Articles

Radial Basis Function Neural Networks Based QSPR for the Prediction of log *P*

YAO, Xiao-Jun[a](姚小军) LIU, Man-Cang[a](刘满仓) ZHANG, Xiao-Yun[a](张晓昀)
ZHANG, Rui-Sheng[a](张瑞生) HU, Zhi-De*,[a](胡之德) FAN, Bo-Tao[b](范波涛)
[a] *Department of Chemistry, Lanzhou University, Lanzhou, Gansu 730000, China*
[b] *Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France*

Quantitative structure-property relationship (QSPR) method is used to study the correlation models between the structures of a set of diverse organic compounds and their log *P*. Molecular descriptors calculated from structure alone are used to describe the molecular structures. A subset of the calculated descriptors, selected using forward stepwise regression, is used in the QSPR models development. Multiple linear regression (MLR) and radial basis function neural networks (RBFNNs) are utilized to construct the linear and non-linear correlation model, respectively. The optimal QSPR model developed is based on a 7-17-1 RBFNNs architecture using seven calculated molecular descriptors. The root mean square errors in predictions for the training, predicting and overall data sets are 0.284, 0.327 and 0.291 log *P* units, respectively.

Keywords radial basis function neural network, QSPR, molecular descriptor, log *P*

Introduction

Octanol-water partition coefficient is an important fundamental property of an organic compound. Its logarithm (log *P*) has been widely used to measure the hydrophobicity (or lipophilicity) of chemicals, which is of great importance in toxicology, pharmaceutical and environmental science. Many reports have shown its correlation with numerous physical and biological processes. As a result of the importance of this property, it would be very useful to develop predictive models for it. Many previous works have aimed at predicting log *P* and several excellent reviews are available.[1-5] The most promising method is to use QSPR, in which descriptors derived from molecular structure alone to predict log *P* have been employed. The advantage of this approach is that the descriptors used can be calculated from molecular structure alone and are not dependent on any experiment properties. To develop a QSPR, molecules must be described using molecular structural descriptors and retain as much structural information as possible. In recent years there has been a shift from purely empirical parameters to calculated descriptors, such as quantum chemistry and topological descriptors. After the calculation of molecular descriptors, linear methods, such as multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) or non-linear methods, such as neural networks can be used in the development of a mathematical relationship between the structural descriptors and the property. Neural networks are particularly useful in cases where it is difficult to specify an exact mathematical model, which describes a specific structure-property relationship.[6,7] They have been widely used to predict physico-chemical properties.[8-17] Most of the previous models to calculate log *P* have been derived using MLR, and only limited work used neural networks trained by the back-propagation algorithm. Compared with BP neural networks, the parameters of radial basis function neural networks (RBFNNs) can be adjusted by fast linear methods. The optimization of its topology and learning parameters are easy to implement.[18,19] Many problems in chemistry and chemical engineering have been successful-

* E-mail: huzd@lzu.edu.cn
Received December 6, 2001; revised March 20, 2002; accepted March 27, 2002.
Project supported by the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) (PRA SI 00-05).

ly solved by the use of RBFNNs: multivariate calibration,[20,21] QSPR[22,23] and classification.[24,25]

The goal of the present work is to extend our previous work[26,27] and establish a QSPR model that can be used for the prediction of log P of organic compounds from their molecular structures. MLR and RBFNNs are utilized to establish quantitative linear and non-linear relationship between log P and molecular descriptors, respectively.

Methodology

All log P data of 271 compounds in the present investigation were taken from the literature.[5] A complete list of the compounds with experimental log P values is shown in Table 1. The data set was divided into two subsets: a training set of 232 compounds and a predicting set of 39 compounds (marked with asterisk). The training set was used to adjust the parameters of the RBFNNs and the predicting set was used to evaluate its prediction ability.

Table 1 Compounds and the predicted results of log P

No.	Compound	log P	MLR	RBFNNs
1	Acetic acid	−0.17	−0.271	−0.154
2[a]	2-Propanone	−0.24	0.034	0.047
3	Acetophnone	1.63	2.048	1.942
4	Allyl alcohol	0.17	0.293	0.151
5	Allyamine	0.03	−0.164	0.030
6	Aniline	0.90	0.667	0.837
7	Anisole	2.11	1.939	1.777
8	Benzaldehyde	1.48	1.637	1.601
9[a]	Benzene	2.13	2.262	2.397
10	Benzeneacetaldehyde	1.78	2.149	1.992
11	Benzeneacetic acid	1.41	1.920	2.072
12	Benzeneethanamine	1.41	1.591	1.667
13	Benzeneethanol	1.36	2.172	2.021
14	Benzenepropanol	1.88	2.721	2.484
15	Benzoic acid	1.88	1.533	1.719
16[a]	Benzyl acetate	1.96	2.348	2.486
17	Benzyl alcohol	1.05	1.747	1.583
18	Benzylamine	1.09	1.136	1.282
19	Benzyl methyl ether	1.35	2.321	2.179
20	Bromobenzene	2.99	2.722	2.835
21	2-Bromobenzoic acid	2.20	1.807	2.073
22	1-Bromobutane	2.75	2.750	2.829
23[a]	Bromochloromethane	1.41	1.520	1.598
24	Bromocyclohexane	3.20	3.297	3.326
25	Bromoethane	1.60	1.520	1.598
26	1-Bromoheptane	4.36	4.609	4.754
27	1-Bromohexane	3.80	3.986	4.106
28	Bromomethane	1.19	0.930	1.056
29	Benzylbromide	2.92	3.255	3.339
30[a]	1-Bromopropane	4.89	5.234	5.397
31	1-Bromopentane	3.37	3.371	3.467
32	1-Bromopropane	2.10	2.134	2.205
33	2-Bromopropane	1.90	1.953	2.015
34	3-Bromopropene	1.79	1.683	1.656
35	1,3-Butadiene	1.99	1.413	1.515
36	Butanal	0.88	0.733	0.598
37[a]	Butanoic acid	0.79	0.496	0.697
38	1-Butanol	0.84	1.162	0.877
39	2-Butanol	0.65	0.987	0.723
40	2-Butanone	0.29	0.579	0.481
41	*cis*-2-Butene	2.33	1.717	1.729
42	*trans*-2-Butene	2.31	1.709	1.710
43	Butyl acetate	1.82	1.490	1.703
44[a]	Butylamine	0.86	0.752	0.850
45	*tert*-Butylamine	0.40	0.431	0.566
46	Butylbenzene	4.26	4.373	4.363
47	Butyl methacrylate	2.88	2.190	2.455
48	*p*-Butylphenol	3.65	3.241	3.024
49	Chlorobenzene	2.84	2.722	2.835
50	1-Chlorobutane	2.64	2.750	2.829
51[a]	Chloroethane	1.43	1.520	1.598
52	Chloroethene	1.38	1.068	1.096
53	1-Chloroheptane	4.15	4.609	4.754
54	Chloromethane	0.91	0.930	1.056
55	1-Chloropropane	2.04	2.134	2.205
56	2-Chloropropane	1.90	1.953	2.015
57	*o*-Chlorotoluene	3.42	3.225	3.314
58[a]	*m*-Chlorotoluene	3.28	3.212	3.292
59	*p*-Chlorotoluene	3.33	3.229	3.314
60	*trans*-Cinnamic acid	2.13	2.382	2.542
61	Coronene	1.98	1.783	1.626
62	*m*-Cresol	1.98	1.779	1.617
63	*p*-Cresol	1.97	1.782	1.626
64	1,4-Cyclohexadiene	2.30	2.229	2.328
65[a]	Cyclohexane	3.44	2.780	2.858
66	Cyclohexanol	1.23	1.567	1.363
67	Cyclohexanone	0.81	1.231	1.178
68	Cyclohexene	2.86	2.445	2.414
69	2-Cyclohexen-1-one	0.61	1.129	1.090
70	Cyclohexylemine	1.49	1.147	1.223
71	Cyclooctane	4.45	3.939	3.944

Continued

No.	Compound	log *P*	MLR	RBFNNs
72[a]	Cyclopentane	3.00	2.191	2.287
73	Decane	6.25	5.859	6.030
74	Decanoic acid	4.09	3.478	3.949
75	1-Decanol	4.57	4.508	4.356
76	2-Decanone	3.77	3.961	3.646
77	Dibutyl ether	3.21	3.566	3.443
78	*o*-Dichlorobenzene	3.38	3.234	3.165
79[a]	*m*-Dichlorobenzene	3.48	3.212	3.292
80	*p*-Dichlorobenzene	3.38	3.229	3.314
81	Dichlorodifluoromethane	2.16	1.566	2.006
82	1,1-Dichloroethane	1.79	1.953	2.015
83	*cis*-1,2-Dichloroethene	1.86	1.713	1.710
84	*trans*-1,2-Dichloroethene	1.93	1.708	1.710
85	Dichloromethane	1.25	1.520	1.598
86[a]	1,2-Dichloropropane	2.00	2.134	2.205
87	*cis*-1,3-Dichloropropene	2.03	2.330	2.293
88	Diethylamine	0.58	0.760	0.850
89	Diethylcarbonate	1.21	1.302	1.738
90	Diethyl ether	0.89	1.340	1.043
91	Difluoromethane	0.20	0.809	0.202
92	Diisopropyl ether	1.52	2.097	1.872
93[a]	Dimethylamine	-0.38	-0.206	0.090
94	3,3-Dimethyl-2-butanol	1.48	1.770	1.575
95	Dimethyl ether	0.10	0.276	0.037
96	*N*, *N*-Dimethylformamide	-1.01	-1.286	-0.929
97	2,2-Dimethyl-1-propanol	1.31	1.339	1.121
98	Dipropylamine	1.67	1.799	1.847
99	Dipropyl ether	2.03	2.384	2.195
100[a]	Dodecanoic acid	4.60	4.623	5.081
101	1-Dodecanol	5.13	5.663	5.547
102	Epichlorohydrin	0.30	0.716	0.535
103	Ethanol	-0.30	0.080	-0.069
104	Ethyl acetate	0.73	0.508	0.689
105	Eethyl acrylate	1.32	0.866	1.054
106	Ethylamine	-0.13	-0.225	0.055
107[a]	*p*-Ethylaniline	1.96	1.505	1.561
108	Ethylbenzene	3.15	3.256	3.339
109	Ethyl benzoate	2.64	2.289	2.498
110	Ethylene oxide	-0.30	-0.250	-0.299
111	Ethyl methacrylate	1.94	1.629	1.963
112	Ethylmethylamine	0.15	0.286	0.424
113	*o*-Ethylphenol	2.47	2.213	2.077
114[a]	*m*-Ethylphenol	2.50	2.211	2.077
115	*p*-Ethylphenol	2.50	2.212	2.077
116	Ethyl propanoate	1.21	0.985	1.176
117	*o*-Ethyltoluene	3.53	3.767	3.827

Continued

No.	Compound	log *P*	MLR	RBFNNs
118	*p*-Ethyltoluene	3.63	3.766	3.827
119	Ethyl vinyl ether	1.04	1.010	0.766
120	Fluorobenzene	2.27	2.217	2.422
121[a]	Fluromethane	0.51	0.145	0.589
122	1-Fluoropentane	2.33	2.435	2.328
123	Formic acid	-0.54	-0.596	-0.642
124	Heptane	4.50	3.986	4.106
125	1-Heptanol	2.62	2.745	2.554
126	2-Heptanol	2.31	2.559	2.315
127	3-Heptanol	2.24	2.544	2.305
128[a]	4-Heptanol	2.22	2.545	2.305
129	2-Heptanone	1.98	2.204	1.983
130	1-Heptene	3.99	3.573	3.528
131	Heptylamine	2.57	2.341	2.381
132	Hexachlorobenzene	5.47	5.400	5.459
133	Hexachloroethane	4.00	4.069	4.053
134	Hexadecanoic acid	7.17	6.802	7.201
135[a]	1,5-Hexadiene	2.80	2.704	2.663
136	Hexanal	1.78	1.827	1.626
137	Hexanoic acid	1.92	1.460	1.715
138	1-Hexanol	2.03	2.199	1.970
139	2-Hexanol	1.76	2.001	1.751
140	3-Hexanol	1.65	2.002	1.751
141	2-Hexanone	1.38	1.659	1.451
142[a]	1-Hexene	3.40	2.947	2.904
143	Hexylamine	2.06	1.796	1.847
144	Hexylbenzene	5.52	5.529	5.407
145	1-Hexyne	2.73	2.580	2.403
146	5 * Hexyn-2-one	0.58	1.352	1.209
147	Iodobenzene	3.28	3.105	3.238
148	1-Iodobutane	3.00	2.912	2.956
149[a]	Iodoethane	2.00	1.910	2.038
150	1-Iodoheptane	4.70	4.494	4.569
151	Iodomethane	1.50	1.547	1.718
152	1-Iodopropane	2.50	2.348	2.462
153	Isobutylbenzene	4.01	4.312	4.319
154	Isopropylamine	0.26	0.070	0.230
155	Isopropylbenzene	3.66	3.745	3.779
156[a]	Isopropyl benzoate	3.18	2.615	2.905
157	1-Isopropyl-4-methylbenzene	4.10	4.267	4.262
158	Methacrylic acid	0.93	0.269	0.562
159	Methanol	-0.74	-0.403	-0.433
160	Methyl acetate	0.18	0.130	0.232
161	4-Methylacetophenone	2.19	2.419	2.320
162	Methyl acrylate	0.80	0.389	0.580
163[a]	Methylamine	-0.57	-0.539	-0.175

Continued

No.	Compound	log *P*	MLR	RBFNNs
164	*o*-Methylaniline	1.32	1.062	1.190
165	*m*-Methylaniline	1.40	1.046	1.153
166	*p*-Methylaniline	1.39	1.062	1.190
167	3-Methylanisole	2.66	2.334	2.220
168	4-Methylanisole	2.81	2.337	2.220
169	2-Methylbenzaldehyde	2.26	2.039	1.937
170[a]	α-Methylbenzeneacetic acid	1.80	2.284	2.456
171	3-Methylbenzenemethanol	1.60	2.141	2.005
172	4-Methylbenzenemethanol	1.58	2.141	2.005
173	Methyl benzoate	2.20	1.884	2.083
174	2-Methyl-2-butanol	0.89	1.325	1.115
175	3-Methyl-1-butanol	1.28	1.491	1.231
176	3-Methyl-2-butanol	1.28	1.392	1.150
177[a]	3-Methyl-2-butanone	0.56	0.981	0.906
178	Methyl tert-butyl ether	0.94	1.489	1.239
179	Methylcyclopentane	3.37	2.715	2.759
180	Methyl decanoate	4.41	4.124	4.510
181	5-Methyl-2-hexanone	1.88	2.086	1.881
182	Methyl methacrylate	1.38	0.769	0.974
183	5-Methyl-2-octanone	2.92	3.182	2.934
184[a]	Methyloxirane	0.03	0.221	0.086
185	4-Methyl-2-pentanone	1.31	1.558	1.372
186	4-Methyl-1-pentene	2.50	2.767	2.689
187	4-Methylphenyl acetate	2.11	2.332	2.447
188	2-Methyl-1-propanol	0.76	0.990	0.729
189	2-Methyl-2-propanol	0.35	0.870	0.653
190	2-Methyltetrahydrofuran	1.85	1.181	1.031
191[a]	2-Nonanone	4.02	3.946	3.760
192	1-Nanene	3.16	3.313	3.086
193	Octane	5.15	4.817	4.773
194	Octanoic acid	3.05	2.492	2.807
195	1-Octanol	3.07	3.382	3.156
196	2-Octanol	2.90	3.194	2.880
197	4-Octanol	2.68	3.181	2.867
198[a]	2-Octanone	2.37	2.754	2.528
199	1-Octene	4.57	4.190	4.143
200	Octylbenzene	6.30	6.719	6.423
201	Pentachlorobenzene	5.03	4.769	4.737
202	1,4-Pentadiene	2.48	1.992	2.030
203	Pentamethylbenzene	4.56	4.769	4.737
204	1-Pentanol	1.51	1.667	1.406
205[a]	2-Pentanol	1.25	1.486	1.223
206	3-Pentanol	1.21	1.473	1.216
207	2-Pentanone	0.84	1.053	0.956
208	3-Pentanone	0.82	1.084	0.963
209	1-Pentene	2.20	2.309	2.272

Continued

No.	Compound	log *P*	MLR	RBFNNs
210	Pentylamine	1.49	1.264	1.330
211	Pentylbenzene	4.90	4.943	4.882
212[a]	1-Pentyne	1.98	1.882	1.846
213	Phenetole	2.51	2.379	2.250
214	Phenol	1.48	1.349	1.230
215	Phenyl acetate	1.49	1.938	2.066
216	1-Phenylethanol	1.42	2.137	2.005
217	Phenyl formate	1.26	1.761	1.689
218	1-Phenyl-1-propanone	2.19	2.456	2.338
219[a]	Piperidine	0.84	0.705	0.834
220	Propanal	0.59	0.185	0.133
221	Propanoic acid	0.33	0.103	0.250
222	1-Propanol	0.25	0.597	0.375
223	2-Propanol	0.05	0.431	0.258
224	Propargyl alcohol	−0.38	0.126	0.056
225	Propyl acetate	1.24	0.981	1.182
226[a]	Propylamine	0.48	0.282	0.424
227	Propylbenzene	3.69	3.804	3.837
228	Propyl formate	0.83	0.677	0.850
229	2-Propylphenol	2.93	2.736	2.507
230	4-Propylphenol	3.20	2.749	2.533
231	Octadecanoic acid	8.23	7.902	8.092
232	Styrene	3.05	3.198	3.229
233[a]	1,2,3,4-Tetrachlorobenzene	4.55	4.257	4.291
234	1,2,3,5-Tetrachlorobenzene	4.65	4.244	4.266
235	1,2,4,5-Tetrachlorobenzene	4.51	4.253	4.266
236	1,1,2,2-Tetrachloroethane	2.39	3.068	3.088
237	Tetrachloroethene	2.88	2.783	2.804
238	Tetrachloromethane	2.64	2.462	2.536
239	Tetradecanoic acid	6.1	5.684	6.176
240[a]	1,2,3,4-Tetramethylbenzene	4.00	4.257	4.291
241	1,2,3,5-Tetramethylbenzene	4.10	4.244	4.266
242	1,2,4,5-Tetramethylbenzene	4.10	4.253	4.266
243	Toluene	2.73	2.722	2.835
244	*o*-Toluic acid	2.32	1.809	2.069
245	*m*-Toluic acid	2.37	1.817	2.069
246	*p*-Toluic acid	2.34	1.814	2.069
247[a]	Tribromomethane	2.38	1.953	2.015
248	1,2,3-Trichlorobenzene	4.04	3.728	3.779
249	1,2,4-Trichlorobenzene	3.98	3.731	3.779
250	1,3,5-Trichlorobenzene	4.02	3.721	3.779
251	1,1,1-Trichloroethane	2.49	2.462	2.536
252	1,1,2-Trichloroethane	2.38	2.551	2.583
253	Trichloroethene	2.53	2.158	2.129
254[a]	Trichloromethane	1.97	1.953	2.015
255	1,2,3-Trichloropropane	2.63	3.167	3.188

Continued

No.	Compound	log *P*	MLR	RBFNNs
256	2,2,3-trifluoro-3-methylbutane	3.16	3.255	3.211
257	Triethylamine	1.45	1.606	1.604
258	Trimethylamine	0.16	0.157	0.357
259	1,2,3-Trimethylbenzene	3.60	3.731	3.779
260	1,2,4-Trimethylbenzene	3.63	3.731	3.779
261[a]	1,3,5-Trimethylbenzene	3.42	3.721	3.779
262	2,3,6-Trimethylbenzene	2.67	2.673	2.477
263	2-Undecanone	4.09	4.505	4.192
264	Vinyl acetate	0.73	0.429	0.547
265	*o*-Xylene	3.12	3.225	3.314
266	*m*-Xylene	3.20	3.213	3.292
267	*p*-Xylene	3.15	3.229	3.314
268[a]	2,4-Xylenol	2.35	2.175	2.049
269	2,5-Xylenol	2.34	2.181	2.060
270	2,6-Xylenol	2.36	2.175	2.049
271	3,5-Xylenol	2.35	2.178	2.049

[a] Predicting set.

Three types of molecular descriptors are calculated to represent molecular structures: constitutional, topological and quantum chemistry descriptors. Constitutional descriptors are basically related to the number of atoms and bonds in each molecule. Topological descriptors include valence and non-valence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule, encoding information about the size, the composition and the degree of branching of a molecule. Quantum chemical descriptors include information about binding and formation energies, partial atom charges, dipole moment and energy levels in the molecule orbital. The calculation of quantum chemistry descriptors was implemented with semi-empirical PM3 Hamilton. Software 3D QSAR/WHIM[29] was used to calculate constitutional and topological descriptors. A full list of the calculated descriptors can be seen from Table 2.

Once descriptors were generated, a forward stepwise regression method was used to develop the linear model of the property of interest, which takes the form below:

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

In Eq. (1), Y is the property, that is, the dependent variable, X_1—X_n represent the specific descriptors, while b_1—b_n represent the coefficients of those descriptors, and b_0 is the intercept of this equation.

After the development of a linear model, RBFNNs are used to develop of non-linear model. RBFNNs can be described as a three-layer feedforward structure. As presented schematically in Fig. 1, the RBFNNs consist of three layers: input layer, hidden layer and output layer. The input layer does not process the information; it only distributes the input vectors to the hidden layer. The hidden layer of RBFNNs consists of a number of RBF units (n_h) and bias (b_k). Each hidden layer unit represents a single radial basis function, with associated center position and width. Each neuron on the hidden layer employs a radial basis function as non-linear transfer function to operate on the input data. The most often used RBF is Gaussian function that is characterized by a center (c_j) and width (r_j). The RBF functions by measuring the Euclidean distance between input vector (x) and the radial basis function center (c_j) and performs the non-linear transformation with RBF in the hidden layer, which is given as follows:

Table 2 Descriptors, coefficients, standard error and *T*-values for the linear model

Chemical meaning	Descriptor	Coefficient	Standard error	Standardized coefficients	*T*
Intercept	b_o	7.382	1.735		4.256
Sum of atomic polarizabilities	SP	0.910	0.024	2.538	37.777
Index of atomic composition	IAC	-0.390	0.015	-2.004	-25.839
Number of F atom	NF	1.893	0.123	0.349	15.334
Number of O atom	NO	1.119	0.113	0.564	9.885
Kier flexibility index	PHI	0.0997	0.013	0.161	7.782
Mean atomic Sanderson electronegativity	ME	-8.073	1.766	-0.129	-4.571
Subpolarity parameter	SPP	-1.107	0.300	-0.165	-3.696
R (correlation coefficient)	0.974				
RMS (root mean square error)	0.346				

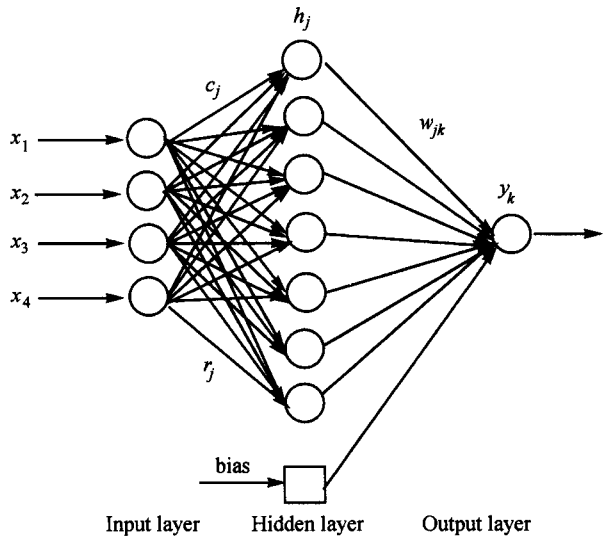

Fig. 1 Typical architecture of the RBFNNs.

$$h_j(x) = \exp(-\|x - c_j\|^2/r_j^2) \quad (2)$$

In which, h_j is the notation for the output of the jth RBF unit. For the jth RBF c_j and r_j are the center and width respectively. The operation of the output layer is linear, which is given in Eq. (3)

$$y_k(x) = \sum_{j=1}^{n_h} w_{kj}h_j(x) + b_k \quad (3)$$

Where y_k is the kth output unit for the input vector x, w_{kj} is the weight connection between the kth output unit and the jth hidden layer unit, and b_k is the bias.

From Eq. (1) and Eq. (2), It can be seen that designing an RBFNN involves selecting centers, number of hidden layer unints, width and weights. The widths of the radial basis function can either be chosen the same for all the units or be chosen different for each unit. In this paper, considerations were limited to the Gaussian functions with a constant width, which was the same for all units. Forward subset selection routine[30,31] was used to select the centers from training set samples. The adjustment of the connection weight between hidden layer and output layer was performed using a least-squares solution after the selection of centers and width of radial basis functions.

The overall performance of RBFNNs is evaluated in terms of root mean squared error (RMS) according to the equation below:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - \hat{y}_k)^2}{n_s}} \quad (4)$$

Where y_k is the desired output and $\hat{y}_k$ is the actual output of the network, and n_s is the number of compounds in analyzed set.

All calculation programs were written in MATLAB M-file and compiled using MATCOM compiler running Redhat Linux 6.0 operating system on a Pentium 266 PC with 128 M RAM.

Results and discussion

Firstly, stepwise regression routine was used to develop the linear model for the prediction of log P by using calculated structural descriptors. The best linear model contains seven molecular descriptors. The regression coefficients of the descriptors and their physico-chemical meaning are listed in the Table 2. This model produced a RMS error of 0.336 and a correlation coefficient of 0.974 for the training set compounds. The external predicting set had a RMS error of 0.346 using leave-one-out cross-validation. The predicted log P using MLR are shown in Table 1 and Fig. 2.

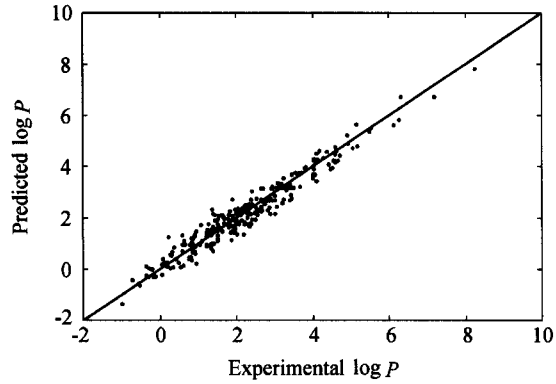

Fig. 2 Predicted vs. experimental log P (MLR).

The log P of a compound is determined by its partition between two phases of octanol-water and is a direct consequence of the difference between the Gibbs free energies of solvation of studied molecule in these phases. The difference of the Gibbs free energy between two solvated states is due to the differences in favorable and unfavorable interactions between the solvent and the solute. These interactions include dispersion interaction, dipole-

dipole interaction, dipole-induced interaction and hydrogen bonding interaction. The descriptors involved in the present equation can represent these interactions and gain some insight into the relative contributions of the different interactions in the partition process. The dispersion interaction is mainly determined by the molecular size, as described by one constitutional descriptor: sum of atomic polarizabilities and two topological indices (index of atomic composition and Kier flexibility index). The dipole-dipole interaction was described by mean atomic Sanderson electronegativity and subpolarity parameter. The hydrogen bonding information was contained in one constitutional descriptor: the number of O atom.

After the establishment of a linear model, radial basis function network was used to develop a non-linear model based on the same subset of descriptors. The RBFNN has seven inputs (a set of seven molecular descriptors), one output layer unit (log P) and one hidden layer of n_h units. Such a RBFNN can be designed as $7-n_h-1$ net to indicate the number of unit in input, hidden layer and output layer, respectively. A RBFNN is completely specified by choosing the following parameters: (a) the number n_h of radial basis functions; (b) the center c_j and width r_j of each radial basis function, and (c) the connection weights w_{kj} between jth hidden layer unit and kth output unit.

The number of radial basis functions (the hidden layer units) n_h greatly influences the performance of a RBFNN. In this paper, the radial basis functions were added one by one and terminated if no performance of the networks was improved by adding a new basis function. The centers of RBFNNs are determined with forward subset selection method. The advantages of this method over other center selection methods are that it can determine the number of hidden layer units simultaneously and there is no need to fix the number of hidden layer units in advance. This method goes through a process of selecting a subset of radial basis functions from a larger set of candidates (training set samples). The model starts empty, the radial basis function to add is the one, which reduces the sum of squared errors most. This process of adding hidden units and increasing the model complexity is continued till some criterion such as GCV stops increasing. The criterion of the selection used here is an approximation of the leaving-one-out (LOO) cross-validation methods, according to the equation below:

$$\sigma^2_{LOO} = \frac{\hat{y}P[\mathrm{diag}(P)]^{-2}P\hat{y}}{P} \tag{5}$$

Where $\hat{y}$ is the output of the network, P is the projection matrix, which can be computed by $P = I_p - ZZ^t$ from the outputs matrix Z of hidden layer units and the unit matrix I with dimension p, p is the pattern number in training sets. The LOO cross-validation method was used to prevent the network from overfitting.

After the selection of the centers and number of hidden layer units, the connection weights can be easily calculated by linear least square methods.

$$w = yZ'(ZZ')^{-1} \tag{6}$$

Where y is the matrix of training example targets, Z is the matrix of hidden layer unit outputs, Z' is the transpose of matrix Z and w is the weight matrix connection hidden layer and output layer.

The optimal width was selected by experimenting with a number of trials and selecting the one most favored by the model selection criterion: a width smaller than 1 gives poor prediction ability, varying the width indicates that width has little effect on the performance of RBFNNs if width exceeds 3.0. So the optimal width from 1.0 to 3.0 every 0.1 is selected. Each minimum error on LOO cross-validation was plotted versus the width (Fig. 3) and the minimum was chosen as the optimal conditions. In this case: $r = 1.6$ and $n_h = 17$.

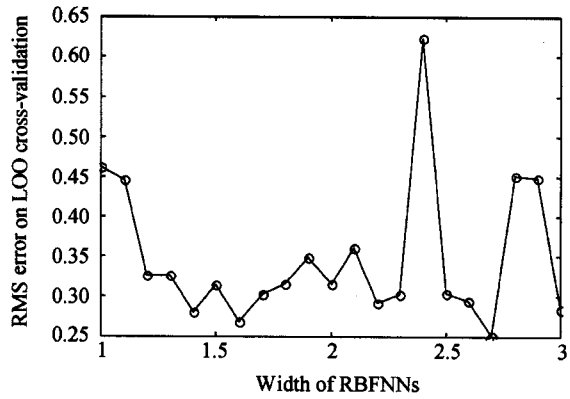

Fig. 3 Width of RBFNNs vs. RMS error on LOO cross-validation.

Through the above process, the best number of hidden layer units and the optimum width are selected as 17 and 1.6, respectively. The selected centers and their

distributions among training samples are listed in Tables 3 and 4. From the best network, the inputs in the predicting set were presented with it and the results with RBFNNs were obtained, which are shown in Table 1 and Fig. 4. The network gives RMS of 0.327 for the predicting set. The performance of RBFNN is better than that obtained by multiple linear regression. Analysis of the results obtained indicates that the model proposed correctly represents structural-property relationships of these compounds.

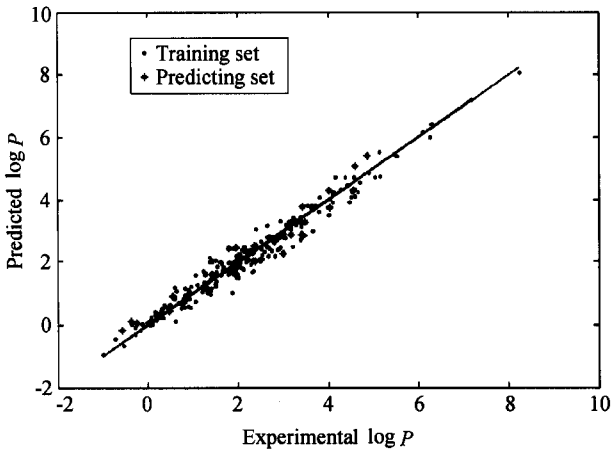

Fig. 4 Predicted vs. experimental log P (RBFNNs).

Table 3 Full list of centers selected for RBFNNs

No.	Compound
231	Octadecanoic acid
89	Diethyl carbonate
200	Octylbenzene
77	Dibutyl ether
73	Decane
131	Heptylamine
106	Ethylamine
122	1-Fluoropentane
221	Propionic acid
258	Trimethylamine
12	Benzeneethanamine
99	Dipropyl ether
119	Ethyl vinyl ether
91	Difluoromethane
132	Hexachlorobenzene
81	Dichlorodifluoromethane
178	Methyl *tert*-butyl ether

Table 4 Full list of descriptors used

Descriptor	Chemical meaning
LUMO	LUMO energy level
HOMO	HOMO energy level
Xdip	Dipole moment in the x axis
Ydip	Dipole moment in the y axis
Zdip	Dipole moment in the z axis
Dipole	Dipole moment
HOF	Heat of formation
ET	Total energy
Electron	Electronic energy
Core	Core-core interaction energy
Qtot	Total charge
SPP	Subpolarity parameter
Ldip	Local dipole moment
SP	Sum of atomic polarizabilities (scaled on C atom)
SE	Sum of atomic Sanderson electronegativities (scaled on C atom)
SV	Sum of atomic van der Waals volumes (scaled on C atom)
ME	Mean atomic Sanderson electronegativity (scaled on C atom)
NF	Number of F atoms
NC	Number of C atoms
NO	Number of O atoms
NCl	Number of Cl atoms
TPC	Total path count
NH	Number of H atom
NHA	Number of hydrogen bonding acceptor
NHD	Number of hydrogen bonding donor
MW	Molecular weight
TPC	Totoal path count
Chi0	Randic index of 0 order
Chi1	Randic index of 1 order
Chi2	Randic index of 2 order
IAC	Index of atomic composition
PHI	Kier flexibility index
IDM	Mean information on magnitude of distance
ICEN	Centric information index
IDE	Mean information on distance equality
IDDE	Information on equality of distance degrees
IDDM	Information on equality of distance magnitude
IDMT	Total information on magnitude of distance
IDET	Total information on distance equality
ZM1	Zegreb index 1
ZM2	Zegreb index 2

Conclusions

QSPR models for the prediction of log *P* for a set of diverse organic compounds using multiple linear regression and RBFNNs based on descriptors calculated from molecular structure alone have been developed. Satisfactory results are obtained with the proposed method. The models proposed can also identify and provide some insight into what structural features are related to log *P* of these compounds. Additionally, using RBFNNs based on these same sets of descriptors produced better models with good predictive ability. RBFNNs prove to be a useful tool in the prediction of log *P* and exhibited a high speed of leaning when compared with multi-layered feedforward neural networks trained with BP algorithm. The training procedure is also much simple when using RBFNNs because there are fewer parameters having to be optimized: the width of radial basis function and the number of units in the hidden layer. Furthermore the proposed method can also be extended to other QSPR investigation.

Acknowledgements

Thanks Prof. Todeschini and other members in Milano Chemometrics and QSAR research group for providing WHIM-3D/QSAR package for use in this research.

References

1 Leo, A. J. *Chem. Rev.* **1993**, *93*, 1281.
2 Sangster, J. *Octanol-Water Partition Coefficients: Fundamentals and Physical Chemistry*, Vol. 2, John Wiley Chichester, **1997**.
3 Mannhold, R.; Rekker, R. F.; Dross, K.; Bijloo, G.; de Vries, G. *Quant. Struct.-Act. Relat.* **1998**, *17*, 517.
4 Buchwald, P.; Bodor, N. *Curr. Med. Chem.* **1998**, *5*, 353.
5 David, R. L. *CRC Handbook of Chemistry & Physics*, 81st ed., CRC Press, **2000—2001**.
6 Zupan, J.; Gasteiger, J. *Anal. Chim. Acta* **1991**, *248*, 1.
7 Gasteiger, J.; Zupan, J. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503.
8 Heravi, M. J.; Fatemi, M. H. *Anal. Chim. Acta* **2000**, *415*, 95.
9 Anker, S. L.; Jurs, P. C. *Anal. Chem.* **1992**, *64*, 1157.
10 Egolf, L. M.; Wessel, M. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 947.
11 Hall, L. H.; Story, C. T. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004.
12 Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. *J. Mol. Model.* **1997**, *3*, 142.
13 Beck, B.; Breindl, A.; Clark, T. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046.
14 Sutter, J. M.; Peterson, T. A.; Jurs, P. C. *Anal. Chim. Acta* **1997**, *342*, 113.
15 Pompe, M.; Razinger, M.; Novic, M.; Veber, M. *Anal. Chim. Acta* **1997**, *348*, 215.
16 Zhang, R. S.; Yan, A. X.; Liu, M. C.; Liu, H.; Hu, Z. D. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 113.
17 Yan, A. X.; Zhang, R. S.; Liu, M. C; Hu, Z. D.; Hooper, M. A.; Zhao, Z. F. *Comput. Chem.* **1998**, *22*, 405.
18 Chen, S.; Cowan, C. F. N.; Grant, P. M. *IEEE Trans. Neur. Net.* **1991**, *2*, 302.
19 Walczak, B.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 179.
20 Fischbacher, C.; Jageman, K. U.; Danzer, K.; Müller. U. A.; Papenkordt, L.; Schüler. J. *Fresenius' J. Anal. Chem.* **1997**, *359*, 78.
21 Li, Q. F.; Yao, X. J.; Chen, X. G.; Liu, M. C.; Zhang, R. S.; Zhang, X. Y.; Hu, Z. D. *Analyst* **2000**, *125*, 2049.
22 Lohniger, H. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 736.
23 Tetteh, J.; Metcalfe, E.; Howells, S. L. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 177.
24 Pulido, A.; Ruisanchez, I.; Rius, F. X. *Anal. Chim. Acta* **1999**, *388*, 273.
25 Stubbings, T.; Hutter, H. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 163.
26 Yao, X. J.; Zhang, X. Y.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *Comput. Chem.* **2001**, *25*, 475.
27 Yao, X. J.; Zhang, X. Y.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *Comput. Chem.* **2002**, *26*, 159.
28 *HyperChem 4.0*, Hypercube, Inc, **1994**.
29 Todeschini, R. *WHIM-3D/QSAR* software for the calculation of the WHIM descriptors, Rel. 2.1 for Windows, Talete, Milan, **1996**.
30 Orr, M. J. L. *Introduction to Radial Basis Function Networks*, *Centre for Cognitive Science*, Edinburgh University, **1996**.
31 Orr, M. J. L. *MATLAB Routines for Subset Selection and Ridge Regression in Linear Neural Networks*, Centre for Cognitive Science, Edinburgh University, Edinburgh, **1996**.

(E0112064 LI, L. T.; HUANG, W. Q.)