DOI 10.5012/bkcs.2011.32.7.2397

Development of *k*NN QSAR Models for 3-Arylisoquinoline Antitumor Agents

Alexander Tropsha,[†] Alexander Golbraikh,[†] and Won-Jea Cho[*]

College of Pharmacy and Research Institute of Drug Development, Chonnam National University, Gwangju 500-757, Korea
[*]E-mail: wjcho@chonnam.ac.kr
[†]Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599-7360, USA
Received April 4, 2011, Accepted June 1, 2011

Variable selection *k* nearest neighbor QSAR modeling approach was applied to a data set of 80 3-arylisoquinolines exhibiting cytotoxicity against human lung tumor cell line (A-549). All compounds were characterized with molecular topology descriptors calculated with the MolconnZ program. Seven compounds were randomly selected from the original dataset and used as an external validation set. The remaining subset of 73 compounds was divided into multiple training (56 to 61 compounds) and test (17 to 12 compounds) sets using a chemical diversity sampling method developed in this group. Highly predictive models characterized by the leave-one out cross-validated R^2 (q^2) values greater than 0.8 for the training sets and R^2 values greater than 0.7 for the test sets have been obtained. The robustness of models was confirmed by the Y-randomization test: all models built using training sets with randomly shuffled activities were characterized by low $q^2 \leq 0.26$ and $R^2 \leq 0.22$ for training and test sets, respectively. Twelve best models (with the highest values of both q^2 and R^2) predicted the activities of the external validation set of seven compounds with R^2 ranging from 0.71 to 0.93.

Key Words : Antitumor agents, 3-Arylisoquinolines, *k* nearest neighbor QSAR

Introduction

A large number of anticancer chemotherapeutic agents have been developed over the last decades. Still, there is no doubt that further research into the design of novel anticancer compounds with low toxicity and higher selectivity is needed.[1] Isoquinoline analogues of natural or synthetic antitumor agents have attracted considerable interest as potent anticancer agents. It has been established that they act *via* the inhibition of topoisomerases I (topo I) or II (topo II), or as DNA intercalators.[2]

The synthesis, biological evaluation, and 3D-QSAR studies of 3-arylisoquinolines were carried out in one of our laboratories previoulsy.[3] These compounds have been shown to be highly cytotoxic against several types of human tumor cells. Interestingly, some (but not all) of these compounds also showed topo I inhibitory activity.[4] Indenoisoquinoline derivatives were thoroughly investigated by the Cushman group and were also reported to have topo I inhibitory activity.[5] Although these studies have identified important structural patterns within isoquinolines responsible for their cytotoxicity, it proved to be difficult to rationalize the relationships between their topo I inhibitory activity and cytotoxicity. In the previous paper, a hypothetical pharmacophore model of 3-arylisoquinolines active against human lung tumor cell (A-549) was proposed based on results of the Comparative Molecular Field Analysis (CoMFA).[6] Subsequent research focused on introducing the various amines in the C-1 position replacing *N*-methylpiperazine in 3-arylisoquinolines and conducting additional studies towards the antitumor cytotoxicity and topo I inhibition activity of isoquinolinamines. As a result, water-soluble 3-arylisoquinolinamines were identified as potent cytotoxic agents.[7]

The important insights into rational design of novel potent comounds could be obtained by studying the relationships between structure and cytotoxicity of 3-arylisoquinoline compounds, which could be achieved by using reliable and robust QSAR approaches. Over the last two decades many QSAR approaches have been developed. Based on the type of descriptors used, they can be divided into those utilizing physicochemical properties of molecules, two-dimensional (2D) and three-dimensional (3D) QSAR approaches.

One of the most popular 3D-QSAR methods, CoMFA developed in mid-eighties[8] and other CoMFA-like methods (some of them are discussed in references[9] require spatial alignment of molecules. CoMFA and many other 3D-QSAR methods have several shortcomings.[10] In many cases, it is impossible to precisely define a pharmacophore model (i.e., specific and characteristic 3D arrangements of chemical functional groups responsible for biological activity.[11] If all molecules in a dataset are flexible, their unique structural alignment is impossible. If a spatial structure of a binding site is known, docking procedures such as DOCK,[12] AutoDock,[13] Gold,[14] FlexiDock,[15] FlexX[16] and FlexE,[17] *etc.* can provide the alignment of molecules, which could be used in 3D-QSAR. However, docking algorithms are not accurate enough to rank the binding energies of molecules.[18] This can lead to non-optimal alignment of ligands, and eventually can introduce errors in QSAR analysis. On the other hand, QSAR analysis based on descriptors calculated using mole-

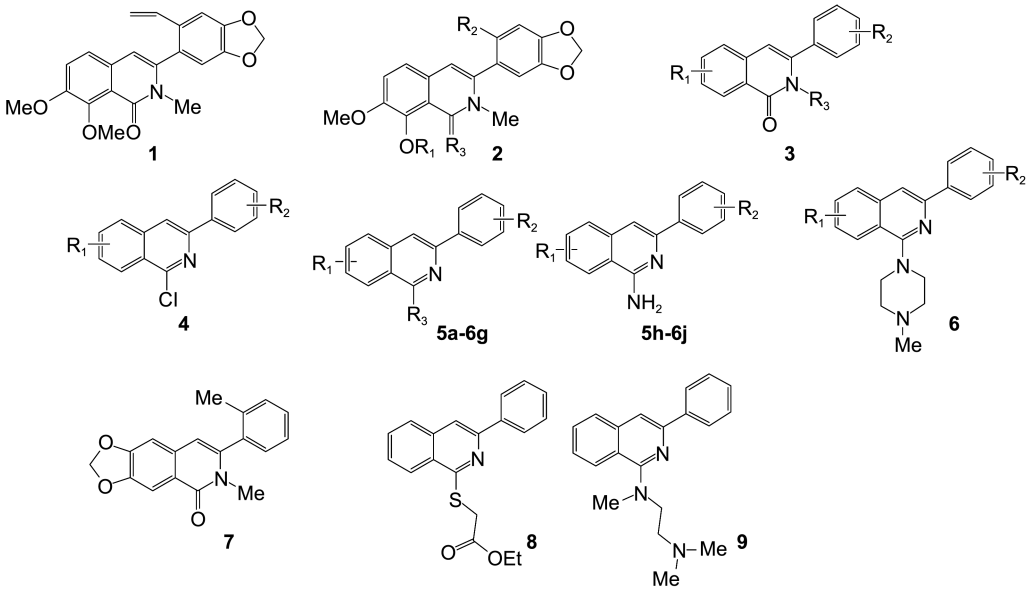
cular graphs, i.e. structural formulas of compounds (2D-QSAR) provides an appealing alternative to 3D-QSAR since the former methods neither require conformational search nor structural alignment, are faster and easier to implement in an automated fashion, and are typically characterized by models with the same or better statistical significance and predictive power. These features also make 2D-QSAR approaches easily adaptable to the task of database searching, or virtual screening.[19]

We have developed several stochastic variable selection QSAR procedures including genetic algorithm or simulated annealing partial least squares (GA/SA-PLS)[20] and *k*-nearest-neighbor (*k*NN) QSAR.[10,21] As opposed to the conventional 3D-pharmacophore, we have defined a descriptor pharmacophore as a subset of molecular descriptors that give rise to the most statistically significant QSAR models.[19] Typically, we use these methods with multiple descriptors derived from 2D molecular topology, which eliminates the conformational and alignment ambiguities inherent in most 3D-QSAR approaches. Stochastic optimization algorithms such as GA or SA are used to develop a robust QSAR model that is characterized by the highest value of cross validated R^2 (q^2). By default, the descriptor pharmacophore corresponds to a variable selection model with a local maximum of q^2. These methods are computationally efficient and automated and they have been used to generate predictive models that are comparable to, or better than, those obtained with CoMFA.[22]

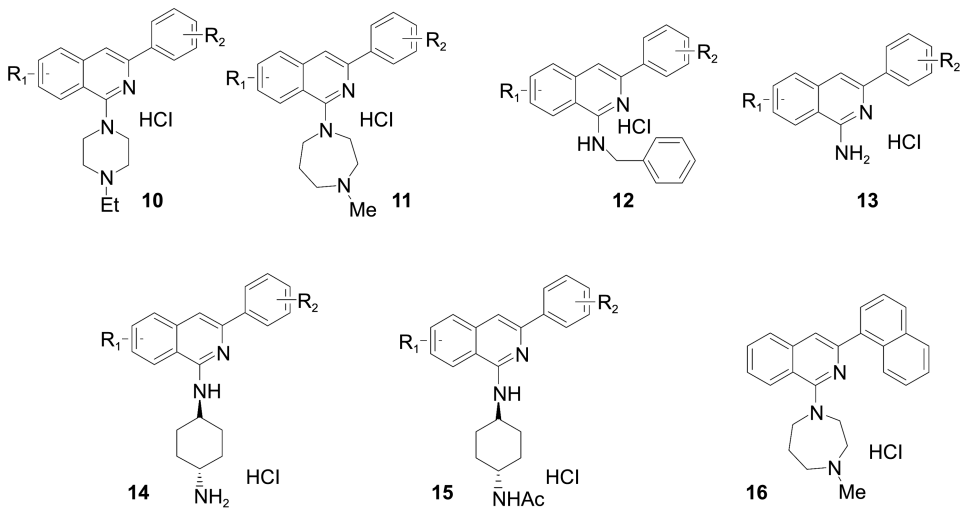
In this paper, we have applied the *k*NN QSAR approach and rigorous model validation procedures to a dataset of 80 antitumor 3-arylisoquinolines that were synthesized and evaluated previously.[6] Our results demonstrate the effectiveness of the *k*NN QSAR approach and provide rationale for further design and synthesis of novel potent antitumor agents.

Methods

Cytotoxicity Measurements. The structures of 80 compounds used for building *k*NN models are shown in Schemes 1[6]



Scheme 1. Structure of the training set compounds (**1**).



Scheme 2. Structure of the training set compounds (**2**).

Table 1. Structure and *in vitro* cytotoxicity (pIC_{50} values for inhibiting A549 tumor cell line growth) for 57 compounds included in the training set

No	Compound	R1	R2	R3	PIC_{50}
1	**1**	-	-	-	8.05
2	**2a**	Me	Et	H_2	5.89
3	**2b**	Me	Vinyl	H_2	5.72
4	**2c**	Me	$CH_2CH(OMe)_2$	O	6.49
5	**2d**	Me	Et	O	7.57
6	**2e**	H	Et	O	4.93
7	**2a**	5-NMe_2	4-OMe	H	4.90
8	**3b**	6-Me	4-Cl	H	5.24
9	**3c**	H	2-PMB	H	4.12
10	**3d**	H	3-Me	H	4.39
11	**3e**	H	4-Br	H	4.21
12	**3f**	H	4-Me	H	4.52
13	**3g**	H	4-OMe	H	3.81
14	**3h**	H	2-CHO	Me	3.87
15	**3i**	H	2-CH_2OH	Me	3.84
16	**3j**	H	2-CH_2OPMB	Me	4.67
17	**3k**	H	2-$CH_2CH(OMe)_2$	Me	3.69
18	**3l**	H	2-vinyl	Me	4.22
19	**3m**	H	H	Me	4.07
20	**4a**	5-NMe_2	4-Cl	-	3.84
21	**4b**	6-Me	2-Cl	-	3.88
22	**4c**	6-Me	2-Me	-	3.58
23	**4d**	6-Me	4-Me	-	3.95
24	**4e**	6-Me	H	-	3.80
25	**4f**	H	3-Me	-	3.77
26	**4g**	H	4-Br	-	3.95
27	**4h**	H	H	-	3.82
28	**5a**	6-Me	2-Me	Bn	4.56
29	**5b**	H	2-Me	Bn	3.85
30	**5c**	H	H	Bn	3.76
31	**5d**	6-Me	3-Me	Bn	6.13
32	**5e**	H	H	morpholine	3.91
33	**5f**	6-Me	2-Me	NH_2	6.76
34	**5g**	H	2-Me	NH_2	5.50
35	**5h**	H	H	NH_2	5.56
36	**5i**	H	H	NH-PMB	3.61
37	**5j**	H	H	piperidine	3.83
38	**6a**	5-NMe_2	2-Me	-	4.67
39	**6b**	5-NMe_2	3-Me	-	5.03
40	**6c**	5-NMe_2	4-Br	-	5.04
41	**6d**	5-NMe_2	4-Cl	-	5.13
42	**6e**	5-NMe_2	4-OMe	-	4.72
43	**6f**	5-NMe_2	H	-	4.83
44	**6g**	6-Me	2-Me	-	5.03
45	**6h**	6-Me	4-Cl	-	5.41
46	**6i**	6-Me	4-Me	-	4.95
47	**6j**	6-Me	H	-	5.13
48	**6k**	H	2-Me	-	4.23
49	**6l**	H	3-Me	-	4.67
50	**6m**	H	4-Br	-	4.73
51	**6n**	H	4-Cl	-	4.68
52	**6o**	H	4-Me	-	4.90
53	**6p**	H	4-OMe	-	4.75
54	**6q**	H	H	-	4.47
55	**7**	-	-	-	4.63
56	**8**	-	-	-	3.69
57	**9**	-	-	-	4.60

Table 2. Structure and *in vitro* cytotoxicity (pIC_{50} values for inhibiting A549 tumor cell line growth) for 23 compounds included in the training set (**2**)

No	Compound	R1	R2	pIC_{50}
1	**10a**	H	H	4.26
2	**10b**	6-Me	H	4.74
3	**10c**	6-Me	2'-Me	4.68
4	**10d**	6-Me	4'-Me	4.57
5	**10e**	6-Me	2'-Cl	4.87
6	**11a**	H	H	4.96
7	**11b**	H	3'-Me	5.09
8	**11c**	H	4'-Me	5.91
9	**11d**	H	4'-Cl	4.93
10	**11e**	H	4'-Br	5.22
11	**11f**	6-Me	2'-Me	6.36
12	**12a**	H	H	3.75
13	**12b**	H	3'-Me	3.99
14	**12c**	6-Me	3'-Me	3.77
15	**12d**	H	4'-Me	3.77
16	**12e**	6-Me	H	3.91
17	**13a**	H	H	5.56
18	**13b**	6-Me	2'-Me	6.77
19	**13c**	H	2'-Me	5.50
20	**14a**	H	H	5.10
21	**14b**	H	2'-Me	4.99
22	**15**	H	H	3.84
23	**16**	-	-	6.28

and 2.[4a] In Table 1, the cytotoxicities of 57 compounds (1-9) are listed and those for the remaining 23 compounds (10-16) are listed in Table 2. Cytotoxicity (IC_{50} value) was obtained by the National Cancer Institute protocol based on the sulforhodamine B (SRB) method.[23] In brief, tumor cells were cultured to keep logarithmic growth by changing the medium 24 h before cytotoxicity assays. On the day of the assay, the cells were harvested by trysinization, counted, diluted in media and added to 96-well plates. The concentration of tumor cells (A-549) used was 1×10^4 cells per ml. The cells were then preincubated for 24 h in a 5% CO_2 incubator at 37 °C. The compounds dissolved in DMSO were added to the wells in six 2-fold dilutions starting from the highest concentrations, and incubated for 48 h in a 5% CO_2 incubator at 37 °C. The final DMSO concentration was 0.05%. At the termination of the incubation, the culture medium in each well was removed, and the cells were fixed with cold 10% trichloroacetic acid (TCA) for 1 h at room temperature. The microplates were washed, dried, and stained with 0.4% SRB in 1% acetic acid for 30 minutes at room temperature. The cells were washed again and the bound stain was solubilized with 10 mM Tris base solution (pH 10.5), and the absorbances were measured spectrophotometrically at 520 nm on a microtiter plate reader (Molecular Devices, Sunnyvale, CA). The data was transformed into Lotus-123 format and survival fractions were calculated by regression analysis (plotting the cell viability versus the concentration of the test compound). The IC_{50}

values represent the concentrations of the compounds that inhibit 50% of cell growth. All data represents the average values for a minimum of three wells.

Computational Details

MolconnZ 4.05 Descriptors. MolconnZ descriptors were calculated for all compounds in the dataset. They included valence, path, cluster, path/cluster and chain molecular connectivity indices,[24] kappa molecular shape indices,[25] topological[26] and electrotopological[27] state indices, differential connectivity indices,[24b,28] graph's radius and diameter,[29] Wiener[30] and Platt[31] indices, Shannon[32] and Bonchev-Trinajstic[33] information indices, counts of different vertices, counts of paths and edges between different types of vertices. Descriptors were normalized by range-scaling, so that they had values within the interval [0,1]. Normalization was required to prevent unequal descriptor weighting during the QSAR model generation process.

Dataset Division into Training, Test, and Independent Validation Sets. It is generally accepted that the internal validation of QSAR models built for the training set is sufficient to establish their predictive power.[34] However, our previous studies as well as those conducted by other groups have demonstrated that there exists no correlation between leave-one-out (LOO) cross-validated R^2 (q^2) and the correlation coefficient R^2 between the predicted and observed activities for a test set.[9c,35] Our group has advocated repeatedly the importance of the external validation that requires an independent test set for the validation of the model.[36] We have implemented a rational approach based on a sphere-exclusion algorithm for dividing the dataset into multiple training and test sets for internal and external validation, respectively.[20b,36] If possible, validation requirements must include not only test sets, but also a second external test set (an independent validation set) for the additional validation.[37] The independent validation test set should be selected randomly from the entire dataset in the beginning of the calculations. It should be used to simulate the use of QSAR models for new compounds resulting from ongoing experimental projects.

The dataset of 80 compounds was divided into three subsets (Figure 1). The first subset of seven compounds was selected randomly. The remaining 73 compounds were divided rationally into multiple training and test sets with the diversity sampling Sphere Exclusion (SE) algorithm. The

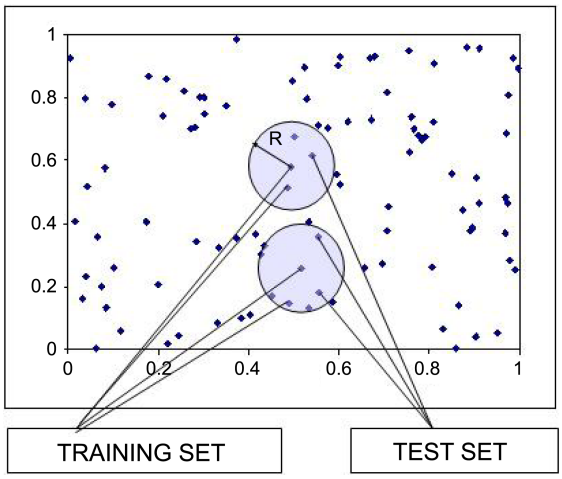


Figure 2. Division of a dataset into training and test sets using sphere-exclusion algorithm.

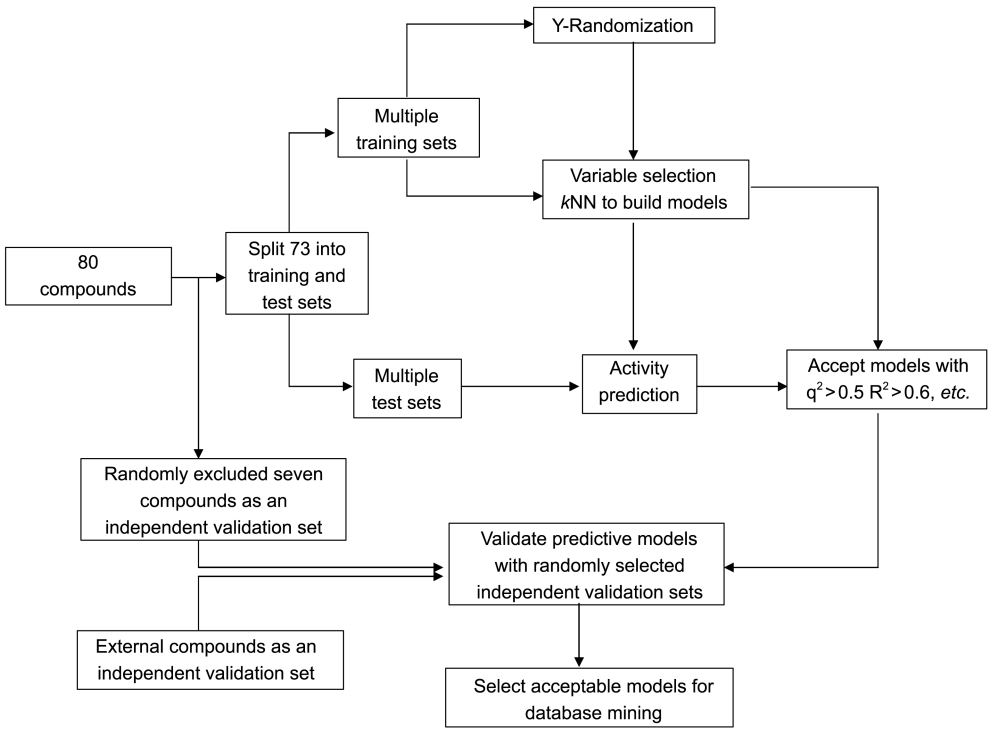


Figure 1. QSAR modeling workflow as applied to the dataset of 80 3-arylisoquinolines.

overall modeling workflow incorporating database division and model building and validation steps is shown in Figure 2.[36a]

This procedure starts with the calculation of the pairwise distance matrix D between all compounds represented by their descriptors. Let D_{min} and D_{max} be the minimum and maximum elements of D, respectively. N probe sphere radii are defined by the following formulas. $R_{min}=R_1=D_{min}$, $R_{max}=R_N=D_{max}/4$, $R_i=R_1+(i-1)*(R_N-R_1)/(N-1)$, where i=2, ..., N−1. Each probe sphere radius corresponds to one division into the training and test set. An SE algorithm used in this study consisted of the following steps. (i) select a compound randomly. (ii) include it in the training set. (iii) construct a sphere around this compound. (iv) select compounds from this sphere and include them alternatively into test or training sets. (v) exclude compounds considered in the previous step from further analysis (vi) if no more compounds left, stop. Otherwise let m be the number of probe spheres constructed and n be the number of remaining compounds. Let d_{ij} (i=1,..., m; j=1,..., n) be the distances between the remaining compounds and the probe sphere centers. Select a compound corresponding to the lowest d_{ij} value and go to step (ii). The training sets were used to build models and the test sets were used for model testing. The independent set of seven compounds was used for an additional external validation.

***k*-Nearest Neighbor (*k*NN) QSAR with Variable Selection.** We have described this approach elsewhere[21] and present here only a brief overview. *k*NN QSAR is a stochastic variable selection procedure where the model optimization is driven by simulated annealing, as illustrated in Figure 3. The *k*NN procedure is aimed at the development of the model with the highest leave-one-out (LOO) cross-validated correlation coefficient R^2 (q^2) for the training set.

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

where N and $\bar{y}$ are the number of compounds and the average observed activity of the training set, respectively, and y_i and $\hat{y}_i$ are the observed and predicted activities of the *i-th* compound, respectively.

The procedure starts with the random selection of a predefined number of descriptors from all descriptors. Activity of a compound y_i excluded in the LOO cross-validation procedure is predicted as the weighted average of activities of its nearest neighbors according to the following formula:

$$\bar{y}_i = \frac{\sum_{j=1}^{k} y_j \exp\left(-d_{ij}/\sum_{l=1}^{k} d_{il}\right)}{\sum_{i=1}^{N} \exp\left(-d_{ij}/\sum_{l=1}^{k} d_{il}\right)} \quad (2)$$

where d_{ij} are distances between the *i*-th compound and its *k* nearest neighbors (j=1,..., k). The optimal number of nearest neighbors that yields the highest q^2 value is defined as part of the LOO cross-validation process as well. After each run of the LOO procedure, a predefined number of descriptors are randomly changed, and the new value of q^2 is calculated. If q^2 (new) > q^2(old), the new set of descriptors is accepted. If q^2 (new) ≤ q^2(old), the new set of descriptors is accepted with probability p = exp(q^2(new) − q^2(old))/T, and rejected with probability (1−p), where T is a simulated annealing parameter, "temperature. During the process, T is decreasing until the predefined value, and when this value is achieved the optimization process is terminated.

For each of the six splits into the training and test sets, the number of descriptors selected in the kNN-QSAR procedure varied from 6 to 40 with step 2 (18 values). The number of models built for each pair (split, number of descriptors selected) was 10. Thus, the total number of models was 10 × 6 × 18 = 1080.

Model Validation and the Applicability Domain. QSAR models were initially validated using test sets. The models were considered acceptable if the following conditions were satisfied. (i) $q^2 > 0.5$ for the training set; (ii) correlation coefficient between predicted and observed activities of the test set $R^2 > 0.6$; (iii) $[R^2-R_0^2]/R^2 < 0.1$ and $0.85 < k < 1.15$ or $[R^2-R_0'^2]/R^2 < 0.1$ and $0.85 < k' < 1.15$, where R_0^2 and $R_0'^2$ are the coefficients of determination for regressions through the origin between predicted and observed, and observed and predicted IC_{50} values of cytotoxicity for the test set, respectively, k and k′ are the corresponding slopes, and (iv) $|R_0^2-R_0'^2| < 0.3$. The *k*NN-QSAR procedure as described (Figure 3) has been successfully used in our laboratory for many datasets.[21,38]

The activity of the test set compounds was predicted only if these compounds were within the applicability domain of the respective training set models. We define this domain[36a] as a threshold distance in multidimensional descriptor space between a test set compound and its closest nearest neighbor in the training set. If the distance is beyond the threshold, the prediction is not made since it is considered unreliable. This threshold distance is calculated as $D^2_{cutoff} = \langle D^2_{nn} \rangle +$

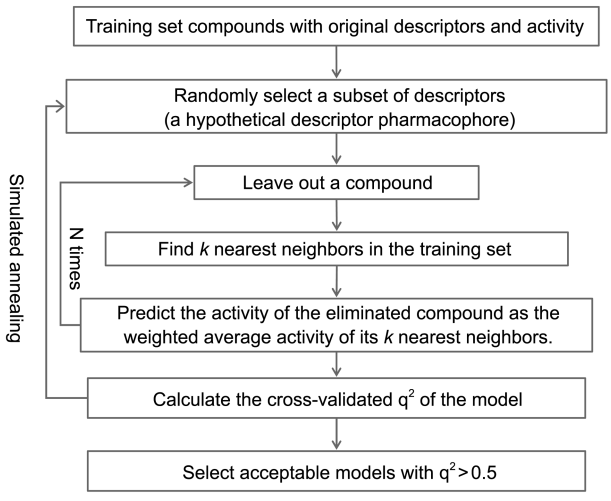


Figure 3. Flow chart of *k*NN-QSAR with Variable Selection.

Z*VAR, where $<D^2_{nn}>$ is the squared mean distance between all training set compounds and their *k* nearest neighbors in the descriptor subspace defined by the descriptors selected (descriptor pharmacophore), VAR is the variance of D_{nn}, and Z is a user-defined parameter (in this studies we used Z=0.5).

Training set models that passed our validation criteria (i)-(iv) were used for the prediction of the independent validation set of seven randomly selected compounds.

For our ongoing experimental investigations, we rely on the consensus prediction, which implies averaging the binding affinities of each compound predicted by individual models with best statistics.[9d]

Y-randomization Test. This is a generally used technique to establish, if the model is robust, and to exclude the possibility of overfitting and chance correlation.[39] The Y-randomization test was carried out after dividing the dataset into training and test sets. The robustness of the models was examined by comparing the statistics of models developed for the original dataset with those obtained by using randomized antitumor activity values of the training set. It is expected that the latter models should have low q^2 values for the training set and low R^2 values for the test set. The Y-randomization test was repeated five times for each splitting of the dataset into training and test sets.

Results and Discussion

We now discuss the results of kNN QSAR modeling of 80 3-arylisoquinolines in terms of the optimized q^2 values, variable selection, observed *vs.* predicted activities, statistical significance and experimental validation.

Development of validated *k*NN QSAR Models. To establish reliability and the true predictive power of QSAR models, it is necessary to demonstrate that they can accurately predict activities of compounds of external test sets. Generally, we accept models with q^2 values for the training set greater than 0.5 and R^2 values for predicted *vs.* observed activities of the test set compounds greater than 0.6; other characteristics should satisfy conditions described in the Methods section. From the entire dataset of 80 compounds a subset of seven compounds was randomly excluded as an independent validation set and the remaining 73 compounds were divided into six different training and test sets containing 61 and 12, 57 and 16, 60 and 13, 59 and 14, 56 and 17, and 58 and 15 compounds in training and test sets, respectively (Table 3). The 12 best models were selected from multiple *k*NN models. All these models showed good correlation between q^2 (0.81-0.88) and R^2 (0.70-0.86). The models with the highest predictive power were obtained for the test sets with 12 to 17 compounds, with the optimal number of descriptors ranging between 10 and 28.

To evaluate the external predictive power of our models, the predicted activities of the seven compounds set aside as external validation set were compared with their observed activities. We used all models that passed our validation criteria to calculate the predicted activity. The results clearly demonstrate the high prediction accuracy with R^2 values ranging from 0.71 to 0.93 (Table 4). Model 5 showed best R^2

Table 4. Twelve best *k*NN models (see Table 3): statistics for activity prediction for external validation set

No	Compounds within AD[a]	q^2	R^2	K1	K2	R_{01}^2	R_{02}^2
1	7	0.82	0.71	1.07	0.93	0.71	0.56
2	7	0.82	0.76	1.01	0.98	0.75	0.63
3	7	0.83	0.88	0.99	1.01	0.81	0.87
4	7	0.88	0.91	1.00	1.00	0.87	0.90
5	7	0.81	0.93	1.01	0.99	0.91	0.93
6	7	0.84	0.89	0.99	1.00	0.84	0.88
7	7	0.83	0.82	1.01	0.98	0.80	0.82
8	7	0.81	0.85	1.01	0.99	0.83	0.85
9	7	0.81	0.87	1.01	0.99	0.79	0.86
10	7	0.81	0.86	1.02	0.98	0.85	0.86
11	6	0.81	0.76	0.98	1.01	0.75	0.75
12	4	0.82	0.92	1.01	0.99	0.91	0.87

[a]AD – Applicability Domain.

Table 3. Twelve Best kNN Models: statistics for the training and test sets

No	training set	test set	descriptors	q^2	R^2	k1[a]	k2	[b]R_{01}^2	R_{02}^2
1	61	12	12	0.82	0.73	1.04	0.95	0.73	0.54
2	57	16	14	0.82	0.73	1.02	0.97	0.73	0.59
3	60	13	10	0.83	0.71	0.96	1.04	0.51	0.70
4	60	13	10	0.88	0.71	0.98	1.01	0.62	0.71
5	59	14	14	0.81	0.73	0.99	1.01	0.60	0.73
6	56	17	12	0.84	0.72	1.02	0.97	0.72	0.66
7	56	17	14	0.83	0.71	1.02	0.97	0.71	0.65
8	56	17	18	0.81	0.70	1.03	0.96	0.69	0.65
9	58	15	18	0.81	0.76	1.01	0.99	0.69	0.76
10	56	17	22	0.81	0.72	1.02	0.97	0.72	0.66
11	58	15	26	0.81	0.86	0.99	1.01	0.76	0.84
12	57	16	28	0.82	0.84	0.98	1.02	0.73	0.83

[a]k1 and k2 are slopes for regressions through the origin between predicted and observed, and observed and predicted activities of the test set. [b]R_{01}^2, R_{02}^2: the coefficients of determination for regressions through the origin between predicted and observed, and observed and predicted activities of the test set.

Table 5. Predicted and observed pIC_{50} values for the external test set compounds using 12 best models (see Table 3)

		Model 1		Model 2		Model 3		Model 4	
Compd	obs[a]	pre[b]	error[c]	pre	error	pre	error	pre	error
3e	4.21	5.24	1.03	4.52	0.31	4.52	0.31	4.52	0.31
4b	3.88	3.95	0.07	3.84	-0.04	3.84	-0.04	3.84	-0.04
10c	4.68	5.03	0.35	4.57	-0.11	4.57	-0.11	4.57	-0.11
4f	3.77	3.80	0.03	3.80	0.03	3.80	0.03	3.80	0.03
11e	5.22	5.91	0.69	5.91	0.69	5.07	-0.15	5.07	-0.15
6c	5.04	5.13	0.09	4.67	-0.37	4.67	-0.37	4.83	-0.21
6e	4.72	4.67	-0.10	4.57	-0.15	4.67	-0.05	4.83	0.11

		Model 5		Model 6		Model 7		Model 8	
Compd	obs	pre	error	pre	error	pre	error	Pre	error
3e	4.21	4.52	0.31	4.52	0.31	4.52	0.31	4.52	0.31
4b	3.88	3.95	0.07	3.95	0.07	3.95	0.07	3.95	0.07
10c	4.68	4.57	-0.10	4.57	-0.11	4.57	-0.11	4.57	-0.11
4f	3.77	3.8	0.03	3.8	0.03	3.80	0.03	3.80	0.03
11e	5.22	5.07	-0.20	4.93	-0.29	4.93	-0.29	4.93	-0.29
6c	5.04	5.13	0.09	5.13	0.09	5.13	0.09	5.13	0.09
6e	4.72	4.83	0.11	4.57	-0.15	5.13	0.41	5.03	0.31

		Model 9		Model 10		Model 11		Model 12	
Compd	obs	pre	error	Pre	error	pre	error	Pre	error
3e	4.21	4.52	0.31	4.52	0.31	3.95	-0.26	-	-
4b	3.88	3.95	0.07	3.95	0.07	3.84	-0.04	3.84	-0.04
10c	4.68	4.87	0.19	4.57	-0.11	4.74	0.06	4.57	-0.11
4f	3.77	3.95	0.18	3.80	0.03	3.80	0.03	3.80	0.03
11e	5.22	4.93	-0.30	5.07	-0.15	4.73	-0.49	-	-
6c	5.04	5.13	0.09	5.03	-0.01	-	-	-	-
6e	4.72	4.57	-0.20	5.13	0.41	5.03	0.31	5.03	0.31

[a]obs: observed activities. [b]pre: predicted activities. [c]error: standard errors of prediction

(0.93), R_{01}^2 (0.91) and R_{02}^2 (0.93).

As discussed in Methods, we also performed the Y-randomization test. Activities of the training set compounds were randomized and all the calculations were repeated five times. The highest training set q^2 was 0.26 and the highest test set R^2 was 0.22. These results demonstrate that the models developed for the actual dataset were robust, and their high q^2 and R^2 values could not be explained by over-fitting or chance correlation.

In summary, the results of activity prediction for compounds in datasets successfully paralleled those obtained in model building and validation using internal training and test sets. This observation agrees with our general experience in developing QSAR models with the confirmed predictive power. It emphasizes that before attempting to predict target properties of untested compounds one should exhaustively validate QSAR models developed for the training sets both internally and externally such that only models that pass rigorous validation criteria should be used for the prediction. It was interesting to analyze the performance of QSAR models with respect to distinctive chemical modifications implied in the design of the external set of new compounds.

Conclusions

In this study, we have developed and thoroughly validated QSAR models for a series of 3-arylisoquinoline compounds that have been studied as potential anticancer agents. We have demonstrated that the validated *k*NN QSAR modeling workflow was successful in generating models with high internal and external accuracy. These models can be further exploited for the design and discovery of new, potent anti-tumor agents by the means of virtual screening of available chemical databases as discussed in the recent review.[40] These virtual screening studies towards the identification and subsequent testing of novel compounds predicted to have high antitumor activity are in progress.

Acknowledgments. This work was supported by Korea Research Foundation grant (KRF-2009-0071379).

References

1. (a) Cook, S. J.; Wakelam, M. *Curr. Opin. Pharmacol.* **2005**, *5*, 341. (b) Fan, Q. L.; Zou, W. Y.; Song, L. H.; Wei, W. *Cancer*

Chemother. Pharmacol. **2005**, *55*, 189. (c) Inagawa, H.; Nishizawa, T.; Honda, T.; Nakamoto, T.; Takagi, K.; Soma, G. *Anticancer Res.* **1998**, *18*, 3957.
2. (a) Le, T. N.; Gang, S. G.; Cho, W. J. *Tetrahedron Lett.* **2004**, *45*, 2763. (b) Nakanishi, T.; Masuda, A.; Suwa, M.; Akiyama, Y.; Hoshino-Abe, N.; Suzuki, M. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2321. (c) Vogt, A.; Tamewitz, A.; Skoko, J.; Sikorski, R. P.; Giuliano, K. A.; Lazo, J. S. *J. Biol. Chem.* **2005**, *280*, 19078.
3. (a) Cho, W. J.; Park, M. J.; Chung, B. H.; Lee, C. O. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 41. (b) Cho, W. J.; Park, M. J.; Imanishi, T.; Chung, B. H. *Chem. Pharm. Bull.* **1999**, *47*, 900.
4. (a) Cho, W. J.; Min, S. Y.; Le, T. N.; Kim, T. S. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 4451. (b) Lee, S. H.; Van, H. T. M.; Yang, S. H.; Lee, K. T.; Kwon, Y.; Cho, W. J. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 2444. (c) Van, H. T. M.; Le, Q. M.; Lee, K. Y.; Lee, E. S.; Kwon, Y.; Kim, T. S.; Le, T. N.; Lee, S. H.; Cho, W. J. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 5763. (d) Cho, W. J.; Le, Q. M.; Van, H. T. M.; Lee, K. Y.; Kang, B. Y.; Lee, E. S.; Lee, S. K.; Kwon, Y. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 3531.
5. (a) Ioanoviciu, A.; Antony, S.; Pommier, Y.; Staker, B. L.; Stewart, L.; Cushman, M. *J. Med. Chem.* **2005**, *48*, 4803. (b) Staker, B. L.; Feese, M. D.; Cushman, M.; Pommier, Y.; Zembower, D.; Stewart, L.; Burgin, A. B. *J. Med. Chem.* **2005**, *48*, 2336. (c) Xiao, X.; Antony, S.; Pommier, Y.; Cushman, M. *J. Med. Chem.* **2005**, *48*, 3231.
6. Cho, W. J.; Kim, E. K.; Park, I. Y.; Jeong, E. Y.; Kim, T. S.; Le, T. N.; Kim, D. D.; Leed, E. S. *Bioorg. Med. Chem.* **2002**, *10*, 2953.
7. (a) Kim, K. E.; Cho, W. J.; Chang, S. J.; Yong, C. S.; Lee, C. H.; Kim, D. D. *Int. J. Pharm.* **2001**, *217*, 101. (b) Kim, K. E.; Cho, W. J.; Kim, T. S.; Kang, B. H.; Chang, S. J.; Lee, C. H.; Kim, D. D. *Drug. Dev. Ind. Pharm.* **2002**, *28*, 889.
8. (a) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959. (b) Marshall, G. R.; Cramer, R. D. *Trends Pharmacol. Sci.* **1988**, *9*, 285.
9. (a) Cho, S. J.; Tropsha, A. *J. Med. Chem.* **1995**, *38*, 1060. (b) Klebe, G. *Comparative Molecular Similarity Indices Analysis-CoMSIA. In 3D QSAR in Drug Design*; Kluwer/ESCOM: Dodrecht, 1988. (c) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. *J. Med. Chem.* **1998**, *41*, 2553. (d) Pérez, C.; Pastor, M.; Ortiz, A.; Gago, F. *J. Med. Chem.* **1988**, *41*, 836.
10. Golbraikh, A.; Bonchev, D.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147.
11. Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, S. W. D.; Rodgers, J. R.; Watson, D. G. *Acta Crystallogr., Sect B: Struct., Crystallogr. Cryst. Chem.* **1979**, *B 35*, 2331.
12. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. *J. Comput. Aided Mol. Des.* **2001**, *15*, 411.
13. (a) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J. Comput. Aided Mol. Des.* **1996**, *4*, 293. (b) Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. *Proteins* **2002**, *46*, 34.
14. (a) Jones, G.; Willett, P.; Glen, R. C. *J. Mol. Biol.* **1995**, *245*, 43. (b) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003**, *52*, 609.
15. (a) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M. D.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; deSolms, S. J.; Gaffin, N.; Ghosh, A. K.; Giuliani, E. A.; Graham, S. L.; Guare, J. P.; Hungate, R. W.; Lyle, T. A.; Sanders, W. M.; Tucker, T. J.; Wiggins, M.; Wiscount, C. M.; Woltersdorf, O. W.; Young, S. D.; Darke, P. L.; Zugay, J. A. *J. Med. Chem.* **1995**, *38*, 305. (b) Judson, R. *Genetic Algorithms and Their Use in Chemistry. In: Reviews in Computational Chemistry*; VCH: 1997.
16. Kramer, B.; Rarey, M.; Lengauer, T. *Proteins* **1999**, *37*, 228.
17. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. *J. Mol. Biol.* **2001**, *27*, 377.
18. (a) Cho, S. J.; Serrano, M. G.; Bier, J.; Tropsha, A. *J. Med. Chem.* **1996**, *39*, 5064. (b) Pilger, C.; Bartolucci, C.; Lamba, D.; Tropsha, A.; Fels, G. *J. Mol. Graph. Model.* **2001**, *19*, 288.
19. Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. *J. Med. Chem.* **2004**, *47*, 2356.
20. (a) Oloff, S.; Mailman, R. B.; Tropsha, A. *J. Med. Chem.* **2005**, *48*, 7322. (b) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. *J. Med. Chem.* **2002**, *45*, 2811.
21. (a) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. *J. Med. Chem.* **1999**, *42*, 3217. (b) Zheng, W.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185.
22. (a) Cho, S. J.; Zheng, W.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259. (b) Cho, S. J.; Zheng, W.; Tropsha, A. *Pac. Symp. Biocomput.* **1998**, 305. (c) Zheng, W.; Cho, S. J.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251.
23. Rubinstein, L. V.; Shoemaker, R. H.; Paull, K. D.; Simon, R. M.; Tosini, S.; Skehan, P.; Scudiero, D. A.; Monks, A.; Boyd, M. R. *J. Natl. Cancer Inst.* **1990**, *82*, 1113.
24. (a) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1986. (b) Kier, L. B. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976. (c) Randic, M. *J. Am.Chem. Soc.* **1975**, *97*, 6609.
25. (a) Kier, L. B. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109. (b) Kier, L. B. *Quant. Struct-Act. Relat.* **1987**, *6*, 8.
26. Hall, L. H.; Kier, L. B. *Quant. Struct.-Act. Relat.* **1990**, *9*, 115.
27. (a) Hall, L. H.; Mohney, B. K.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76. (b) Hall, L. H.; Mohney, B. K.; Kier, L. B. *Quant. Struct.-Act. Relat.* **1991**, *10*, 43. (c) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. *J. Comput. Aided Mol. Des.* **1996**, *10*, 513. (d) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: 1999.
28. Kier, L. B.; Hall, L. H. *Quant. Struct.-Act. Relat.* **1991**, *10*, 134.
29. Petitjean, M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331.
30. Wiener, H. J. *J. Am. Chem. Soc.* **1947**, *69*, 17.
31. Platt, J. R. *J. Chem. Phys.* **1947**, *15*, 419.
32. Shannon, C.; Weaver, W. In *Mathematical Theory of Communication*; University of Illinois: Urbana, 1949.
33. Bonchev, D.; Mekenyan, O.; Trinajstic, N. *J. Comput. Chem.* **1981**, *2*, 127.
34. (a) Basak, S. C.; Mills, D. *SAR QSAR Environ. Res.* **2001**, *12*, 481. (b) Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A. *Chem. Rev.* **2000**, *100*, 3697. (c) Cronin, M. T.; Dearden, J. C.; Duffy, J. C.; Edwards, R.; Manga, N.; Worth, A. P.; Worgan, A. D. *SAR QSAR Environ. Res.* **2002**, *13*, 167. (d) Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. *J. Med. Chem.* **2001**, *44*, 3254. (e) Girones, X.; Gallegos, A.; Carbo-Dorca, R. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1400. (f) Moss, G. P.; Dearden, J. C.; Patel, H.; Cronin, M. T. *Toxicol. In Vitro* **2002**, *16*, 299. (g) Randic, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 899. (h) Suzuki, T.; Ide, K.; Ishida, M.; Shapiro, S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 718. (i) Trohalaki, S.; Gifford, E.; Pachter, R. *Comput. Chem.* **2000**, *24*, 421. (j) Wang, X.; Yin, C.; Wang, L. *Chemosphere* **2002**, *46*, 1045.
35. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269.
36. (a) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. *J. Comput. Aided Mol. Des.* **2003**, *17*, 241. (b) Golbraikh, A.; Tropsha, A. *J. Comput. Aided Mol. Des.* **2002**, *16*, 357.
37. (a) Pintore, M.; Piclin, N.; Benfenati, E.; Gini, G.; Chretien, J. R. *Qsar & Comb. Sci.* **2003**, *22*, 210. (b) Zhang, S.; Golbraikh, A.; Tropsha, A. *J. Med. Chem.* **2006**, *49*, 2713.
38. Golbraikh, A.; Bonchev, D.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 769.
39. Wold, S. a. E. L. *Statistical Validation of QSAR Results. In Chemometrics Methods in Molecular Design*; VCH: Weinheim, Germany, 1995.
40. Tropsha, A.; Golbraikh, A. *Curr. Pharm. Des.* **2007**, *13*, 3494.