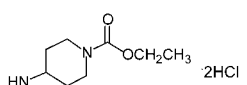
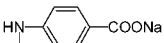
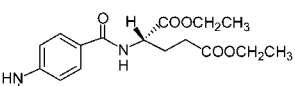



Compound	R	Structure	Cellular Protein-DNA	
		Type	Complex Formation (%)	Ref.
32		1	190	16
33		1	183	16
34		1	83	16
35		1	172	16
36		1	77	16
37		1	140	16
38		1	203	16
39		1	183	16
40		1	186	16
41		1	179	16
42		1	17	16
43		1	138	16
44		1	6.9	16
45		1	83	16
46		1	151	9
47		1	211	9
48		1	115	9

Table 1. (Continued)


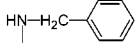
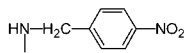
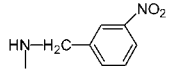
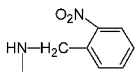
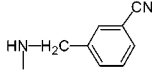
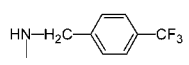
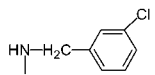
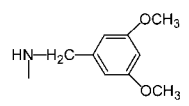
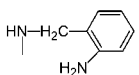
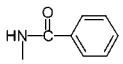
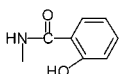
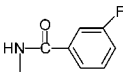
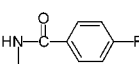
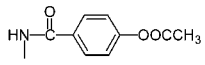
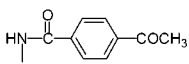
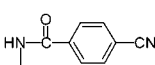
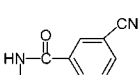
Compound	R	Structure		Ref.
		Type	Cellular Protein-DNA Complex Formation (%)	
49		1	32	9
50		1	181	10
51		1	216	10
52		1	130	10
53		1	144	10
54		1	225	10
55		1	99	10
56		1	159	10
57		1	144	10
58		1	184	10
59		1	177	13
60		1	160	13
61		1	116	13
62		1	117	13
63		1	137	13
64		1	124	13
65		1	159	13
66		1	149	13

Table 1. (Continued)

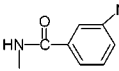
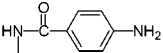
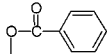
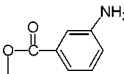
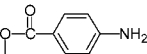
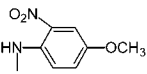
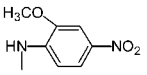
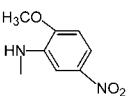
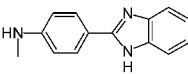
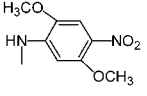
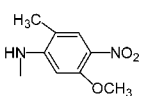
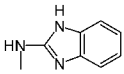
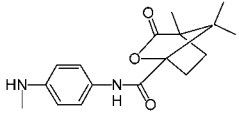
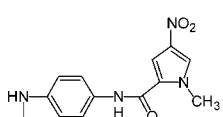
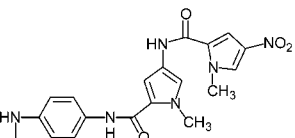
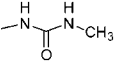
Compound	R	Structure	Cellular Protein-DNA	
		Type	Complex Formation (%)	Ref.
67		1	149	13
68		1	120	13
69		1	94	13
70		1	100	13
71		1	94	13
72		1	15	18
73		1	83	18
74		1	12	18
75		1	128	18
76		1	4.4	18
77		1	3.5	18
78		1	58	18
79		1	88	18
80		1	100	17
81		1	26	17
82		1	81	15

Table 1. (Continued)

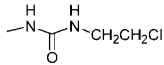
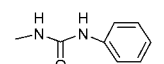
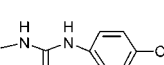
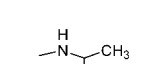
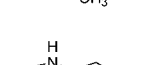
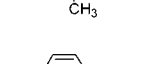
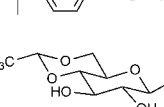
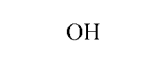
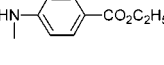
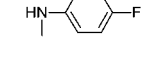
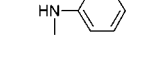
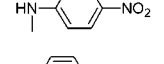
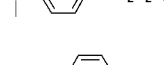
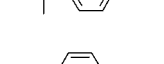
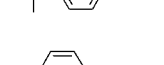

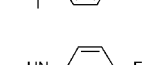
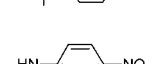
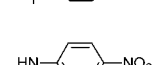
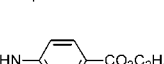
Compound	R	Structure	Cellular Protein-DNA	
		Type	Complex Formation (%)	Ref.
83		1	143	15
84		1	148	15
85		1	125	15
86		1	109	15
87		1	73	15
88		1	207	14
89		2	6.1	11
90	OH	2	15.6	11
91		3	22	12
92		3	11	12
93		4	4	12
94		4	99	12
95		4	138	12
96		4	52	12
97		5	75	12
98		5	127	12
99		5	125	12
100		5	108	12
101		3	23	12
102		6	8	12
103		6	9	12

Table 1. (Continued)

Compound	R	Structure	Cellular Protein-DNA	
		Type	Complex Formation (%)	Ref.
104		6	12	12
105		6	8	12
106		7	117	14
107		7	105	14
108		7	96	14
109		7	69	14
110		7	119	14
111		7	94	14
112		7	175	14
113		7	146	14
114		7	109	14
115		7	75	14
116		7	200	14
117		8	41	15
118		8	7	15
119		9	1	15

kNN QSAR Modeling. Generation of Molecular Descriptors. All chemical structures were generated using Sybyl software.⁵⁰ MCI^{46,47} and MOE⁴⁸ descriptors were generated for the variable selection QSAR study. Molecular topological indices were obtained with the MolConnZ program (MZ descriptors),⁵² and MOE descriptors were generated with the QuaSAR-descriptor option implemented in MOE molecular modeling software.⁴⁸

MolConnZ Descriptors. Over 400 different topological indices were generated using MolConnZ 3.50.⁵² Many of these descriptors characterize chemical structure, but some depend on the arbitrary numbering of atoms in a molecule and are introduced solely for bookkeeping purposes. After descriptors containing no structural information (bookkeeping descriptors, as well as descriptors with zero value or zero variance) were eliminated, only 150 chemically relevant descriptors were

Table 2. Structure and Activity of Test Set Molecules

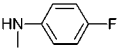
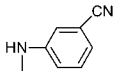
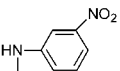
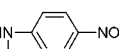
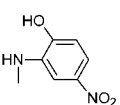
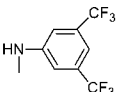
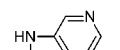
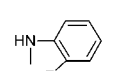
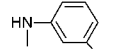
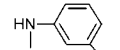
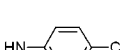
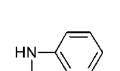

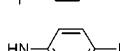
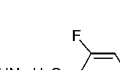
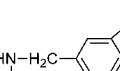
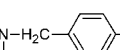
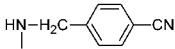
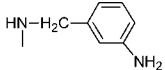
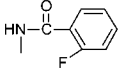
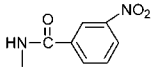
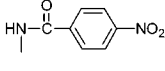
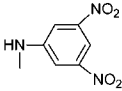
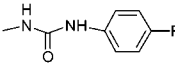

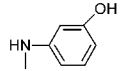
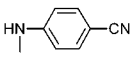
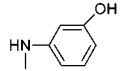
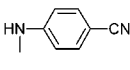

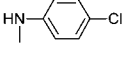
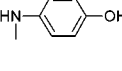
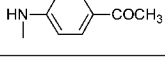
Compound	R	Structure	Cellular Protein-DNA	
		Type	Complex Formation (%)	Ref.
120	-NH ₂	1	36.4	7
121	-NHCH ₂ CH ₂ OH	1	121.4	7
122	-NHCH ₂ CH ₂ CH ₃	1	69.7	7
123	-NHCH ₂ CH ₂ CH ₂ OH	1	89.2	7
124		1	213	8
125		1	137	8
126		1	230	8
127		1	323	8
128		1	15	8
129		1	21	8
130		1	148	8
131		1	121	9
132		1	158	9
133		1	51	9
134		1	99	9
135		1	62	9
136		1	179	9
137		1	64	9
138		1	126	10
139		1	216	10
140		1	169	10

Table 2. (Continued)

Compound	R	Structure	Cellular Protein-DNA	
		Type	Complex Formation (%)	Ref.
141		1	284	10
142		1	191	10
143		1	128	13
144		1	86	13
145		1	160	13
146		1	20	18
147		1	118	17
148		3	9	12
149		3	4	12
150		4	62	12
151		4	18	12
152		3	33	12
153		7	128	14
154		7	77	14
155		7	83	14
156		7	147	14

eventually used in this study. Because the absolute scales for MCI descriptors can differ by orders of magnitude, all descriptors were range-scaled prior to distance calculations to avoid disproportional weightings in multidimensional MCI descriptor space.

MOE Descriptors. MOE is a unique and flexible software system designed specifically for molecular computing.⁴⁸ In addition to 2D topological indices (e.g., Kier and Hall connectivity and κ shape indices), the MOE system also generates descriptors representing physical properties (e.g., molecular refractivity, logP), pharmacophore feature descriptors (e.g.,

counts of hydrogen bond acceptor/donor atoms), and 3D molecular descriptors that are insensitive to 3D alignment (e.g., potential energy descriptors). In this study, 191 MOE descriptors were calculated for the original data set using the QuaSAR-descriptor option implemented in the MOE system.

kNN QSAR Algorithm. The kNN QSAR method,⁴⁵ uses the kNN classification principle⁵³ and a variable selection procedure. Briefly, pairwise similarity is characterized by Euclidean distance between any two compounds in the multidimensional descriptor space. A subset of n_{var} descriptors is selected randomly (n_{var} is a number between 1 and the total

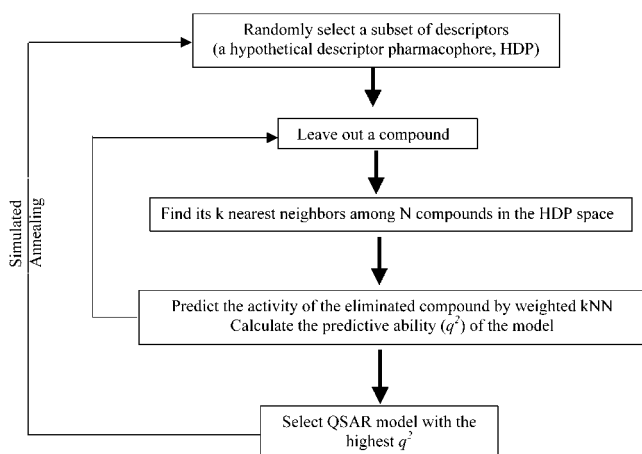


Figure 2. Flowchart of the modified *k*NN method.

number of the original descriptors) as a hypothetical descriptor pharmacophore (HDP). The n_{var} is set to different values in several different runs, and the HDP is validated by a standard leave one out cross-validation procedure, where one compound is eliminated from the training set and its biological activity is predicted as the average activity of k most similar molecules in the training set ($k = 1-5$). The optimization process is driven by a generalized SA technique using q^2 as the objective function (eq 1) to find the best descriptor pharmacophore that maximizes the q^2 value of the *k*NN model.

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (1)$$

where y_i and \hat{y}_i are the actual and predicted activities of the i th compound, respectively, and \bar{y} is the average activity of all compounds in the training set.

Additional details of the *k*NN method implementation, including the description of the SA procedure used for stochastic sampling of the descriptor space, are given elsewhere.⁴⁵

We enhanced recently the original *k*NN method by using weighted molecular similarity⁵⁴ as follows. In the original method,⁴⁵ the activity of each compound was predicted as the algebraic average activity of its *k*NN compounds in the training set. However, in general, the topological distances between a compound and each of its *k*NNs are not the same in the descriptor space. Thus, the weight principle was applied in this QSAR method to give a higher weight to the neighbor with a smaller distance from the unknown compound and the activity of the unknown was calculated as follows

$$W_i = \frac{e^{-d_i} \sum_{i=1}^K d_i}{\sum_{j=1}^K e^{-d_j} \sum_{i=1}^K d_i} \quad (2)$$

$$Y_{\text{unknown}} = \sum_{i=1}^K W_i Y_i \quad (3)$$

where K is the number of *k*NNs in the *k*NN QSAR model; d_i is the Euclidean distance between the compound and its i th nearest neighbor; W_i is the weight for the i th nearest neighbor; Y_i is the actual activity value for the i th nearest neighbor; and Y_{unknown} is activity value of the unknown compound predicted from the activities of its *k*NNs.

The *k*NN QSAR algorithm generates an optimal k value and an optimal n_{var} subset of descriptors, which provide a QSAR model with the highest value of q^2 . Figure 2 shows the overall flowchart of the *k*NN method.

Algorithm for Selection of Training and Test Sets. As we have demonstrated earlier,⁵⁵ a high q^2 is necessary, but not sufficient criteria of a good model. To obtain a robust model, both training set q^2 and predictive (test set) R^2 should be applied to validate the model. Two independent algorithms were used in this study to split the original data set into training and test sets. Models were developed for different training sets and validated using $q^2 > 0.4$ and predictive $R^2 > 0.6$ as a fitness measure to generate the best possible model.

An algorithm similar to the stochastic cluster analysis (SCA) method developed earlier by Reynolds and co-workers⁵⁶ was used to select training and test sets based on the diversity sampling in the MCI descriptor space. The volume of the original descriptor space is normalized to 1, and then, the volume corresponding to an individual point representing a compound is defined as $1/N$, where N is the number of representative points in the descriptor space. The most active compound is selected and included into the training set. A sphere is constructed with its center at the representative point of this compound and with radius $R = c(V/N)^{1/K}$. Here, K denotes the number of descriptors, and c refers to the dissimilarity level. Compounds corresponding to representative points within this sphere, other than the center, are included in the test set, and all of the points within this sphere are then excluded from the initial set. The point with the smallest distance from the first representative point is selected as the second center, and the above procedure is repeated until all of the compounds are assigned to either training or test set. This algorithm allows construction of training sets that cover the whole descriptor space occupied by representative points. The dissimilarity level can be varied to construct models with training and test sets of different sizes. A higher dissimilarity level c gives a smaller training set and a larger test set. As anticipated, the distribution of compounds between test and training sets is sensitive to the types of descriptors used in the calculation.

In addition to the diversity sampling method described above, the training and test sets were also selected using the SAGE method developed recently in our group.⁴⁹ The major objective of SAGE is to obtain a subset of M points (molecules) that are optimally diverse and representative in the descriptor space. Because an exhaustive evaluation of all combinations of M points from an available pool of N objects is computationally prohibitive, the SA algorithm^{58,59} was adapted as an efficient stochastic optimization technique. Briefly, compounds in a data set are represented as points in the multidimensional descriptor space. The diversity of a subset of points is measured by a specially designed diversity function,⁴⁹ and a desired number of optimally diverse points (compounds) are selected using SA algorithm as the optimization tool.⁴⁹ In this study, the original data set was subdivided into three categories based on the biological diversity (the inactive with $\ln(\text{PCPDCF})$ less than 3.0, the moderate active with $\ln(\text{PCPDCF})$ between 3.0 and 4.5, and the highly active with $\ln(\text{PCPDCF})$ more than 4.5). An individual diversity sampling was performed in each category to generate the final training and test sets.

Robustness of QSAR Models. The robustness of these QSAR models was established by comparing the q^2 values for the models derived from experimental training sets with those from so-called random data sets, which are generated by random shuffling of compound activities. The statistical significance of QSAR models for training sets was evaluated with the standard hypothesis testing method.⁵⁷ In this approach, two alternative hypotheses are formulated as follows: (i) $H_0: h = \mu$; (ii) $H_1: h > \mu$, where μ is the average value of q^2 for random data sets and h is the q^2 value for the actual data set. The decision-making is based on a standard one tail test, in which a Z score is calculated and compared with the tabular critical values of Z_c at different levels of significance (α)⁵⁷ to determine the level at which H_0 should be rejected.

Table 3. Results of *k*NN Modeling Using MCI Descriptors for Different Training and Test Sets

test set size	model no.	n_{var}	k value	q^2	R_{pred}^2	$q_{\text{ran}}^{2\text{a}}$
19	1	10	3	0.56	0.68	0.21
	2	20	2	0.63	0.62	0.24
	3	30	2	0.61	0.64	0.12
	4	40	2	0.58	0.64	0.17
	5	50	2	0.58	0.70	0.09
28	6	10	2	0.42	0.53	0.28
	7	20	2	0.61	0.58	0.23
	8	30	2	0.58	0.64	0.19
	9	40	4	0.50	0.70	0.13
	10	50	2	0.59	0.65	0.12
37	11	10	2	0.62	0.38	0.25
	12	20	2	0.48	0.32	0.19
	13	30	3	0.61	0.60	0.20
	14	40	3	0.54	0.61	0.16
	15	50	2	0.53	0.41	0.12
41	16	10	4	0.46	0.37	0.29
	17	20	4	0.54	0.38	0.17
	18	30	2	0.65	0.45	0.30
	19	40	2	0.61	0.50	0.14
	20	50	2	0.53	0.41	0.08
46	21	10	2	0.48	0.19	0.21
	22	20	2	0.60	0.57	0.24
	23	30	4	0.52	0.49	0.17
	24	40	1	0.56	0.49	0.18
	25	50	2	0.63	0.51	0.10
60	26	10	2	0.54	0.40	0.27
	27	20	2	0.53	0.26	0.19
	28	30	2	0.63	0.26	0.11
	29	40	2	0.53	0.35	0.11
	30	50	2	0.63	0.38	0.13

^a q_{ran}^2 is the best value of 10 independent calculations.

Table 4. Results of the One Tail Hypothesis Testing

model no.	13	14	33	34
Z score	13.7	11.4	13.4	11.9
Z_c ($\alpha = 0.001$)	4.31	4.30	4.28	4.25

Results and Discussion

Generation of Training and Test Sets. Because the results of the *k*NN method are sensitive to the selected number of variables, n_{var} , and optimal n_{var} are not known a priori, multiple models should be generated to examine the relationship between q^2 and n_{var} . In this study, the original data set was divided into several selections of training and test sets, and different models were constructed using the *k*NN QSAR method and MCI descriptors. The robustness of these models was validated by comparing q^2 values derived from the actual data set and those from data sets with randomized activity values. The 30 best models are described in Table 3, which also reports the best q^2 values for models generated for activity-shuffled data sets. Higher q^2 values were obtained consistently for the actual data set as compared to data sets with randomized activities. These results were examined with the one tail hypothesis test, and models for the actual data set were found to be significantly better than those for data sets with randomized activities in terms of Z scores (Table 4). As the n_{var} increased from 10 to 50 with a step of 10, the values of q^2 did not change significantly. However, with an increased number of compounds in the test sets (and decreased size of corresponding training sets), the test set R^2 values also decreased, as could be expected because of a decreasing (thus, insufficient) size of the training set. These calculations allowed us to identify the smallest training set (120 compounds) that would

Table 5. Comparison of QSAR Models Obtained with Training and Test Sets of 120 and 37 Epipodophyllotoxin Derivatives, Respectively

model no.	n_{var}	k value	q^2	R_{pred}^2	q_{ran}^2	
<i>k</i> NN/MCI descriptors	31	10	2	0.63	0.31	0.24
	32	20	1	0.60	0.30	0.21
	33	30	2	0.60	0.62	0.21
	34	40	3	0.56	0.62	0.14
	35	50	2	0.58	0.52	0.20
<i>k</i> NN/MOE descriptors	36	10	2	0.52	0.42	0.16
	37	20	2	0.55	0.47	0.22
	38	30	3	0.45	0.62	0.15
	39	40	3	0.43	0.72	0.16
	40	50	4	0.36	0.60	0.14
CoMFA			0.41	0.36		
q^2 -GRS			0.43	0.39		

still afford reasonable prediction of the corresponding test set. To compare the performances of CoMFA and *k*NN QSAR, a molecular diversity sampling tool, SAGE,⁴⁹ was applied to split the 157 epipodophyllotoxin derivatives into training (120 compounds, Table 1) and test (37 compounds, Table 2) sets on the basis of their chemical diversity in the CoMFA descriptor space.

CoMFA/ q^2 -GRS Modeling. Both conventional and q^2 -GRS CoMFA were as applied to model the 157 epipodophyllotoxin derivatives. As shown in Table 5, the q^2 and predictive R^2 of conventional and q^2 -GRS CoMFA were 0.41 and 0.36, and 0.43 and 0.39, respectively. Despite the higher diversity and larger number of compounds included in this calculation, the results were comparable to those obtained previously.²⁴ However, the low q^2 and R^2 values make it impossible to use these models for reliable predictions.

***k*NN Modeling.** *k*NN analysis selects only the most informative descriptors to build QSAR models for the training set molecules. Both MCI and MOE descriptors were applied in this study, and the number of variables (n_{var}) was set to 10, 20, 30, 40, and 50 for both types of descriptors. At each predefined n_{var} , three models were generated using each type of descriptor, respectively. The best 10 models obtained with the *k*NN analysis are presented in Table 5 in comparison with CoMFA models. To evaluate the robustness of these models, both q^2 and predictive R^2 were calculated. The MOE descriptors provided models with the best q^2 of 0.55, which was slightly better than CoMFA models. The models derived from 2D topological indices (MCI) gave consistently higher q^2 values (0.56–0.63), regardless of the number of descriptors used. Several of these models also afforded acceptable prediction accuracy ($R^2 > 0.6$), especially when more than 20 descriptors were used. The robustness of these models was further validated by the activity-shuffling experiment as described above. The best q^2 values for models derived from the data sets with randomly shuffled activity data are also included in Table 5. The robustness of these models was examined and confirmed by the one tail hypothesis test. Z scores and the corresponding critical values (Z_c) at the significance level (α) of 0.001 for the best two models are shown in Table 4.

A frequency analysis of the topological indices implicated in the 10 best models demonstrated a significant convergence in specific variables as well as variable classes. Nineteen variables were found in most of the 10 best models (with frequency above 5 out of 10), and

nine specific variable classes were consistently employed in all of the 10 models. A closer examination of these selected variables and variable classes suggests that a cluster of specific properties is responsible for successful models. These specific properties include connectivity valence path indices; difference connectivity simple path indices; Shannon information index; electrotopological state index values for hydroxyl oxygens, amino nitrogens, ether oxygens, or aromatic carbons; number of hydrogen bond donors; electrotopological state indices of sp^3 carbon bonded to other carbon atoms; and electrotopological state descriptors of potential internal hydrogen bond strength. Although most of the properties are hard to interpret in the physicochemical sense, some of them (e.g., atom type or bond type electrotopological state indices and hydrogen bond-related descriptors) did imply that chemical features such as hydroxyl, amino, aromatic ring, and hydrogen bond-forming atoms are relevant to biological activities. This implication shall facilitate our understanding of the SARs for epipodophyllotoxin derivatives.

Data shown in Table 5 indicate that the models obtained with 2D descriptors (MCI descriptors) had higher quality with respect to q^2 and predictive R^2 values as compared to models obtained with 3D descriptors (CoMFA) or with the combination of 2D and 3D descriptors (MOE descriptors). Because preliminary SAR studies have already demonstrated that the (2*R*,3*R*,4*S*) stereochemistry is essential to maintain the antitopoisomerase activity of epipodophyllotoxin derivatives,⁶⁰ these three chiral centers were left untouched in all of our synthetic modifications. In addition, the bulky pendant C-1 group is always α -oriented. Thus, for all epipodophyllotoxin derivatives involved in this QSAR study, the skeletal stereochemistry is basically fixed, as would also be true for future design of novel derivatives. In this context, 2D structural information appears sufficient to formulate informative QSAR equations for these epipodophyllotoxins. Applying MCI or MOE descriptors in QSAR formulation, rather than force field descriptors, would eliminate the dependence on the alignment rules and overall orientation, simplify the preparation of data sets (exempted from operations such as minimization, alignment, and optimization), and shorten the computational time. Because of these desirable features, these descriptor types are recommended for future QSAR studies, library design, or database mining related to the epipodophyllotoxin derivatives.

The observed vs predicted activities generated with the best model 33 (cf. Table 5) for the test are shown in Figure 3. A careful analysis of outliers (i.e., molecules with poorly predicted activities) indicates that these molecules can be classified into two groups: (i) compounds with meta substitution on the aromatic ring (e.g., **125**, **128**, **129**, **133**, **135**, **146**, **149**, **150**, and **151**) and (ii) compounds with the highest (**127**, **141**) or lowest (**146**, **148**, **149**, and **151**) activities. The inaccurate activity prediction of meta-substituted molecules may result from a possibility that a target other than topoisomerase II mediates the activities of these molecules.^{51,60} Molecules with extreme activities are typically poorly predicted by the *k*NN QSAR method (which interpolates activities of nearest neighbors of a given compound). Further theoretical developments are un-

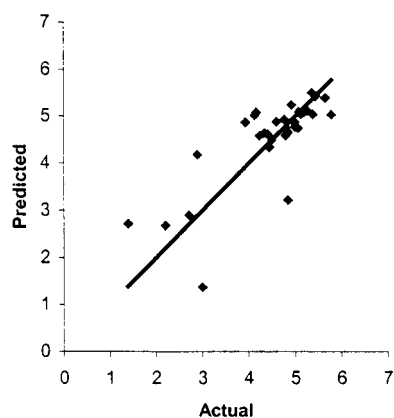


Figure 3. Actual vs predicted $\ln(\text{PCPDF})$ of test set molecules using modified *k*NN approach with MCI descriptors (model 3).

derway in our group to improve the extrapolating power of QSAR methods.

Summary and Future Studies

In this study, we used a recently developed variable selection *k*NN QSAR method⁴⁵ to obtain QSAR models for 157 epipodophyllotoxin derivatives and compared results from this approach to those obtained with the more popular CoMFA method. Using topological (MCI) or a combination of topological and topographical (MOE) descriptors of chemical structures, we have built several robust QSAR models with high values of q^2 (for training sets) and predictive R^2 (for test set); our efforts to build similar quality models with the CoMFA method were unsuccessful. The ambiguity of physicochemical interpretation of topological descriptors makes the *k*NN QSAR models not applicable to direct specific chemical modifications of existing molecules. However, the high predictive ability of the models allows virtual screening of chemical databases or virtual libraries determined by either synthetic feasibility or commercial availability of starting materials to prioritize the synthesis of the most promising candidates. Therefore, these models should facilitate the rational design of novel derivatives, guide the design of focused libraries based on the epipodophyllotoxin skeleton, and facilitate the search for related structures with similar biological activity from large databases.

Acknowledgment. This investigation was supported by grants UNC No. 49849 from the Elsa U. Pardee Foundation and CA 17625 in part from the National Cancer Institute awarded to K.H.L. and an NIH grant MH60328 awarded to A.T. We acknowledge Tripos, Inc. for the software grant and thank Dr. Susan L. Morris-Natschke for her critical reading of the manuscript.

References

- (1) For Antitumor Agents. 212., see Hayashi, K.; Nakanishi, Y.; Bastow, K. F.; Cragg, G.; Nozaki, H.; Lee, K. H. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 345–348.
- (2) Jardine, I. Podophyllotoxins. In *Anticancer Agents Based on Natural Products Models*; Cassidy, J. M., Douros, J., Eds.; Academic Press: New York, 1980; pp 319–351 and references therein.
- (3) Issell, B. F. The Podophyllotoxin Derivatives, VP-16-213 and VM-26. *Cancer Chemother. Pharmacol.* **1982**, *7*, 73–80.

- (4) Van Maanen, J. M. S.; Retel, J.; De Vries, J.; Pinedo, H. M. Mechanism of Action of Antitumor Drug Etoposide: A review. *J. Natl. Cancer Inst.* **1988**, *80*, 1526–1533.
- (5) Hainsworth, J. D.; Williams, S. D.; Einhorn, L. H.; Birch, R.; Greco, F. A. Successful Treatment of Resistant Germinal Neoplasms with VP-16 and Cisplatin: Results of a Southeastern Cancer Study Group Trial. *J. Clin. Oncol.* **1985**, *3*, 666–671.
- (6) Shah, J. C.; Chen, J. R.; Chow, D. Preformulation Study of Etoposide: Identification of Physicochemical Characteristics Responsible for the Low and Erratic Oral Bioavailability of Etoposide. *Pharm. Res.* **1989**, *6*, 408–412.
- (7) Lee, K. H.; Imakura, Y.; Haruna, M.; Beers, S. A.; Thurston, L. S.; Dai, H. J.; Chen, C. H. Antitumor Agents. 107. New Cytotoxic 4-Alkylamino Analogues of 4'-Demethylepipodophyllotoxin as Inhibitors of Human DNA Topoisomerase II. *J. Nat. Prod.* **1989**, *52*, 606–613.
- (8) Wang, Z. Q.; Kuo, Y. H.; Schnur, D.; Bowen, J. P.; Liu, S. Y.; Han, F. S.; Chang, J. Y.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 113. New 4-Arylamino Derivatives of 4'-O-Demethylepipodophyllotoxin and Related Compounds as Potent Inhibitors of Human DNA Topoisomerase II. *J. Med. Chem.* **1990**, *33*, 2660–2666.
- (9) Lee, K. H.; Beers, S. A.; Mori, M.; Wang, Z. Q.; Kuo, Y. H.; Li, L.; Liu, S. Y.; Chang, J. Y.; Han, F. S.; Cheng, Y. C. Antitumor Agents. 111. New 4-Hydroxylated and 4-Halogenated Anilino Derivatives of 4'-Demethylepipodophyllotoxin as Potent Inhibitors of Human DNA Topoisomerase II. *J. Med. Chem.* **1990**, *33*, 1364–1368.
- (10) Zhou, X. M.; Wang, Z. Q.; Chang, J. Y.; Chen, H. X.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 120. New 4-Substituted Benzylamine and Benzyl Ether Derivatives of 4'-O-Demethylepipodophyllotoxin as Potent Inhibitors of Human DNA Topoisomerase II. *J. Med. Chem.* **1991**, *34*, 3346–3350.
- (11) Hu, H.; Wang, Z. Q.; Liu, S. Y.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 123. Synthesis and Human DNA Topoisomerase II Inhibitory Activity of 2'-Chloro Derivatives of Etoposide and 4 β -(Arylamino)-4'-O-demethylpodophyllotoxins. *J. Med. Chem.* **1992**, *35*, 866–871.
- (12) Wang, Z. Q.; Hu, H.; Chen, H. X.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 124. New 4 β -Substituted Anilino Derivatives of 6,7-O-Demethylpodophyllotoxin and Related Compounds as Potent Inhibitors of Human DNA Topoisomerase II. *J. Med. Chem.* **1992**, *35*, 871–877.
- (13) Zhou, X. M.; Wang, Z. Q.; Chen, H. X.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 125. New 4-Benzoylamino Derivatives of 4'-O-Demethyl-4-deoxypodophyllotoxin and 4-Benzoyl Derivatives of 4'-O-Demethylpodophyllotoxin as Potent Inhibitors of Human DNA Topoisomerase II. *Pharm. Res.* **1993**, *10*, 214–219.
- (14) Wang, Z. Q.; Shen, Y. C.; Chen, H. X.; Chang, J. Y.; Guo, X.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 126. Novel 4 β -Substituted Anilino Derivatives of 3',4'-O-Didemethylpodophyllotoxin as Potent Inhibitors of Human DNA Topoisomerase II. *Pharm. Res.* **1993**, *10*, 343–350.
- (15) Miyahara, M.; Kashiwada, Y.; Guo, X.; Chen, H. X.; Cheng, Y. C.; Lee, K. H. Nitrosourea Derivatives of 3',4'-Didemethoxy-3',4'-Dioxo-4-deoxypodophyllotoxin and Urea Derivatives of 4'-O-Demethylpodophyllotoxin as Potent Inhibitors of Human DNA Topoisomerase II. *Heterocycles* **1994**, *39*, 361–369.
- (16) Zhang, Y. L.; Guo, X.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 148. Synthesis and Biological Evaluation of Novel 4-Amino Derivatives of Etoposide with Better Pharmacological Profiles. *J. Med. Chem.* **1994**, *37*, 446–452.
- (17) Ji, Z.; Wang, H. K.; Bastow, K. F.; Zhu, X. K.; Cho, S. J.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 177. Design, Syntheses, and Biological Evaluation of Novel Etoposide Analogues Bearing Pyrrolicarboxamidino Group as DNA Topoisomerase II Inhibitors. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 607–612.
- (18) Zhu, X. K.; Guan, J.; Tachibana, Y.; Bastow, K. F.; Cho, S. J.; Cheng, H. H.; Cheng, Y. C.; Gurwith, M.; Lee, K. H. Antitumor Agents. 194. Synthesis and Biological Evaluations of 4 β -Mono-, -Di-, and -Trisubstituted Anilino-4'-demethyl-Podophyllotoxin and Related Compounds with Improved Pharmacological Profiles. *J. Med. Chem.* **1999**, *42*, 2441–2446.
- (19) Osheroff, N.; Zechiedrich, E. L.; Gale, K. C. Catalytic Function of DNA Topoisomerase II. *BioEssays* **1991**, *13*, 269–275.
- (20) Alton, P. A.; Harris, A. L. Annotation. *Br. J. Haematol.* **1993**, *85*, 241–245.
- (21) Liu, S. Y.; Hwang, B. D.; Haruna, M.; Imakura, Y.; Lee, K. H.; Cheng, Y. C. Podophyllotoxin Analogues: Effects on DNA Topoisomerase II, Tubulin Polymerization, Human Tumor KB Cells, and Their VP-16-Resistant Variants. *Mol. Pharmacol.* **1989**, *36*, 78–82.
- (22) Zhang, Y. L.; Tropsha, A.; McPhail, A. T.; Lee, K. H. Antitumor Agents. 152. In Vitro Inhibitory Activity of Etoposide Derivative NPF Against Human Tumor Cell Lines and a Study of Its Conformation by X-ray Crystallography, Molecular Modeling, and NMR Spectroscopy. *J. Med. Chem.* **1994**, *37*, 1460–1464.
- (23) VanVliet, D. S. Design and Synthesis of Podophyllotoxin and Etoposide Analogues via Traditional and Combinatorial Methodology. Ph.D. Dissertation, The University of North Carolina at Chapel Hill, 1999.
- (24) Cho, S. J.; Tropsha, A.; Suffness, M.; Cheng, Y. C.; Lee, K. H. 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.
- (25) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis CoMFA. 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (26) Cho, J. S.; Tropsha, A. Cross-Validated R²-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method To Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060–1066.
- (27) MacDonald, T. L.; Lehnert, E. K.; Loper, J. T.; Chow, K. C.; Ross, W. E. On the Mechanism of Interaction of DNA Topoisomerase II with Chemotherapeutic Agents. In *DNA Topoisomerase in Cancer*, Potmesil, M., Kohn, K. W., Eds.; Oxford University Press: New York, 1991; pp 119–214.
- (28) Baguley, B. C. DNA Intercalating Antitumor Agents. *Anti-Cancer Drug Des.* **1991**, *6*, 1–35.
- (29) Cho, S. J.; Kashiwada, Y.; Bastow, K. F.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 164. Podophenazine, 2'',3''-Dichloro-podophenazine, Benzopodophenazine, and Their 4 β -p-Nitroaniline Derivatives as Novel DNA Topoisomerase II Inhibitors. *J. Med. Chem.* **1996**, *39*, 1396–1402.
- (30) Cramer, R. D., III; DePriest, S. A.; Patterson, D. E.; 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; pp 443–485.
- (31) Kim, K. H.; Brusniak, M. Y. K.; Pearlman, R. S. UniSur-CoMFA: for stable and consistent 3D-QSAR. *Alfred Benzon Symp.* **1998**, *42* (Rational Molecular Design in Drug Research), 67–86.
- (32) Waller, C. L.; Oprea, T. I.; Giolitti, A.; Marshall, G. R. Three-Dimensional QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. 1. A CoMFA Study Employing Experimentally-Determined Alignment Rules. *J. Med. Chem.* **1993**, *36*, 4152–4160.
- (33) Cho, S. J.; Garsia, M. L. S.; Bier, J.; Tropsha, A. Structure Based Alignment and Comparative Molecular Field Analysis of Acetylcholinesterase Inhibitors. *J. Med. Chem.*, in press.
- (34) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure–Activity Relationships and Quantitative Structure–Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
- (35) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vols. I, II.
- (36) Hansen, P. J.; Jurs, P. C. Chemical Applications of Graph Theory II: Isomer Enumeration. *J. Chem. Educ.* **1988**, *65*, 574.
- (37) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct. Act. Relat.* **1994**, *13*, 285–294.
- (38) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure–Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77–84.
- (39) Fogel, D. B. Applying Evolutionary Programming to Selected Traveling Salesman Problems. *Cybern. Syst. (USA)* **1993**, *24*, 27–36.
- (40) Fogel, D. B.; Fogel, L. J.; Porto, V. W. Evolutionary Methods for Training Neural Networks. *IEEE Conf. Neural Networks Ocean Eng.* (Catal. No. 91CH3064-3) **1991**, 317–327.
- (41) Goldberg, D. E. *Genetic Algorithm in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
- (42) Forrest, S. Genetic Algorithms: Principles of Natural Selection Applied to Computation. *Science* **1993**, *261*, 872–878.
- (43) Bohachevsky, I. O.; Johnson, M. E.; Stein, M. L. Generalized Simulated Annealing for Function Optimization. *Technometrics* **1986**, *28*, 209–217.
- (44) Kalivas, J. H. Generalized Simulated Annealing for Calibration Sample Selection from an Existing Set and Orthogonalization of Undesigned Experiments. *J. Chemom.* **1991**, *5*, 37–48.
- (45) Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure–Property Relationship Approach Based on the *k*-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.
- (46) Kier, L. B.; Hall, L. H. *Molecular Structure Description—The Electrotopological State*; Academic Press: San Diego, 1999.
- (47) Hall, L. H.; Kier, L. B. *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure–Property Modeling*. In *Reviews in Computational Chemistry II*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; p 367.
- (48) Information on MOE is available at <http://www.chemcomp.com/feature/descr.htm>

- (49) Zheng, W.; Cho, S. J.; Waller, C. L.; Tropsha, A. Rational Combinatorial Library Design. 3. Simulated Annealing Guided Evaluation (SAGE) of Molecular Diversity: A Novel Computational Tool for Universal Library Design and Database Mining. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 738–746.
- (50) The Sybyl program (version 6.0) is available from Tripos Associates, 1699 South Hanley Road, St. Louis, MO 63144.
- (51) Tachibana, Y.; Zhu, X. K.; Krishnan, P.; Lee, K. H.; Bastow, K. F. Characterization of human lung cancer cells resistant to 4'-O-demethyl-4 β -(2''-nitro-4''-fluoroanilino)-4-desoxypodophyllotoxin, a unique compound in the epipodophyllotoxin antitumor class. *Anti-Cancer Drugs* **2000**, *11*, 19–28.
- (52) Molconn-Z, version 3.5; Hall Associates Consulting: Quincy, MA; information on Molconn Z is available at <http://www.eslc.vabio-tech.com/molconn/>.
- (53) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; John Wiley & Sons: New York, 1986.
- (54) Xiao, Y. D.; Shen, M.; Tropsha, A. A Fast Feature Selection Approach to QSAR/QSPR Studies of Large Data Sets Based on Weighted *k*-Nearest-Neighbor Principle with Restricted Similarity Cutoff (RW-kNN QSAR). Manuscript in preparation.
- (55) Golbraikh, A.; Tropsha, A. Beware of q^2 ! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
- (56) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305–312.
- (57) Gilbert, N. *Statistics*; W. B. Saunders, Co.: Philadelphia, PA, 1976.
- (58) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
- (59) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680.
- (60) Zhang, Y.; Lee, K. H. Recent Progress in the Development of Novel antitumor Etoposide Analogues. *Chin. Pharm. J.* **1994**, *46*, 319–369.

JM0105427