

Article

## High-Dimensional Descriptor Selection and Computational QSAR Modeling for Antitumor Activity of ARC-111 Analogues Based on Support Vector Regression (SVR)

Wei Zhou <sup>1,2,†</sup>, Zhijun Dai <sup>1,†</sup>, Yuan Chen <sup>1</sup>, Haiyan Wang <sup>3</sup> and Zheming Yuan <sup>1,2,\*</sup>

<sup>1</sup> Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Changsha 410128, China; E-Mails: mengrzhou@163.com (W.Z.); daizhijun@foxmail.com (Z.D.); chenyan0510@126.com (Y.C.)

<sup>2</sup> Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, College of Bio-Safety Science & Technology, Hunan Agricultural University, Changsha 410128, China

<sup>3</sup> Department of Statistics, Kansas State University, Manhattan, KS 66506, USA; E-Mail: hwang@ksu.edu

† These authors contributed equally to this work.

\* Author to whom correspondence should be addressed; E-Mail: zhmyuan@sina.com; Tel.: +86-731-84618163; Fax: +86-731-84618163.

*Received: 3 November 2011; in revised form: 9 January 2012 / Accepted: 17 January 2012 /*

*Published: 20 January 2012*

---

**Abstract:** To design ARC-111 analogues with improved efficiency, we constructed the QSAR of 22 ARC-111 analogues with RPMI8402 tumor cells. First, the optimized support vector regression (SVR) model based on the literature descriptors and the worst descriptor elimination multi-roundly (WDEM) method had similar generalization as the artificial neural network (ANN) model for the test set. Secondly, seven and 11 more effective descriptors out of 2,923 features were selected by the high-dimensional descriptor selection nonlinearly (HDSN) and WDEM method, and the SVR models (SVR3 and SVR4) with these selected descriptors resulted in better evaluation measures and a more precise predictive power for the test set. The interpretability system of better SVR models was further established. Our analysis offers some useful parameters for designing ARC-111 analogues with enhanced antitumor activity.

**Keywords:** ARC-111 analogues; QSAR; support vector regression; high-dimensional descriptor selection nonlinearly (HDSN) method; worst descriptor elimination multi-roundly (WDEM) method; RPMI8402

---

## 1. Introduction

Topoisomerase I (TOP I) is a clinical target for the treatment of cancer [1]. Camptothecin (CPT) and several CPT derivatives (e.g., CPT-11, topotecan) have been developed for clinical use due to CPT-induced TOP I inhibition, referred to as a cleavage complex. Despite their potential, CPTs are chemically unstable, and are substrates for the ATP-binding cassette (ABC) transporter breast cancer resistance protein (BCRP) known to be expressed in many human tumors, which bind to human serum albumin (HAS) in their carboxylate forms, leading to reduced potency in humans compared to mice [2]. So it is necessary and important to develop alternative TOP I targeting agents. 8,9-Dimethoxy-5-(2-*N,N*-dimethylaminoethyl)-2,3-methylenedioxy-5*H*-dibenzo[*c,h*][1,6] naphthyridin-6-one (ARC-111) is a promising new TOP I-targeting antitumor drug with a different drug resistance profile [2]. Cytotoxicity of ARC-111 in RPMI 8402 tumor cells has been proved to be correlated with TOP I-targeting activity, so ARC-111 is thought to be one of the assessment indicators for antitumor activities [3].

The quantitative structure-activity relationship (QSAR) is a powerful approach used for studying the relationship between drug activities and molecular structures, and it is helpful to explain how structural features determine drug activities. Especially, an acceptable QSAR has the advantages of higher-speed and lower-costs than experimental testing for drug activity evaluation. Yu *et al.* have compared QSAR modeling of antitumor activity of ARC-111 analogues using stepwise multiple linear regression (stepwise MLR), partial least squares (PLS) and artificial neural network (ANN), and the results showed the ANN model was the most powerful for the test [4]. However, the ANN model still had an obvious defect in the reliability of structural information because its independent variables had to be selected by linear techniques from only 15 molecular descriptors, so the QSAR of an increasing number of ARC-111 analogues possessing antitumor activities are still not well understood. Therefore, selecting more effective molecular features from the high-dimensional ones of ARC-111 analogues using new methods will possibly provide more useful information for the design of new antitumor drugs. Parameter Client provides an interface for different programs that calculate several groups of descriptors with a total number of >3000 [5]. For each ARC-111 analogue, its high-dimensional descriptors could be calculated freely and quickly. Because many of these descriptors are redundant and sometimes irrelevant, models for nonlinear selection of the most useful subset of descriptors are needed for theoretical analysis and for practical applications.

The support vector machine (SVM) is a class of learning-based nonlinear modeling technique with proven performance in a wide range of practical applications [6]. Originally, SVMs were developed for classification or qualitative modeling problems. With the introduction of a  $\epsilon$ -insensitive loss function, SVM has been extended to solve nonlinear regression (or quantitative modeling) problems. To select reasonable features, we employed two in-house developed methods based on SVM

regression (SVR): the worst descriptor elimination multi-roundly (WDEM) [7] and the high-dimensional descriptors selection nonlinearly (HDSN) [8], and then constructed QSAR models of ARC-111 analogues based on the SVR technique in this study.

The objectives of this work were: (1) to test the effectiveness of the SVR model on ARC-111 analogues by comparing them with other chemometric tools including stepwise MLR, PLS, and ANN; (2) to construct and evaluate QSAR models using SVR with selection of descriptors from high-dimensional features of ARC-111 analogues; (3) to analyze the explanatory power of the SVR models; and (4) to predict the activities of several theoretical drugs based on our model and thus provide specific parameters for future drug development.

## 2. Results and Discussion

### 2.1. Comparative QSAR Modeling with the Low-Dimensional Literature Descriptors Using Stepwise MLR, PLS, ANN and SVR Techniques

To verify the generalization ability of QSAR constructed using SVR technique, a low-dimensional literature dataset with 9 descriptors was adopted. The 9 descriptors were the combined set of features from stepwise MLR and PLS in [4]. To further eliminate the redundant descriptors from this literature dataset, every available descriptor were gradually removed one by one from the model using our WDEM method (10-fold cross-validation) until the model with the lowest *MSE* was obtained. Six key descriptors [*MW*, *Dipole*, *MolPol*, *JGT*, *E(H-bond)* and  $\Delta H_f^0$ ] were reserved by 3 rounds of nonlinear selection. Then the two low-dimensional datasets with 9 and 6 descriptors, respectively, were trained by leave-one-out (LOO) cross-validation and modeled in five Kernel functions ( $t = 0$ ;  $t = 1, d = 2$ ;  $t = 1, d = 3$ ;  $t = 2$ ;  $t = 3$ ). The results of the independent test showed (1) the SVR1 model ( $t = 1, d = 3$ ) with all literature features had higher predictive ability than stepwise MLR and PLS; and (2) the SVR2 model ( $t = 2$ ) with *MSE* of 0.061,  $R^2$  of 0.950 and  $R^2_{pred}$  of 0.918 for the test set had comparable predictive ability with the ANN (the number of units in hidden layers was four and the number in the training set was ten [4]) model even though SVR2 used less descriptors (Table 1). It indicated the SVR model was also a powerful technique for a given set of low-dimensional descriptors.

**Table 1.** Comparative quantitative structure-activity relationship (QSAR) modeling of the independent test, based on the literature dataset.

	Stepwise MLR	PLS	ANN	SVR1	SVR2
Number of descriptors	5	7	9	9	6
<i>MSE</i>	0.201	0.167	0.050	0.141	0.061
$R^2$	0.910	0.890	0.962	0.937	0.950
$R^2$	0.730	0.775	0.933	0.811	0.918

The SVR model with 6 descriptors (SVR2) produced better results than the SVR model with all 9 descriptors (SVR1). We noted that the 6 descriptors were obtained with the WDEM from the 9 descriptors. This showed that the WDEM method might be effective to choose relevant descriptors for more accurate prediction of the activities of ARC-111 analogues. This property will be helpful for the modeling with high-dimensional features. Considering nonlinear function, predictive ability and

computing time, the Radial Basis Function ( $t = 2$ ) and 10-fold cross-validation will be adopted in future feature selecting, and the Radial Basis Function ( $t = 2$ ) and LOO cross-validation will be adopted in independent tests.

## 2.2. QSAR Modeling with the High-Dimensional Descriptors Using SVR Technique

To improve drug design of ARC-111 analogues, the analysis of high-dimensional descriptors may result in better prediction. Using the software, PCLIENT, 2,923 molecular descriptors were calculated. Then the high-dimensional dataset containing the independent variables (all 2,923 descriptors) and the dependent variables [ $\text{pIC}_{50}$  (expt.) values] was used for modeling. Because the high-dimensional descriptors had more redundant information, we focused on how to select nonlinearly less but more critical descriptors using SVR. We have developed two novel methods that could select important descriptors from thousands of them. By initial coarse screening using the HDSN method to filter out irrelevant features, the data set would switch from high-dimensional into low-dimensional. Then further careful screening using the WDEM method would turn the data set with low-dimensional features into one with only important descriptors. Throughout the process, the descriptors in modeling with higher  $MSE$  values were removed gradually and nonlinearly until the model with the lowest  $MSE$  value was obtained. Finally, the SVR models for the test set based on the obtained descriptors were developed and evaluated.

In feature screening, the Radial Basis Function ( $t = 2$ ) and 10-fold cross-validation were adopted. Based on our HDSN method, descriptors of 18 ARC-111 analogues in SVR3 (and SVR4) model were reduced from 2,923 to 9 (and 13) by 9 (and 8) rounds of nonlinear screening. Furthermore, based on our WDEM method, descriptors were further reduced to 7 (and 11) by 2 rounds of nonlinear screening (Table 2). In the independent test, five Kernel functions and LOO cross-validation were adopted. Finally, the effective SVR3 and SVR4 models were obtained only by the Radial Basis Function ( $t = 2$ ). The results of the independent test (Table 2) showed the SVR3 (and SVR4) models had similar or better predictive power with  $MSE$  of 0.032 (and 0.028),  $R^2$  of 0.964 (and 0.971) and  $R^2_{pred}$  of 0.957 (and 0.962) for the test set than stepwise MLR, PLS and ANN techniques. By nonlinear screening using our HDSN and WDEM methods, the SVR model with the obtained features from high-dimensional features of ARC-111 analogues had stronger generalization ability than all reference models for antitumor activity prediction in RPMI 8402. Furthermore, based on the SVR4 model,  $\text{pIC}_{50}$  (pred.) values of 12 theoretical ARC-111 analogues were predicted for drug activity evaluation. The results showed no drug with higher antitumor activity appeared in these theoretical designs, and suggested utilizing other substituents or other positions to design more effective drugs.

**Table 2.** Comparative QSAR modeling of the independent test based on the high-dimensional descriptors selection using support vector regression (SVR).

	Stepwise MLR	PLS	ANN	SVR3	SVR4
Number of descriptors	5	7	9	7	11
$MSE$	0.201	0.167	0.050	0.032	0.028
$R^2$	0.910	0.890	0.962	0.964	0.971
$R^2_{pred}$	0.730	0.775	0.933	0.957	0.962

The SVR3 and SVR4 models predicted that the antitumor activity of ARC-111 analogues depends on 7 and 11 molecular factors, respectively. According to the interpretability analysis of the SVR model we have established [9], the significance of the regression model and the importance of single indicator was obtained based on SVR and *F*-test. The results showed the nonlinear regression of the SVR3 model ( $R^2 = 0.947$ ) was highly significant because its *F* value (21.017) was greater than  $F_{0.01}(7, 10)$  value, and the nonlinear regression of the SVR4 model ( $R^2 = 0.947$ ) was significant because its *F* value (7.310) was greater than  $F_{0.05}(11, 6)$  value. The five most important descriptors in SVR3 were *c6A* (highly significant), *ATSIv* (highly significant), *nCIC* (highly significant), *MATS3e* (highly significant) and *nCrs* (significant), and the only one most important descriptor in SVR4 was *BELv2* (significant) (Table 3).

**Table 3.** The retained descriptors by the high-dimensional descriptor selection nonlinearly (HDSN) and worst descriptor elimination multi-roundly (WDEM) methods and their *F*-test values.

Model	Group name	Descriptor name	<i>F</i> -value
SVR3	GSFRAG	<i>c6A</i> : Number of fragments Cyc6[A]	26.555 **
	2D autocorrelations	<i>ATSIv</i> : Broto-Moreau autocorrelation of a topological structure - lag 1 / weighted by atomic van der Waals volumes	25.175 **
	Constitutional descriptors	<i>nCIC</i> : Number of rings	12.210 **
	2D autocorrelations	<i>MATS3e</i> : Moran autocorrelation - lag 3 / weighted by atomic Sanderson electronegativities	12.114 **
	Functional group counts	<i>nCrs</i> : Number of ring secondary C(sp3)	5.898 *
	Topological charge indices	<i>GGI5</i> : Topological charge index of order 5	3.687
	Geometrical descriptors	<i>QYYv</i> : Qyy COMMA2 value / weighted by atomic van der Waals volumes	2.387
	SVR4	BCUT descriptors	<i>BELv2</i> : Lowest eigenvalue n. 2 of Burden matrix / weighted by atomic van der Waals volumes
GSFRAG-L		<i>p3-2N</i> : Number of fragments Path3 with label N on atom 2	3.771
Randic molecular profiles		<i>SP20</i> : Shape profile no. 20	3.511
Eigenvalue-based indices		<i>SEigZ</i> : Eigenvalue sum from Z weighted distance matrix (Barysz matrix)	2.456
Constitutional descriptors		<i>nN</i> : Number of Nitrogen atoms	2.456
RDF descriptors		<i>RDF040v</i> : Radial distribution function - 4.0 / weighted by atomic van der Waals volumes	2.435
Walk and path counts		<i>TWC</i> : Total walk count	2.425
RDF descriptors		<i>RDF040p</i> : Radial distribution function - 4.0 / weighted by atomic polarizabilities	2.398
Topological descriptors		<i>ZM1V</i> : first Zagreb index by valence vertex degrees	2.084
RDF descriptors		<i>RDF040e</i> : Radial distribution function - 4.0 / weighted by atomic Sanderson electronegativities	1.304
GETAWAY descriptors		<i>HATS0u</i> : Leverage-weighted autocorrelation of lag 0 / unweighted	0.599

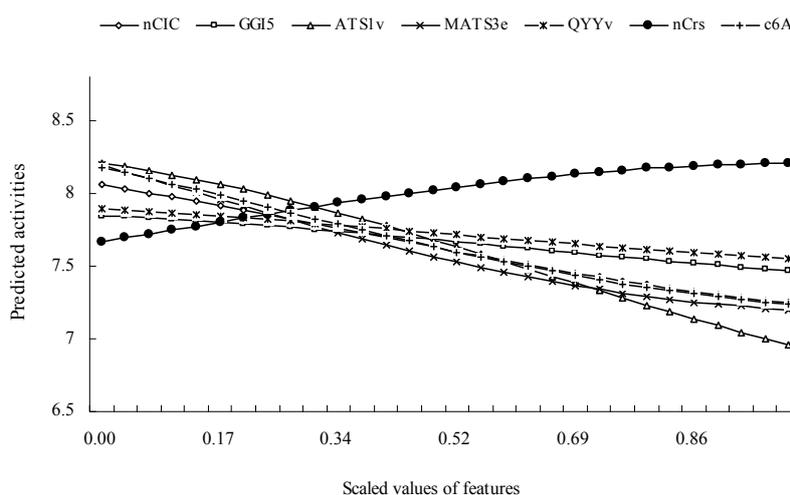
\*  $p < 0.05$ ; \*\*  $p < 0.01$ ;  $F_{0.05}(1,10) = 4.96$ ;  $F_{0.01}(1,10) = 10.04$ ;  $F_{0.05}(1,6) = 5.99$ ;  $F_{0.01}(1,6) = 13.74$ ;  $F_{0.05}(7,10) = 3.14$ ;  $F_{0.01}(7,10) = 5.2$ ;  $F_{0.05}(11,6) = 4.03$ ;  $F_{0.01}(11,6) = 7.8$ .

The *F*-test values of the independent variables showed that GSFRAG, 2D autocorrelations and constitutional descriptors, played important roles in describing anticancer activities. According to the analysis of single indicator importance, *c6A*, *ATSIv*, *nCIC*, *MATS3e* and *nCrs* in the SVR3 model and *BELv2* in the SVR4 model appeared to be the most significant descriptors of ARC-111 analogues.

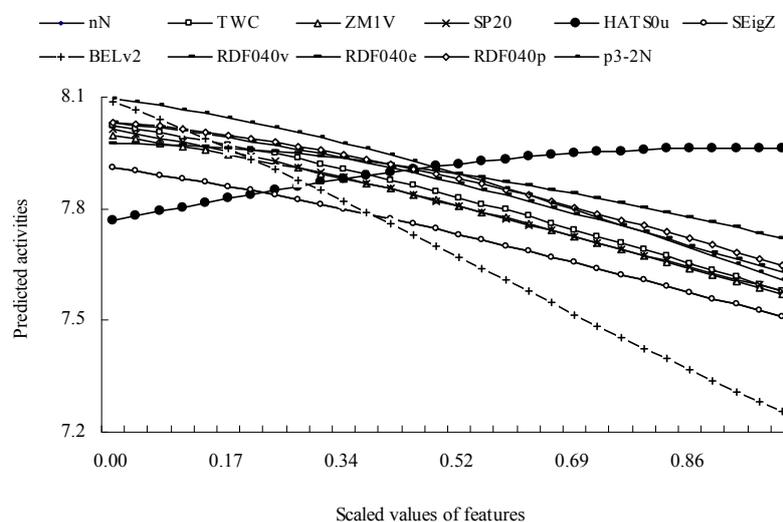
*ATSIv* [10], *nCIC* [11], *MATS3e* [12–16], *nCrs* [17–23] and *BELv2* [10,24–26] have been previously reported in different literature models, respectively. To our knowledge, *c6A* has never been reported as a critical descriptor, so it is unclear what new information is added as an important descriptor. Previous works have shown the physical and biological significance of several significant descriptors founded in our analysis. *nCIC*, as one of the highly significant descriptors, appears to have an influence on binding. It is likely that the active site of a possible target possesses more than one binding site, therefore the number of rings could be important for fitting into a hydrophobic pocket [11]. *MATS3e*, as one of the highly significant descriptors, are weighted by atomic Sanderson electronegativities, and might partly influence the drug aqueous solubility [15]. *BELv2*, weighted by atomic van der Waals volumes of Burden matrix, contribute to decrease the affinity of the ligands [25].

For all descriptors, the analysis of single-factor effects showed that the antitumor activity was positively correlated with *nCrs* values but negatively correlated with a further 6 descriptor values in the SVR3 model, and antitumor activity was positively correlated with *HATS0u* values but was negatively correlated with the values of a further 10 descriptors in the SVR4 model (Figure 1).

**Figure 1.** Single-factor effects of features in the SVR3 (A) and SVR4 (B) models.



(A)



(B)

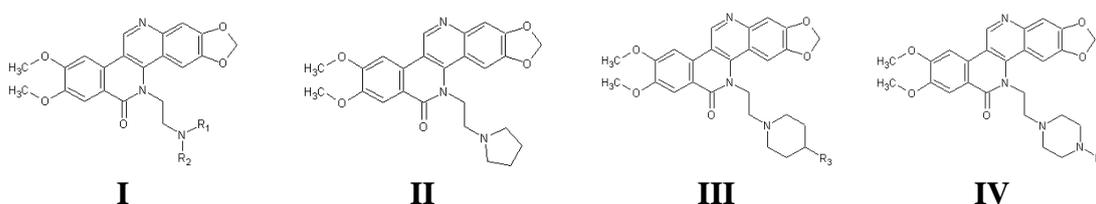
Perhaps, starting from a descriptor pool and then revealing the physico-chemical properties of a limited number of selected descriptors, as seen in some papers, can lead to a compromise between both approaches. In most of the models for prediction, theoretical molecular descriptors were used. Experimental chromatographic descriptors could be useful but are tedious to determine and therefore less popular [10]. Therefore, our results can be helpful to explain how descriptors could determine the antitumor activities of ARC-111 analogues, and improve drug design for new drug development. In addition to anticancer bioactivity [27], the structure activity relationship analysis can be applied to toxicology [28–30], *etc.* Therefore, a good QSAR model has broad application prospects.

### 3. Materials and Methods

#### 3.1. Structures and Activities

According to the types and roles of ARC-111 substituents reported in literature [3], 12 theoretical ARC-111 analogues were designed and evaluated. The structures of these 12 theoretical analogues and 22 experimental ones from [4] were divided into four types (Figure 2) and listed in Table 4.  $IC_{50}$  ( $\mu M$ ), the concentration of compounds causing 50% cell growth inhibition against tumor cell lines [3], are converted to negative logarithms of  $IC_{50}$  ( $pIC_{50}$ ) [4]. The collected 22 experimental  $pIC_{50}$  [ $pIC_{50}$  (expt.)] values against RPMI8402 tumor cells ranged from 6.071 to 9.523. To obtain statistically robust QSAR models and compare with the results of MLR, PLS, and ANN in [4], the experimental data sets in Table 4 were partitioned into the training set with 18 compounds and the test set with 4 compounds as in [4].

**Figure 2.** Four types of ARC-111 analogues structures.



**Table 4.** Substituents and activities of 34 ARC-111 analogues.

Experimental drugs					Theoretical drugs								
Compound	Type	Substituent				$pIC_{50}$ (expt.)	Compound	Type	Substituent				$pIC_{50}$ (pred.) <sup>b</sup>
		R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>				R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	
1	I	Me	Me			8.699	1	I	Me	Et			8.651
2		Me	Bn			7.276	2		Me	<i>t</i> -Bu			8.172
3		Et	Bn			7.114	3		Et	<i>t</i> -Bu			7.876
4		<i>i</i> -Pr	Bn			6.523	4		<i>t</i> -Bu	<i>t</i> -Bu			7.388
5		<i>t</i> -Bu	Bn			6.071	5		<i>t</i> -Bu	<i>i</i> -Pr			6.908
6		Bn	Bn			6.420	6	III			Bn		6.668
7		Et	Et			8.222	7				Et		7.208
8		<i>i</i> -Pr	<i>i</i> -Pr			8.097 <sup>a</sup>	8				<i>t</i> -Bu		6.904
9		H	Me			9.523	9				<i>i</i> -Pr		6.617

Table 4. Cont.

Experimental drugs					Theoretical drugs								
Compound	Type	Substituent				pIC <sub>50</sub> (expt.)	Compound	Type	Substituent				pIC <sub>50</sub> (pred.) <sup>b</sup>
		R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>				R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	
10		H	Et			8.699 <sup>a</sup>	10	IV			Et	6.248	
11		H	<i>i</i> -Pr			8.523	11				<i>t</i> -Bu	6.102	
12		H	<i>t</i> -Bu			8.699	12				<i>i</i> -Pr	6.100	
13		H	Bn			7.796							
14		H	H			8.398							
15		Me	<i>i</i> -Pr			8.097 <sup>a</sup>							
16		Et	<i>i</i> -Pr			8.301							
17	II					8.523							
18	III			H		8.155							
19				Me		7.523							
20	IV				Bn	6.398 <sup>a</sup>							
21					H	7.046							
22					Me	6.523							

<sup>a</sup> Four experimental compounds in the test set; <sup>b</sup> predicted values of 12 theoretical compounds by the SVR4 model.

### 3.2. Calculation of Molecular Descriptors

First, to understand the QSAR reliability of modeling ARC-111 analogue activities using SVR technique, 4 electronic [*Dipole*, *E(H-bond)*,  $\Delta H_f^0$  and  $E_T$ ], 2 spatial (*MW*,  $R_g$ ) and 1 physicochemical (*MolPol*) descriptors as well as different topological parameters (*JGT*, *Wiener*) from the literature were adopted to construct QSAR models. The 9 descriptors were obtained by molecule energy optimization using MM2 ChemOffice 2005, and then were calculated by MODEL and ChemOffice 2005 [4].

Second, to develop a better QSAR model based on high-dimensional data sample using SVR technique, molecular structures were represented by about 3,000 molecular descriptors that encoded much more structural information. These descriptors were generated by the software PCLIENT (<http://www.vcclab.org/lab/pclient/>) and classified under 24 groups (Table 5) [5]. The calculation process of the descriptors involved the following steps: the structures of the compounds were drawn using JME Editor of Peter Ertl and saved as SMILES files, and then the SMILES files as a task were added to the software PCLIENT for calculating all of the descriptors in the default state.

Table 5. Group and count of descriptors from the software PCLIENT.

Group No.	Group of descriptors	Count	Group No.	Group of descriptors	Count
1	Constitutional descriptors	48	13	RDF descriptors	150
2	Topological descriptors	119	14	3D-MoRSE descriptors	160
3	Walk and path counts	47	15	WHIM descriptors	99
4	Connectivity indices	33	16	GETAWAY descriptors	197
5	Information indices	47	17	Functional group counts	121
6	2D autocorrelations	96	18	Atom-centered fragments	120
7	Edge adjacency indices	107	19	Charge descriptors	14

Table 5. Cont.

Group No.	Group of descriptors	Count	Group No.	Group of descriptors	Count
8	BCUT descriptors	64	20	Molecular properties	28
9	Topological charge indices	21	21	ET-state Indices	>300
10	Eigenvalue-based indices	44	22	ET-state Properties *	3
11	Randic molecular profiles	41	23	GSFRAG Descriptor	307
12	Geometrical descriptors	74	24	GSFRAG-L Descriptor	886
					Total: >3000

\* This group of descriptors did not exist in the default state.

### 3.3. Model Development

To reduce dimensionality and improve model robustness in QSAR analysis, high-dimensional features would be screened coarsely and nonlinearly into low-dimensional features with lower mean squared error (*MSE*) by our HDSN method [8], and then low-dimensional features would be further screened nonlinearly by our WDEM method [31].

### 3.4. Model Evaluation

The selection of descriptors and the optimization of Kernel functions parameters were examined by 10-fold or LOO validation with the minimum *MSE*; the predictive capacity of the models was assessed based on *MSE*, the squared multiple correlation coefficient ( $R^2$ ) and the squared predictive correlation coefficient ( $R_{pred}^2$ ) values calculated by the following equations:

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

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

$$R_{pred}^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_{training})^2} \quad (3)$$

Here  $y_i$ ,  $\hat{y}_i$ ,  $\bar{y}$ ,  $\bar{\hat{y}}$  and  $n$ , respectively, represented the experimental values, the predicted values, the mean values of the experimental values, the mean values of the predicted values and the number of compounds of the test set, and  $\bar{y}_{training}$  was the mean activity value of the training set. Generally, an acceptable QSAR model was considered to have a higher predictive power only having the lower *MSE* [31], the higher  $R^2$  [8] and the higher  $R_{pred}^2$  (at least >0.6) [32] for the test set.

## 4. Conclusions

In our QSAR analysis, the structural information of 34 ARC-111 analogues was described using 2923 molecular descriptors obtained. Two groups of more important descriptors were obtained using two nonlinear descriptor selection methods, and then used to model the activities of these ARC-111 analogues based on SVR. The two SVR models demonstrated consistently better performance than

reference models in terms of prediction accuracy for the test data. Our results offer new theoretical tools for drug design and development.

### Acknowledgments

The authors acknowledge financial support from the Science Foundation for Distinguished Young Scholars of Hunan Province, China (No. 10JJ1005), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20114320120005), the Research Foundation of Education Bureau of Hunan Province, China (No. 09C502) and the Science Foundation for Talents from Hunan Agricultural University of China (No. 07YJ05). The authors thank Mingshun Chen at USDA-ARS and Department of Entomology, Kansas State University, Manhattan, Kansas, USA for his comments and suggestions during the manuscript preparation. The authors also thank Gang Qian, Lifeng Wang and Yijun Su at Department of Bioinformatics, Hunan Agricultural University, Changsha, China for their help.

### References

1. Rasheed, Z.A.; Rubin E.H. Mechanisms of resistance to topoisomerase I-targeting drugs. *Oncogene* **2003**, *22*, 7296–7304.
2. Li, T.K.; Houghton, P.J.; Desai, S.D.; Daroui, P.; Liu, A.A.; Hars, E.S.; Ruchelman, A.L.; LaVoie, E.J.; Liu L.F. Characterization of ARC-111 as a novel Topoisomerase I-targeting anticancer drug. *Cancer Res.* **2003**, *63*, 8400–8407.
3. Ruchelman, A.L.; Houghton, P.J.; Zhou, N.; Liu, A.; Liu, L.F.; LaVoie, E.J. 5-(2-Aminoethyl)dibenzo[*c,h*][1,6]naphthyridin-6-ones: variation of *N*-alkyl substituents modulates sensitivity to efflux transporters associated with multidrug resistance. *J. Med. Chem.* **2005**, *48*, 792–804.
4. Yu, Y.J.; Su, R.X.; Wang, L.B.; Qi, W.; He, Z.M. Comparative QSAR modeling of antitumor activity of ARC-111 analogues using stepwise MLR, PLS, and ANN techniques. *Med. Chem. Res.* **2010**, *19*, 1233–1244.
5. Tetko, I.V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.A.; Radchenko, E.V.; Zefirov, N.S.; Makarenko, A.S.; *et al.* Virtual computational chemistry laboratory - design and description. *J. Comput. Aid. Mol. Des.* **2005**, *19*, 453–463.
6. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
7. Tan, X.S.; Yuan, Z.M.; Zhou, T.J.; Wang, C.J.; Xiong, J.Y. Multi-KNN-SVR combinatorial forecast and its application to QSAR of fluorine-containing compounds. *Chem. J. Chin. Univ.* **2008**, *29*, 95–99.
8. Dai, Z.J.; Zhou, W.; Yuan, Z.M. A novel method of nonlinear rapid feature selection for high-dimensional features and its application in peptide QSAR modeling based on support vector machine. *Acta Phys. Chim. Sin.* **2011**, *27*, 1654–1660.
9. Wang, L.F.; Tan, X.S.; Bai, L.Y.; Yuan, Z.M. Establishing an interpretability system for support vector regression and its application in QSAR of organophosphorus insecticide. *Asian J. Chem.* **2011**, *24*, in press.

10. Geerts, T.; Vander Heyden, Y. *In-silico* predictions of ADME-Tox properties: Drug absorption. *Comb. Chem. High Throughput Screen.* **2011**, *14*, 339–361.
11. Marković, V.; Erić, S.; Juranić, Z.D.; Stanojković, T.; Joksović, L.; Ranković, B.; Kosanić, M.; Joksović, M.D. Synthesis, antitumor activity and QSAR studies of some 4-aminomethylidene derivatives of edaravone. *Bioorg. Chem.* **2011**, *39*, 18–27.
12. Prabhakar, Y.S.; Rawal, R.K.; Gupta, M.K.; Solomon, V.R.; Katti, S.B. Topological descriptors in modeling the HIV inhibitory activity of 2-Aryl-3-pyridyl-thiazolidin-4-ones. *Comb. Chem. High Throughput Screen.* **2005**, *8*, 431–437.
13. Riahi, S.; Ganjali, M.R.; Pourbasheer, E.; Norouzi, P. QSRR study of GC retention indices of essential-oil compounds by multiple linear regression with a genetic algorithm. *Chromatographia* **2008**, *67*, 917–922.
14. Goudarzia, N.; Fatemib, M.H.; Samadi-Maybodib, A. Quantitative structure-properties relationship study of the <sup>29</sup>Si-NMR chemical shifts of some silicate species. *Spectrosc. Lett.* **2009**, *42*, 186–193.
15. Cao, D.S.; Xu, Q.S.; Liang, Y.Z.; Chen, X.; Li, H.D. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. *J. Chemometr.* **2010**, *24*, 584–595.
16. Gozalbes, R.; Pineda-Lucena, A. QSAR-based solubility model for drug-like compounds. *Bioorgan. Med. Chem.* **2010**, *18*, 7078–7084.
17. Lapinsh, M.; Veiksina, S.; Uhlén, S.; Petrovska, R.; Mutule, I.; Mutulis, F.; Yahorava, S.; Prusis, P.; Wikberg, J.E.S. Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes. *Mol. Pharmacol.* **2005**, *67*, 50–59.
18. Liang, G.Z.; Yang, S.B.; Zhou, Y.; Zhou, P.; Li, Z.L. Using scores of amino acid topological descriptors for quantitative sequence-mobility modeling of peptides based on support vector machine. *Chin. Sci. Bull.* **2006**, *51*, 2700–2705.
19. Bansal, R.; Karthikeyan, C.; Moorthy, N.S.H.N.; Trivedi, P. QSAR analysis of some phthalimide analogues based inhibitors of HIV-1 integrase. *Arkivoc* **2007**, *XV*, 66–81.
20. Malik, A.; Singh, H.; Andrabi, M.; Husain, S.A.; Ahmad, S. Databases and QSAR for cancer research. *Cancer Inform.* **2006**, *2*, 99–111.
21. Lapins, M.; Wikberg, J.E. Proteochemometric modeling of drug resistance over the mutational space for multiple HIV protease variants and multiple protease inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 1202–1210.
22. Tanabe, K.; Lučić, B.; Amić, D.; Kurita, T.; Kaihara, M.; Onodera, N.; Suzuki, T. Prediction of carcinogenicity for diverse chemicals based on substructure grouping and SVM modeling. *Mol. Divers.* **2010**, *14*, 789–802.
23. Gharagheizi, F. An accurate model for prediction of autoignition temperature of pure compounds. *J. Hazard. Mater.* **2011**, *189*, 211–221.
24. Fedorowicz, A.; Zheng, L.; Singh, H.; Demchuk, E. QSAR study of skin sensitization using local lymph node assay data. *Int. J. Mol. Sci.* **2004**, *5*, 56–66.
25. González, M.P.; Terán, C.; Teijeira, M.; Besada, P.; González-Moa, M.J. BCUT descriptors for predicting affinity toward A<sub>3</sub> adenosine receptors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3491–3495.
26. Pasha, F.A.; Srivastava, H.K.; Srivastava, A.; Singh, P.P. QSTR study of small organic molecules against *Tetrahymena pyriformis*. *QSAR Comb. Sci.* **2007**, *26*, 69–84.

27. Putz, M.V.; Putz, A.-M.; Lazea, M.; Ienciu, L.; Chiriac, A. Quantum-SAR extension of the spectral-SAR algorithm. Application to polyphenolic anticancer bioactivity. *Int. J. Mol. Sci.* **2009**, *10*, 1193–1214.
28. Putz, M.V.; Lacrămă, A.-M. Introducing spectral structure activity relationship (S-SAR) analysis. Application to ecotoxicology. *Int. J. Mol. Sci.* **2007**, *8*, 363–391.
29. Lacrămă, A.-M.; Putz, M.V.; Ostafe, V. A spectral-SAR model for the anionic-cationic interaction in ionic liquids: application to *vibrio fischeri* ecotoxicity. *Int. J. Mol. Sci.* **2007**, *8*, 842–863.
30. Chicu, S.A.; Putz, M.V. Köln-Timișoara molecular activity combined models toward interspecies toxicity assessment. *Int. J. Mol. Sci.* **2009**, *10*, 4474–4497.
31. Chen, Y.; Yuan, Z.M.; Zhou, W.; Xiong, X.Y. A novel QSAR model based on geostatistics and support vector regression. *Acta Phys. Chim. Sin.* **2009**, *25*, 1587–1592.
32. Zhang, S.X.; Wei, L.Y.; Bastow, K.; Zheng, W.F.; Brossi, A.; Lee, K.H.; Tropsha, A. Antitumor agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *J. Comput. Aid. Mol. Des.* **2007**, *21*, 97–112.

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).