Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Pharmaceutical and Biomedical Analysis 38 (2005) 497–507
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

Study of probabilistic neural networks to classify the active compounds in medicinal plants

C.X. Xue [a], X.Y. Zhang [a], M.C. Liu [a], Z.D. Hu [a,*], B.T. Fan [b]

[a] *Department of Chemistry, Lanzhou University, Lanzhou, Gansu 73000, PR China*
[b] *Université Paris 7-denis Diderot, ITODYS, 1 rue Guy de La Brosse, 75005 Paris, France*

Accepted 20 January 2005
Available online 17 March 2005

Abstract

Probabilistic neural networks (PNNs) were utilized for the classifications of 102 active compounds from diverse medicinal plants with anticancer activity against human rhinopharyngocele cell line KB. Molecular descriptors calculated from structure alone were used to represent molecular structures. A subset of the calculated descriptors selected using factor correlation analysis and forward stepwise regression was used to construct the prediction models. Linear discriminant analysis (LDA) was also utilized to construct the classification model to compare the results with those obtained by PNNs. The accuracy of the training set, the cross-validation set, and the test set given by PNNs and LDA were 100, 92.3, 90.9% and 71.8, 92.3, 54.5%, respectively, which indicated that the results obtained by PNNs agree well with the experimental values of these compounds and also revealed the superiority of PNNs over LDA approach for the classification of anticancer activities of compounds. The models built in this work would be of potential help in the design of novel and more potent anticancer agents.

Keywords: Probabilistic neural networks; Rhinopharyngocele; Classification; Linear discriminant analysis; Medicinal plants

1. Introduction

Plants have been demonstrated to be a very viable source for the development of clinically relevant anticancer compounds. Cancer remains to be a major threat to the public health. Many medicinal plants have remarkable cancer-resistant effect and little side effect on patients. Therefore, medicinal plants therapy is well suited for many of the patients suffering from cancer. A lot of natural products have been found to exhibit cytotoxic activity against human tumor cell lines [1]. Plants that are known to have anticancer activity are worth investigating. However, structural factors that are required for the anticancer activity on these compounds are still unknown. One of the useful tools in rational drug design is by the use of a quantitative structure–activity relationship (QSAR) analysis, especially when the structure and property of the bioreceptor remain unclear. Recently, artificial neural networks have found a widespread use for classification tasks and function approximation in many fields of chemistry and bioinformatics [2]. There are various types of neural networks that can be used for these problems. Among them, the probabilistic neural networks (PNNs) provide a very general and powerful classification paradigm when there is adequate data of known classification. The main advantage of PNNs is that it can be effectively used to sparse data [3,4].

In the present work, a variety of 102 active compounds extracted from medicinal plants were further screened for preliminary in vitro testing against human rhinopharyngocele cell line KB. The aim of the present paper is that for the first time to develop a prediction model for these 102 compounds by the use of multiple linear regression (MLR) and PNNs, and also to find the essential structural features for anticancer agents against human rhinopharyngocele cell line KB. Linear discriminant analysis (LDA) was also used to establish a classification model to compare the results with that obtained by PNNs.

* Corresponding author. Tel.: +86 931 891 2578; fax: +86 931 891 2582.
E-mail addresses: huzd@lzu.edu.cn, snowmoun@21cn.com (Z.D. Hu).

0731-7085/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2005.01.035

2. Experimental

2.1. Data sets

All the anticancer activity values used in this work were collected from handbook [1]. A complete list of the compounds name and corresponding anticancer activity was listed in Table 1 . The molecular structures were shown in Fig. 1. The majority of the tested compounds are efficient antitumor agents showing ED_{50} (the dose that inhibited 50% control growth of KB cells) values from 0.000026 to 26.0 μg/ml. Compared with previous work [5–10], the compounds studied in our investigation were more diverse. It is difficult to build a QSAR model simply by their activity value because there is a very low similarity of the complex structure. So, the compounds were divided into four classifications according to their anticancer activity: higher, high, moderate and low activity anticancer agent with ED_{50} values from 0.001 to 0.1, 0.1 to 1.0, 1.0 to 10.0 and over 10 which were represented by '++++', '+++', '++' and '+', respectively. The data was randomly divided into the training set, the cross-validation set and the test set. The training set and cross-validation set were used to adjust the parameters of PNNs. The test set was used to evaluate the performance of the trained network.

2.2. Quantum chemical and topological descriptors

To develop a QSAR, molecular structures are often represented using molecular descriptors, which encode much structural information. In recent years there has been a shift from empirical parameters to purely calculated descriptors, such as quantum chemical descriptors and topological indices. The advantage of these calculated descriptors over other empirical descriptors is the possibility to calculate descriptors solely from molecular structure and apply them to sets of structurally diverse compounds.

All molecules were drawn into Hyperchem [11] and preoptimized using MM+ molecular mechanics force field. A more precise optimization was done with semi-empirical PM3 method in Hyperchem and thereafter quantum chemical descriptors were obtained. All calculations were carried out at restricted Hartree Fock level with no configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient was 0.001. The quantum chemical descriptors include information about binding and formation energies, dipole moment, and molecular orbital energy levels. Topological descriptors include valence and non-valence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule by TOPIX [12], encoding information about the size, composition and the degree of branching of a molecule.

2.3. Theory of MLR

In MLR analysis, the descriptors in the regression equation must be independent variable. To reduce the number of the descriptors and minimize the information overlap in the descriptors, the concept of non-redundant descriptors (NRD) [13] was used. The linear correlation coefficients value of the two descriptors should be less than 0.9.

Once descriptors were generated, a forward stepwise regression method was used to develop the linear model of the property of interest, which is shown as follows:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \quad (1)$$

where, Y is the property, that is, the dependent variable, $X_1 - X_n$ represent the specific descriptor, while $b_1 - b_n$ represent the coefficients of those descriptors, and b_0 the intercept of this equation.

2.4. Theory of LDA

The basic theory of linear discriminant analyze is to classifies the dependent by dividing an n-dimensional descriptor space into two regions that are separated by a hyperplane which defined by a linear discriminant function [14], for more than two groups, a set of discriminant functions are generated. The regions formed by the hyperplane correspond to the classes to which individual compounds are predicted to belong.

2.5. Theory of PNNs

PNNs was developed by Specht and has been well described in Refs. [3,4]. Here, we only give a brief description of its principle. The PNNs architecture is distinct from that of a standard back-propagation neural network and provides superior performance in classification applications [3,4,15]. The PNNs operates by defining a probability density function (PDF) for each data class based on the training set data and an optimized kernel width parameter (σ) [16–19]. The basic architecture of PNNs is shown in Fig. 2. It consists of an input layer, a pattern layer, a summation layer and an output layer. At each neuron in the pattern layer, the dot product distance, d, is computed and then processed through a nonlinear transfer function as:

$$\text{output} = \exp - \left(\frac{(1-d)}{\sigma^2} \right) \quad (2)$$

The summation layer sums the outputs from all hidden neurons of each respective data class. The products of the summation layer are forwarded to the output layer.

The calculation programs implementing PNNs were written in M-file based on basis MATLAB [20] script for probabilistic neural networks. All computation was performed on a Pentium IV computer with 256 MB RAM working under MS Windows XP.

Table 1
Studied compounds and the data used in this work

No.	Name	E_{LOMO}	Chi3	Chi4	NrBR	NrRI	DiEM	Pola	ED_{50}	$\log(1/ED_{50})$
1	Acanthamolide	−0.58	8.89	6.73	10	2	4.10	43	2.2000	−0.34
2	Acanthoglabrolide	−0.48	10.44	7.47	12	2	4.18	51	3.1000	−0.49
3	Acantholide	−0.48	8.89	6.73	10	2	4.10	43	2.2000	−0.34
4	Acanthospermolide	−0.51	9.13	7.04	10	2	3.93	45	0.5400	0.27
5	3β-Acetoxynorerthrosuamine	−0.28	13.30	11.27	12	3	5.39	72	0.0030	2.52
6	Acetylglaucarubinone	−0.55	16.64	14.43	17	5	4.91	92	0.0010	3.00
7	Ailanthinone	−0.38	15.40	13.17	16	5	4.52	84	0.0010	3.00
8	Allamardicin	−0.33	9.26	8.76	10	4	3.98	44	10.0000	−1.00
9	Allamandin	−0.36	9.42	8.98	10	4	3.90	45	2.1000	−0.32
10	Allamdin	−0.45	8.36	7.26	8	3	3.96	38	10.0000	−1.00
11	Amaralin	−0.30	8.38	7.22	10	4	3.34	39	4.9000	−0.69
12	Arnebin	−1.08	8.72	7.66	11	2	4.61	44	25.0000	−1.40
13	Aromaticin	−0.33	7.77	6.23	8	3	3.2	36	2.0000	−0.30
14	Autumnolide	−0.03	9.03	7.86	11	4	3.29	44	3.1000	−0.49
15	Baccharin	−0.26	16.99	13.62	15	7	5.12	84	0.0001	4.00
16	Baileyin	−0.30	7.07	5.77	8	3	3.50	30	16.0000	−1.20
17	Baileyolin	−0.61	13.07	11.56	11	5	4.77	67	0.0280	1.55
18	Bersenogenin	−0.52	13.4	11.83	10	5	4.67	70	0.0046	2.34
19	Bruceantin	−0.03	16.8	14.46	18	5	4.94	93	0.0010	3.00
20	Bruceantinol	−0.76	17.91	15.73	19	5	5.36	100	0.0010	3.00
21	Chaparrinone	−0.72	12.78	11.52	13	5	3.61	70	0.1420	0.85
22	Chelerythrine methanolate	−0.77	11.72	10.81	12	5	4.52	57	4.0000	−0.60
23	Chrysin	−1.07	6.93	6.42	7	3	3.94	31	13.0000	−1.11
24	Cissampareine	−0.47	17.96	15.73	17	7	5.49	88	1.1000	−0.04
25	Cnicin	−0.18	9.61	7.07	10	2	4.55	45	3.4000	−0.53
26	Costunolide	−0.07	5.94	4.56	6	2	3.36	26	0.2600	0.59
27	Cryptopleurine	−0.58	11.47	10.07	11	5	4.43	55	2.6e−5	−4.59
28	Cucurbitacin B	0.050	16.56	14.06	16	4	5.54	91	0.0050	2.30
29	Cucurbitacin D	−0.56	15.68	12.89	15	4	5.19	86	0.0050	2.30
30	Cucurbitacin E	−0.51	16.56	14.06	16	4	5.54	91	0.0100	2.00
31	Cucurbitacin I	−0.66	15.68	12.89	15	4	5.19	86	0.0050	2.30
32	Cucurbitacin L	0.01	15.68	12.89	15	4	5.19	86	0.0100	2.00
33	Cucurbitacin P	0.41	15.68	12.89	15	4	5.19	86	0.5400	0.27
34	Cucurbitacin Q	−0.09	16.56	14.06	16	4	5.54	91	0.0320	1.49
35	Cymarin	−0.24	17.14	14.59	14	6	5.74	85	1.0000	0.00
36	Damsin	−0.12	7.93	6.37	8	3	3.02	38	0.3200	0.49
37	Deacetyleupaserrin	−0.38	8.82	7.60	10	2	4.32	43	0.2900	0.54
38	Demethyldeoxypodophyllotoxin	−0.18	11.63	10.29	12	5	4.31	52	0.0012	2.92
39	Deoxypodophyllotoxin	−0.23	11.83	10.74	12	5	4.43	54	20.0000	−1.30
40	3-Desmethylcolchicine	−0.60	10.11	8.99	11	3	4.27	57	0.0240	1.62
41	Dihydroacanthospermal	−0.23	10.76	7.53	12	2	4.23	53	2.6000	−0.42
42	Elephantin	−0.92	10.01	8.73	12	4	4.27	45	0.2800	0.55
43	Elephantopin	−0.77	10.24	8.06	12	4	3.99	45	0.2800	0.55
44	3-Epiberscillogenin	−0.91	13.07	11.56	11	5	4.77	67	0.6200	0.21
45	10-Epieupatoroxin	−0.36	12.31	10.12	13	5	4.19	61	2.6000	−0.42
46	Epitulipindide	−0.19	6.85	5.83	8	2	3.64	33	2.1000	−0.32
47	Epitulipinolide diepoxide	−0.22	8.82	6.97	10	4	3.62	38	0.3400	0.47
48	Eremantholide A	−0.56	11.36	9.28	11	4	3.71	57	2.0000	−0.30
49	Eupachlorin	−0.19	11.54	9.42	12	3	4.16	60	0.2100	0.68
50	Eupachlorin acetate	−0.30	12.15	9.94	13	3	4.47	64	0.1800	0.74
51	Eupachloroxin	−0.31	12.62	10.21	13	4	4.19	64	0.2100	0.68
52	Eupacunin	−0.33	10.13	7.50	12	2	4.65	49	2.1000	−0.32
53	Eupacunolin	−0.51	10.45	7.84	12	2	4.66	51	3.7000	−0.57
54	Eupacunoxin	−0.08	11.13	8.12	13	3	4.7	51	2.1000	−0.32
55	Eupafolin	−1.12	8.95	7.57	10	3	4.51	42	18.0000	−1.26
56	Euparotin	−0.21	11.23	9.32	12	4	4.16	57	0.2100	0.68
57	Euparotin acetate	−0.30	11.83	9.85	13	4	4.48	61	0.2100	0.68
58	Eupaserrin	−0.22	9.41	7.94	11	2	4.8	46	0.2300	0.64
59	Eupatilin	−1.17	9.54	8.39	10	3	4.75	46	45.0000	−1.65
60	Eupatin	−0.90	10.24	8.62	11	3	4.71	50	4.6000	−0.66
61	Eupatocunin	0.04	10.30	7.51	12	2	4.64	50	0.1100	0.96
62	Eupatocunoxin	0.06	11.24	8.36	13	3	4.65	52	1.7000	−0.23
63	Eupatorin	−1.18	9.55	8.32	10	3	4.74	46	4.2000	−0.62

Table 1 (*Continued*)

No.	Name	E_{LOMO}	Chi3	Chi4	NrBR	NrRI	DiEM	Pola	ED_{50}	$\log(1/ED_{50})$
64	Eupatoroxin	−0.19	12.31	10.12	13	5	4.19	61	2.8000	−0.45
65	Eupatundin	−0.17	11.71	9.59	13	4	4.21	58	0.3900	0.41
66	Fastigilin A	−0.24	11.06	8.30	12	3	3.90	55	3.9000	−0.59
67	Fastigilin B	−0.20	10.35	8.60	12	3	4.00	53	14.0000	−1.15
68	Fastigilin C	−0.32	10.35	8.6	12	3	4.00	53	1.0000	0.00
69	Gaillardin	−0.28	8.29	7.32	10	3	3.96	41	0.8000	0.10
70	Genistein	−0.82	7.50	6.62	8	3	4.18	34	7.4000	−0.87
71	Glabratolide	−0.33	8.31	6.19	9	2	3.99	39	2.3000	−0.36
72	Glaucarubinone	−0.23	16.01	13.18	16	5	4.58	87	0.0250	1.60
73	Glaziovine	−0.69	9.24	8.02	9	4	3.73	44	2.6000	−0.42
74	Heliotrine	0.97	8.60	6.31	8	2	4.36	37	15.0000	−1.18
75	9-Hydroxyglabratolide	−0.28	8.92	6.63	10	2	3.95	43	2.0000	−0.30
76	Isochamanetin	−0.56	10.35	8.93	10	4	4.98	47	5.3000	−0.72
77	Isocucurbitacin D	0.12	15.68	12.89	15	4	5.19	86	0.0240	1.62
78	Isopicropodophyllone	−0.99	12.4	11.17	13	5	4.44	58	3.2000	−0.51
79	Jacaranone	−0.85	4.02	2.65	3	1	3.40	16	2.1000	−0.32
80	Jatrophone	−0.18	9.03	7.2	9	3	3.58	43	0.1700	0.77
81	Lipiferolide	−0.10	7.85	6.44	9	3	3.64	36	0.1600	0.80
82	Nobiletin	−1.03	10.88	9.7	11	3	4.82	56	3.0000	−0.48
83	Odoratin	−0.37	8.26	6.57	9	3	3.37	39	4.0000	−0.60
84	Pleniradin	−0.16	7.68	6.79	9	3	3.40	37	14.0000	−1.15
85	Provincialin	−0.56	12.24	9.74	13	2	5.62	60	3.5000	−0.54
86	Psorospermin	−0.67	10.55	8.89	11	5	4.22	48	0.1000	1.00
87	Quassimarin	−0.74	16.67	14.39	17	5	4.90	92	0.0100	2.00
88	Radiatin	−0.28	10.58	7.94	12	3	3.68	53	1.6000	−0.20
89	Simalikalactone D	−0.78	15.43	13.14	16	5	4.52	84	0.0010	3.00
90	Taxodione	−1.53	9.10	8.65	10	3	3.84	49	3.0000	−0.48
91	Taxodone	−1.18	9.10	8.65	10	3	3.84	49	0.6000	0.22
92	Tripdiolide	−0.76	13.03	12.43	14	7	4.16	68	0.0042	2.38
93	Triptolide	−0.47	12.75	11.76	13	7	4.15	65	0.0017	2.77
94	Triptonide	−0.61	12.75	11.76	13	7	4.15	65	0.0001	4.00
95	Tulipalin	−0.21	2.29	1.33	2	1	1.09	5	16.0000	−1.20
96	Tulipinolide	−0.04	7.00	5.78	8	2	3.59	34	0.4600	0.34
97	Vernodalin	−0.60	10.58	8.24	11	3	4.26	51	1.8000	−0.26
98	Vernolepin	−0.33	8.46	7.17	9	3	3.33	41	1.7000	−0.23
99	Vernolide	0.01	10.35	7.99	11	4	3.97	50	2.0000	−0.30
100	Vernomenin	−0.62	8.54	7.23	9	3	3.43	41	20.0000	−1.30
101	Vernomygdin	−0.22	10.3	7.83	11	4	4.02	49	1.5000	−0.18
102	Xerantholide	−0.40	7.57	6.04	9	3	3.41	34	1.5030	−0.18

Definition of the descriptors were given in Tables 2 and 3.

3. Results and discussion

3.1. Results of MLR

More than 40 non-empirical molecular descriptors, which encode the essential structural features of the molecules, were calculated for each compound. The full list of the descriptors calculated is given in Table 2. Forward stepwise regression routine was used to develop the linear model for the prediction of $\log(1/ED_{50})$ using calculated structural descriptors. The best linear model contains seven molecular descriptors. Of them, one is quantum chemical, one is electrochemical and five are topological descriptors. The best seven-descriptor correlation model was shown in detail in Table 3. This model produced a correlation coefficient of 0.773 for the compounds.

From Table 3, it can be seen that there is no simple linear correlation between the anticancer of medicinal plants and the input parameters. Therefore, LDA and PNNs were applied to classify the anticancer values based on the same subset of descriptors. For the purposes of modeling, a value of 1, 2, 3 and 4 was assigned to compounds with low, moderate, high, higher anticancer activities, respectively.

3.2. PNNs structure optimization

The most important parameter that needs to be determined to obtain an optimal PNNs is the spread parameter (σ) of the random variables. The selection of this value is crucial because it determines the shape of the Gaussian function. A large radius possesses a smooth shape and has the advantage of interpolation, and a small radius leads to a sharp shape and reduces the overlap between adjacent samples [21]. But too small a spread cannot generalize well, because unknown samples only lie in the region that Gaussian function enclosing can be generalized. To optimize the radius, 13 samples were used as a cross-validation set. A trial and error method was used to find the best radius. The absolute error (AE) was

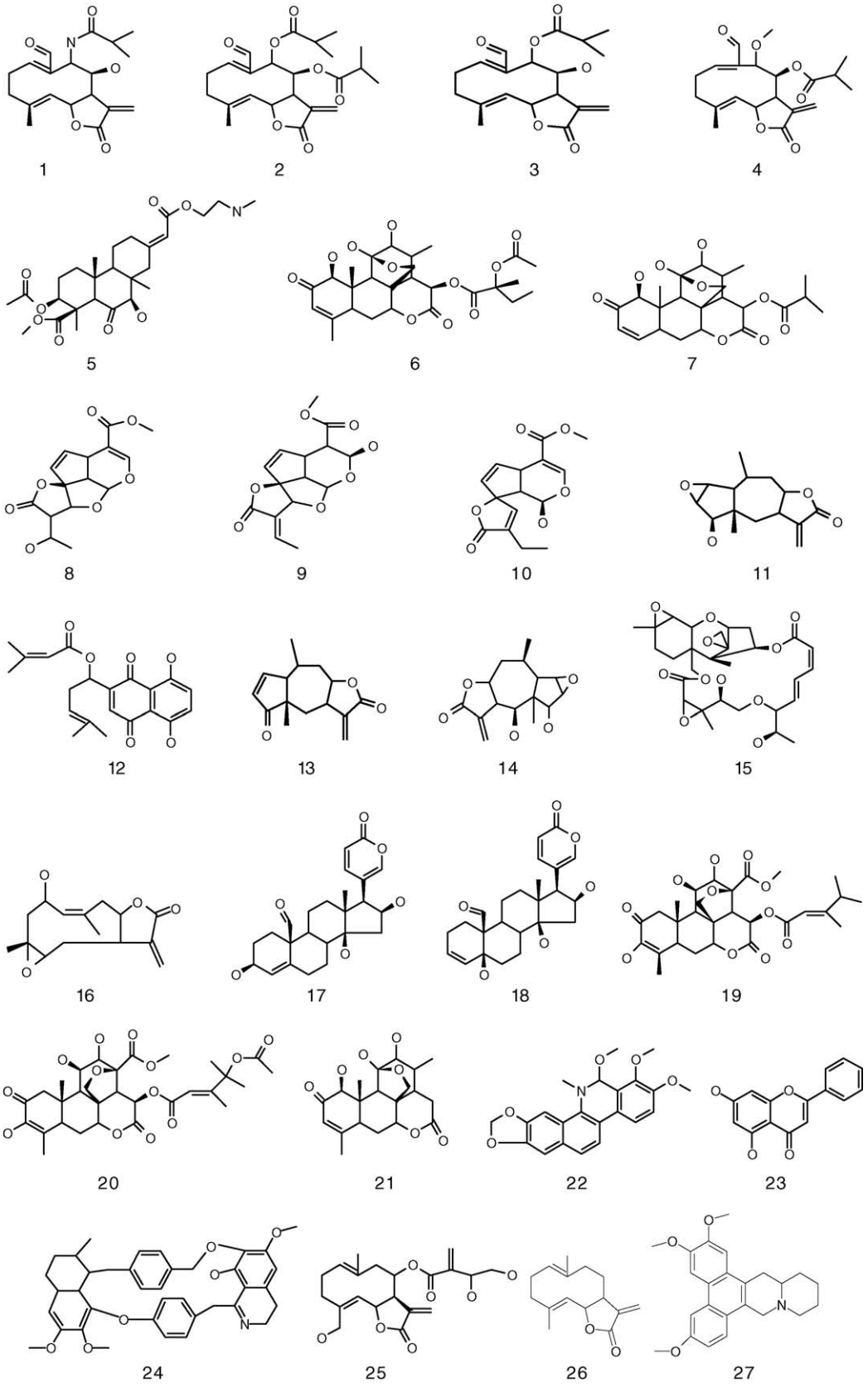

Fig. 1. Structures of 102 active compounds.

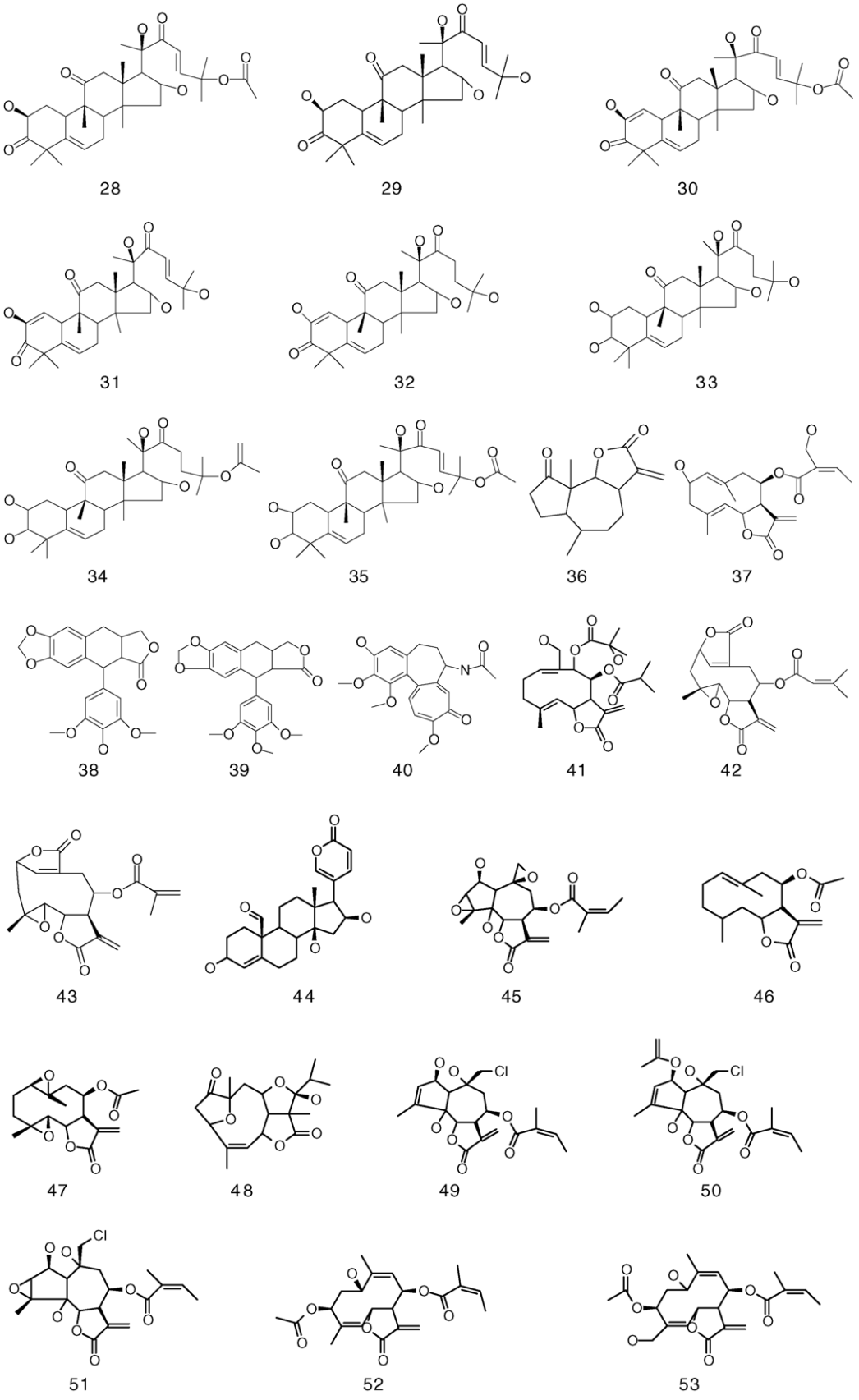

Fig. 1. (*Continued*).

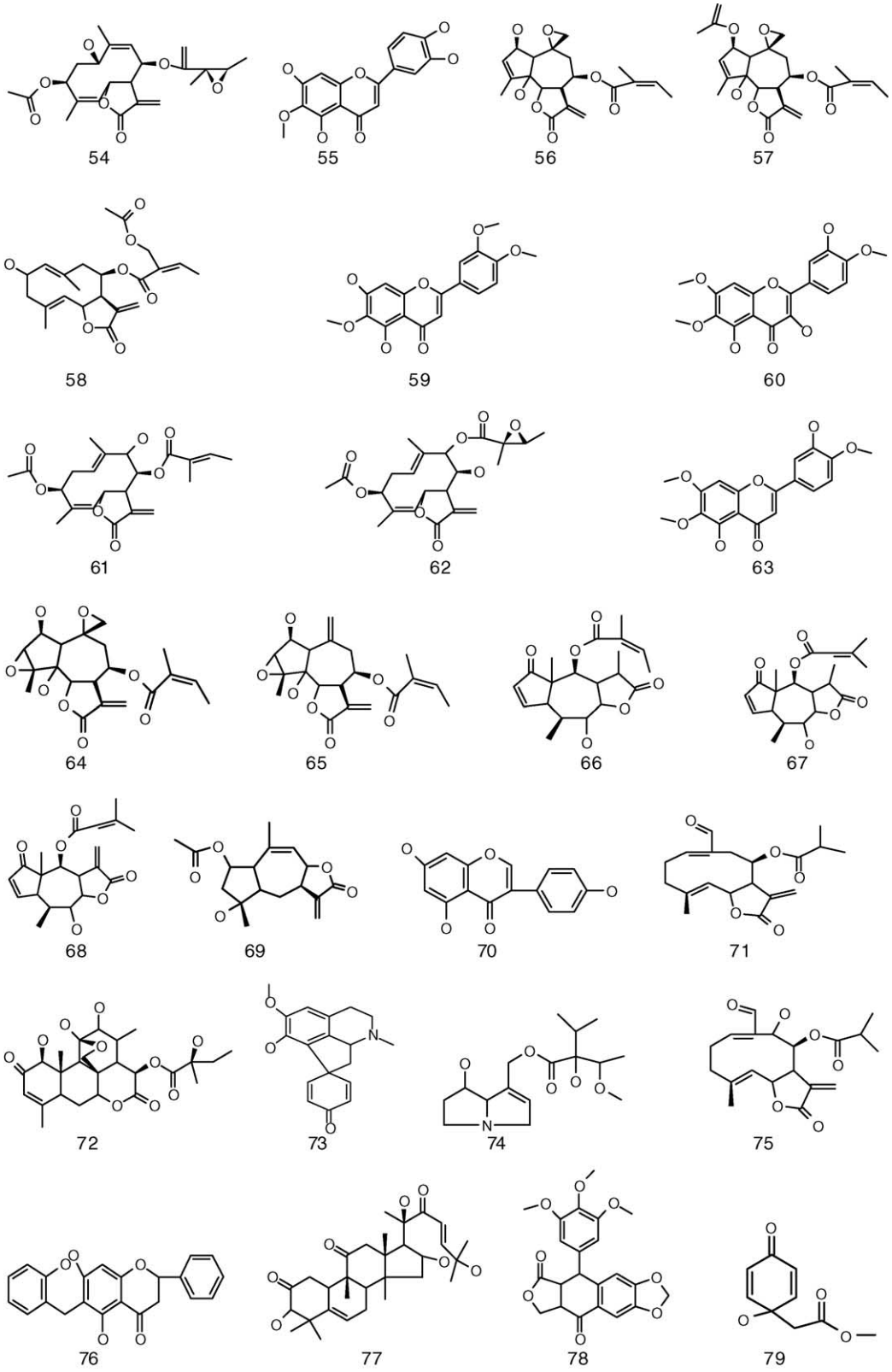

Fig. 1. (*Continued*).

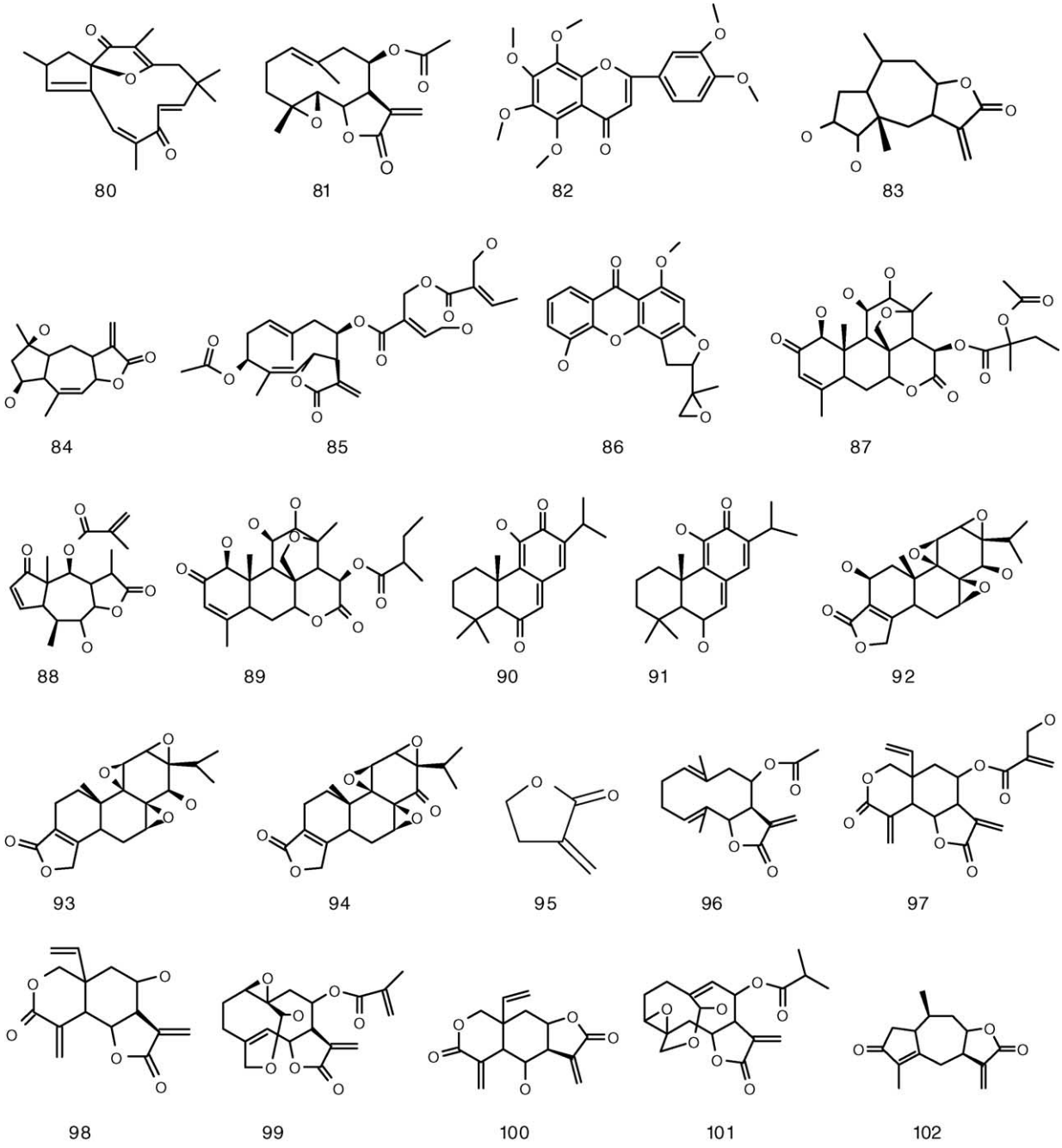

Fig. 1. (*Continued*).

used as the error function, and it is computed according to the following:

$$\mathrm{AE} = \mathrm{sum}(\mathrm{abs}(d_i - o_i)) \quad (3)$$

where d_i is the desired output in the cross-validation set and o_i the actual output.

To obtain the optimal radius, the neural networks with different radii were trained, the spread varying from 1.00 to 1.45. The AE was calculated on different radii, according to the generalization ability on the cross-validation set in order to determine the optimal radius. The curve of AE versus the radius is shown in Fig. 3. From Fig. 3, the optimal radius was found as 1.30.

3.3. Results of PNNs and LDA

From the above discussion, the radius of hidden layer nodes was fixed to 1.30. The predicted results of the optimal neural network were shown in Table 4. The number of compounds which were misclassified in the training set, the cross-validation set, and the test set are 0, 1, and 1, respectively, and the corresponding accuracy are 100, 92.31, and

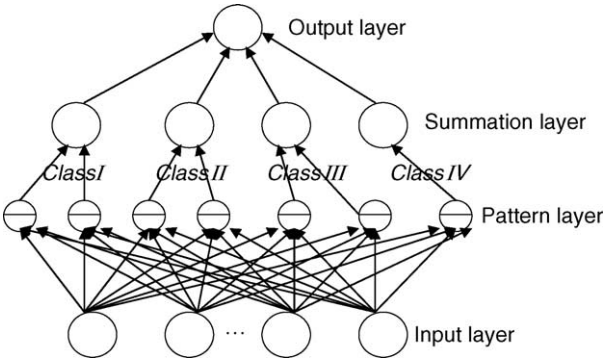

Fig. 2. Structure of probabilistic neural networks.

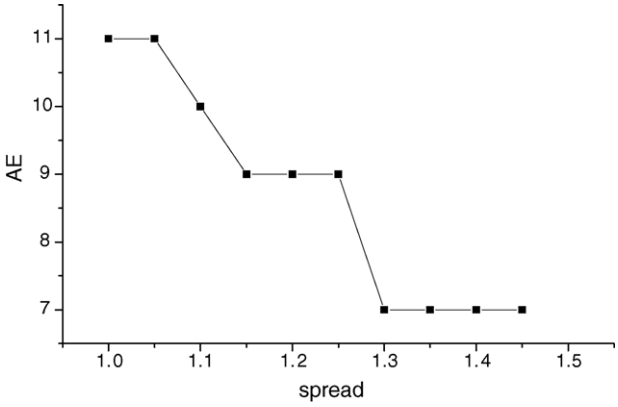

Fig. 3. The spread vs. AE error on cross-validation set.

Table 2
Full list of the descriptors calculated in this study

Descriptor	Descriptor
Total energy	Number of branches
Binding energy	Number of rings
Isolated atomic energy	Wiener index
Electronic energy	Information wiener
Core-core interaction	Distance equality mean
Heat of formation	Distance equality total
Dipole	Polarity
Surface area	Wiener index on distance code
Volume	Modified randic index
Hydration energy	First zagreb index
log p	Second zagreb index
Refractivity	Balanba
Polarizability	Dimension index
Molecular weight	Number of C atoms
Homo energy level	Number of O atoms
Lumo energy level	Number of N atoms
Chi0 index	Number of single bonds
Chi1 index	Number of double bonds
Chi2 index	Number of triple bonds
Chi3 index	Number of aromatic bonds
Chi4 index	

90.91%. In Table 4, we can remark that the misclassification of nos. 84 and 96 are relatively important. These two predicted results are not completely agreeing with the experiment results, but they lie on the borders (nos.: 84 (+/++); 96 (+++/++)).

After the establishment of PNNs model, LDA was used to build another classification model to compare the results with that obtained by PNNs. LDA was performed using the SPSS statistical software [22]. In this work, the prior probabilities were computed from group size. The predicted accuracy for the training set, the cross-validation set, and the test set was 71.8, 92.3, and 54.5%, respectively. By comparison of the results obtained by PNNs and LDA, it could be seen that the results obtained by PNNs were better than that obtained by LDA.

3.4. Discussion of the descriptors

By interpreting of the descriptors used in this work, it is possible to gain some insight into factors that are likely to govern the anticancer activity of the active compounds in medicinal plants. (1) The standardized regression coefficients reveal the significance of an individual descriptor presented in the regression model. Obviously, in Table 3, the effect of number of rings (NoRI) on the activity of the anticancer is more significant than that of the other descriptors. From Table 1, it can be seen that the compounds which have 6 and 7 rings are all the highest anticancer agents, the ones which have 5 rings are almost the highest anticancer agents. (2) The drugs take effect on organism by the molecular interactions between them. These interactions commonly

Table 3
MLR results on the correlation between input parameters and the activity values, $\log(1/ED_{50})$

Chemical meaning	Descriptor	Coefficient	S.E.	Standardized coefficients	*T*-value
Intercept	Constant	−1.311	0.722		−1.815
Lumo energy level	E_{LOMO}	0.333	0.280	0.093	1.190
Chi3 index	Chi3	−0.844	0.302	−2.017	−2.798
Chi4 index	Chi4	−0.577	0.265	−1.243	−2.175
Number of branches	NrBR	−4.228E−02	0.093	−.093	−0.453
Number of rings	NrRI	0.753	0.205	0.760	3.683
Distance equality mean	DiEM	0.328	0.306	0.174	1.074
Polarity	Pola	0.232	0.049	3.356	4.706
R (correlation coefficient)	0.773				
F-value	196.65				

Table 4
Predicting results of training set, cross-validation set and test set by PNNs model

No.	Class	Predicted	No.	Class	Predicted	No.	Class	Predicted
1	++	++	35	++++	++++	69[a]	++	++
2	++	++	36[b]	+++	+++	70	++	++
3	++	++	37	+++	+++	71	++	++
4[b]	+++	+++	38	++++	++++	72	++++	++++
5[a]	++++	++++	39	+	+	73	++	++
6	++++	++++	40	++++	++++	74	+	+
7	++++	++++	41[a]	++	++	75	++	++
8	+	+	42	+++	+++	76[b]	++	++
9	++	++	43	+++	+++	77	++++	++++
10	+	+	44[b]	+++	+++	78[a]	++	++
11	++	++	45	++	++	79	++	++
12[b]	+	+	46	++	++	80	+++	+++
13	++	++	47	+++	+++	81	+++	+++
14[a]	++	++	48	++	++	82	++	++
15	++++	++++	49	+++	+++	83	++	++
16	+	+	50[a]	+++	+++	84[b]	+	++
17	++++	++++	51	+++	+++	85	++	++
18	++++	++++	52[b]	++	++	86	++++	++++
19	++++	++++	53	++	++	87[a]	++++	++++
20[b]	++++	++++	54	++	++	88	++	++
21	+++	+++	55	+	+	89	++++	++++
22	++	++	56	+++	+++	90	++	++
23[a]	+	+	57	+++	+++	91	+++	+++
24	++	++	58	+++	+++	92[b]	++++	++++
25	++	++	59[a]	+	+	93	++++	++++
26	+++	+++	60[b]	++	++	94	++++	++++
27	++++	++++	61	+++	+++	95	+	+
28[b]	++++	++++	62	++	++	96[a]	+++	++
29	++++	++++	63	++	++	97	++	++
30	++++	++++	64	++	++	98	++	++
31	++++	++++	65	+++	+++	99	++	++
32[a]	++++	++++	66	++	++	100[b]	+	+
33	+++	+++	67	+	+	101	++	++
34	++++	++++	68[b]	+++	+++	102	++	++

[a] Test set.
[b] Cross-validation set.

include the bond of charge transfer, H-bond and dispersion interaction [23]. E_{LUMO} is the energy of the lowest unoccupied orbital and describes the electrophilicity ability of a molecule and also the ability of a molecule to accept electrons. According to frontier molecule orbital (FMO) theory, frontier orbital energies control chemical reactivity. E_{LUMO} can be considered as a measure of a compound's susceptibility to nucleophilic attack [24]. The positive coefficient of E_{LUMO} in MLR model shows that increasing the energy of LUMO causes the anticancer activity increase in this work. (3) The positive coefficient of polarity (Pola) also shows that increasing the polarity of molecule results in the anticancer activity increases, it is probably due to the fact that the molecule with high polarity can easily accommodate its charge when it reacts to other molecule, consequently, results in the high anticancer activity. (4) Meanwhile, the negative relationship between anticancer activity and Chi3 index (Chi3), Chi4 index (Chi4) and number of branches (NoBR) reveals that increasing these parameters of molecule decreases the activity.

4. Conclusion

A multiple linear regression study was conducted on 102 diverse active compounds extracted from medicinal plants. More than 40 descriptors were calculated for each molecule. The best set of calculated descriptors was selected by factor correlation analysis and forward stepwise regression. Probabilistic neural networks and linear discriminant analysis were then applied to classify the anticancer values based on the same subset of descriptors. The optimization of PNNs structure is easier and faster compared with back-propagation (BP) neural networks, because there is only one adjustable parameter. The predictive results of PNNs are consistent with the experimental data and better than that obtained by LDA. Some conclusions were drawn to give insight into the local molecular features that determine the anticancer activity of these compounds. Therefore the model developed in this paper is a good and simple approach for predicting the expected anticancer classification of molecules and is very helpful to

search and screen potent anticancer drug from medicinal plants.

Acknowledgments

This work was supported by the Ministry of Sciences and Technology of China and the Ministry of Foreign Affairs of France (AFCRST PRA SI 00-05).

References

[1] Handbook of actived compounds of medicinal plants. Traditional Chinese Medicinal information center of national medical bureau, Beijing, Publication of people's health, 1986.
[2] G. Schneider, P. Wrede, Prog. Biophys. Mol. Biol. 70 (1998) 175–222.
[3] D.F. Specht, Probabilistic neural networks for classification, mapping, or associative memory, in: Proceedings of the IEEE International Conference on Neural Networks, vol. 1, San Diego, 1988, pp. 523–525.
[4] D.F. Specht, Probabilistic neural networks, Neural Netw. 3 (1990) 109–118.
[5] E. Mongelli, S. Pampuro, J. Coussio, H. Salomon, G. Ciccia, J. Ethnopharm. 71 (2000) 145–151.
[6] J. Popoca, A. Aguilar, D. Alonso, M.L. Villarreal, J. Ethnopharmacol. 59 (1998) 173–177.
[7] A. Ankli, M. Heinrich, P. Bork, L. Wolfram, P. Bauerfeind, R. Brun, C. Schmid, C. Weiss, R. Bruggisser, J. Gertsch, M. Wasescha, O. Sticher, J. Ethnopharmacol. 79 (2002) 43–52.
[8] T. Yanagisawa, M. Urade, Y. Takahashi, H. Kishimoto, K. Sakurai, Oral Oncol. 34 (1998) 30–38.
[9] E. Mongelli, C. Desmarchelier, J.T. Rodríguez, J. Coussio, G. Ciccia, J. Ethnopharmacol. 58 (1997) 157–163.
[10] A. Santa Maria, A. Lopez, M.M. Diaz, J. Albán, A. Galán de Mera, J.A. Vicente Orellana, J.M. Pozuelo, J. Ethnopharmacol. 57 (1997) 183–187.
[11] HyperChem 4.0, Hypercube Inc., 1994.
[12] D. Svozil, H. Lohninger, Topix Version 1.2, 1999 http://www.lohninger.com/topix.html.
[13] A.R. Katritzky, V. Gordeeva, J. Chem. Inf. Comput. Sci. 33 (1993) 835–857.
[14] S.K. Kachigan, Statistical Analysis, Radius Press, New York, 1986.
[15] J.R. Long, H.T. Mayfield, M.V. Henley, Anal. Chem. 63 (1991) 1256–1261.
[16] M. Hajmeer, I. Basheer, J. Mic. Meth. 51 (2002) 217–226.
[17] R.E. Shaffer, S.L. Rose-Pehrsson, R.A. McGill, Anal. Chim. Acta 384 (1999) 305–317.
[18] Y. Chtioui, D. Bertrand, D. Barba, Chemom. Intell. Lab. Syst. 35 (1996) 175–186.
[19] G.R. Magelssen, J.W. Elling, J. Chromatogr. A 775 (1997) 231–242.
[20] MATLAB 5.2, The Mathworks, Natick, MA, USA, 1998.
[21] R.E. Shaffer, S.L. Rose-Pehrsson, Anal. Chem. 71 (1999) 4263–4271.
[22] SPSS Version 10.0.5 for Windows SPSS Inc., Chicago, IL, http://www.spss.com/spss10/.
[23] L.S. Wang, S.K. Han, Molecule Structure, Property and Activity, Publication of Chemistry and Industry, Beijing, 1997, Chapter 2, p. 59.
[24] M.M. Lynam, J. Dombarsky, J. Koca, P. Adriaens, Environ. Toxicol. Chem. 17 (1998) 988–997.