Prediction of Carcinogenicity of Noncongeneric Chemicals Using an Artificial Neural Network

Takahiro SUZUKI[a]

Abstract

The discriminant method of two-value regression and three-layer artificial neural network (ANN) modeling with back-propagation have been applied to develop quantitative structure-toxicity relationships. The training set contains 323 diverse chemicals and their carcinogenicity data were obtained from the US National Toxicity Program. The data provide yes/no response (carcinogens or non-carcinogens) as the endpoints. Two sets of molecular descriptors, DRAGON and BCI Fingerprints, were explored for discriminating potential. To solve the problems in training the ANNs, various conditions of the network such as the training cycles and neuron numbers of the intermediate layer were optimized. The optimum ANN model with 25 kinds of DRAGON descriptors gave best prediction performance for the test set of 185 compounds not included in the training set. The model showed the prediction accuracy of 64.8% for test set chemicals.

Key words: carcinogenicity, molecular descriptors, neural network, quantitative structure-activity relationships

1. Introduction

Numerous chemicals of natural and synthetic origin have been produced. There are more than 80,000 chemicals registered for use in commerce in Japan, and an estimated 2,000 new ones are supplied annually for our use such as foods, personal care products, various drugs, house-hold cleaners, and agricultural chemicals. However, the adverse effects of most of those chemicals on human health and ecosystems are not known.

For carcinogenicity only limited data are available and rodent bioassays are very laborious, time-consuming (3-5 years), and costly (>1 million U.S. dollars per chemical). According to guidelines of the Food and Drug Administration (FDA), carcinogenicity is required against multiple biological models. Thousands of chemicals currently in commerce and in the environment have not undergone carcinogenicity testing. Therefore, a reliable tool for predicting carcinogenicity would be highly desirable for virtual screening of compound libraries of both pharmaceuti-

[a] Natural Science Laboratory, Toyo University, 11-10, Oka 2, Asaka-shi, Saitama 351-8510, Japan

cally and other commercially interesting molecules.

The quantitative structure–activity relationship (QSAR) approaches based on the assumption that the structure of a molecule should contain the features responsible for its physical, chemical, and biological properties have been applied to the prediction of toxicity (Dunn & Wold, 1981; Klopman et al., 2004). QSARs can be employed for estimating the activity of other chemicals not tested experimentally. The Predictive Toxicology Challenge (PTC) 2000–2001 workshop was held in three yeas ago (Helma et al 2001). Its aim was to obtain models for predicting the outcome of biological tests for the carcinogenicity of chemicals using information on chemical structure. The learning data contained rodent carcinogenicity for 417 compounds from the National Toxicology Program (NTP) database (which has a preponderance of industrial chemicals with relatively simple chemical structures), and test data contained of 185 compounds independently tested by FDA database (which has a preponderance of pharmaceuticals with multiple ring systems). On the basis of 7 sets of available descriptors (total: >7,000), 14 research teams submitted about 30 models. However, the exercise showed that models performed poorly in predicting the results from the FDA database. One of the reasons might come from the fact that most models used linear relationships between chemical structure and carcinogenicity, which apply well within congeneric chemical classes.

In this study, both multiple linear regression (two-value regression) and artificial neural network (ANN) modeling were applied to make the prediction scheme for chemical carcinogenicity. The prediction performance was evaluated through simulated external validation employing complementary subsets. ANN simulates the functioning of human neurons and can be used to model complex phenomena where noise and nonlinear processes may be present, such as in this problem. The ANN modeling has been used in limited cases of the modeling of toxicity of chemicals. Villemin et al. (1994) used ANN to model polycyclic aromatic hydrocarbons in carcinogenic classes, obtaining good results. A modeling of aromatic nitrogen compounds in carcinogenic activity by ANN gave fairly good results (Gini et al, 1999). Their regression analysis to develop models proved to be unsuccessful. In a study using 280 compounds of various kinds (Benigni & Richard, 1996), it was concluded that BPNN models fitted training sets but had no general applicability.

2. Materials and Methods

2.1. Data set

The female rat data were the PTC data set containing 323 compounds by the National Toxicology Program (NTP) and test data containing 185 compounds, downloaded from the public domain (Helma et al., 2001), with the carcinogenicity

of each labeled "+" for positive or "−" for negative. Although the NTP and FDA used different rodent strains, the combined data set contain chemical carcinogenicities from experimental measurements on a single 2-year carcinogenicity study to identify trans-species tumorigens.

2.2. Molecular descriptors

Cancer is not a single disease and several mechanisms are involved in the various processes leading to the different tumors. From the mechanistic point of view, there are basically two types of carcinogens: genotoxic and epigenetic/nongenotoxic. Genotoxic carcinogens (DNA-reactive carcinogens) are chemicals that directly interact with DNA as either parent chemicals or reactive metabolites. Epigenetic carcinogens are agents that act through a secondary mechanism that does not involve direct DNA damage. In reality, the demarcation is seldom absolute. Evidence suggests that some carcinogens act via genotoxic mechanisms in one set of targets but via nongenotoxic mechanisms in another set of targets.

From a variety of available molecular descriptors, two sets of descriptors, DRAGON and BCI fingerprints (Helma et al., 2001), were employed here. The first set of descriptors is a system for calculating molecular descriptors for 3D-QSARs developed by the Milano Chemometrics and QSAR Research Group (Todeschini & Consonni, 2000) and 839 kinds of descriptors were prepared.

The other set of descriptors is keyed; that is, a dictionary gives the relationship between substructures and bits. The fingerprint of a molecule is a string of bits that shows whether certain molecular fragments are present in a molecule. Total of 57240 BCI fingerprints were prepared and used for modeling.

2.3. Classification modeling

For categorical modeling the carcinogenicity, the following statistical methods can be applied: linear discriminant analysis as implemented in STATISTICA and binary logistic regression as available in SPSS (Mazzatorta et al., 2004). By preference, the discriminant method of two-value regression analysis was employed in this study because of several advantages over linear discriminant analysis (Hatch & Magee, 1998).

2.4. The artificial neural network (ANN) modeling

The artificial neural networks (ANNs) have an inherent ability to provide nonlinear and cross product terms for QSAR modeling. The ANNs may be especially useful when a rigid theoretical basis and/or mathematical relationship to describe a phenomenon to be modeled is not available in advance.

From many ANN approaches, both different in architecture and in learning algorithms, the three-layer ANNs with the back-propagation of errors were employed in this study. Since the theory and practical application of the ANN are popular,

an explanation of the methodology can be delegated to the literature (Zupan & Gasteiger, 1999). The most commonly used logsigmoid transfer function and the delta rule for the error correction formula were used in the networks. The ANN calculations were carried out by our in-house program.

3. Results and Discussion

3.1. Significant molecular descriptors

The statistical analysis to find the significant descriptors was carried out using a heuristic method based on the linear regression technique for the training set of 323 compounds. This procedure is based on the scale forward selection technique (Draper & Smith, 1966). The analyses revealed that 25 DRAGON-descriptors (Table 1) were proved to be the significant model parameters for the best two-value regression model obtained:

$$\begin{aligned}\mathrm{CLASS} = {} & 3.45 - 0.012\mathrm{Sp} - 0.265\mathrm{Ms} - 0.084\mathrm{nOH} - 0.122\mathrm{nNH} + 0.046\,\chi_0 + 0.260\mathrm{CIC} \\ & + 0.172\mathrm{SRW01} + 0.001\mathrm{SRW08} - 1.194\mathrm{BENe1} + 151.3\mathrm{JGI10} - 0.158\mathrm{MATS7e} \\ & - 0.166\mathrm{MATS8p} + 0.105\mathrm{GATS2e} - 0.288\mathrm{Mor32m} - 0.361\mathrm{Mor16v} + 0.068\mathrm{Mor03e} \\ & + 0.325\mathrm{Mor12e} + 0.558\mathrm{Mor11p} + 0.086\mathrm{Mor13p} + 0.857\mathrm{Mor19p} - 0.608\mathrm{E3u} \\ & - 0.332\mathrm{G3p} + 0.638\mathrm{HATS7u} - 1.191\mathrm{H5p} - 0.805\mathrm{R2v} \qquad (1)\end{aligned}$$

In the above equation, the carcinogens are coded Class 1 and non-carcinogens are coded Class 0. As can be seen from the definition of the descriptors in Table 1, most of the descriptors are electrostatic, topological or geometrical (3D-structural) in nature.

The alternative regression model with 25 BCI fingerprints gave slightly inferior fit on the given set of chemicals. Therefore, the DRAGON descriptors were implemented as the inputs for a neural network.

3.2. ANN model

The architecture of the neural network model used in this study is shown in Fig. 1. The network consists of three layers: 25 input nodes plus a bias, a varying number of hidden-layer neurons (between two to ten plus one bias), and a single output neuron corresponding to a compound's carcinogenicity ("1" for carcinogens and "0" for non-carcinogens). The statistical quality of the ANN and two-value regression modeling results for training and test sets were evaluated based on the following parameter:

$$\begin{aligned}&\text{Correct classification rate (\%)} \\ &\quad = \text{correctly classified compounds/total number of test compounds} \qquad (2)\end{aligned}$$

Table 1 List of 25 kinds of DRAGON molecular descriptors used for the modeling of carcinogenicity.

Symbol	Descriptor
Sp	sum of atomic polarizabilities (scaled on carbon atom)
Ms	mean electotopological state
nOH	number of OH groups
nNH	number of NH groups
χ_0	connectivity index chi-0 topological descriptor
CIC	complementary information content (neighborhood symmetry)
SRW01	self-returning walk count of order 01 molecular walk counts
SRW08	self-returning walk count of order 08 molecular walk counts
BENe1	negative Burden engenvalue weighted by atomic Sanderson electronegativities BCUT descriptors
JGI10	mean topological charge idex of order 10 Galvez topological charge indices
MATS7e	Moran autocorrelation weighted by atomic Sanderson electronegativities 2D autocorrelations
MATS8p	Moran autocorrelation weighted by atomic polarizabilities 2D autocorrelations
GATS2e	Geary autocorrelation weighted by atomic Sanderson electronegativities 2D autocorrelations
Mor32m	3D-MoRSE weighted by atomic masses
Mor16v	3D-MoRSE weighted by atomic van der Waals volume
Mor03e	3D-MoRSE weighted by atomic Sanderson electronegativities
Mor12e	3D-MoRSE weighted by atomic Sanderson electronegativities
Mor11p	3D-MoRSE weighted by atomic polarizabilities
Mor13p	3D-MoRSE weighted by atomic polarizabilities
Mor19p	3D-MoRSE weighted by atomic polarizabilities
E3u	3rd component accessibility directional WHIM index
G3p	3rd component symmetry directional WHIM index weighted by atomic polarizabilities
HATS7u	leverage-weighted autocorrelation unweighted GETAWAY descriptor
H5p	H autocorrelation weighted by atomic polarizabilities GETAWAY descriptor
R2v	R autocorrelation weighted by atomic van der Waals volumes GETAWAY descriptors

The best architecture was determined to be (25+1) : (3+1) : 1 (25 input neurons for the 25 descriptors plus a bias, three hidden-layer neurons plus a bias, and one output layer neuron for a total of 82 adjustable parameters) using training data sets. Maximum correct classification rate value was achieved when the network was trained for 30000 epochs. The number of training data points (=323) is about 4 times greater than the number of adjustable parameters. The quality of fitting,

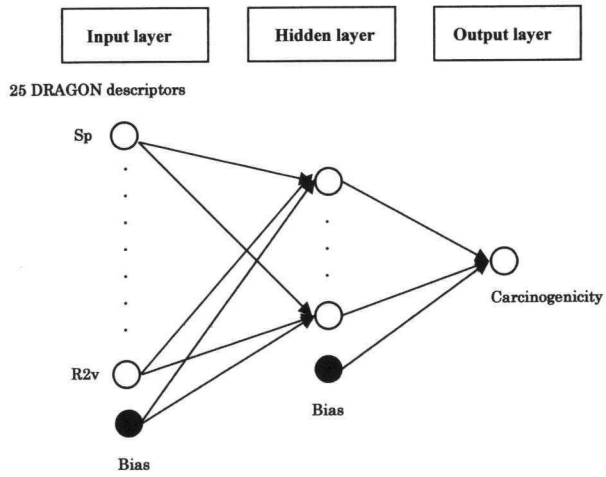

Fig. 1 Architecture of neural network model for chemical carcinogenicity modeling.

the value of correct classification rate of 91.0% with the ANN model, was found to be better than that obtained with the two-value regression model of 81.4%.

The prediction results obtained by the ANN are shown in Table 2 together with the experimental carcinogenicity and those results calculated by previous regression model. The quality of the prediction, the correct classification rate of 64.8% with the ANN model, was found to be better than that obtained with the two-value regression model of 62.2%. An alternative ANN model using 25 kinds of BCI fingerprints gave the correct classification rate of 55.1%.

Within the same structural class, QSAR analysis can be further optimized by classifying the data into more well-defined subclasses. This is because structurally closely related chemicals are expected to behave similarly in the mechanism of action.

4. Conclusions

The present study examined the feasibility of an ANN for predicting carcinogenicity of chemicals of various types. Further developments will be needed to increase the prediction power of the ANN model for practical assessing the chemical carcinogenicity. We will have to add separate models for individual classes of compounds to our system, since cancer is not a single disease and various mechanisms are involved in a variety of processes. The results of this study showed the potential of applying the neural computing technique for predicting toxicity of chemicals.

Table 2 Prediction of carcinogenicity of test compounds by MLR and ANN models.

no.	compound	CAS RN	carcinogenicity*		
			obsd	MLR	ANN
1	acebutolol	37517-30-9	0	0	0
2	acrivastine	87848-99-5	0	0	0
3	acyclovir	59277-89-3	0	0	0
4	allopurinol	315-30-0	0	0	0
5	amiloride	2609-46-3	0	0	0
6	amlodipine	88150-42-9	0	1	1
7	amphetamine	300-62-9	0	0	0
8	ampicillin	69-53-4	0	1	0
9	amrinone	60719-84-8	0	0	0
10	astemizol	68844-77-9	0	1	1
11	atenolol	29122-68-7	1	0	0
12	auranofin	34031-32-8	1	0	0
13	benazepril	86541-75-5	0	0	0
14	bepridil	64706-54-3	1	0	0
15	betaxolol	63659-18-7	0	0	0
16	bisoprolol	66722-44-9	0	0	0
17	bitolterol	30392-40-6	0	1	1
18	brotizolam	57801-81-7	1	0	0
19	budesonide	51333-22-3	1	0	0
20	bunolol	27591-01-1	1	0	0
21	bupropion	31677-93-7	0	0	0
22	buspirone	36505-84-7	0	0	0
23	captopril	62571-86-2	0	0	0
24	carteolol	51781-06-7	0	0	0
25	cetirizine	83881-51-0	0	0	0
26	chlordiazepoxide	58-25-3	0	0	0
27	chlorpheniramine	132-22-9	0	0	0
28	chlorpromazine	50-53-3	1	0	0
29	189		1	1	0
30	ciprofloxacin	85721-33-1	0	0	0
31	cisapride	81098-60-4	0	1	1
32	clemastine	15686-51-8	0	0	0
33	clozapine	5786-21-0	0	0	0
34	cyclobenzaprine	303-53-7	0	0	0
35	dacarbazine	4342-03-2	1	0	0
36	dantrolene	7261-97-4	0	0	0
37	desogen	54024-22-5	1	0	0
38	dexfenfluramine	3239-44-9	0	0	0
39	diazepam	439-14-5	0	0	0
40	diclofenac	15307-86-5	0	1	0
41	didanosine	69655-05-6	0	0	0
42	diflunisal	22494-42-4	0	1	0
43	diltiazem	42399-41-7	0	1	0
44	diphenhydramine	58-73-1	0	0	0
45	dipyridamole	58-32-2	0	1	1

Table 2 (Continued)

no.	compound	CAS RN	carcinogenicity*		
			obsd	MLR	ANN
46	doxazosin	74191-85-8	0	1	1
47	doxylamine	469-21-6	1	0	0
48	enalapril	75847-73-3	0	0	0
49	ephedrine	299-42-3	0	0	0
50	erythromycin	114-07-8	0	1	1
51	estazolam	29975-16-4	0	0	0
52	etodolac	41340-25-4	0	0	0
53	etretinate	54350-48-0	0	0	0
54	famciclovir	104227-87-4	0	0	0
55	famotidine	76824-35-6	0	0	0
56	felbamnate	25451-15-4	1	0	0
57	finasteride	98319-26-7	1	0	0
58	flecainide	54143-55-4	0	1	1
59	fluconazole	86386-73-4	1	0	0
60	flunisolide	77326-96-6	0	0	0
61	fluoxetine	54910-89-3	0	1	1
62	flurazepam	17617-23-1	0	1	0
63	flurbiprofen	5104-49-4	0	0	0
64	fluvastatin	93957-54-1	1	1	1
65	foscarnet	63585-09-1	0	0	0
66	fosinopril	98048-97-6	1	0	0
67	furazolidone	67-45-8	0	0	0
68	gabapentin	60142-96-3	1	0	0
69	gemfibrozil	25812-30-0	1	0	0
70	glipizide	29094-61-9	0	0	0
71	glyburide	10238-21-8	0	0	0
72	granisetron	109889-09-0	1	0	0
73	guanadrel	40580-59-4	1	0	0
74	guanfacine	29110-47-2	0	0	0
75	hydralazine	86-54-4	1	0	1
76	indapamide	26807-65-8	0	1	0
77	indomethacine	53-86-1	0	0	0
78	iodinated glycerol	5634-39-9	1	0	0
79	ipratropium	66985-17-9	0	0	0
80	isosorbide	652-67-5	0	0	0
81	isradipine	75695-93-1	1	1	0
82	itraconazole	84625-61-6	1	0	0
83	ketoconazole	65277-42-1	0	0	0
84	ketoprofen	22071-15-4	0	0	0
85	ketorolac	74103-06-3	0	0	0
86	labetalol	36894-69-6	0	0	0
87	lamotrigine	84057-84-1	0	0	0
88	lansoprazole	103577-45-3	1	0	0
89	levamisole	14769-73-4	0	0	0
90	levomethadone	125-58-6	0	0	0

Table 2 (Continued)

no.	compound	CAS RN	carcinogenicity*		
			obsd	MLR	ANN
91	loratidine	79794-75-5	0	0	0
92	lorazepam	846-49-1	0	0	0
93	lovastatin	75330-75-5	1	1	0
94	mannitol	87-78-5	0	0	0
95	mebendazole	31431-39-7	0	0	0
96	mefloquine	53230-10-7	0	0	0
97	menthol	89-78-1	0	0	0
98	metaproterenol	586-06-1	0	0	0
99	alpha-methyldopa	555-30-6	0	0	0
100	methylphenidate	113-45-1	0	0	0
101	metolazone	17560-51-9	0	0	0
102	metoprolol	37350-58-6	0	1	1
103	metronidazole	443-48-1	1	0	0
104	mexiletine	5370-01-4	0	1	0
105	midazolam	59467-70-8	1	0	0
106	milrinone	78415-72-2	0	0	0
107	minocycline	10118-90-8	1	0	0
108	misoprostol	59122-46-2	0	0	0
109	moricizine	31883-05-3	1	0	0
110	mycophenolate	24280-93-1	0	1	1
111	nabumetone	42924-53-8	0	1	1
112	nadolol	42200-33-9	0	0	0
113	nafenopin	3771-19-5	1	0	0
114	nedocromil	69049-73-6	0	0	0
115	nefazodone	83366-66-9	0	0	0
116	netilmicin	56391-56-1	0	0	0
117	nicardipine	55985-32-5	1	1	0
118	nimodipine	66085-59-4	1	1	1
119	nisoldipine	63675-72-9	0	1	1
120	nizatidine	76963-41-2	0	0	0
121	omeprazole	73590-58-6	1	0	0
122	ondansetron	116002-70-1	0	1	0
123	olsalazine	15722-48-2	0	1	1
124	oxaprozin	21256-18-8	0	0	0
125	oxytetracycline	6153-64-6	0	0	0
126	pamidronic acid	40391-99-9	1	0	0
127	paroxetine	61869-08-7	0	1	0
128	penbutolol	38363-40-5	0	1	1
129	penicillin	69-57-8	0	1	1
130	pentaerythritol	115-77-5	0	0	1
131	pentoxifylline	6493-05-6	0	0	0
132	pergolide	66104-22-1	0	0	0
133	perindopril	82834-16-0	0	0	0
134	permethrin	52645-53-1	0	1	1
135	phenazopyridine	94-78-0	1	1	1

Table 2 (Continued)

no.	compound	CAS RN	carcinogenicity*		
			obsd	MLR	ANN
136	phenformin	114-86-3	0	0	0
137	phenylephrine	59-42-7	0	0	0
138	pimozide	2062-78-4	0	0	0
139	pindolol	13523-86-9	0	0	0
140	pirbuterol	38677-81-5	0	0	0
141	piroxicam	36322-90-4	0	0	0
142	pravastatin	81093-37-0	1	0	0
143	prazepam	2955-38-6	0	0	0
144	procarbazine	671-16-9	1	0	0
145	promethazine	60-87-7	0	0	0
146	propafenone	54063-53-5	0	0	0
147	propranolol	318-98-9	0	0	0
148	pyrilamine	91-84-9	0	1	1
149	quinapril	85441-61-8	0	0	0
150	ramipril	87333-19-5	0	0	0
151	ranitidine	66357-35-5	0	0	0
152	ribavirin	36791-04-5	1	0	0
153	rifabutin	72559-06-9	0	1	0
154	rifampin	13292-46-1	0	1	1
155	ripazepam	26308-28-1	0	0	0
156	risperidone	106266-06-2	1	0	0
157	scopolamine	138-12-5	0	0	0
158	selenium sulfide	7446-34-6	1	0	0
159	sertraline	79617-96-2	1	0	0
160	simvastatin	79902-63-9	1	1	0
161	sodium fluoride	7681-49-4	0	0	0
162	sotalol	3930-20-9	0	0	0
163	sulfadiazine	68-35-9	0	0	0
164	sumatriptan	103628-46-2	0	0	0
165	tamoxifen	10540-29-1	1	1	1
166	temazepam	846-50-4	0	0	0
167	terazosin	63590-64-7	1	1	1
168	terbinafine	78628-80-5	1	1	1
169	terbutaline	23031-25-6	0	0	0
170	terfenadine	50679-08-8	0	0	0
171	tetracycline	60-54-8	0	0	0
172	thiabendazole	148-79-8	1	0	0
173	ticlopidine	55142-85-3	0	0	0
174	timolol	26839-75-8	1	0	0
175	tocainide	59-26-7	0	0	1
176	tolmetin	152-11-4	0	0	0
177	torsemide	56211-40-6	1	0	0
178	tramadol	27203-92-5	0	0	0
179	triprolidine	486-12-4	0	0	0
180	tryptophan	54-12-6	0	0	0

Table 2 (Continued)

no.	compound	CAS RN	carcinogenicity*		
			obsd	MLR	ANN
181	ursodeoxycholic acid	128-13-2	1	0	0
182	valaciclovir	124832-26-4	0	1	0
183	valproic acid	99-66-1	1	0	0
184	venlafaxine	93413-69-5	0	0	0
185	zolpidem	82626-48-0	1	0	0
correct classification rate				62.2	64.8

*: "1" refers to carcinogenic positive; "0" is negative.

References

Benigni, R., Richard, A. M., 1996. QSARS of Mutagens and Carcinogens: Two Case Studies Illustrating Problems in the Construction of Models for Noncongeneric Chemicals, *Mutat. Res.* 371, 29-46.

Draper, N.R., Smith, H., 1966. *Applied Regression Analysis*, Wiley, New York.

Dunn III, W. J., Wold, S., 1981. An Assessment of Carcinogenicity of N-Nitroso Compounds by the SIMCA Method of Pattern Recognition, *J. Chem. Inf. Comput. Sci.*, 21, 8-13.

Hatch, K. L., Magee, P. S., 1998. A Discriminant Model for Allergic Contact dermatitis in Anthraquinone Disperse Dyes, *Quant. Struct.-Act. Relat.*, 17, 20-26.

Helma, C., King, R.D., Kramer, S., Srinivasan A. The Predictive Toxicology Challenge 2000-2001, http://www.informatik.uni-freiburg.de/-ml/ptc/.

Klopman, G., Chakravarti, S. K., Zhu, H., Ivanov, J. M., Saiakhov, R. D., 2004. ESP: A Method to Predict Toxicity and Pharmacological Properties of Chemicals Using Multiple MCASE Databases, *J. Chem. Inf. Comput. Sci.*, 44, 704-715.

Mazzatorta, P., Benfenati, E., Lorenzini, P., Vighi, M., 2004. QSAR in Ecotoxicology: An Overview of Modern Classification Techniques, *J. Chem. Inf. Comput. Sci.*, 44, 105-112.

Todeschini, R, Consonni, V., 2000. *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim.

Villemin, D., Cherqaoui, D., Mesbash, A., 1994. Predicting Carcinogenicity of Polycyclic Aromatic Hydrocarbons from Back-Propagation Neural Network, *J. Chem. Inf. Comput. Sci.*, 34, 1288-1293.

Zupan, J., Gasteiger, J., 1999. *Neural Networks for Chemistry and Drug Design*, 2nd ed., Wiley-VCH, Weinheim.