Journal of Environmental Sciences (China) Vol. 4, No. 4, pp. 106–115, 1992 ISSN 1001-0742

Study on QSAR of nitrated polycyclic aromatic hydrocarbons using a knowledge-based system approach*

Bai Naibin and Du Min

Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China

Abstract A knowledge-based system approach has been suggested to study the relationships between molecular structures and mutagenicities of nitrated polycyclic aromatic hydrocarbons. A series of feature molecular structure descriptors on which the mutagenicities depend closely have been selected by using classic pattern recognition method. Then the feature descriptors are input in an expert system "shell", EXSYS, to build a simple mutagenicity classification model based on decision rules. Basing upon four classes of activities, the direct-acting mutagenicities in *Salmonella Typhimurium* TA98 of 80 nitrated PAHs are distinguished one by one, and the correct recognition rate can reach up to 0.91. This result is better than what was obtained by only using the pattern recognition method based upon two classes.

Keywords: knowledge-based system; nitrated polycyclic aromatic hydrocarbons; pattern recognition method.

INTRODUCTION

In the study of the qualitative relationship between molecular structures of nitrated polycyclic aromatic hydrocarbons (PAHs) and the direct-acting mutagenicities in *Salmonella Typhimurium* strain TA98, it has been found that their effects on TA98 can be enhanced by some special molecular structures, such as pyrene, fluoranthene, 4,4'-dinitrobiphenyl, fluorene and so on (Gilles, 1984). In addition, in almost 80 nitrated PAHs there is also a strong inhibition on TA98 when hydrocarbons are substituted by the nitro group in a L-region or in a bay-region (Peter, 1985). However they are hard to be characterized in numberic forms, so it is impossible to form adequate models using pattern recognition method. A recently developed expert system method may be able to solve this difficult problem, because in expert system frames one can describe objects in human logic language besides in numberic form (Hushon, 1987). In this paper, we suggest a knowledge-based system approach combined with pattern recognition to discuss structure-activity relationships of nitrated PAHs.

* This work is supported by the National Natural Science Foundation of China.

THE MUTAGENIC DATA OF NITRATED PAHs

Using RTECS (register of toxicology effects of chemical substances) as data source, the mutagenic data of nitrated PAHs are first retrieved on line from the Chemical Substance Toxicity Database (Wang, 1988). Based on the mutagenic data obtained, the direct-acting mutagenicities TA98 of 80 nitrated PAHs are selected through original consulting scientific and technical literatures. For some of these compounds, data show significant divergence; for a few of them, only regional values are available. Thus, data must be classified to compare their activities on the same base: if TA98 < 10, they are non-active or low active, degree is 1; if TA98 < 1000, they are moderately active, degree is 2; if TA98 < 10000, they are highly active, degree is 3; if TA98 > 10000, they are most highly active, degree is 4.

Table 1 shows the activities of compounds.

Table 1 The relationship between molecular structure codes and the direct-acting mutagenicities in *Salmonella Typhimurium* TA98

No.	Compounds	Molecular structure codes								Activity code		
		1	2	3	4	5	6	7	8	TA98, REV/ nmole	Class EXP.	Class CAL.
1	Nitrobenzene	3.21	57	123	1	1	0	0	0	0	1	1
2	1, 2-Dinitrobenzene	3.93	58	168	2	1	0	0	0	0	1	1
3	1, 3-Dinitrobenzene	3.92	69	168	2	1	0	0	0	0	1	1
4	1, 4-Dinitrobenzene	3.92	65	168	2	1	0	0	0	0	1	1
5	O-Nitrotoluene	3.62	57	137	1	1	1	0	0	0	1	1
6	M-Nitrotoluene	3.60	62	137	1	1	1	0	0	0	1	1
7	P-Nitrotoluene	3.60	64	137	1	1	1	0	0	0	1	1
8	2, 4-Dinitrotoluene	4.35	73	182	2	1	1	0	0	0.06	1	1
9	3, 4-Dinitrotoluene	4.35	73	182	2	1	1	0	0	0	1	1
10	2, 6-Dinitrotoluene	4.36	81	182	2	1	1	0	0	0.02	1	1
11	4-Nitrodimethy-benzene	4.03	73	151	1	1	2	0	0	0	1	1
12	1-Nitronaphthalene	4.70	79	173	1	2	0	0	0	0.05	1	1
13	2-Nitronaphthalene	4.69	75	173	1	2	0	0	0	0.20	1	1
14	1, 3-Dinitronaphthlene	5.50	90	218	2	2	0	0	0	0.90	1	1

Table 1 (Continued)

15	1, 5-Dinitronaphthalene	5.43	89	218	2	2	0	0	0	3.30	1	1
16	1, 8-Dinitronaphthalene	5.43	77	218	2	2	0	0	0	0	1	1
17	2-Methyl-1-Nitronaphthalene	5.11	89	187	1	2	1	0	0	0	1	1
18	3-Methyl-2-Nitronaphthalene	5.10	75	187	1	2	1	0	0	1.0	1	1
19	1-Methyl-2-Nitro-naphthalene	5.11	88	187	1	2	1	0	0	0.20	1	1
20	1,3,6,8-tetranitro-naphthalene	6.08	98	308	4	2	0	0	0	0.20	1	1
21	5-Nitro-2H-acenaphthene	5.18	75	199	1	2	2	0	0	2.50	1	1
22	2-Nitrofluorene	5.68	91	211	1	2	1	0	P	14	2	2
23	2, 7-Dinitrofluorene	6.40	103	256	2	2	1	0	P	471	2	2
24	2-Nitrobiphenyl	5.71	92	199	1	2	0	1	0	0	1	1
25	3-Nitrobiphenyl	5.69	94	199	1	2	0	0	0	0	1	1
26	4-Nitrobiphenyl	5.69	83	199	1	2	0	0	0	0.51	1	1
27	2, 2'-Dinitrobiphenyl	6.44	92	244	2	2	0	1	0	0	1	1
28	2, 3'-Dinitrobiphenyl	6.42	94	244	2	2	0	1	0	0	1	1
29	2, 4'-Dinitrobiphenyl	6.42	103	244	2	2	0	1	0	0	1	1
30	2, 4-Dinitrobiphenyl	6.42	103	244	2	2	0	1	0	0	1	1
31	3, 3'-Dinitrobiphenyl	6.41	97	244	2	2	0	0	0	0	1	1
32	3, 5-Dinitrobiphenyl	6.41	106	244	2	2	0	0	0	0	1	1
33	4, 3'-Dinitrobiphenyl	6.41	105	244	2	2	0	0	0	1.68	1	1
34	4, 4'-Dinitrobiphenyl	6.41	92	244	2	2	0	0	Q	14.78	2	2
35	2, 4, 2'-Trinitrobiphenyl	7.16	103	289	3	2	0	1	0	0.64	1	1
36	2,4,3'-Trinitrobiphenyl	7.15	105	289	3	2	0	1	0	1.07	1	1
37	2,4,6-Trinitrobiphenyl	7.16	122	289	3	2	0	1	0	0	1	1
38	2,4,4'-Trinitrobiphenyl	7.15	114	289	3	2	0	1	Q	4.53	1	2
39	2,6,2'-Trinitrobiphenyl	7.17	109	289	3	2	0	1	0	0	1	1
40	2,6,3'-Trinitrobiphenyl	7.16	113	289	3	2	0	1	0	0	1	1
41	2,6,4'-Trinitrobiphenyl	7.16	122	289	3	2	0	1	0	0	1	1
42	3,4,3'-Trinitrobiphenyl	7.15	103	289	3	2	0	0	0	83.35	2	2

Table 1 (Continued)

43	3,4,4'-Trinitrobiphenyl	7.15	114	289	3	2	0	0	Q	400.29	2	2
44	2,4,2'4'-Tetranitrobiphenyl	7.89	103	334	4	2	0	1	Q	457.19	2	2
45	2,4,2',6'-Tetranitrobiphenyl	7.89	122	334	4	2	0	1	0	0.86	1	1
46	2,4,3',4'-Tetranitrobiphenyl	7.88	103	334	4	2	0	1	Q	33.44	2	2
47	2,6,2',6'-Tetranitrobiphenyl	7.91	109	334	4	2	0	1	0	0	1	1
48	3,4,3',4'-Tetranitrobiphenyl	7.88	103	334	4	2	0	0	Q	715.72	2	2
49	9-Nitrophenanthrene	6.19	103	223	1	3	0	0	0	0.5	1	1
50	2-Nitrofluoranthene	6.68	113	247	1	3	0	0	S	1030	3	3
51	3,4-Dinitrofluoranthene	7.43	100	292	2	3	0	0	S	4830	3	4
52	3,7-Dinitrofluoranthene	7.43	103	292	2	3	0	0	S	14450	4	4
53	3,9-Dinitrofluoranthene	7.41	122	292	2	3	0	0	S	121500	4	4
54	8-Nitrofluoranthene	6.68	109	247	1	3	0	0	S	11125	4	3
55	7-Nitrofluoranthene	6.69	104	247	1	3	0	0	S	544	2	3
56	3-Nitrofluoranthene	6.69	100	247	1	3	0	0	S	5439	3	3
57	1-Nitrofluoranthene	6.69	113	247	1	3	0	0	S	740	2	3
58	2-Nitrophenanthrene	6.18	99	223	1	3	0	0	0	0.5	1	1
59	9-Nitroanthracene	6.19	98	223	1	3	0	1	0	< 0.5	1	1
60	2-Nitroanthracene	6.17	93	223	1	3	0	0	0	892	2	1
61	7-Nitrobenzo [a] -anthracene	7.35	125	273	1	4	0	1	0	0.30	1	1
62	3-Nitroperylene	8.18	131	297	1	4	0	0	0	< 30	2	2
63	1-Nitropyrene	6.68	110	247	1	4	0	0	T	453	2	3
64	2-Nitropyrene	6.67	116	247	1	4	0	0	T	2225	3	3
65	1, 4-Dinitropyrene	7.41	115	292	2	4	0	0	T	144760	4	4
66	1, 6-Dinitropyrene	7.41	121	292	2	4	0	0	T	183570	4	4
67	1, 8-Dinitropyrene	7.41	116	292	2	4	0	0	T	254000	4	4
68	1, 3, 6-Trinitropyrene	8.15	144	337	3	4	0	0	T	40700	4	4
69	1, 3, 6, 8-Tetranitropyrene	8.68	144	382	4	4	0	0	T	15590	4	4
70	2-Nitrochrysene	7.68	107	273	1	4	0	0	0	< 0.6	1	2

Table 1 (Continued)

71	5-Nitrochrysene	7.68	121	273	1	4	0	1	0	<0.6	1	1
72	6-Nitrochrysene	7.68	112	273	1	4	0	0	0	269	2	2
73	6-Nitrobenzo [a] pyrene	8.18	134	297	1	5	0	1	T	0	1	1
74	1-Nitrobenzo [a] pyrene	8.17	134	297	1	5	0	0	T	1367	3	3
75	3-Nitrobenzo [a] pyrene	8.15	131	297	1	5	0	0	T	1070	3	3
76	1-Nitrobenzo [e] pyrene	9.26	145	297	1	5	0	1	T	39	2	2
77	3-Nitrobenzo [e] pyrene	9.35	145	297	1	5	0	0	T	890	2	3
78	4-Nitrobenzo [g, h, i] pyrene	8.67	131	321	1	6	0	1	0	<0.6	1	1
79	7-Nitrobenzo [g, h, i] pyrene	8.67	133	321	1	6	0	1	0	<0.3	1	1
80	1-Nitrocoronene	9.15	148	345	1	7	0	0	0	28	2	2

Notes: The marks of molecular structure codes are:

1— index of molecular connection; 2— molecular incumbrance area;
3— molecular weight; 4— number of nitro group;
5— aromatic hydrocarbon cycle number; 6— number of methyl group;
7— stereoscopic block factor; 8— special molecular sub structure factor.

STRUCTURAL DATA OF NITRATED PAHs

Eight molecular structure codes are designed according to molecular structure graphs and corresponding classes in TA98. The first six of the eight codes are numerical: (1) index of molecular connection; (2) molecular incumbrance area; (3) molecular weight; (4) number of nitro group; (5) aromatic hydrocarbon cycle number; (6) number of methyl group.

The other two codes are "true-false" codes and their definition and calculated method appear in the literature (Tokiwa, 1981; Motoshi, 1985; Rosenkranz, 1983; Villiam, 1985; Teruhisa, 1986; Barbar, 1987; Reiko, 1987). They are: (7) stereoscopic block factor; (8) special molecular substructure factor. If nitro substituents are not located in a L-region, or in a bay-region, or in the 2, 6 site of biphenyl, the code (7) is "0", which means false. Otherwise, the code is "1", which means true. The letters *P*, *Q*, *R* and *T* represent four different special molecular substructure code (8) values;

P— fluorene; *Q*— 4, 4'-dinitrobiphenyl;
R— fluoranthene; *T*— pyrene.

If the special molecular substructure does not exist in a compound, then it will belong to the general structure and code (8) will be "0".

Table 1 also shows the structural codes (1)–(8).

EXTRACTION OF MOLECULAR STRUCTUPE CODE

To perform the principal component analysis of pattern recognition, 80×7 structure-activity code matrix (including activity codes but excluding the true-false codes in the molecular structure codes) are weighted and normalized according to their degree of activity. The six eigenvalues and the corresponding eigenfunctions obtained by the principal component analysis method are shown in Table 2. After the eigenvalues are put in order, the first two principal components are selected. The probability value of the classification result by using the first two principal compounds can be expressed in the equation:

$$P = \frac{P_1 + P_2}{\sum_{j=1}^{6} P_j} = 0.932$$

So, the molecular structure codes which represent the first two principal components are considered as the feature molecular structure codes. The first, second and fifth codes of molecular structure (index of molecular connection, molecular incumbrance area and aromatic hydrocarbon cycle number) have higher coefficient weights corresponding to the first principal component. However, the results of correlation matrix analysis in Table 3 show that there are higher correlation coefficients among the 1st, 2nd and 5th codes, in particular, the three factors are not linearly independent. One of them can represent another. Therefore the index of molecular connections may be considered to be the feature structure codes of the first principal component. The fourth code (nitro number) has a higher coefficient weight corresponding to the second principal

Table 2 Eigenvalues and eigenfunctions of molecular structure code matrix

Eigenvalues		1	2	3	4	5	6
		10.840	3.764	0.607	0.284	0.148	0.018
Eigen-functions	1	0.481*	0.116	0.049	−0.181	0.683*	−0.503*
	2	0.510*	0.015	0.155	−0.663	−0.524*	0.034
	3	0.424	0.326	0.088	0.188	0.261	0.776*
	4	0.068	0.800*	0.042	0.338	−0.349	−0.343
	5	0.553*	−0.483	−0.019	0.607*	−0.260	−0.157
	6	− 0.135	−0.081	0.982*	0.094	0.035	0.039

* higher weighted item

component. During the calculating process, four molecular structure codes (molecular weight, aromatic hydrocarbon cycle number, molecular incumbrance area and methyl group number) are deleted by the principal component analysis. Finally, the characteristic space of molecular structure code consists of four factors: (1) index of molecular connection; (2) nitro group number; (3) stereoscopic block factor; (4) special molecular sub-structure factor. These four molecular structure codes are used to build up a structure-activity expert system based on the production rules.

Table 3 The correlation coefficients matrix of molecular structure codes

	1	2	3	4	5	6
1	1.000	0.942	0.942	0.254	0.775	−0.485
2	0.942	1.000	0.866	0.141	0.815	−0.435
3	0.942	0.866	1.000	0.529	0.614	−0.475
4	0.254	0.141	0.529	1.000	−0.297	−0.217
5	0.775	0.816	0.614	−0.297	1.000	−0.351
6	−0.485	−0.435	−0.475	−0.217	−0.351	1.000

THE KNOWLEDGE BASE

To build up a knowledge base means to deduce a reasonable search path and to compile correct rules consisting of knowledge and experience of experts in certain fields. During the deductive process, results will be obtained step by step according to the facts and rules.

In this system, the knowledge base consists of 16 productive rules. The context tree is shown in Fig. 1, and some of the rules are shown in Table 4. The compounds are classified step by step according to the four feature molecular structure codes discussed above and the rules. First, 80 nitrated PAHs are classified as class *A* or *B* by the index of molecular connection representing the first principal component. For class *A*, which has lower activity, the compounds are further classified into three groups according to whether or not the sub-structures *P* and *Q* exist in them. The first two classes show that they have acitivity, the other class shows no activity or low activity. For class *B*, compounds are divided into class *B1* and *B2* according to whether or not the stereo block exists in the compound. For class *B*1 (the less active compounds of class *B*), it is possible to make an additional determination of whether substructure *Q* is present or not: if *Q* is present, then the compound has activity; if *Q* is absent, the compound has no acitvity or low activity. For class *B*2 (the more active compounds of class *B*), compounds are classified into three groups: (1) moderately active, (2) highly active, and (3) the most highly active according to whether or not the substructures *R* and *T* and the nitro group number exist in the compound. All of the reductive results are assigned probabilities values ranging from 0/10 to 10/10.

A value of 0/10 equivalent to "absolutely no". A value of 10/10 equivalent to "absolutely yes". Values of 1/10 to 9/10 represent degrees of certainty ranging from "very probably no" to "very probably yes". Because the correct recognition rate is larger than 0.91, the probability of results are about 9/10.

The rules are loaded into an expert system shell "EXSYS" and run (EXSYS, Inc., 1985). The computer will ask user to input the index of molecular connection, the stereo block code, the special structure code, and the nitro group number of compounds. The reasoning will be carried out according to the rules and facts until the active class of the compound and its probability has obtained. The results will be displayed on the screen and printed.

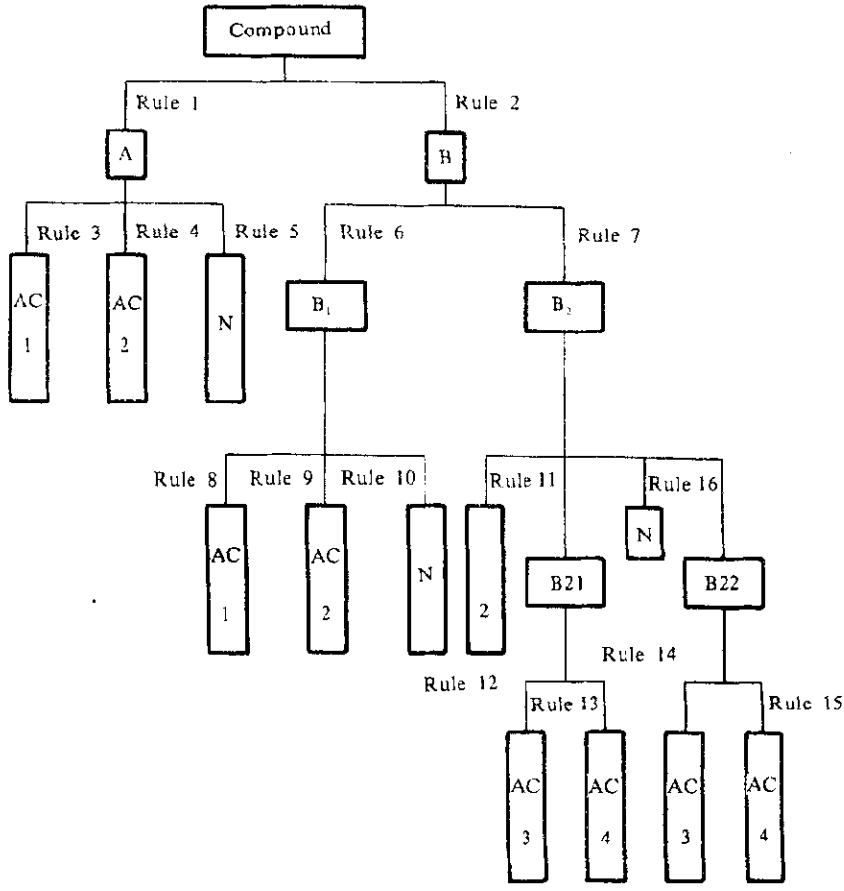

Fig. 1 Context tree

Table 4 Sample of rules in knowledge base

```
.....
RULE NUMBER: 6
IF:
      THE CLASS OF DISCUSSED COMPOUND IS B
   and [E] =1.0
THEN:
      THE CLASS OF DISCUSSED COMPOUND IS B1

RULE NUMBER: 7
IF:
      THE CLASS OF DISCUSSED COMPOUND IS B
   and [E] =0.0
THEN:
      THE CLASS OF DISCUSSED COMPOUNDS IS B2

RULE NUMBER: 8
IF:
      THE CLASS OF DISCUSSED COMPOUND IS B1
   and THE PARTICULAR STRUCTURE IN DISCUSSED MOLECULAR IS GENERAL
THEN:
      THE ACTIVITY OF DISCUSSED COMPOUND IS NON- or LOW- ACTIVE (TA=0 to 10)-
      Probability=9/10

RULE NUMBER: 9
IF:
      THE CLASS OF DISCUSSED COMPOUND IS B1
   and THE PARTICULAR STRUCTURE IN DISCUSSED MOLECULAR IS Q or T
THEN:
      THE ACTIVITY OF DISCUSSED COMPOUND IS MEDIAN ACTIVE (TA=10 to 1000)-
      Probability=9/10

RULE NUMBER: 10
IF:
      THE CLASS OF DISCUSSED COMPOUND IS B1
   and THE PARTICULAR STRUCTURE IN DISCUSSED MOLECULAR IS P or R
THEN:
      THE COMPOUND IS NOT PRISENCE IN CLASS B1-Probability=10/10
.....
```

RESULTS AND DISCUSSION

The system has been carried out on a PC/AT computer. Rule sets of the expert system have been managed by using an expert system tool "EXSYS".

The 80 nitrated PAHs shown in Table 1 are distinguished one by one by using the expert system; the total correct recognition rate can reach 0.91. This result is better than what was obtained only by means of the pattern recognition method. Instead of two classes only, activities of compounds are classified into four groups in our new system. Thus, our results of classification are more accurate and they coincide with experimental data in literature very well (Tokiwa, 1981; Motoshi, 1985; Rosenkranz, 1983; Villiam, 1985; Teruhisa, 1986; Barbar, 1987; Reiko, 1987).

Undoubtedly, the new model is superior because it paid more attention to the obvious effects of the special molecular structure and the stereo block factor on biological activities, and this happens to be something that can not be solved perfectly by using pattern recognitions based on numerical distinguishing method.

REFERENCES

Barbar Zielinska, Mutation Research, 1987, 190(4): 259
EXSYS, Inc., Expert System Development Package, 1985
Gilles Klopment, Mutation Research, 1984, 126(2): 139
Hushon, J. M., Environ. Sci. Technol., 1987, 21(9): 838
Motoshi Nohara, Mutation Research, 1985, 149(1): 9
Peter, P. Fu, Mutation Research, 1985, 143(3): 173
Rosenkranz, H. S., Mutation Research, 1983, 114(3): 217
Reiko Nukagawa., Mutation Research, 1987, 191(2): 85
Tokiwa, H. Mutation Research, 1981, 91(5): 321
Teruhisa Hirayama., Mutation Research, 1986, 163(2); 101
Villiam, A. V., Mutation Research, 1985, 143(3): 213
Wang Zhongcheng, Computer in Chemistry and Application, 1988, 5(3): 218

(Received February 24, 1992)