Journal of Molecular Graphics and Modelling 29 (2011) 826–833

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

An accurate nonlinear QSAR model for the antitumor activities of chloroethylnitrosoureas using neural networks

Yu Qin, Hongfei Deng, Hong Yan*, Rugang Zhong

College of Life Science and Bio-engineering, Beijing University of Technology, Pingleyuan Street No. 100, Chaoyang District, Beijing 100124, China

ARTICLE INFO

Article history:
Received 11 December 2010
Received in revised form 11 January 2011
Accepted 17 January 2011
Available online 1 February 2011

Keywords:
Chloroethylnitrosourea
QSAR
Neural network
Nonlinearity
Molecular descriptor

ABSTRACT

The quantitative structure–activity relationship (QSAR) studies are investigated in a series of chloroethylnitrosoureas (CENUs) acting as alkylating agents of tumors by neural networks (NNs). The QSAR model is described inaccurately by the traditional multiple linear regression (MLR) model for the substitution of CENUs at N-3, whose characteristics play key roles in the biological activity. A nonlinear QSAR study is undertaken by a three-layered NN model, using molecular descriptors that are known to be responsible for the antitumor activity to optimize the input variables of the MLR model. The results demonstrate that NN models present the relationship between antitumor activity and molecular descriptors clearly, and they yield predictions in excellent agreement with the experiment's obtained values ($R^2 = 0.983$). The R^2 value is 0.983 for the 5-8-1 NN model, compared with 0.506 for the MLR model, and the nonlinear model is able to account for 98.3% of the variance of antitumor activities.

1. Introduction

Most tumors are likely to be the result of mutagens, such as tobacco smoke, heat-processed foods and endogenous metabolic products, which generate reactive electrophilic species that alkylate DNA [1–4]. The consequences of alkylation appear to be related to DNA damage in the form of single-strand breaks and cross-linking. Despite their mutagenic and carcinogenic potential, alkylating agents also have cytotoxic properties for the DNA replication of tumor cells and were used in chemotherapy against various tumors [5–8].

Chloroethylnitrosoureas (CENUs) are an extremely active class of alkylating agents, including 1,3-bis (2-chloroethyl)-1-nitrosourea, 1-(2-chloroethyl)-3-cyclohexyl-1-nitrosourea, 1-(2-chloroethyl)-3-methylcyclohexyl-1-nitrosourea, and so on. The CENUs have a wide range of activities against various neoplasms, such as leukemias, encephalomas, lymphomas, melanomas and some solid tumors [9,10]. The CENUs modify nucleosides by transferring chloroethyl groups to nucleophilic sites between the double strands in crosslink DNA and are an attractive possible source of cytotoxicity [11,12].

Quantitative structure–activity relationship (QSAR) studies are mathematical equations that quantitatively correlate chemical structures with biological activity. These relationship models have proved to be helpful in understanding the influence of molecular properties on the biological activity of different compounds, ultimately providing rational clues for the development of new compounds with desirable biological properties. Because they provide valuable information for molecular design and medicinal chemistry, QSAR studies have been widely used in drug design and discovery [13–15].

Significant efforts have been devoted to the QSAR of CENUs, seeking to enhance their antitumor activity with fewer hazards. An estimate of the antileukemic activity of 17 l-(2-haloethyl)-l-nitrosoureas was computed through linear regression equations for the relationships between biological parameters and chemical and physicochemical parameters [16]. The correlation between the antitumor activity and the lipophilic character (Log *P*) of the nitrosoureas, including a rather wide range of chain, circular, aromatic hydrocarbon and glycosyl side groups, indicated that Log *P* of nitrosoureas was the most important parameter determining their antitumor activities [17]. The TOPS-MODE approach was used to predict the rodent carcinogenicity of a set of nitroso-compounds, by establishing the relationship between carcinogenic potential and the length of the alkyl chains via a multiple linear regression (MLR) model, which was able to explain the difference between the nitrosoureas and hydroxyalkyl substituents [18]. By applying an alternative replacement method to a large set of nitroso-compounds, the QSAR approaches were expanded to typically 62 chemicals, and were able to explain 84.3% of the experimental variance [19]. However, the nonlinear QSAR of CENUs, especially considering the influence of the structural characteristics of the N-3 substituents that play the key role in their biological activity, have not yet been reported.

* Corresponding author. Tel.: +86 10 6739 6211; fax: +86 10 6739 2001.
E-mail address: hongyan@bjut.edu.cn (H. Yan).

1093-3263/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2011.01.007

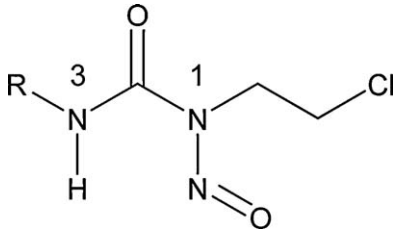


Fig. 1. General structure of CENUs.

Two major concerns in QSAR studies are how to find the optimal molecular descriptors and the optimal statistical methods. For the selection of suitable descriptors, a powerful variable selection method, such as MLR, genetic algorithm (GA) or partial least squares (PLS), is needed. For the statistical method, certain regression-based techniques have been applied to QSAR studies, such as MLR, principal component analysis (PCA) and PLS. These algorithms depend on an assumed linear relationship between the dependent variable and one or more descriptors [20]. Hence, the output is a linear function that is readily understood and easily interpretable. However, in many cases the variables are so complex that they may not be sufficiently precisely emulated by a simple linear regression model to describe the relationships between structure and bioactivity. In contrast to the simple QSAR methods based on regression analysis, neural networks (NNs) have recently been successfully implemented to solve complex nonlinear relationships, and they do not require any prior model of input–output relations [21–24]. The combination of NN and the MLR could be taken as a feature selection method to discover the possible relationship between the input descriptors and the output bioactivity, which can be used for nonlinear phenomena or curved manifolds [25,26]. In this case, the NN acts as a nonlinear regression method, whereas MLR selects the best set of input variables for NN.

We propose, therefore, to use NN to develop a nonlinear QSAR model with better predictive power, using descriptors known to be responsible for the antitumor activity of CENUs with the diversity substitutions on the N-3. In the present work, the correlations with activity of a series of 58 CENU derivatives with different steric features or various hydrophilic congeners are examined by the multiple linear regression and neural network chemometrics methods.

2. Methodology

2.1. Database set

A series of 58 CENU compounds, listed in Table 1, are subjected to QSAR analysis. These compounds were first synthesized by several authors [27,28]. Their general structure is presented in Fig. 1. In Table 1, C is the molal concentration (mol/kg) of CENUs producing a 3-log kill in the viability of leukemia cells (i.e., a 1000-fold reduction in the number of tumor cells). We have collected those claimed to be relevant for describing the antitumor activity variation of the series under investigation. Fifty compounds were selected as the training set for the model generation. Eight compounds (marked with an asterisk in Table 1) were selected as the test set based on the criterion that the test set must represent a wide structural diversity and a range of antiviral activities similar to that of the training set.

2.2. Molecular descriptors

The QSAR technique requires high-quality biological and chemical data to produce a well-trained computational model that can identify the physiochemical and structural properties of the molecule that contribute to a certain biological outcome. In the present work, attempts were made to correlate these properties with a huge number of descriptors encoding the steric, hydrophobic, electronic and structural features of CENUs. Thus, the molecular descriptors were generated using the Gaussian 03 program package and the Hyperchem 7.0 package [29,30]. To avoid including redundant or unnecessary information in this analysis, pairs of variables with a correlation coefficient greater than 0.9 were classified as interrelated, and only one of them was included in the model.

2.3. Regression analysis

A step-wise multiple linear regression procedure has been used for variable selection or model development in biological systems. It is clear that MLR models can be obtained using a step-wise multiple regression procedure; among these models, the best one must be chosen [31,32]. For this purpose, it is common to consider four statistical parameters: the number of descriptors, the square correlation coefficient (R^2), the standard deviation (S) and the F statistic. A reliable MLR model is one that has high R^2 and F values and low S and number of descriptors.

2.4. Neural networks

Because neural networks are artificial systems, they use a large number of interrelated data-processing neurons to emulate the function of brain. Although there are a number of different NN models in use today, the most frequently used type of NN in QSAR, and the one employed in our research, is the three-layered back-propagation neural network. In the back propagation strategy, the neurons are arranged in an input layer, a hidden layer, and an output layer. Each neuron in any layer is fully connected with the neurons of another layer, and there are no connections between neurons in the same layer. The network received a set of inputs as the training set. After training, a nonlinear transfer function was applied to each node in the hidden layer. The goal of training the network is to optimize the weights between the layers so as to minimize the output errors [33,34].

3. Results and discussion

3.1. Variable selection

The correlation between the computed structural parameters and the physicochemical properties was first constructed based on the training set through linear regression analysis. As shown in Table 2, five descriptors, the partition coefficient of lipophilic character (Log P), the energy difference between the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO) (ΔE), the Mulliken charge of N_1 (MUCH), the total dipole (TD), and the single-point energy of CENUs (SPE), were identified and included in the MLR model, and there was no significant correlation between the selected descriptors. The calculated values of the descriptors are shown in Table 3.

3.2. Multiple linear regression model

The mechanism of the biological activity can be interpreted using the proposed linear model. The final correlation equation is the following:

$$\mathrm{Log}\left(\frac{1}{C}\right) = 11.083 - 0.176\,\mathrm{Log}\,P - 18.931\,\Delta E + 21.673\mathrm{MUCH} + 0.088\mathrm{TD} - 0.001\mathrm{SPE} \quad N = 50, \quad R^2 = 0.506, \quad S = 0.594, \quad F = 9.019.$$

Table 1
Chemical structures of the CENUs and their antitumor activity.

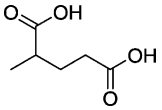
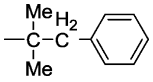
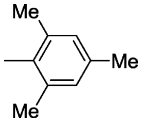
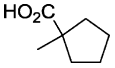
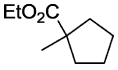
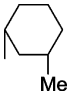
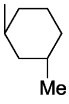
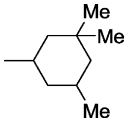
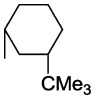
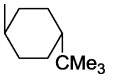
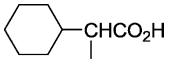
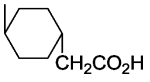
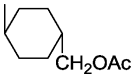
No[a]	*R*	Obsd[b]	Calcd			
			MLR[c]	ΔLog (1/*C*)	NN[d]	ΔLog (1/*C*)
1	-CH(CH$_3$)(CH$_2$)$_4$CH$_3$	0.8530	0.9968	−0.1438	0.8497	0.0033
2	-(CH$_2$)$_2$Cl	1.5900	1.3566	0.2334	1.5951	−0.0051
3	-CH(CH$_3$)CO$_2$H	0.9180	1.3216	−0.4036	0.9158	0.0022
4	-C[(CH$_3$)$_2$]CO$_2$H	0.5910	1.2551	−0.6641	0.5913	−0.0003
5		0.8700	1.5277	−0.6577	0.8700	0.0000
6		0.7370	1.1403	−0.4033	0.7209	0.0161
7		1.5800	1.5035	0.0765	1.5800	0.0000
8		0.7400	1.1338	−0.3938	0.7363	0.0037
9		0.6050	1.3299	−0.7249	0.5999	0.0051
10		0.5620	1.2664	−0.7044	0.5620	0.0000
11		1.2500	1.0746	0.1754	1.2659	−0.0159
12		1.1600	1.1566	0.0034	1.1634	−0.0034
13		1.0900	1.0167	0.0733	1.2012	−0.1112
14		1.0900	1.0315	0.0585	1.2141	−0.1241
15		1.5500	1.0286	0.5214	1.2448	0.3052
16		0.2940	0.9771	−0.6831	0.3143	−0.0203
17		0.7630	0.9286	−0.1656	0.7753	−0.0123
18		0.6840	0.9120	−0.2280	0.6839	0.0001
19		1.2300	1.2610	−0.0310	1.2300	0.0000
20		1.6500	1.4596	0.1904	1.6499	0.0001
21		1.6600	1.7239	−0.0639	1.6604	−0.0004

Table 1 (*Continued*)

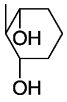
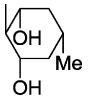
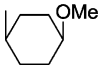
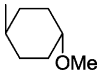
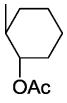
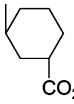
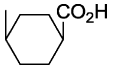
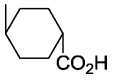
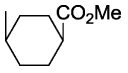
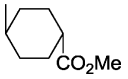
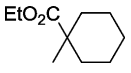
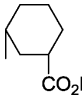
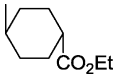
No[a]	*R*	Obsd[b]	Calcd			
			MLR[c]	ΔLog (1/*C*)	NN[d]	ΔLog (1/*C*)
22		1.8200	1.3490	0.4710	1.8202	−0.0002
23		1.2800	1.4524	−0.1724	1.2140	0.0660
24		1.4400	1.5894	−0.1494	1.4348	0.0052
25		1.3600	1.4779	−0.1179	1.3764	−0.0164
26		1.0200	1.4594	−0.4394	1.0694	−0.0494
27		1.9100	1.8952	0.0148	1.9093	0.0007
28		1.8700	1.8459	0.0241	1.8782	−0.0082
29		1.3800	1.3408	0.0392	1.4315	−0.0515
30		1.4700	1.3408	0.1292	1.4313	0.0387
31		1.6500	1.3549	0.2951	1.6504	−0.0004
32		1.5500	1.5447	0.0053	1.5548	−0.0048
33		1.4700	1.2916	0.1784	1.4644	0.0056
34		1.6300	1.3053	0.3247	1.6556	−0.0256
35		1.4700	1.3787	0.0913	1.4700	0.0000
36		1.5100	1.5277	−0.0177	1.5082	0.0018
37		1.6900	1.2882	0.4018	1.6450	0.0450
38		1.0900	1.1608	−0.0708	1.0796	0.0104
39		1.4100	1.5601	−0.1501	1.4099	0.0001
40		1.5100	1.2756	0.2344	1.5439	−0.0339

Table 1 (*Continued*)

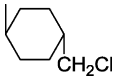
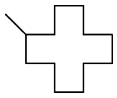
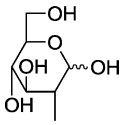
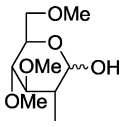
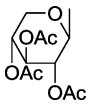
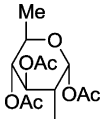
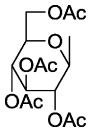
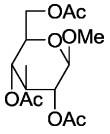
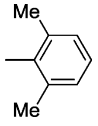
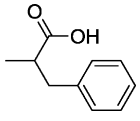
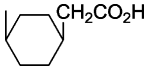
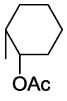
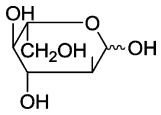
No[a]	*R*	Obsd[b]	Calcd			
			MLR[c]	ΔLog (1/*C*)	NN[d]	ΔLog (1/*C*)
41		1.3400	1.5170	−0.1770	1.3400	0.0000
42		1.2500	1.4247	−0.1747	1.2492	0.0008
43		1.2700	1.4059	−0.1359	1.2701	−0.0001
44		0.5990	0.8091	−0.2101	0.5990	0.0000
45		1.7600	1.8833	−0.1233	1.7600	0.0000
46		1.4800	1.7581	−0.2781	1.4800	0.0000
47		1.9100	1.8287	0.0813	1.9100	0.0000
48		1.9300	1.6326	0.2974	1.9274	0.0026
49		2.1100	2.2566	−0.1466	2.1100	0.0000
50		1.8200	2.2731	−0.4531	1.8115	0.0085
51*		1.2900	1.3749	−0.0849	1.2898	0.0002
52*		1.1000	1.3320	−0.2320	1.0995	0.0005
53*		1.3500	1.3718	−0.0218	1.3500	0.0000
54*		1.3200	1.4049	−0.0849	1.3200	0.0000
55*		1.2400	1.2225	0.0175	1.2398	0.0002

Table 1 (*Continued*)

No[a]	*R*	Obsd[b]	Calcd			
			MLR[c]	ΔLog (1/*C*)	NN[d]	ΔLog (1/*C*)
56*		0.9510	1.3459	−0.3949	0.9509	0.0001
57*		1.2100	1.0699	0.1401	1.2627	−0.0527
58*		1.4600	2.4978	−1.0378	1.4619	−0.0019

[a] The numbers marked by an asterisk are the CENUs in the test set.
[b] Experimental values taken from Refs. [24,25] for CENUs 1–58.
[c] The values predicted from the MLR equation.
[d] The values predicted from the NN′ architecture 5-8-1.

Table 2
Correlation matrix for the five selected descriptors.

	Log *P*	Δ*E*	MUCH	TD	SPE
Log *P*	1.000				
Δ*E*	0.037	1.000			
MUCH	0.544	0.190	1.000		
TD	−0.137	0.130	−0.070	1.000	
SPE	0.238	0.143	0.195	−0.083	1.000

The calculated results of the MLR model for the whole data set are shown in Table 1 and Fig. 2A. The min/max absolute values of ΔLog (1/*C*) were 0.0034/0.7249 and 0.0175/1.0378 for the training and test sets, respectively. In the training set, only 28% of the compounds had a ΔLog (1/*C*) less than 0.1, and no compound had a ΔLog (1/*C*) larger than 1. In the test set, 50% of the compounds showed a ΔLog (1/*C*) less than 0.1, and only compound **58** showed a large ΔLog (1/*C*), i.e., 1.0378, which can be considered an outlier. Generally speaking, the magnitudes of ΔLog (1/*C*) were strongly correlated to the structural characteristics of N-3 substituents of CENUs.

When the substituents of N-3 are chain alkyls (compounds **1**–**5**), the more branched alkyl or carboxyl groups had a higher ΔLog (1/*C*), which were probably relevant to the steric-hindrance effect of branched groups. When the substituents of N-3 were cyclopentyl (compounds **8**–**10**), the ΔLog (1/*C*) were relatively large, probably because of the steric effects and electron-withdrawing groups, such as carboxyl or ester groups. The cyclohexyl substituents on N-3 (compounds **11**–**43**) were the largest classes of compounds chosen for the training set, and the ΔLog (1/*C*) were probably relevant to the properties and steric or locational effects of substituents in the cyclohexyl ring. The 3-methylcyclohexyl CENUs (compounds **13** and **14**) had ΔLog (1/*C*) values of 0.0733 and 0.0585, and the 4-methylcyclohexyls (compounds **15** and **18**) had relatively larger ΔLog (1/*C*) values of 0.5214 and 0.2280, respectively. The cis-3-methylcyclohexyl CENUs (compound **13**) had a ΔLog (1/*C*) of 0.0733, which was larger than that of trans-3-methylcyclohexyl (compound **14**) of the 0.0585 ΔLog (1/*C*). There were two hydroxyls distributed at positions 2 and 6 of cyclohexyl (compounds **27** and **28**), and their ΔLog (1/*C*) values, 0.0148 and 0.0241, were lower than that of one hydroxyl at position 2 (compound **22**), ΔLog (1/*C*) of 0.4710. The cis-4-methoxylcyclohexyl CENUs (compound **29**), had a ΔLog (1/*C*) of 0.0392, which was lower than that of the trans-4-methoxylcyclohexyl (compound **30**), ΔLog (1/*C*) of 0.1292. The ΔLog (1/*C*) values of glycosyl substituents on N-3 were also relevant to the steric effects of substituents in the hexatomic ring (com-

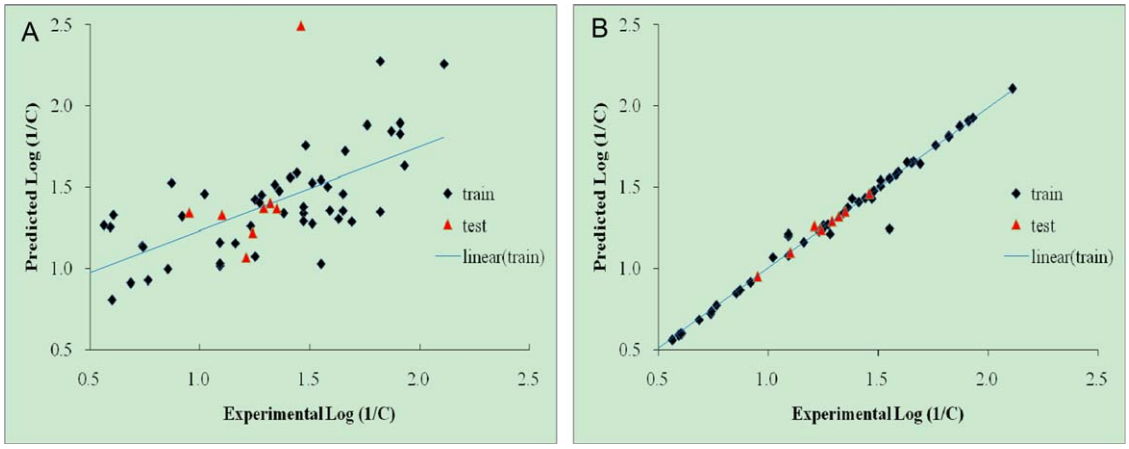


Fig. 2. Plots of predicted versus experimental Log (1/*C*) values of the training set (black dots) and the test set (red triangles) for (A) MLR model and (B) 5-8-1-NN model.

Table 3
The calculated results of the selected molecular descriptors.

No[a]	Log *P*	Δ*E*	MUCH	TD	SPE
1	3.55	0.1841	−0.3461	4.0551	−1167.98
2	1.53	0.1826	−0.3527	2.4177	−1430.99
3	3.11	0.1675	−0.3425	3.6364	−1241.78
4	0.91	0.1822	−0.3499	1.8818	−1199.27
5	0.60	0.1822	−0.3563	3.7932	−1387.84
6	3.42	0.1840	−0.3464	4.1925	−1281.09
7	0.50	0.1815	−0.3549	3.3447	−1159.96
8	2.19	0.1842	−0.3474	4.1408	−1088.13
9	1.82	0.1829	−0.3497	3.7730	−1276.69
10	2.88	0.1831	−0.3483	3.9768	−1355.32
11	2.83	0.1842	−0.3469	4.1780	−1127.46
12	2.45	0.1836	−0.3472	4.3084	−1126.23
13	3.37	0.1842	−0.3470	4.1779	−1166.78
14	3.30	0.1842	−0.3470	4.2064	−1166.77
15	3.30	0.1842	−0.3471	4.1975	−1166.78
16	4.06	0.1842	−0.3468	4.1652	−1245.40
17	4.60	0.1842	−0.3468	4.2472	−1284.72
18	4.66	0.1842	−0.3472	4.2768	−1284.72
19	2.89	0.1765	−0.3555	4.2879	−1355.33
20	2.16	0.1842	−0.3469	4.6238	−1355.34
21	2.35	0.1842	−0.3462	7.3879	−1394.64
22	1.34	0.1828	−0.3560	5.4015	−1202.67
23	1.11	0.1841	−0.3477	4.3520	−1202.67
24	1.00	0.1843	−0.3462	5.3628	−1202.67
25	1.11	0.1841	−0.3475	4.5927	−1202.67
26	1.10	0.1841	−0.3475	4.3626	−1202.67
27	0.16	0.1846	−0.3494	7.1561	−1277.88
28	0.69	0.1846	−0.3494	7.2087	−1317.20
29	2.09	0.1841	−0.3473	4.4989	−1241.98
30	2.09	0.1841	−0.3473	4.4986	−1241.98
31	2.56	0.1822	−0.3567	6.2177	−1355.32
32	1.66	0.1840	−0.3479	4.7941	−1355.32
33	1.53	0.1840	−0.3478	2.0796	−1316.03
34	1.68	0.1840	−0.3477	2.5108	−1316.02
35	1.86	0.1840	−0.3484	3.8775	−1316.02
36	1.89	0.1840	−0.3480	5.0851	−1355.33
37	1.89	0.1840	−0.3476	2.2798	−1355.33
38	3.45	0.1828	−0.3486	3.4795	−1394.64
39	2.20	0.1838	−0.3487	5.7567	−1394.63
40	2.20	0.1840	−0.3476	2.2956	−1394.65
41	2.73	0.1834	−0.3530	5.1130	−1587.05
42	2.66	0.1840	−0.3478	2.7726	−1587.06
43	3.01	0.1841	−0.3520	3.8681	−1626.36
44	5.80	0.1839	−0.3453	3.9628	−1363.31
45	−1.02	0.1821	−0.3604	4.2677	−1503.53
46	0.80	0.1787	−0.3598	4.2660	−1621.43
47	0.82	0.1813	−0.3591	2.9324	−1846.96
48	1.32	0.1829	−0.3550	4.9406	−1503.53
49	1.04	0.1813	−0.3577	4.8542	−2114.13
50	0.67	0.1832	−0.3526	7.3314	−1772.93
51*	2.57	0.1752	−0.3428	3.2728	−1202.46
52*	2.59	0.1842	−0.3502	4.4407	−1391.00
53*	2.22	0.1841	−0.3474	3.8478	−1355.34
54*	1.75	0.1840	−0.3486	5.2923	−1202.68
55*	1.96	0.1821	−0.3589	4.0331	−1355.33
56*	2.73	0.1824	−0.3566	3.8396	−1587.06
57*	2.98	0.1838	−0.3480	4.1769	−1165.55
58*	−0.66	0.1824	−0.3558	6.5529	−1886.27

[a] The numbers marked by an asterisk are the CENUs in the test set.

pounds **47** and **49**) because their main structures were similar to those of cyclohexyls.

By interpreting the descriptors involved in the QSAR model, it is possible to gain some insights into the factors that may affect the Log (1/*C*) values of CENUs. As can be seen in the regression equation, the Log *P* term is negatively correlated with Log (1/*C*), which indicated that the more hydrophilic a molecule, the lower its Log *P* value and the higher its antitumor activity. For example, the activity of compound **11** (Log (1/*C*) = 1.0746) was lower than that of **27** (Log (1/*C*) = 1.8952) with Log *P* reduced (**11**′ Log *P* = 2.83, **27**′ Log *P* = 0.16) by the introduction of two hydroxyls to the cyclohexyl.

The Δ*E* term negatively correlates with Log (1/*C*); a chemical reaction could easily occur with the lowest Δ*E* value. That is to say, the chloroethylation of CENUs occurred between the compounds with lowest Δ*E* and DNA. For example, the activity of compound **1** (Log (1/*C*) = 0.9968) was lower than **3** (Log (1/*C*) = 1.3216), as Δ*E* was reduced (**1**′Δ*E* = 0.1841, **3**′Δ*E* = 0.1675) by the replacement of n-pentyl with carboxyl.

The MUCH of N_1 was positively correlated with Log (1/*C*) and played an important role in the migration of the chloroethyl substitution. Our previous study has showed that the decomposition of CENUs into chloroethyl-diazonium cations is the key step in the alkylation of the DNA base reaction [35]. For example, the activity of compound **19** (Log (1/*C*) = 1.2610) was lower than that of **21** (Log (1/*C*) = 1.7239), and likewise the MUCH values were lower (**19**′MUCH = −0.3555, **21**′MUCH = −0.3462) after the replacement of acetyl with carboxyl.

The other two descriptors, TD and SPE, correlated less with Log (1/*C*) than the three considered above. The TD refers to the dipole moments due to non-uniform distributions of positive and negative charges on the various atoms of CENUs. The electrical charges were uniformly distributed in non-polar molecules, whereas, if the electron density was shared unequally between atoms, as in hydroxyl (–OH), the compound exhibited different polarities. Then, the larger the dipole moment, the higher would be the Log 1/*C* value. For example, the activity of compound **11** (Log (1/*C*) = 1.0746) was lower than that of **28** (Log (1/*C*) = 1.8459), and the corresponding TD value was lower (**11**′TD = 4.1780, **28**′TD = 7.2087) because of the polarity enhancement with the introduction of hydroxyl and methyl to cyclohexyl. In addition, SPE was introduced into the model because this descriptor reflects the conformational stability of a molecule. The larger the SPE value, the lower would be the Log 1/*C* value. For example, the activity of compound **8** (Log (1/*C*) = 1.1338) was lower than that of **10** (Log (1/*C*) = 1.2664), and the SPE was lower (**8**′SPE = −1088.13, **10**′SPE = −1355.32), probably because of the easily hydrolyzable nature of ethyl carboxylate ester.

3.3. Neural network models

The NN models were generated using the five descriptors appearing in the MLR model as their inputs. One neuron, which encoded the antitumor activity, constituted the output layer, and the hidden layer contained a variable number of neurons. The input values were normalized to [−1, 1], the number of neurons in the hidden layer was limited to 4–8, the learning rate interval was set to 0.05, the number of epochs was 10^4, and the goal was 0.01. The training function, the adaption learning function and the transfer function were designated as TRAINLM, LEARNGDM and TANSIG, respectively.

The prediction results from the 5-8-1 NN model are given in Table 1 and Fig. 2B. The min/max absolute values of the residuals for the training and test sets were 0/0.3052 and 0/0.0527, respectively. For the whole dataset, 94.8% of the compounds had residuals less than 0.1. These results clearly show the strong correlations between Log (1/*C*) and the structural characteristics of N-3 substituents of CENUs. They also show some differences (especially for the test set) from the MLR model, which confirms the nonlinear relationship between structural information and antitumor activity.

For example, the ΔLog (1/*C*) value of CENUs with chain alkyls substituents on N-3 (compound **4**) was improved from 0.6641 in MLR to 0.0003 in the NN model. The ΔLog (1/*C*) value of CENUs with cyclohexyl substituents on N-3 (compound **16**) was improved from 0.6831 in MLR to 0.0203 in the NN model. The ΔLog (1/*C*) of value CENUs with glycosyl substituents on N-3 (compound **50**) was improved from 0.4531 in MLR to 0.0085 in the NN model.

Table 4
Statistical results of different NN models and MLR analysis.

Model	R^2	S
5-4-1	0.931	0.282
5-5-1	0.911	0.199
5-6-1	0.919	0.298
5-7-1	0.947	0.224
5-8-1	0.983	0.119
MLR	0.506	0.594

3.4. Comparison of MLR and NN models

The fitting quality of the MLR and NN models is estimated by the square correlation coefficient (R^2) and the standard error of calculation (S) in Table 4. The results are also shown in Fig. 2. As can be seen from the data, the R^2 values of NN models range from 0.919 to 0.983 for the 5-8-1 NN model for the training set, which is significantly higher than the 0.506 value for the MLR model. This means that the NN model is able to account 98.3% of the variance of the antitumor activity. This statistical parameter is also much more than the replacement method (RM), which was an alternative method on the base of elimination method (EM) and was able to explain 84.3% of the experimental carcinogenic potency of 62 typical nitroso-compounds [19]. The high correlation coefficients given by the trained NN models indicated that the Log ($1/C$) value significantly correlated with the five variables adopted in this work.

4. Conclusions

The NN models can be used to establish the QSAR model of CENUs with higher accuracy, taking into account the influence of the structural characteristics of the N-3 substituents. The antitumor activity of CENUs was represented by linear and nonlinear models based on five-parameters (Log P, ΔE, MUCH, TD and SPE). A linear model was obtained by MLR, with R^2 values of 0.506 and S of 0.594 for the training set. The R^2 and S values from the nonlinear 5-8-1 NN model for the training set were 0.983 and 0.119, respectively. The results show that the nonlinear model is reliable and correctly identified the structural factors that play important roles in the determination of antitumor activity. The NN QSAR may be of considerable interest for the design of new antitumor drugs and will be analyzed in future studies to provide medicinal chemists with immediately useful features derived by NN analyses, allowing for more precise control of the antitumor activity of CENU derivatives.

Acknowledgements

This work was financially supported by the Key Projects in the National Science & Technology Pillar Program during the Eleventh Five-Year Plan Period (No. 2008ZX10001–015) and the Natural Sciences Foundation of Beijing (No. 200710005002).

References

[1] H. Bartsch, R. Montesano, Relevance of nitrosamines to human cancer, Carcinogenesis 5 (1984) 1381–1393.
[2] D.M. DeMarini, Genotoxicity of tobacco smoke and tobacco smoke condensate: a review, Mutat. Res. 567 (2004) 447–474.
[3] M. Jagerstad, K. Skog, Genotoxicity of heat-processed foods, Mutat. Res. 574 (2005) 156–172.
[4] L.J. Marnett, P.C. Burcham, Endogenous DNA adducts: potential and paradox, Chem. Res. Toxicol. 6 (1993) 771–785.
[5] B. Kaina, M. Christmann, S. Naumann, W.P. Roos, MGMT: key node in the battle against genotoxicity, carcinogenicity and apoptosis induced by alkylating agents, DNA Repair 6 (2007) 1079–1099.
[6] R. Jain, M. Sharma, Fluorescence postlabeling assay of DNA damage induced by *N*-methyl-*N*-nitrosourea, Cancer Res. 53 (1993) 2771–2774.
[7] B. Sedgwick, Nitrosated peptides and polyamines as endogenous mutagens in O^6-alkylguanine-DNA alkyltransferase deficient cells, Carcinogenesis 18 (1997) 1561–1567.
[8] C.T. Gnewuch, G. Sosnovsky, A critical appraisal of the evolution of N-nitrosoureas as anticancer drugs, Chem. Rev. 97 (1997) 829–1013.
[9] M.J. van den Bent, M.E. Hegi, R. Stupp, Recent developments in the use of chemotherapy in brain tumours, Eur. J. Cancer 42 (2006) 582–588.
[10] E.P. Mitchell, P.S. Schein, Contributions of nitrosoureas to cancer treatment, Cancer Treat. Rep. 70 (1986) 31–41.
[11] R.B. Weiss, B.F. Issell, The nitrosoureas: carmustine (BCNU) and lomustine (CCNU), Cancer Treat. Rev. 9 (1982) 313–330.
[12] S.L. Gerson, MGMT: its role in cancer aetiology and cancer therapeutics, Nat. Rev. Cancer 4 (2004) 296–307.
[13] A. Nargotra, S. Koul, S. Sharma, I.A. Khan, A. Kumar, N. Thota, J.L. Koul, S.C. Taneja, G.N. Qazi, Quantitative structure–activity relationship (QSAR) of aryl alkenyl amides/imines for bacterial efflux pump inhibitors, Eur. J. Med. Chem. 44 (2009) 229–238.
[14] O. Nicolotti, V.J. Gillet, P.J. Fleming, D.V.S. Green, Multiobjective optimization in quantitative structure–activity relationships: deriving accurate and interpretable QSARs, J. Med. Chem. 45 (2002) 5069–5080.
[15] H.Y. Xu, J.Y. Zhang, J.W. Zou, X.S. Chen, QSPR models for the physicochemical properties of halogenated methyl-phenyl ethers, J. Mol. Graphics Model. 26 (2008) 1076–1081.
[16] G.P. Wheeler, B.J. Bowdon, J.A. Grimsley, Interrelations of some chemical, physicochemical, and biological activities of several 1-(2-haloethyl)-1-nitrosoureas, Cancer Res. 34 (1974) 194–200.
[17] C Hansch, A. Leo, C. Schmidt, P.Y.C. Jow, J.A. Montgomery, Antitumor structure–activity relationships. Nitrosoureas vs. L-1210 leukemia, J. Med. Chem. 23 (1980) 1095–1101.
[18] A.M. Helguera, M.P. González, M.N.D.S. Cordeiro, M.A. Cabrera, Quantitative structure carcinogenicity relationship for detecting structural alerts in nitroso-compounds, Toxicol. Appl. Pharmacol. 221 (2007) 189–202.
[19] A.H. Morales, P.R. Duchowicz, M.A. Cabrera, E.A. Castro, M.N.D.S. Cordeiro, M.P. González, Application of the replacement method as a novel variable selection strategy in QSAR: 1. Carcinogenic potential, Chemom. Intell. Lab. Syst. 81 (2006) 180–187.
[20] M.M.C. Ferreira, Multivariate QSAR, J. Braz. Chem. Soc. 13 (2002) 742–753.
[21] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice Hall PTR, Upper Saddle River, NJ, 1999.
[22] J. Zupan, J. Gasteiger, Neural Networks for Chemists: An Introduction, VCH, Weinheim, 1993.
[23] N.X. Tan, H.B. Rao, Z.R. Li, X.Y. Li, Prediction of chemical carcinogenicity by machine learning approaches, SAR QSAR Environ. Res. 20 (2009) 27–75.
[24] W. El-Deredy, S.M. Ashmore, N.M. Branston, J.L. Darling, S.R. Williams, D.G.T. Thomas, Pretreatment prediction of the chemotherapeutic response of human glioma cell cultures using nuclear magnetic resonance spectroscopy and artificial neural networks, Cancer Res. 57 (1997) 4196–4199.
[25] G.N. Zahra, A.R. Behzad, Modeling the antileishmanial activity screening of 5-nitro-2-heterocyclic benzylidene hydrazides using different chemometrics methods, Eur. J. Med. Chem. 45 (2010) 719–726.
[26] F.T. Simona, I. Daniela, S. Takahiro, A tentative quantitative structure–toxicity relationship study of benzodiazepine drugs, Toxicol. In Vitro 24 (2010) 184–200.
[27] T.P. Johnston, G.S. McCaler, P.S. Opliger, W.R. Laster, J.A. Montgomery, Synthesis of potential anticancer agents. 38. N-nitrosoureas. 4. Further synthesis and evaluation of haloethyl derivatives, J. Med. Chem. 14 (1971) 600–614.
[28] J.A. Montgomery, Chemistry and structure–activity studies of the nitrosoureas, Cancer Treat Rep. 60 (1976) 651–664.
[29] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery, T. Vreven Jr., K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, Gaussian 03, Revision A. 01, Gaussian, Inc., Pittsburgh, PA, 2003.
[30] Hyperchem Release 7.0, Hypercube, Inc., 115 NW 4th Street, Gainsville, FL 32601, USA, 2004.
[31] M.K. Gupta, R. Sagar, A.K. Shaw, Y.S. Prabhakar, CP-MLR directed QSAR studies on the antimycobacterial activity of functionalized alkenols-topological descriptors in modeling the activity, Bioorg. Med. Chem. 13 (2005) 343–351.
[32] J.T. Leonard, K. Roy, QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques, Bioorg. Med. Chem. 14 (2006) 1039–1046.
[33] L. Douali, D. Villemin, D. Cherqaoui, Neural networks: accurate nonlinear QSAR model for HEPT derivatives, J. Chem. Inf. Comput. Sci. 43 (2003) 1200–1207.
[34] T.A. Andrea, H. Kalayeh, Application of neural networks in quantitative structure–activity relationships of dihydrofolate reductase inhibitors, J. Med. Chem. 34 (1991) 2824–2836.
[35] T.T. Liu, L.J. Zhao, R.G. Zhong, Researches on the mechanism of single strand break following the alkylation of DNA base induced by chloroethylnitrosoureas, Chem. J. Chin. Univ. 31 (2010) 957–963.