Acta Pharmacol Sin 2008 Mar; 29 (3): 385–396

Full-length article

Modeling resistance index of taxoids to MCF-7 cell lines using ANN together with electrotopological state descriptors[1]

Pei-pei DONG[2,3], Yan-yan ZHANG[2,3], Guang-bo GE[2,3], Chun-zhi AI[2,3], Yong LIU[2], Ling YANG[2,5], Chang-xiao LIU[4,5]

[2]Laboratory of Pharmaceutical Resource Discovery, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China; [3]Graduate School of Chinese Academy of Sciences, Beijing 100049, China; [4]Tianjin Key Laboratory of Pharmacodynamics and Pharmacokinetics, Tianjin Institute of Pharmaceutical Research , Tianjin 300193, China

Key words

artificial neural network model; taxoids; multidrug resistance; resistance index; electrotopological state indices; principle component analysis; quantitative structure-activity relationship

[1]Project supported by the National Natural Science Foundation of China (No 30640066 and 30630075), and the Innovation Youth Foundation of Dalian Institute of Chemical Physics (No S200612).
[5]Correspondence to Prof Ling YANG and Prof Chang-xiao LIU.
Phn 86-411-8437-9317.
Fax 86-411-8467-6961.
E-mail yling@dicp.ac.cn (Ling YANG)
Phn 86-22-2300-6863.
Fax 86-22-2300-6860.
E-mail liuchangxiao@vip.163.com (Chang-xiao LIU)

Received 2007-07-16
Accepted 2007-09-25

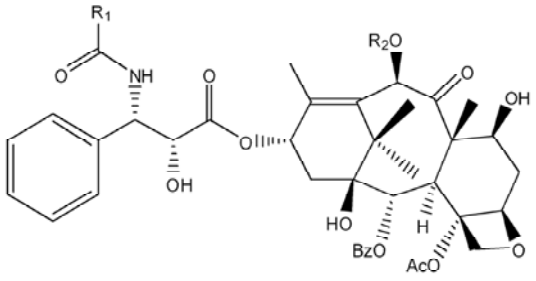
doi: 10.1111/j.1745-7254.2008.00746.x

Abstract

Aim: To develop an artificial neural network model for predicting the resistance index (RI) of taxoids. **Methods:** A dataset of 63 experimental data points were compiled from published studies and randomly subdivided into training and external test sets. Electrotopological state (E-state) indices were calculated to characterize molecular structure together with a principle component analysis to reduce the variable space and analyze the relative importance of E-state indices. Back propagation neural network technique was used to build the models. Five-fold cross-validation was performed and 5 models with different compound composition in training and validation sets were built. The independent external test set was used to evaluate the predictive ability of models. **Results:** The final model proved to be good with the cross-validation Q^2_{cv}0.62, external testing R^2 0.84, and the slope of the regression line through the origin for the testing set at 0.9933. **Conclusion:** The quantitative structure–activity relationship model can predict the RI to a relative nicety, which will aid in the development of new anti-multidrug resistance taxoids.

Introduction

Paclitaxel (taxol, Bristol–Myers Squibb, New York, New York, USA[1]) and docetaxel (taxotere, Sanofi–Aventis, Paris, Paris, France[2]; Figure 1) are arguably two of the most effective and clinically successful anticancer agents widely used for the administration of several solid tumors, such as breast and ovarian cancers. Both agents have a unique anticancer mechanism known as microtubule-stabilizing activity. They act by accelerating the polymerization of tubulin and inhibiting the depolymerization of microtubules, thus leading to cell apoptosis[3–5]. Although both drugs possess strong antitumor activity, chemotherapy is usually limited by the presence of multidrug resistance (MDR). MDR is the cross-re-



1 Paclitaxel: R_1=Ph, R_2=Ac

2 Docetaxel: R1=tBuO, R_2=H

Figure 1. Structure of paclitaxel and docetaxel.

sistance of tumor cell lines to several structurally and functionally unrelated chemotherapeutic agents after exposure to a single cytotoxic drug[6,7]. Therefore, it is urgent to develop a new generation of anti-MDR taxoids.

Extensive research has been conducted to better understand the mechanism of MDR, and until now, several targets have been recognized to be associated with MDR, such as the overexpression of the ATP-binding cassette (ABC) transporter proteins and the mutations on tubulin[8,9]. The ABC transporter proteins include (but are not limited to) the P-glycoprotein, the multidrug resistance protein (MRP) 1, MRP2, and the breast cancer resistance protein[8]. For tubulin, it has been proven that the point mutation at the β-tubulin within or near the paclitaxel binding site and the expression of the β-tubulin isotypes, which are less sensitive to taxoid inhibition, usually lead to MDR[10]. For the complexity of the receptor targets relative to MDR, it is difficult to make use of receptor-based methods in exploring MDR problems. Since the last decade, a lot of taxoids have been synthesized, and their cytotoxicity activities to different cell lines have been evaluated, so we can now explore the problem of MDR from the perspective of ligands, that is, exploring the quantitative structure–activity relationship (QSAR) of taxoids and their anti-MDR activities. An important parameter in evaluating the anti-MDR activity of compounds is the resistance index (RI), which is the ratio of IC_{50} of the resistance cell lines to that of the sensitive ones.

Since the last decade, there has been a lot QSAR-based research about taxoids[11-15]. Most of these studies made use of 3-D methods, such as the comparative molecular field analysis (COMFA) or the comparative molecular similarity indices analysis (COMSIA); another character of these researches is that the activity they adopted is IC_{50} of taxoids to inhibit the disassembly of microtubule or growth of tumor cell lines instead of the RI of anti-MDR properties. MDR is a common and serious problem that hinders the application of taxoids; good IC_{50} activity alone can not satisfy the clinical demand. The next generation of taxoids should conquer the problem of MDR. Until now, we have found only 1 article that depicts the QSAR model of the RI. In this study, Monti *et al* adopted multilinear regression (MLR) to mimic the relationship between the RI and the structure of *cis*-platinum complexes. Four descriptors were adopted in their final models, and there are 16 compounds in the whole dataset[16]. As for taxoids, until now, there is no such model to predict the RI, so to obtain the RI, many cytotoxicity evaluation experiments should be conducted. Experimental methods are usually time and money consuming and they are not consistent with the basic drug development strategy of "fail early, fail cheap"[17,18], especially to millions of candidate molecules. So it is necessary for us to build a QSAR model to predict the RI for taxoids.

Molecular descriptors are one of the key factors to a successful QSAR model, and they should encode the most useful physicochemical information on structure features that are relative to the activities to be modeled. Electrotopological state (E-state) indices are widely used in QSAR modeling, including recent cancer-related research[19,20]. The large amount of variables in E-state indices can fully represent the structure characters of molecules, such as information about non-covalent interactions, which may be important to the occurrence of anti-MDR activity. The artificial neural network (ANN), used as a modeling technique, has recently become a popular and powerful chemometric tool[21-23]. Compared with classical statistical methods, ANN-based approaches do not require preliminary knowledge of the mathematical form of the relationship between the variables[24], which makes the ANN suitable for extrapolating the complex and unsure relationships between the biological phenomenon and the structure of the compounds. Several successful QSAR models in our previous studies have proven the feasibility of the combination of the E-state index and the ANN[20,25] to build models.

The purpose of this article was to build a QSAR model combining the E-state indices and the ANN to predict the RI for taxoids. Structure and cytotoxicity data of 63 taxoids, including paclitaxel and docetaxel, were collected from published studies[26-30]. Compared with the RI model of *cis*-platinum complexes, we enlarged the chemical space of our models by collecting 63 compounds synthesized by different laboratories at different times; moreover, more than 4 descriptors were adopted, and the ANN was used as a modeling technique as it does not have to suppose a linear relationship between structure and activity as in MLR. In order to determine the optimal composition of compounds in the training and validation sets, 5-fold cross-validation was performed. The robustness and generalization of our models were still evaluated by an external, independent testing set. The final model was statistically proven to be stable and predictive. This model will aid in filtering drug candidates and accelerate the design and development of new generation anti-MDR taxoids.

Material and methods

Dataset In order to build a reliable QSAR model, 63 taxoids with diverse structures were collected from published studies[26-30], which represented most of the structure modifications since the last decade to improve the clinical perfor-

mance of paclitaxel and docetaxel. According to the modification positions, these compounds are categorized into 6 classes[31], as shown in Figure 2, and the substitution information of the compounds in each class are listed in Table 1. The data about the inhibitory effects (IC_{50}) of these compounds to drug-sensitive human breast carcinoma (MCF-7S) and multidrug-resistant human breast carcinoma (MCF-7R) cell lines were also collected to calculate the RI. Cytotoxicity experiments were conducted following the same *in vitro* assay protocol developed by Skehan *et al*[32].The reason we chose MCF-7(*S* and *R*) cell lines was because they are widely used in biological activity evaluations of taxoids, which will aid in the collection of compounds. All the MCF-7R cell lines were induced by doxorubicin to ensure that they had the same MDR mechanisms. The anti-MDR activity of different taxoids was expressed as a relative value of the RI (taxoid)/RI (paclitaxel), and the values of –log (RI [taxoid]/RI [paclitaxel]) were used for analysis in the back propagation neural network (BPNN) model, which covered a large range, with nearly 3 orders of magnitude from –0.57 to +2.28.

The most reliable way to validate the generalization ability of a model is by external validation[33], that is, to assess the adequacy of the model by the dataset, which is not used in model building. So we randomly selected 14 compounds as an independent external testing set. Five-fold cross-validation was performed on the remaining 49 taxoids to evaluate the internal stability of models and to optimize the composition of compounds in the training and validation sets, so 49 compounds were randomly categorized into 5 groups. One group was selected as the validation set each time, and the remaining 4 groups as the training set; 5 different training and validation datasets could be used to construct different models[19]. The detailed grouping information of the datasets for each model together with the activities for each compound was given as supporting information.

Descriptor generation We used the Molconn-z program in the SYBYL software package (Tripos Associates, St Loius, MO, USA) to calculate molecular structure descriptors known as E-state indices, whose availability has been proven in a lot of QSAR models[20,34]. In total, 248 standard descriptors were calculated included in the molecular connectivity Chi indices, Kappa shape indices, E-state indices, hydrogen E-state indices, atom type and bond type E-state indices, topological equivalence indices and total topological index, counts of graph paths, atoms, atom types, bond types, and so on (Molconn-Z manual), which can sufficiently represent the structural characters of molecules. The E-state indices are shown to contain information reflecting the intermolecular accessibility of atoms and groups in a molecule, especially the electron accessibility, which is encoded into a numerical value. The advantage of these kinds of descriptors is that they encode not only the topological environment of an atom, but also the electronic interactions from other at-

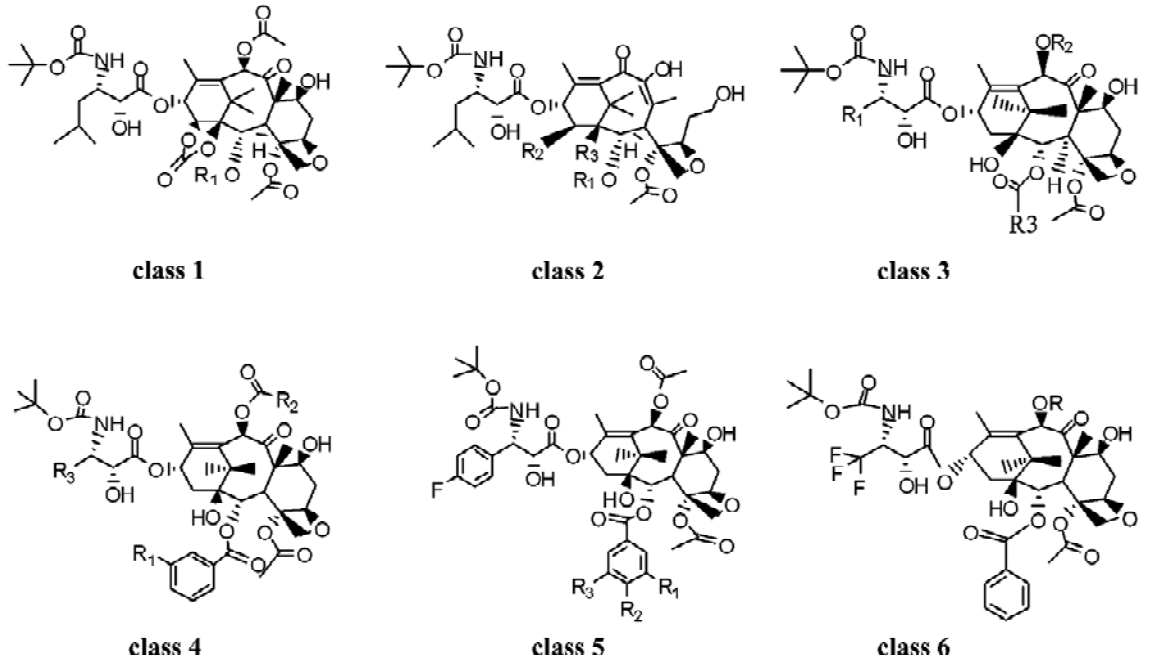


Figure 2. Structure of 6 classes of taxoids.

Table 1. Substituent information for all 63 taxoids.

Name[a]	R_1	R_2	R_3
Paclitaxel	Ph	Ac	
Docetaxel	tBuO	H	
		Class 1	
IDN5109	Bz		
MEO/IDN5109	m-MeOBz		
		Class 2	
IDN5390	Bz	H	OH
MEO/IDN5390	m-MeOBz		1,14-carbonate
		Class 3(1)	
4a	$(CH_3)_2C=CH$	CH_3CH_2-CO	Ph
4b	$(CH_3)_2C=CH$	Cyclopropane-CO	Ph
4c	$(CH_3)_2C=CH$	$(CH_3)_2N$-CO	Ph
4d	$(CH_3)_2C=CH$	CH_3O-CO	Ph
4e	$(CH_3)_2C=CH$	CH_3	Ph
4f	$(CH_3)_2C=CH$	$CH_3(CH_2)_3$-CO	Ph
4g	$(CH_3)_2C=CH$	$CH_3(CH_2)_4$-CO	Ph
4h	$(CH_3)_2C=CH$	$(CH_3)_2CHCH_2$-CO	Ph
4i	$(CH_3)_2C=CH$	$(CH_3)_3CCH_2$-CO	Ph
4j	$(CH_3)_2C=CH$	Cyclohexane-CO	Ph
4k	$(CH_3)_2C=CH$	$CH_3CH=CH$-CO	Ph
4l	$(CH_3)_2C=CH$	$(CH_3CH_2)_2N$-CO	Ph
4m	$(CH_3)_2C=CH$	Morpholine-4-CO	Ph
4n	$(CH_3)_2C=CH$	CH_3NH-CO	Ph
4o	$(CH_3)_2C=CH$	CH_3CH_2NH-CO	Ph
4p	$(CH_3)_2C=CH$	$CH_3CH_2CH_2NH$-CO	Ph
4q	$(CH_3)_2C=CH$	$(CH_3)_2CHNH$-CO	Ph
4r	$(CH_3)_2C=CH$	$CH_2=CHCH_2NH$-CO	Ph
4s	$(CH_3)_2C=CH$	Cyclohexyl-NH-CO	Ph
5a	$(CH_3)_2CH$-CH_2	CH_3CH_2-CO	Ph
5b	$(CH_3)_2CH$-CH_2	Cyclopropane-CO	Ph
5c	$(CH_3)_2CH$-CH_2	$(CH_3)_2N$-CO	Ph
5d	$(CH_3)_2CH$-CH_2	CH_3O-CO	Ph
5e	$(CH_3)_2CH$-CH_2	CH_3	Ph
5s	$(CH_3)_2CH$-CH_2	Cyclohexyl-NH-CO	Ph
sb-t-1102	$(CH_3)_2CH$-CH_2	CH_3CO	Ph
sb-t-1212	$(CH_3)_2C=CH$	CH_3CO	Ph
		Class 3 (2)	
7	Ph	Ac	Cyclohexyl
8	Cyclohexyl	Ac	$(CH_3)_2C=CH$
9	$(CH_3)_2C=CH$	Ac	Cyclohexyl
10	$(CH_3)_2C=CH$	Ac	$(CH_3)_2C=CH$
11	$(CH_3)_2CH$-CH_2	Ac	$(CH_3)_2CH$-CH_2
13	$(CH_3)_2C=CH$	Ac	$(CH_3)_2CH$-CH_2
14	$(CH_3)_2CH$-CH_2	Ac	Cyclohexyl
15	$CH_3CH=CH$	Ac	Cyclohexyl
16	$CH_3CH=CH$	Ac	Cyclohexyl
		Class 4	
7e	MeO	Et	$CH_2CH(CH_3)_2$
7f	MeO	Et	$CH_2CH(CH_3)_2$
7g	MeO	Et	CF_2H
7h	MeO	Et	$CH_2CH_2CH=CH_2$
7i	MeO	Et	$CH_2CH=CH_2$
7j	MeO	Et	(S)-2,2-Dimethyl-cyclopropyl
7l	MeO	Me	$CH=C(CH_3)_2$
7n	N_3	Et	$CH_2CH(CH_3)_2$
7o	N_3	Et	$CH=C(CH_3)_2$
7q	Me	Me	$CH=C(CH_3)_2$
		Class 5	
8c	H	F	H
8f	F	H	F
		Class 6	
11a	H		
11b	Ac		
11c	Me_2N-CO		
11d	Cyclopropane-CO		
11e	MeO-CO		
11f	Morpholine-4-CO		
11g	Et-CO		
11h	$CH_3(CH_2)_3$-CO		
11i	$(CH_3)_3CCH_2$-CO		

[a]Name or number of compounds in References.

oms in the molecules, as depicted in its formula[35]:

$S_i = I_i + \delta I_i$ (1),

where S_i is the E-state of atom i, I_i is the intrinsic state, and dI_i is the perturbations due to the atoms around it. Moreover, most of the descriptors have been proven to be well associated with non-covalent interactions, which are important for bioactivity[36]. Thus, E-state indices can represent the structure information, which may also be relative to the anti-MDR properties for taxoids.

Feature reduction Not all of the 248 descriptors contribute to the bioactivity; some measures were taken to eliminate the noise (uninformative descriptors): eliminating the descriptors with constant values and more than 90% zero values because they offered little discriminating information for the construction of model. After this procedure, 84 descriptors remained, as shown in Table 2. In order to further reduce the variable space and the chance of correlation between the descriptors, a principle component analysis (PCA) was performed on the remaining 84 variables. The 11 derived principle component vectors (PC) were used for model building. The calculation of PCA was done by free data mining software, Tanagra 1.1 (http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html).

ANN In order to build reliable and predictive QSAR models, we adopted the ANN technique, which has been proven to have outstanding non-linear approximation ability[22,23,37]. A typical ANN consists of an input layer, a hidden layer, and an output layer. In the ANN, signals are propagated from the input neurons through the hidden layer to the output neuron, and then the error is calculated and back propagated to iteratively adjust weights and biases in order to minimize the error in prediction; this is the most distinct character of typical back propagation (BP) algorithm.

The ANN program used was the neural network software package of MATLAB 7.0.1 developed by Math Works (Natick, MA, USA). Some fully connected 3-layer BP neural networks with sigmoid transfer function were constructed. The number of neurons in the input layer equaled the number of PC. Before the net training, all of the input and output values were normalized to between –1 and 1, and the outputs were transferred back to the same units as the original outputs for comparison purpose. The Levenberg–Marquardt algorithm was adopted to optimize weights and biases because it was significantly faster than other algorithms based on gradient descent[38]. In each of the 5 different datasets, the training sets were used to determine the architecture of the ANN model; the validation sets were adopted to tune the ANN parameters to prevent overtraining[39], and the independent external testing set was used to evaluate the predictive ability of the models. In order to determine the optimal number of neurons in the hidden layer, we adopted some empirical rules. For example, the number of neurons in the hidden layer can be confirmed by the formula: $m=\log 2\, n+\alpha$, where m is the number of neurons in hidden layer, n is the number of input variables, and α is the integer between 0 and 10[40–42]. The early-stopping method was adopted to help prevent overtraining. For the 5 datasets with different compounds in the training and validation sets, we trained the models separately.

Model evaluation The following parameters were calculated to evaluate the performance of the ANN and the predictive ability of the model: Q^2_{cv} (cross-validation correlation coefficient), RMSE (residual mean square error), R^2 (square correlation coefficients for the regression line for calculated and experimentally-derived activity of the external testing set), R_0^2 (square correlation coefficients for the regression line through the origin for calculated and experimentally-derived activity of the external testing set), and K (the slope of regression line through the origin for testing sets). The residuals between the predicted and experimentally-derived activities were also calculated for the best model. The definitions of Q^2_{cv}[43] and RMSE[33] are listed below:

$$Q^2_{cv} = 1 - PRESS / SD \quad (2)$$

$$S^2_{res}(RMSE) = \frac{\Sigma(\tilde{y}_i - y_i^r)^2}{n-2} \quad (3),$$

where PRESS is the sum of squared deviations between the predicted and measured biological activity values for each compound in the validation set, and SD is the sum of the squared deviations between the measured activities of the compounds in the validation set and the mean activity of the training set compounds. $\tilde{y}_i$ and y_i are the predicted and actual activities, respectively, and y_i^r corresponds to the equation of regression $y_i^r = a\tilde{y}_i + b$. The propositional criteria necessary for the high predictive ability of a model are high Q^2_{cv} (at least >0.5), high R^2 for the external testing set (at least >0.6), $(R^2-R_0^2)/R^2<0.1$, and $0.85 \le K \le 1.15$[33,47].

Results

Molecular descriptors The remaining 84 molecular descriptors after the feature reduction were compressed and analyzed by PCA, resulting in 11 PC for network building. The number of components was determined by the maximum variance described by the PC and the eigenvalues. Eleven PC were sufficient to explain nearly 95% of the variance, and all of their eigenvalues were greater than 1. The coefficients of variables to each PC are described in Table 3. PC1 and PC2 explained 23% and 19% of the total variance, respectively. In each, the molecular connectivity and mo-

Table 2. Electrotopological state indices used in this work.

Variable	Definition	Variable	Definition
Xv0, Xv1, Xv2,	Valence Chi indices:	nsCH3	Number of group atom: $-CH_3$
Xvp3, Xvp4, Xvp5,	Based on δ^v, connection matrix,	nssCH2	$-CH_2-$
Xvp6, Xvp7, Xvp8,	atom type, and count of Hs bonded	ndsCH	=CH-
Xvp9,Xvp10,	to each atom	naaCH	:CH:
Xvc3, Xvc4	Connectivity valence cluster indices	nsssCH	>CH
Xvpc4	simple path/cluster index	ndssC	-=C<
Xvch6	Valence chain indices	naasC	:C:-
ka1,ka2,ka3	Kappa alpha shape indices	nssssC	>C<
phia	Flexibility index	nssNH	-NH-
sumdelI	Sum of delta-I values	ndO	=O
sumI	Sum of intrinsic state values	nssO	-O-
Qv	General polarity descriptor.	naOm	$:O^{-0.5}$
	Extreme atom level E-state value	nsF	-F
	in molecule:	nHssNH	Number of Hs on: -NH-
Hmax	Maximum H E-state	nHdsCH	=CH-
Gmax	Maximum E-state	nHaaCH	:CH:
Hmin	Minimum H E-state	nHCsats	CHn (saturated)
Gmin	Minimum E-state	nHCsatu	CHn (unsaturated)
nvx	Number of non hydrogen atoms	ntrifluoromethyl	Number of group:CF_3
nedges	Number of edges(bonds)	ncarbamate	C(=O)OR
nrings	Number of rings in molecule	ncarboxylate	COO
nHBd,	Number of strong H-bond donors,		Sum of E-states for this
nHBa	Number of strong H-bond acceptors	Strifluoromethyl	type of group: CF_3
nwHBa	Number of weak H-bond acceptors	Sketone	RC(=O)R
	Sum of E-states value for	Scarbamate	NC(=O)OR
SHBd,	H-bond donors,	Scarboxylate	COO
SHBa,	H-bond acceptors		
SwHBa	weak H-bond acceptors		Sum of atom type E-states
SHBint3	Sum of E-state descriptors of strength	SsCH3, $SssCH_2$,	$-CH_3$, $-CH_2-$,
SHBint4	for potential Internal H bonds. Internal	SdsCH, SaaCH,	=CH-, :CH:,
SHBint5	hydrogen bond descriptor is the product	SsssCH, SdssC,	>CH-, =C<
SHBint6	of H E-state value and E-state value.	SaasC, SssssC,	:C:-, >C<,
	Sum of H E-states for atom type:	SssNH,SsOH,SdO	-NH-, -OH, =O,
SHsOH, SHssNH,	Hs on –OH, -NH-,	SssO SaOm SsF	-O-, $:O^{-0.5}$, -F
SHdsCH, SHaaCH	=CH-, :CH:		
SHCsats	Sum of H E-states for Hs on	SHCsatu	Sum of H E-states for
	C sp3 bonded to saturated C		Hs on CHn(unsaturated)

lecular shape indices played important parts. The PC2 mainly consists of the E-state descriptors, which encode the topological and the electronic information about each atom and the interaction deriving from the environment. PC3, with 10% of the variances explained, represents the information of the H-bond interaction derived from the information about the H-bond donor and acceptors. PC4 was dominated by the information about the atom type aaCH:, that is, :CH:, including the number of atoms of this kind, number of H on these atoms, and the total E-state values and HE-state values, and it encodes 9.23% of the variance. The most important descriptor in PC5 is the ndssC, which counts the number of atoms of this kind =C<. Interestingly, the atom O descriptor also accounts for a large part in PC5, which totally depicted 8.4% of the total variance. Although only 6.6% of the variance was explained, PC6 contained important descriptors, mainly about the atom N, such as NH– and the group NC (=O) OR. The remaining 5 PC can contribute to 16.6% of the total variance and each one was dominated by important descriptors.

QSAR modeling As for the 5 different training and validation sets, 5 QSAR models were built separately. Eleven

Table 3. Component score coefficient matrix of 84 descriptors to 11 PC[a].

Descriptor	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Xv0	**0.040**	0.029	0.015	0.004	0.042	-0.013	-0.035	0.055	0.052	0.021	0.026
Xv1	0.039	0.034	0.011	0.015	0.000	0.005	-0.032	0.013	0.086	0.039	0.059
Xv2	0.039	0.032	-0.008	-0.013	-0.002	0.009	-0.004	0.085	0.038	-0.004	0.178
Xvp3	0.030	0.042	0.005	0.022	-0.051	0.002	-0.001	-0.019	0.054	0.011	0.049
Xvp4	0.034	0.036	-0.011	0.027	-0.049	0.004	-0.001	-0.047	0.016	-0.013	0.114
Xvp5	0.032	0.036	-0.006	0.022	-0.060	-0.005	0.018	-0.055	-0.006	-0.035	0.038
Xvp6	0.031	0.037	-0.010	0.017	-0.032	-0.040	0.045	-0.021	-0.100	-0.064	-0.065
Xvp7	0.023	0.045	-0.002	0.006	-0.035	-0.034	0.060	-0.016	-0.090	-0.014	-0.094
Xvp8	0.032	0.039	-0.007	0.003	-0.024	-0.023	0.059	0.004	-0.095	0.004	-0.022
Xvp9	0.028	0.042	-0.001	0.000	-0.034	-0.026	0.067	-0.022	-0.069	0.002	-0.072
Xvp10	0.026	0.043	0.000	-0.004	-0.035	-0.027	0.070	-0.027	-0.061	-0.006	-0.092
Xvc3	0.016	0.014	-0.019	-0.055	0.032	-0.008	0.028	***0.196***	-0.053	-0.048	**0.260**
Xvc4	0.002	0.020	-0.003	-0.053	0.022	-0.011	0.048	**0.180**	-0.014	-0.005	***0.356***
Xvpc4	0.017	0.038	0.001	-0.016	-0.032	-0.028	0.056	**0.122**	-0.107	-0.051	-0.003
Xvch6	0.019	0.028	0.005	0.012	**-0.078**	0.037	0.022	-0.113	0.099	0.002	0.137
ka1	0.023	0.036	0.025	-0.026	0.068	-0.019	-0.051	0.036	0.076	0.016	-0.023
ka2	0.032	0.019	0.034	-0.004	0.046	-0.006	-0.077	-0.014	**0.127**	0.056	-0.067
ka3	0.027	0.015	0.028	-0.045	0.062	0.007	-0.054	0.042	**0.146**	0.026	0.027
phia	0.032	0.010	0.025	-0.036	0.055	-0.004	-0.072	0.000	**0.130**	0.052	-0.098
nvx	0.018	**0.046**	0.031	0.028	0.044	-0.022	-0.044	0.014	0.047	0.011	0.038
nedges	0.017	**0.048**	0.030	0.041	0.027	-0.019	-0.038	0.005	0.037	0.010	0.063
nrings	0.005	0.038	0.013	0.078	-0.047	-0.004	-0.002	-0.032	-0.018	0.003	0.136
sumdelI	-0.026	***0.049***	0.003	-0.014	0.042	-0.012	0.001	-0.020	0.010	-0.056	-0.009
sumI	-0.024	**0.047**	0.019	-0.002	0.053	-0.021	-0.019	-0.017	0.022	-0.046	0.007
Qv	***0.044***	-0.029	-0.008	-0.007	-0.011	0.007	-0.002	0.064	0.001	0.041	-0.006
nHBd	0.003	-0.003	**0.077**	-0.044	-0.032	0.061	-0.019	-0.020	-0.086	-0.056	0.070
nHBa	-0.027	0.041	-0.002	-0.022	0.059	-0.003	-0.034	-0.048	-0.036	-0.029	-0.076
nwHBa	-0.006	0.000	0.061	**0.092**	0.031	-0.027	-0.007	0.032	0.009	-0.006	0.031
SHBd	-0.006	0.003	***0.078***	-0.048	-0.039	0.048	0.001	-0.010	-0.062	-0.048	0.052
SHBa	-0.033	0.043	-0.003	-0.018	0.039	-0.005	0.002	-0.034	0.015	-0.064	-0.011
SwHBa	0.005	-0.014	**0.064**	0.064	-0.019	-0.014	0.004	0.040	-0.042	**0.218**	-0.003
Hmax	-0.019	0.027	**0.066**	-0.036	0.005	-0.019	0.019	0.034	0.067	-0.035	**-0.180**
Gmax	0.030	0.041	0.026	-0.003	0.026	-0.018	0.035	0.033	-0.033	0.072	-0.157
Hmin	-0.029	0.005	0.033	0.037	0.003	-0.044	0.030	-0.043	-0.105	0.001	-0.073
Gmin	**0.040**	-0.029	0.003	0.036	-0.014	-0.021	-0.027	0.019	-0.013	-0.117	-0.031
SHBint3	-0.016	**0.047**	0.018	0.042	0.029	-0.024	0.003	-0.039	-0.004	-0.018	0.166
SHBint4	-0.040	0.031	-0.001	-0.029	0.014	0.015	0.027	-0.027	0.011	0.099	0.020
SHBint5	0.000	0.018	**0.067**	-0.040	-0.034	-0.011	0.033	0.053	0.116	-0.018	**-0.210**
SHBint6	0.001	0.000	0.060	-0.047	–0.051	0.009	0.062	0.064	0.115	-0.054	**-0.198**
nHssNH	0.016	0.004	0.045	-0.006	0.028	**0.123**	-0.029	-0.049	**-0.125**	-0.052	0.066
nHdsCH	0.022	-0.024	0.034	-0.015	0.053	-0.011	**0.098**	-0.072	0.018	0.043	-0.004
nHaaCH	-0.016	0.008	0.043	***0.099***	-0.008	0.020	0.003	0.079	0.011	0.058	0.020
nHCsats	0.023	0.023	-0.051	-0.037	-0.045	0.047	-0.066	0.033	0.042	0.042	-0.059
nHCsatu	0.026	-0.024	0.026	-0.014	0.036	-0.054	0.089	-0.074	0.029	-0.044	0.125
nsCH3	0.027	-0.021	-0.025	-0.042	0.061	-0.018	0.005	**0.122**	-0.071	-0.042	-0.047
nssCH2	0.025	0.022	-0.016	-0.033	-0.061	0.030	-0.022	-0.095	***0.174***	0.066	0.058
ndsCH	0.022	-0.024	0.034	-0.015	0.053	-0.011	**0.098**	-0.072	0.018	0.043	-0.004
naaCH	-0.016	0.008	0.043	***0.099***	-0.008	0.020	0.003	0.079	0.011	0.058	0.020
nsssCH	0.023	0.020	-0.054	-0.004	-0.050	0.021	-0.016	-0.004	-0.121	-0.039	-0.065
ndssC	0.022	-0.018	-0.003	0.017	***0.100***	0.003	0.027	-0.055	0.045	0.033	0.118
naasC	-0.012	0.012	0.025	0.078	0.018	-0.038	-0.028	0.056	0.052	**-0.264**	-0.029
nssssC	-0.022	0.036	-0.002	-0.060	0.011	0.005	0.067	0.086	-0.045	0.063	0.154

Continue

Descriptor	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
nssNH	0.016	0.004	0.045	-0.006	0.028	**0.123**	-0.029	-0.049	**-0.125**	-0.052	0.066
ndO	0.019	0.031	-0.029	0.022	**0.069**	-0.038	0.010	-0.060	-0.102	0.000	-0.119
nssO	0.001	0.001	0.029	-0.035	-0.003	-0.118	**-0.109**	-0.040	-0.080	0.035	0.024
naOm	0.003	0.006	-0.060	0.048	0.051	0.086	0.051	0.020	0.049	-0.006	-0.065
nsF	**-0.041**	0.032	-0.001	-0.022	0.015	0.013	0.025	-0.018	0.045	-0.069	0.018
SHsOH	-0.017	-0.003	0.054	-0.039	**-0.071**	0.025	0.064	0.056	0.049	-0.039	-0.031
SHssNH	0.010	0.008	0.047	-0.008	0.031	**0.125**	-0.024	-0.055	-0.122	-0.045	0.078
SHdsCH	0.023	-0.024	0.034	-0.015	0.054	-0.012	**0.099**	-0.070	0.016	0.042	0.007
SHaaCH	-0.017	0.009	0.043	**0.099**	-0.004	0.014	-0.001	0.079	0.016	0.021	0.016
SHCsats	0.013	0.035	-0.049	-0.037	-0.035	0.038	-0.069	0.028	0.003	0.069	-0.114
SHCsatu	0.023	-0.017	0.030	-0.007	0.040	-0.062	0.090	-0.089	0.036	-0.063	0.156
SsCH3	0.034	-0.024	-0.020	-0.034	0.046	-0.011	-0.003	**0.127**	-0.051	-0.017	-0.017
SssCH2	0.030	0.015	-0.022	-0.009	-0.066	0.044	0.007	-0.121	0.103	0.010	0.167
SdsCH	0.023	-0.024	0.035	-0.016	0.053	-0.010	0.097	-0.072	0.020	0.043	-0.006
SaaCH	-0.013	0.004	0.042	**0.096**	-0.014	0.026	0.006	0.079	-0.006	**0.155**	0.023
SsssCH	0.030	-0.041	0.007	0.019	-0.021	0.003	-0.036	0.041	0.072	-0.028	0.043
SdssC	0.035	-0.033	0.025	-0.005	-0.048	-0.016	0.029	0.035	0.005	-0.002	0.008
SaasC	0.009	-0.015	-0.006	-0.006	-0.016	0.010	0.003	0.009	-0.083	***0.458***	-0.005
SssssC	0.036	-0.034	0.000	0.040	-0.016	-0.016	-0.041	0.004	0.011	-0.098	-0.040
SssNH	0.031	-0.006	0.039	0.012	0.013	0.100	-0.040	-0.031	-0.106	-0.079	0.058
SsOH	-0.007	-0.005	0.056	-0.036	**-0.078**	0.023	0.068	0.066	0.049	-0.048	-0.053
SdO	0.024	0.033	-0.023	0.027	0.066	-0.036	0.002	-0.046	-0.089	0.018	-0.097
SssO	0.001	0.000	0.030	-0.036	-0.005	-0.118	***-0.109***	-0.041	-0.083	0.035	0.032
SaOm	0.005	0.001	-0.061	0.053	0.053	0.085	0.028	0.016	0.071	0.004	-0.016
SsF	**-0.041**	0.032	-0.001	-0.022	0.015	0.013	0.025	-0.018	0.045	-0.068	0.018
ntrifluoromethyl	-0.039	0.028	-0.005	-0.036	0.011	0.025	0.030	-0.019	0.012	0.136	0.062
ncarbamate	0.010	0.010	0.015	-0.014	0.043	***0.137***	0.006	0.020	-0.016	0.003	-0.160
ncarboxylate	0.002	0.007	-0.064	0.051	0.018	0.025	**0.099**	0.032	0.066	-0.033	-0.047
Strifluoromethyl	-0.039	0.028	-0.005	-0.036	0.011	0.025	0.030	-0.018	0.013	**0.137**	0.062
Sketone	0.037	0.031	0.022	-0.001	0.008	-0.020	0.038	0.042	-0.036	0.079	-0.150
Scarbamate	0.015	0.008	0.017	-0.010	0.040	**0.136**	0.000	0.020	-0.019	-0.008	-0.149
Scarboxylate	0.005	0.006	-0.063	0.053	0.017	0.025	0.097	0.036	0.069	-0.026	-0.039
Eigenvalue	19.318	15.936	10.073	7.752	7.022	5.516	5.101	3.228	2.488	1.828	1.261
%VE[b]	22.998	18.971	11.992	9.228	8.359	6.566	6.073	3.843	2.962	2.176	1.502
TVE[c]	22.998	41.969	53.961	63.190	71.549	78.115	84.188	88.031	90.993	93.169	94.670

[a]Important descriptors in each PC are in bold with the most important ones in italics. [b]Percentage of variance explained. [c]Total percentage of variance explained.

PC served as input variables for each model. There are no rigorous theoretical principles for determining the structure for ANN, so different numbers of neurons in the hidden layer and various numbers of epochs were tried in order to prevent overfitting and overtraining. As weights and biases are optimized by the back propagation iterative procedure, training errors typically decrease, but validation errors first decrease and subsequently begin to rise, revealing a progressive worsening of the generalization ability of the network. Thus, when RMSE (transferred back) for training and validation sets both reached comparatively small values, the optimized number of neurons and epochs was confirmed. After the structure of the ANN was chosen, repeated training was done to optimize the weights and biases to find the best predictive models. The architecture of each model and the results of the cross-validation Q^2_{cv} and $RMSE_{(T, V)}$ are summarized in Table 4.

Model evaluation The external independent testing set composing of 14 compounds was used to evaluate the predictive ability of the 5 models with the results shown in Table 5. Although the Q^2_{cv} values of model 3 was >0.5, and both the values of $RMSE_T$ and $RMSE_V$ were less (0.003), the generalization ability of this model is poor, as demonstrated by the results of R_0^2 and the values of $(R^2–R_0^2)/R^2$. The statisti-

Table 4. Statistical results of 5-fold cross-validation.

Model	Neuron	Q^2_{cv}	$RMSE_T$[a]	$RMSE_V$
1	7	0.57	0.018	0.064
2	7	0.620	0.002	0.022
3	4	0.514	0.003	0.003
4	8	0.553	0.0002	0.049
5	5	0.562	0.007	0.051

[a] T, training sets; V, validation sets.

Table 5. Results for external testing set of each model.

Model	R^2	R_0^2	$(R^2–R_0^2)/R^2$	$RMSE_P$[b]	K
1	0.832	0.817	0.018	0.007	0.9746
2[a]	0.840	0.810	0.036	0.0135	0.9933
3	0.695	0.410	0.410	0.144	0.604
4	0.700	0.425	0.393	0.124	0.9613
5	0.795	0.794	0.001	0.001	0.9677

[a]Statistical results of the best model are in bold. [b]P, prediction set.

cal results of model 4 also did not satisfy the criteria for a good model. Although the evaluation results of models 1, 2, and 5 all satisfied the referred criteria necessary for predictive models, we selected model 2 as our final model as it had the highest value of Q^2_{cv} and R^2, allowing us to determine the most stable and predictive model for the RI.

The residuals between the predicted and experimentally-derived activities for compounds in the training, validation, and prediction sets by model 2 are shown in Table 1. We can see that the activities of all 63 taxoids were predicted within 1.007 log units of their experimentally-derived activities with an average absolute error of 0.213 log units. The predictive results of all 63 compounds are plotted in Figure 3. The statistical results of the testing set found that the greatest deviation was 0.54 log units with an average absolute error of 0.226 log units. The predicted results are also plotted in Figure 4.

Discussion

A successful descriptor should represent the key structure information of molecules, influences activity, and then can be useful in the prediction of activity for unknown compounds. According to some structure activity studies[26–30], substitution by definite atoms or groups can influence anti-MDR activity; for example, F-substituted taxoids at different positions usually alter the anti-MDR activity differently, and the –OH and groups including N atoms also play an important part in the change of activity. As discussed earlier, the E-state indices had fully encoded these kinds of structure information; for example, F, N, =C< and :CH: descriptors were all embodied in different PC. Moreover, the reported mechanisms about MDR of taxoids are relative to ABC transporter proteins and tubulin[10]. As for ABC transporter proteins, intermolecular H bonds are key factors for the recognition of taxoids by those proteins[25]. For tubulin, it has been proven that specific conformation, such as the T- taxol for taxoids, should be maintained, and taxoids can act on some definite isotypes of tubulin, which are also relative to the non-covalent interaction intra or inter molecules[44–46]. So maybe the anti-MDR activities of taxoids have some relationship to non-covalent interactions. Topological-based E-state indices comprised H-bond descriptors for inter and intra molecules,

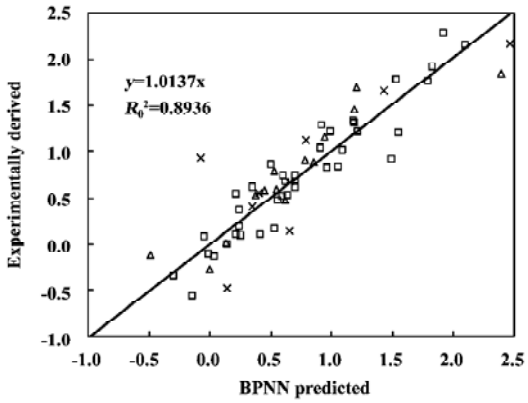


Figure 3. Plot of predicted –log (activity) values versus experimentally-derived ones for all 63 taxoids. (□) training set; (×) validation set; (△) testing set.

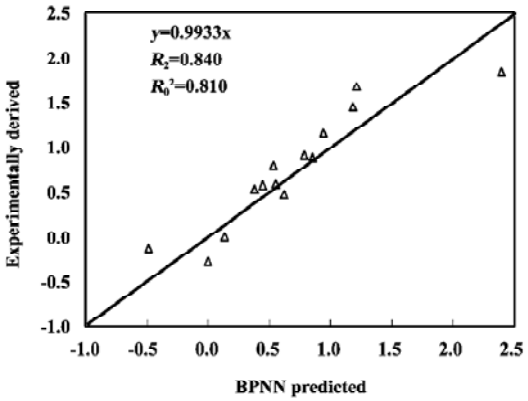


Figure 4. Plot of predicted versus experimentally-derived –log (activity) for testing set.

Table 6. Predicted activities and residual information of taxoids[a].

Name[b]	BPNN (activity)[c]	Exp (activity)[d]	Residuals[e]
	Training set		
Paclitaxel	0.141	0.000	−0.141
IDN5390	1.088	1.021	−0.068
MEO/IDN5109	1.790	1.759	−0.032
MEO/IDN5390	0.702	0.609	−0.093
4b1214	0.984	1.224	0.240
4d	0.698	0.667	−0.031
4f	0.554	0.467	−0.087
4g	0.348	0.609	0.261
4h	1.488	0.918	−0.570
4i	0.965	0.826	−0.139
4m	0.410	0.103	−0.307
4n	0.217	0.546	0.329
4p	0.037	−0.138	−0.175
4q	0.246	0.095	−0.151
4s	−0.045	0.082	0.128
5a	1.179	1.342	0.163
5b	1.184	1.319	0.135
5c	0.908	1.038	0.130
5e	−0.142	−0.567	−0.424
sb-t-1102	0.918	1.291	0.373
7	−0.013	−0.114	−0.102
10	0.239	0.364	0.126
11	0.591	0.516	−0.074
13	1.054	0.845	−0.209
15	0.601	0.745	0.143
16	0.216	0.101	−0.115
7f	1.919	2.284	0.365
7g	1.209	1.215	0.006
7i	1.551	1.210	−0.340
7l	1.525	1.788	0.263
7n	1.828	1.924	0.097
7o	2.099	2.160	0.060
8c	0.211	0.099	−0.112
8f	0.243	0.187	−0.056
11a	−0.299	−0.346	−0.047
11d	0.702	0.740	0.038
11e	0.635	0.525	−0.110
11f	0.532	0.166	−0.366
11h	0.619	0.675	0.056
11i	0.498	0.865	0.367
	Validation set		
4c	0.655	0.669	0.014
4j	0.406	0.560	0.154
4k	0.794	1.129	0.334
4o	0.140	−0.473	−0.613
5s	−0.076	0.931	1.007
8	0.653	0.140	−0.513
7e	2.461	2.170	−0.291
7h	1.428	1.661	0.233
11c	0.353	0.400	0.047
	Test set		
Docetaxel	−0.484	−0.126	0.358
IDN5109	1.206	1.688	0.482
4a1213	0.943	1.158	0.215
4e	−0.004	−0.276	−0.273
4l	0.548	0.587	0.039
4r	0.137	0.008	−0.129
5d	0.852	0.886	0.034
sb-t-1212	0.780	0.906	0.126
9	0.623	0.474	−0.149
14	0.372	0.529	0.157
7j	2.393	1.849	−0.545
7q	1.185	1.451	0.266
11b	0.534	0.793	0.259
11g	0.446	0.576	0.130

[a]Statistical results and the compound subset information are only about model 2. [b]Name or number of compounds in References. [c]Activities predicted by BPNN model 2, which were expressed as –log (RI[taxoids]/RI[paclitaxel]). [d]Activities derived from experimental data. [e]Residuals which equal to Exp(activity)–BPNN (activity).

which represented the non-covalent interactions. According to the above analysis, we can see that E-state indices can represent important attributes of molecular structure, especially those associated with the interaction between taxoids and receptors. So it seems reasonable for us to choose E-state indices as our descriptors for exploring the relationship between the RI and the structure.

As for the statistical results of the 5 ANN models, although each model was with the good internal cross-validation results(Q^2_{cv}>0.5), we can't conclude that all of them have good generalization abilities. The results of model 3 indirectly indicated that only the independent external testing rather than the internal validation could evaluate the predictive ability of a model. The results of 5-fold cross-validation and external testing also ensured that the compound composition of the training and validation sets had important influence on the architecture and performance of models, especially on the predictive ability for the external testing sets. Five-fold cross-validation could help us to find out the optimal combination of compounds that may be useful for obtaining the most predictive model.

According to the results of model 2, in Figure 3, all of the samples distributed closely around the line, and the value of R_0^2 was 0.8936, together with the K (the slope of regression line through the origin) was 1.0137, which further proved

the closeness of the predicted and experimentally-derived activity. The results also indicated that the E-state indices did correlate well with the –log (RI/P). The statistical results of the testing set further confirmed the predictive ability of this model.

As for the complexity of the receptor proteins associated with MDR, we derived a ligand-based QSAR model to predict the values of the RI for different taxoids. E-state indices were used to represent the structure of molecules; BPNN was used to explore the relationship between descriptors and RI activity. During the construction of the models, 5-fold cross-validation was performed to determine the best composition of compounds in the training and validation sets. The predictive ability of the models was also evaluated by an independent testing set. The best model had the statistical results of R^2=0.84, R_0^2=0.835, K=0.9933, and $RMSE_P$=0.014, indicating the excellent robustness and generalization of our model. The results also proved that E-state indices have some relationship to anti-MDR activity, and the BPNN modeling technique can fully emulate this kind of non-linear relationship. Our model can predict the values of the RI for taxoids just from its structure even before it was synthesized, so it will aid in the filter of anti-MDR drug candidates and accelerate the design and development of taxoids with good clinical performance to drug resistance cell lines.

References

1 Wani MC, Taylor HL, Wall ME, Coggon P, McPhail AT. Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. J Am Chem Soc 1971; 93: 2325–7.

2 Gueritte-Voegelein F, Guenard D, Mangatal L, Potier P, Guilhem J, Cesario M, *et al*. Structure of a synthetic taxol precursor: N-tert-butoxycarbonyl-10-deacetyl-N-debenzoyltaxol. Acta Crystallogr C 1990; 46: 781–4.

3 Kingston DGI. Recent advances in the chemistry of taxol. J Nat Prod 2000; 63: 726–34.

4 Miller ML, Ojima I. Chemistry and chemical biology of taxane anticancer agents. Chem Rec 2001; 1: 195–211.

5 Kingston DGI, Newman DJ. Taxoids: cancer-fighting compounds from nature. Curr Opin Drug Discov Devel 2007; 10: 130–44.

6 Edwards P. Peptoid positional scanning libraries for identification of multidrug resistance reversal agents. Drug Discov Today 2006; 11: 669–70.

7 Burchenal JH, Holmberg EA. The utility of resistant leukaemias in screening for chemotherapeutic activity. Ann N Y Acad Sci 1958; 76: 826–9.

8 Leslie EM, Deeley RG, Cole SPC. Multidrug resistance proteins: role of P-glycoprotein, MRP1, MRP2, and BCRP (ABCG2) in tissue defense. Toxicol Appl Pharmacol 2005; 204: 216–37.

9 Orr GA, Verdier-Pinard P, McDaid H, Horwitz SB. Mechanisms of taxol resistance related to microtubules. Oncogene 2003; 22: 7280–95.

10 Ojima I, Ferlini C. New insights into drug resistance in cancer. Chem Biol 2003; 10: 583–4.

11 Cunningham SL, Cunningham AR, Day BW. CoMFA, HQSAR and molecular docking studies of butitaxel analogues with beta-tubulin. J Mol Model 2005; 11: 48–54.

12 Czaplinski KHA, Grunewald GL. A comparative molecular-field analysis derived model of the binding of taxol analogs to microtubules. Bioorg Med Chem Lett 1994; 4: 2211–6.

13 Mohanraj S, Doble M. 3-d QSAR studies of microtubule stabilizing antimitotic agents towards six cancer cell lines. QSAR Comb Sci 2006; 25: 952–60.

14 Pineda O, Farras J, Maccari L, Manetti F, Botta M, Vilarrasa J. Computational comparison of microtubule-stabilising agents laulimalide and peloruside with taxol and colchicine. Bioorg Med Chem Lett 2004; 14: 4825–9.

15 Roy K, Pal DK, De AU, Sengupta C. Hansch analysis of anticancer activities of C-2-modified paclitaxel analogs against human ovarian carcinoma 1A9, human colon carcinoma HCT116 and human Burkitt lymphoma CA46 cell lines. Indian J Chem Sect B-Org Chem Incl Med Chem 1999; 38: 1194–202.

16 Monti E, Gariboldi M, Maiocchi A, Marengo E, Cassino C, Gabano E, *et al*. Cytotoxicity of *cis*-platinum (II) conjugate models. The effect of chelating arms and leaving groups on cytotoxicity: A quantitative structure–activity relationship approach. J Med Chem 2005; 48: 857–66.

17 van de Waterbeemd H, Gifford E. ADMET in silico modelling: Towards prediction paradise? Nat Rev Drug Discov 2003; 2: 192–204.

18 Yu HS, Adedoyin A. ADME-Tox in drug discovery: integration of experimental and computational technologies. Drug Discov Today 2003; 8: 852–61.

19 Helguera AM, Rodriguez-Borges JE, Garcia-Mera X, Fernandez F, Natalia M, Cordeiro DS. Probing the anticancer activity of nucleoside analogues: A QSAR model approach using an internally consistent training set. J Med Chem 2007; 50: 1537–45.

20 Wang YH, Li Y, Li YH, Yang SL, Yang L. Modeling K-m values using electrotopological state: Substrates for cytochrome P450 3A4-mediated metabolism. Bioorg Med Chem Lett 2005; 15: 4076–84.

21 Habibi-Yangjeh A, Danandeh-Jenagharad M, Nooshyar M. Application of artificial neural networks for predicting the aqueous acidity of various phenols using QSAR. J Mol Model 2006; 12: 338–47.

22 Siu FM, Che CM. Quantitative structure–activity (affinity) relationship (QSAR) study on protonation and cationization of alpha-amino acids. J Phys Chem A 2006; 110: 12 348–54.

23 Su Q, Zhou L. QSAR modeling of AT1 receptor antagonists using ANN. J Mol Model 2006; 12: 869–75.

24 Aoyama T, Suzuki Y, Ichikawa H. Neural networks applied to pharmaceutical problems. III. Neural networks applied to quantitative structure–activity relationship (QSAR) analysis. J Med Chem 1990; 33: 2583–90.

25 Wang YH, Li Y, Yang SL, Yang L. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. J Chem Inf Model 2005; 45: 750–7.

26 Barboni L, Ballini R, Giarlo G, Appendino G, Fontana G, Bombardelli E. Synthesis and biological evaluation of methoxylated

analogs of the newer generation taxoids IDN5109 and IDN5390. Bioorg Med Chem Lett 2005; 15: 5182–6.
27 Ojima I, Inoue T, Chakravarty S. Enantiopure fluorine-containing taxoids: potent anticancer agents and versatile probes for biomedical problems. J Fluor Chem 1999; 97: 3–10.
28 Ojima I, Kuduk SD, Pera P, Veith JM, Bernacki RJ. Synthesis and structure-activity relationships of nonaromatic taxoids: Effects of alkyl and alkenyl ester groups on cytotoxicity. J Med Chem 1997; 40: 279–85.
29 Ojima I, Slater JC, Michaud E, Kuduk SD, Bounaud PY, Vrignaud P, *et al*. Syntheses and structure-activity relationships of the second-generation antitumor taxoids: Exceptional activity against drug-resistant cancer cells. J Med Chem 1996; 39: 3889–96.
30 Ojima I, Wang T, Miller ML, Lin SN, Borella CP, Geng XD, *et al*. Synthesis and structure-activity relationships of new second-generation taxoids. Bioorg Med Chem Lett 1999; 9: 3423-8.
31 Zhu QQ, Guo ZR, Huang N, Wang MM, Chu FM. Comparative molecular field analysis of a series of paclitaxel analogues. J Med Chem 1997; 40: 4319–28.
32 Skehan P, Storeng R, Scudiero D, Monks A, McMahon J, Vistica D, *et al*. New colorimetric cytotoxicity assay for anticancer-drug screening. J Natl Cancer Inst 1990; 82: 1107–12.
33 Golbraikh A, Tropsha A. Beware of q^2!. J Mol Graph Model 2002; 20: 269–76.
34 Kier LB, Hall LH. The prediction of ADMET properties using structure information representations. Chem Biodivers 2005; 2: 1428–37.
35 Hall LH, Kier LB. Electrotopological state indexes for atom types––a novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 1995; 35: 1039–45.
36 Hall LH, Kier LB. The E-state as the basis for molecular structure space definition and structure similarity. J Chem Inf Comput Sci 2000; 40: 784–91.
37 Votano JR, Parham M, Hall LM, Hall LH, Kier LB, Oloff S, *et al*. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. J Med Chem 2006; 49: 7169-81.
38 Hagan MT, Menhaj MB. Training feedforward networks with the Marquardt algorithm. Neural Networks, IEEE Transactions on 1994; 5: 989–93.
39 Tetko IV, Livingstone DJ, Luik AI. Neural-network studies .1. Comparison of overfitting and overtraining. J Chem Inf Comput Sci 1995; 35: 826–33.
40 Berry MJA, Linoff G. Data mining techniques. NY: John Wiley & Sons; 1997.
41 Han LQ. The principle, design and application of artificial neural network. Beijing: Chemical Industry Publishing Company; 2002.
42 Chen LJ, Lian GP. Prediction of human skin permeability using artificial neural network (ANN). Acta Pharmacol Sin 2007; 28: 591–600.
43 Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 1988; 110: 5959–67.
44 Ganesh T, Yang C, Norris A, Glass T, Bane S, Ravindra R, *et al*. Evaluation of the tubulin-bound paclitaxel conformation: Synthesis, biology, and SAR studies of C-4 to C-3' bridged paclitaxel analogues. J Med Chem 2007; 50: 713–25.
45 Snyder JP, Nettles JH, Cornett B, Downing KH, Nogales E. The binding conformation of taxol in beta-tubulin: a model based on electron crystallographic density. Proc Natl Acad Sci USA 2001; 98: 5312–6.
46 Vander Velde DG, Georg GI, Grunewald GL, Gunn CW, Mitscher LA. “Hydrophobic collapse” of taxol and taxotere solution conformations in mixtures of water and organic solvent. J Am Chem Soc 1993; 115: 11 650–1.
47 Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 2003; 22: 69–77.