

Prediction of mutagenic toxicity by combination of Recursive Partitioning and Support Vector Machines

Quan Liao · Jianhua Yao · Shengang Yuan

Received: 13 October 2006 / Accepted: 6 February 2007 / Published online: 11 April 2007
© Springer Science+Business Media B.V. 2007

Abstract The study of prediction of toxicity is very important and necessary because measurement of toxicity is typically time-consuming and expensive. In this paper, Recursive Partitioning (RP) method was used to select descriptors. RP and Support Vector Machines (SVM) were used to construct structure–toxicity relationship models, RP model and SVM model, respectively. The performances of the two models are different. The prediction accuracies of the RP model are 80.2% for mutagenic compounds in MDL's toxicity database, 83.4% for compounds in CMC and 84.9% for agrochemicals in in-house database respectively. Those of SVM model are 81.4%, 87.0% and 87.3% respectively.

Keywords Prediction · Mutagenic toxicity · Substructural descriptor · Recursive Partitioning · Support Vector Machines

Abbreviations

SVM Support Vector Machines
RP Recursive Partitioning

Introduction

Mutagenic toxicity, as the capacity of a substance to cause genetic mutations, is of high public concern because it has a close relationship with carcinogenicity and other health problems [1]. In experiments, mutagenic toxicity can be assessed by varied test systems [2]. In drug/pesticide discovery, people hope to know compounds which cannot be candidates because of their carcinogenicity or mutagenicity as early as possible, even before they are synthesized. Many computational models based on structure–mutagenicity relationships had been developed [3–12]. Some of them had been incorporated into commercial programs, such as DEREK [5] and TOPKAT [7].

This paper introduces the work about prediction of mutagenic toxicity, which was based on public data [10–12], substructure descriptors [13, 14], and reliable machine learning methods: Recursive Partitioning (RP) [15, 16] and combination of RP and Support Vector Machines (SVM) [17–19].

Materials and methods

Data sets

A training set and three test sets were used in this work.

Q. Liao · J. Yao (✉) · S. Yuan
Department of Computer Chemistry and
Chemoinformatics, Shanghai Institute of Organic
Chemistry, Chinese Academy of Sciences, 354,
Fenglin Road, Shanghai 200032, China
e-mail: yaojh@mail.sioc.ac.cn

Training set

In the training set, there were 4,734 chemicals (2,079 non-mutagens and 2,655 mutagens) derived from the three recently published references authored by Kazius et al. [10] (4,337 compounds), Feng et al. [11] (1,863 compounds) and Helma et al. [12] (684 compounds). All these data were assessed by Ames test system. In this study, only organic chemicals containing no elements other than C, H, N, O, F, S, P, Cl, Br and I were used. Duplicates, ionic chemicals, mixtures, enantiomers and diastereoisomers were not included in the set.

Test sets

The test sets were consisted of compounds from: (1) MDL Toxicity Database [20] (version 2003.3), 2,199 mutagens; (2) MDL Comprehensive Medicinal Chemistry (CMC) database (version 2003.1) [21], 3,789 oral drugs (considered as non-mutagen) and (3) Agricultural Chemicals Database (in-house), 497 agrochemicals (considered as non-mutagen). All compounds in these sets are neither same, nor ionic chemicals, mixtures, enantiomers and diastereoisomers.

Generation of substructural descriptor

All substructural descriptors were generated by the methods described in our former publications [13, 14]. In brief, four kinds of descriptors were generated according to the definition of descriptors for every compound in the training set:

Atom: Every single atom was considered as an Atom.

Star: Each atom with connectivity more than two was considered as the center of a star, starting from this atom, a substructure with one layer acted as a Star.

Path: Each atom was selected as the starting atom. Any paths with 1–4 bonds (2–5 atoms) was generated as a Path.

Ring: Every ring was picked out as a Ring.

Recursive Partitioning (RP)

Classification and regression trees [15] are modern statistical techniques ideally suited for both exploring and prediction. Recursive Partitioning (RP) is a process in classification trees. Now it is usually employed in identifying complex structure–activity relationships (SAR) in large sets [22]. Many algorithms of RP and their applications have been published [23–28].

In a RP process, the entire data set is firstly put into one root node, and then split into two subsets (as two nodes here) by a single descriptor. The splitting procedure is recursively repeated for each new stage. The best descriptor for splitting a node is selected from all potential descriptors by all examining results. When a node t is split by a substructural descriptor v , the chemicals without descriptor v are put into its left-child-node l , and the chemicals with descriptor v are put into its right-child-node r . The performance of the split is judged by a reduction of impurity function [15] (ΔI) (Eq. 1)

$$\Delta I = I_t - \frac{N_l}{N_t} I_l - \frac{N_r}{N_t} I_r \quad (1)$$

where the N_t , N_l and N_r are the number of chemicals in node t , l and r , respectively, the I_t , I_l and I_r are the impurity function of node t , l and r , respectively. The descriptor with the maximal ΔI was selected. The Gini impurity index [15] (Eq. 2) was used as the impurity function in this study:

$$I_i = \sum_{a \in S} \sum_{b \in S, a \neq b} P_{ai} P_{bi} \quad (2)$$

where S is the set of all classes, a and b are different classes of S , P_{ai} and P_{bi} are the proportion of class a and b in node i . Herein, only two classes were involved: mutagens and non-mutagens. So the Gini index was expressed as Eq. (3):

$$I_i = 2PM_i \times PN_i \quad (3)$$

where PM_i is the proportion of mutagens at the node i and PN_i is the proportion of non-mutagens at the node i .

To decrease possibility of overfitting of the tree, two approaches, pre-pruning (to stop developing subtrees during the tree building process) or post-pruning (to prune sub-trees after a whole tree grown) are often used [29]. In this work, pre-pruning method was employed. The “stopping criteria” used were set as: (1) A node wouldn't be further split if the ΔI reached minimum (0); (2) the number of chemicals in any of its children nodes was less than a predefined value which was 6 in this work.

Support Vector Machines (SVM)

SVM is a very promising machine learning method developed by Vapnik et al. [17–19]. It has several advantages, such as global optimum, reducing over-fitting and dimension independence. Several works published [30–35] have proved its effectivity in classification and regression.

In this work, the program LibSVM 2.6 [36] was employed to construct the SVM model and ran on a Pentium IV PC with 512M RAM. Classification model was obtained by the C-SVC method in LibSVM. Following steps were included: (1)

Preparing the data; (2) Using 10-fold cross-validation to find best parameters (the capacity parameter C , kernel function type and its corresponding parameters), and employing the mean correct classification rate (CC%) of mutagens and non-mutagens as cost function; (3) Using the best parameters to train the whole training set; (4) testing.

Results and discussion




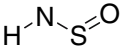





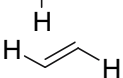
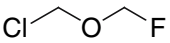

Substructural descriptors

7,444 unique descriptors were generated based on 4,734 chemicals in the training set. Some of the substructures were listed in Table 1. The most frequent descriptor was “**H**”, an Atom-type, denoted for hydrogen. The 2,924 descriptors whose occurrences were more than 5 were used as the initial descriptors.

Classification tree by RP (RP Model)

The procedure of recursive partitioning is shown in Fig. 1. At the beginning, all chemicals in the

Table 1 Examples of descriptors derived from the training set

Type	Descriptor	Occurrence	Type	Descriptor	Occurrence
Atom	H	4,697	Atom	Br	177
Ring		3,041	Path		76
Path		1,939	Path		40
Star		1,160	Atom	I	15
Path		885	Ring		7
Star		741	Star		4
Path		462	Path		1
Ring		219			

training set were put in the root node. When splitting was finished, they were moved into corresponding leaf nodes. There were 133 nodes in the classification tree, 67 leaf nodes and 66 non-leaf nodes. Table 2 listed the information of all the nodes. The PM value of the root node was 0.561. If a leaf node's PM value was greater than that of the root node (0.561), all chemicals in this node were classified as mutagenic, otherwise as non-mutagenic. In this tree (RP model), there were 40 mutagenic leaf nodes and 27 non-mutagenic leaf nodes. For the training set, the 85.2% mutagenic chemicals and 83.0% non-mutagenic chemicals were correctly classified by RP model.

The 66 non-leaf nodes were split by the 66 descriptors listed in Table 3. The ΔPM_{iv} calculated by Eq. (4) was employed to validate the effect of a descriptor on mutagenicity:

$$\Delta PM_{iv} = PM_{ir} - PM_{il} \quad (4)$$

where PM_{ir} is the proportion of mutagenic chemicals in right-child-node r (including descriptor v) of node i , and PM_{il} is the proportion of mutagenic chemicals in left-child-node l (excluding descriptor v) of node i . ΔPM_{iv} was used to detect whether the descriptor v acted on mutagenicity or not at node i . If ΔPM_{iv} was greater than zero, the descriptor v had greater effectivity on

Fig. 1 Part of the classification tree derived from the training set. (Structure a, b and c are examples which contain descriptor 1, 2 and 5, respectively. In a descriptor, C¹ expresses a sp² carbon, double dashed lines express aromatic bonds)

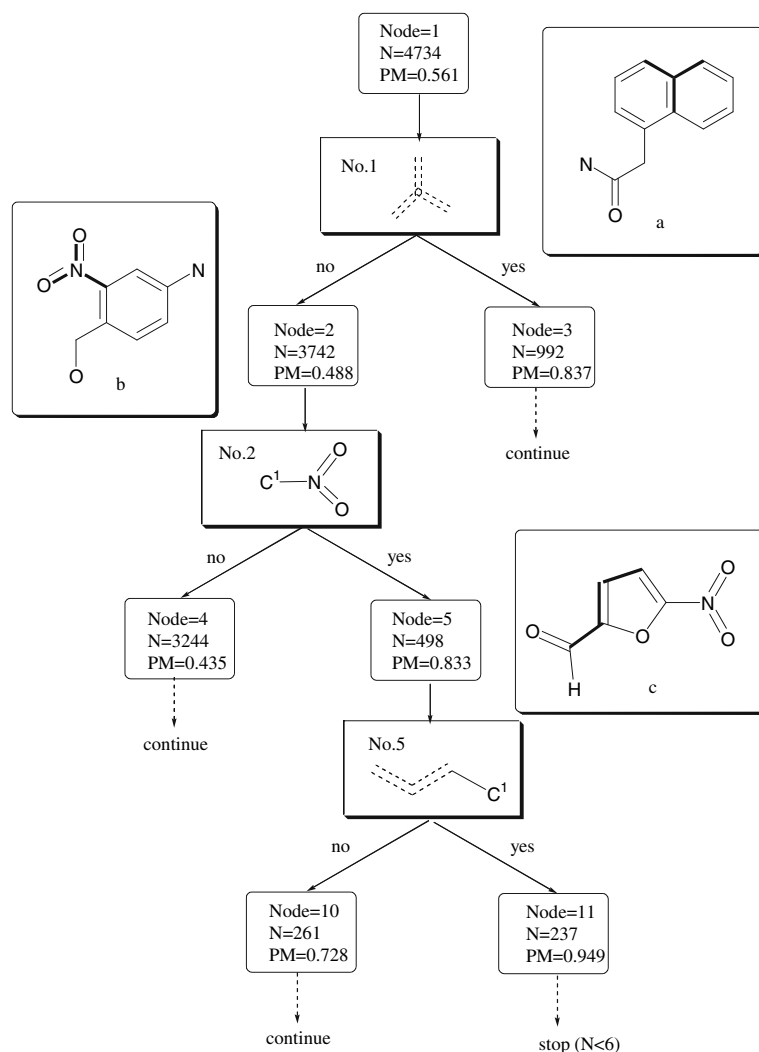


Table 2 All 133 nodes in the classification tree derived from the training set

Node	N ^a	PM ^b	Node	N ^a	PM ^b	Node	N ^a	PM ^b
1	4734	0.561	46	1992	0.288	91 [#]	7	0.286
2	3742	0.488	47	190	0.6	92 [*]	7	0.714
3	992	0.837	48	44	0.318	93 [#]	12	0.25
4	3244	0.435	49 ^{ast}	7	1	94 [#]	141	0.128
5	498	0.833	50	261	0.785	95	65	0.431
6	827	0.886	51 [#]	10	0	96 [#]	11	0.545
7	165	0.588	52	257	0.335	97 [*]	6	1
8	3063	0.405	53 [*]	30	0.867	98	1788	0.231
9	181	0.939	54	120	0.825	99 [*]	35	0.743
10	261	0.728	55 [#]	7	0.286	100 [#]	46	0.304
11 [*]	237	0.949	56	1963	0.278	101 [*]	19	0.737
12 [*]	817	0.892	57 [*]	29	0.966	102	1770	0.224
13 [#]	10	0.4	58	170	0.659	103 [*]	18	0.889
14	119	0.454	59 [#]	20	0.1	104	1757	0.219
15 [*]	46	0.935	60	35	0.4	105 [*]	13	1
16	2412	0.359	61 [#]	9	0	106	1741	0.213
17	651	0.573	62	247	0.81	107 [*]	16	0.875
18 [*]	174	0.96	63 [#]	14	0.357	108	1731	0.208
19 [#]	7	0.429	64	237	0.295	109 [*]	10	1
20	207	0.676	65 [*]	20	0.8	110	1706	0.202
21	54	0.926	66 [*]	112	0.857	111 [*]	25	0.64
22	93	0.366	67 [#]	8	0.375	112	1694	0.197
23 [*]	26	0.769	68	1890	0.262	113 [*]	12	0.833
24	2266	0.334	69	73	0.685	114	1687	0.194
25	146	0.753	70 [*]	159	0.698	115 [*]	7	1
26	300	0.703	71 [#]	11	0.091	116	1627	0.184
27	351	0.462	72	23	0.261	117	60	0.467
28	135	0.763	73 [*]	12	0.667	118	1620	0.18
29	72	0.514	74 [*]	62	0.629	119 [*]	7	1
30 [*]	42	1	75 [*]	185	0.87	120	32	0.75
31 [*]	12	0.667	76	223	0.26	121 [#]	28	0.143
32 [#]	37	0.135	77 [*]	14	0.857	122	1605	0.175
33 [#]	56	0.518	78	1849	0.25	123 [*]	15	0.733
34	2182	0.315	79	41	0.805	124 [*]	22	0.909
35 [*]	84	0.821	80	54	0.778	125 [#]	10	0.4
36 [*]	95	0.937	81	19	0.421	126	1589	0.17
37	51	0.412	82 [#]	11	0.545	127	16	0.688
38	271	0.756	83 [#]	12	0	128 [#]	1547	0.162
39 [#]	29	0.207	84	206	0.223	129	42	0.476
40	287	0.39	85	17	0.706	130 [#]	7	0.429
41 [*]	64	0.781	86	1823	0.241	131 [*]	9	0.889
42	127	0.795	87 [*]	26	0.923	132 [#]	33	0.364
43 [#]	8	0.25	88 [#]	12	0.333	133 [*]	9	0.889
44 [#]	15	0.267	89 [*]	29	1			
45 [*]	57	0.579	90	47	0.851			

* Mutagenic leaf nodes; # Non-mutagenic leaf nodes; ^a Total of molecules; ^b Proportion of mutagens at a node

mutagenicity than that on non-mutagenicity at the node *i*.

All descriptors in Table 3 were compared with the “toxicophores” published [10] and the results showed that some of the front, such as aromatic nitro (no. 2 and 33), aromatic amine (no. 8, 9, 11,

28, 31 and 37), three-membered heterocycles (no. 12 and 54), nitroso (no. 4), unsubstituted heteroatom-bonded heteroatom (no. 42, 56, 59, 63, and 65), azo-type (no. 53), aliphatic halide (no. 34, 38, 40 and 55) and polycyclic aromatic system (no. 1 and 7), tally with the later.

Table 3 Descriptors used in splitting

No	Type	Descriptor ^a	Node ^b	ΔPM	N ₊ ^c	N ₋ ^d
1	Star		1	0.349	830	162
2	Star		2	0.399	603	87
3	Path		3	-0.298	348	500
4	Path	$N=O$	4	0.534	187	12
5	Path		5	0.221	854	460
6	Path		6	-0.492	74	70
7	Ring		7	0.481	386	27
8	Path		8	0.214	673	315
9	Path		9	-0.531	154	89
10	Star		10	0.25	439	179
11	Path		14	0.404	553	228
12	Ring		16	0.419	180	39
13	Atom	O^1	17	-0.242	1593	1227
14	Star		20	-0.249	803	504
15	Path	$O-H$	21	-0.333	702	772
16	Path		22	0.383	212	205
17	Path		24	0.506	142	26
18	Path		25	-0.525	166	403
19	Path		26	-0.55	232	216
20	Star		27	0.391	220	112

Table 3 continued

No	Type	Descriptor ^a	Node ^b	ΔPM	N ₊ ^c	N ₋ ^d
21	Path		28	-0.545	17	86
22	Path		29	0.312	678	432
23	Path		34	0.312	184	84
24	Path		37	0.682	27	11
25	Path		38	-0.785	135	138
26	Path		40	0.532	295	62
27	Path		42	-0.539	44	87
28	Path		46	0.687	42	3
29	Path		47	-0.559	2	19
30	Path		48	-0.4	49	119
31	Path		50	-0.453	12	25
32	Star		52	0.505	133	49
33	Path		54	-0.482	8	11
34	Path		56	0.422	78	33
35	Atom	F	58	-0.607	53	70
36	Path		60	0.406	158	104
37	Path		62	0.241	463	201
38	Path		64	0.597	164	65
39	Path		68	0.554	106	14
40	Star		69	-0.357	18	15

Table 3 continued

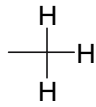
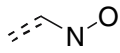
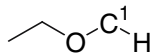
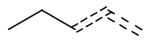


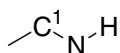

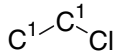

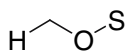




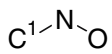
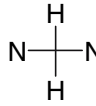
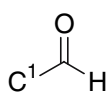
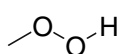
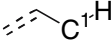
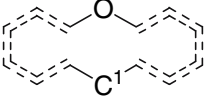
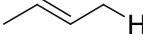

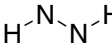
No	Type	Descriptor ^a	Node ^b	ΔPM	N ₊ ^c	N ^{-d}
41	Star		72	-0.545	425	735
42	Path		76	0.483	58	8
43	Path		78	0.682	29	2
44	Path		79	0.667	334	222
45	Star		80	-0.565	322	548
46	Ring		81	-0.464	1885	1156
47	Path		84	0.303	130	116
48	Path		85	0.455	789	914
49	Path		86	0.512	42	11
50	Path		95	0.432	506	259
51	Path		98	0.665	23	3
52	Path		102	0.781	21	0
53	Path		104	0.662	96	32
54	Ring		106	0.792	34	0
55	Path		108	0.438	52	26
56	Path		110	0.636	70	11
57	Star		112	0.806	7	1
58	Star		114	0.283	49	35
59	Path		116	0.82	8	0

Table 3 continued

No	Type	Descriptor ^a	Node ^b	ΔPM	N ₊ ^c	N ₋ ^d
60	Path		117	-0.607	196	93
61	Ring		118	0.558	21	5
62	Path		120	-0.509	46	114
63	Star		122	0.518	18	6
64	Path	N ² =O	126	0.315	641	114
65	Path		127	0.46	22	11
66	Path	=O	129	0.525	942	997

^a In a descriptor, C¹: sp² carbon, N¹: sp² nitrogen, N²: nitrogen other than sp³ and sp² nitrogen, O¹: sp² oxygen, double dashed line: aromatic bonds; ^b Number of a node (same as that in Table 2); ^c Occurrence in mutagens in the training set; ^d Occurrence in non-mutagens in the training set

SVM model

SVM model was constructed based on the 66 descriptors (selected by RP and used in the RP model) listed in Table 3. It is known that SVM can model very complex decision boundaries by mapping the input descriptors into a higher-

dimension feature space using a kernel function. It is very important to select a suitable kernel function. In LibSVM, four kernels (linear, polynomial, radial basis function (RBF) and sigmoid) were included and the RBF kernel was suggested as a reasonable first choice [37]. We firstly tried RBF kernel and used grid search method

Fig. 2 Mean CC% of cross-validation versus different C and γ values (RBF kernel); the optimized subregion was drawn in dashed line. When CC% values are similar, the smaller C values are preferred in order to reduce over-fitting

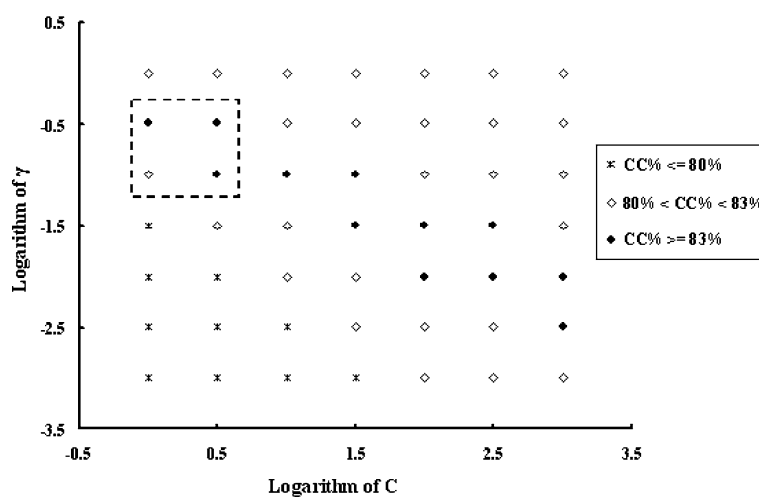


Table 4 Comparison of 10-fold cross-validation results for the four kernels

Kernel ^a	Optimal parameters	Mutagen CC%	Non-mutagen CC%	Mean CC%
RBF	$C = 1, \gamma = 0.3$	84.7	82.9	83.8
Linear	$C = 100$	84.1	76.2	80.2
Polynomial	$C = 100, \gamma = 0.03, r = 0.2, d = 2$	84.5	82.9	83.7
Sigmoid	$C = 300, \gamma = 0.001, r = 0.1$	84.7	76.0	80.4

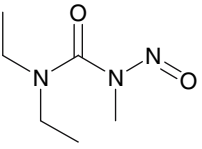
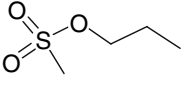
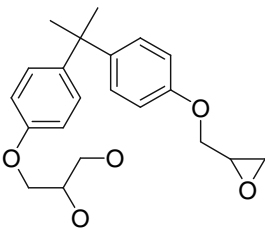
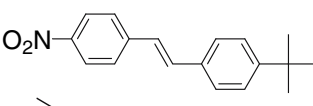
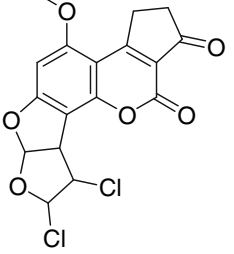
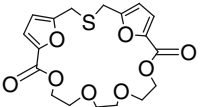
^aRBF: $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$; Linear: $K(x_i, x_j) = x_i^T x_j$; Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$; Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$

Table 5 Performance of prediction by RP and SVM models for the three sets

Model	Mutagens (2,199) ^a	Drugs (3,789) ^b	AC (497) ^c
RP _{correct}	80.2%	83.4%	84.9%
SVM _{correct}	81.4%	87.0%	87.3%
Both _{correct}	77.2%	81.4%	81.7%
Both _{incorrect}	15.6%	11.0%	9.5%
SVM _{correct} & RP _{incorrect}	4.2%	5.6%	5.6%
RP _{correct} & SVM _{incorrect}	3.0%	2.0%	3.2%

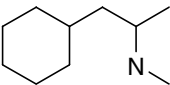
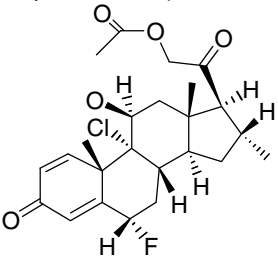
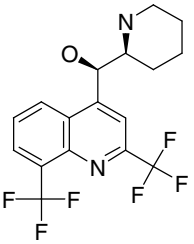
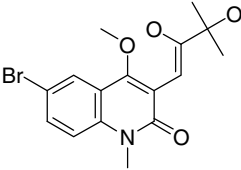
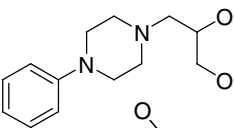
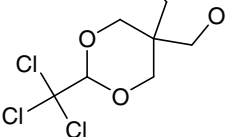
^a 2,199 mutagenic compounds from MDL Toxicity Database; ^b 3,789 non-mutagenic compounds from MDL CMC Database; ^c 497 non-mutagenic compounds from Agricultural chemicals database (in-house)

Table 6 Test examples (compounds in MDL Toxicity database)

No.	Structure	No. of descriptors ^a	Node (PM) ^b	RP ^c	SVM ^d
1		4, 13, 41, 66	18 (0.96)	+	+
2		13, 41, 45, 51	103 (0.889)	+	+
3		12, 14, 15, 36, 41, 44, 46, 48	36 (0.937)	+	+
4		2, 5, 13, 14, 41, 44, 46, 60, 64	11 (0.949)	+	-
5		5, 13, 14, 23, 29, 38, 44, 46, 48, 66	59 (0.1)	-	+
6 ^e		5, 13, 22, 48, 66	128 (0.162)	-	-

^a Same as that in Table 3; ^b Same as that in Table 2; ^c Prediction results by RP model; ^d Prediction results by SVM model; ^e This compound was predicted incorrectly by the two models; +: For mutagenic; -: For non-mutagenic

Table 7 Test examples (compounds in CMC)

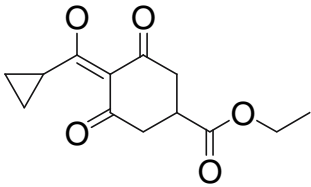
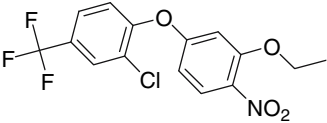
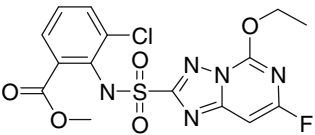
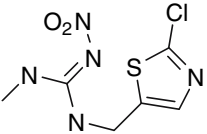
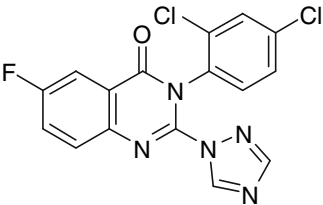
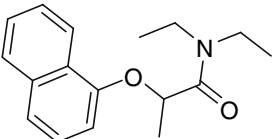
No.	Structure	No. of descriptors ^a	Node (<i>PM</i>) ^b	RP ^c	SVM ^d
1		18, 41, 45	128(0.162)	-	-
2		13, 15, 20, 23, 35, 41, 45, 48, 66	59(0.1)	-	-
3		1, 6, 14, 15, 18, 22, 35, 44, 45, 46	13(0.4)	-	-
4		3, 5, 8, 9, 13, 15, 19, 41, 46, 48, 66	94(0.128)	-	-
5		6, 8, 9, 15, 46	74(0.629)	+	-
6 ^e		15, 23, 36	70(0.698)	+	+

^a Same as that in Table 3; ^b Same as that in Table 2; ^c Prediction results by RP model; ^d Prediction results by SVM model; ^e This compound was predicted incorrectly by the two models; +: For mutagenic, -: For non-mutagenic

to find the optimal parameters C and γ , with C varies from 10^0 to 10^3 , and γ varies from 10^0 to 10^{-3} . As shown in Fig. 2, the optimized sub-region (in dashed line) was found and further search was performed in this sub-region. The best mean CC% of cross-validation was 83.8% (84.7% for positive and 82.9% for negative) when optimal parameters were found: $C = 1$ and $\gamma = 0.3$.

Similarly, other three kernels were also checked. The comparison of them was listed in Table 4. The performance of RBF kernel is similar with Polynomial kernel, and significantly better than linear and sigmoid kernels. We use the RBF kernel to build the final SVM model, which has a mean CC% of 88.0% (88.3% for mutagens and 87.7% for non-mutagens).

Table 8 Test examples (compounds in agricultural chemicals database (in-house))

No.	Structure	No. of descriptors ^a	Node (<i>PM</i>) ^b	RP ^c	SVM ^d
1		3, 13, 15, 41, 45, 48, 66	128(0.162)	-	-
2		2, 13, 14, 35, 41, 46, 48, 64	44(0.267)	-	-
3		5, 8, 11, 13, 27, 35, 41, 46, 48, 66	94(0.128)	-	-
4		13, 22, 25, 64	132(0.364)	-	+
5		5, 8, 13, 19, 21, 32, 35, 46, 66	65(0.8)	+	-
6 ^e		1, 13, 41, 46, 48, 66	12(0.892)	+	+

^a Same as that in Table 3; ^b Same as that in Table 2; ^c Prediction results by RP model; ^d Prediction results by SVM model; ^e This compound was predicted incorrectly by the two models; +: For mutagenic, -: For non-mutagenic

Test

The two models, RP model and SVM model, were tested by the three test sets: 2,199 mutagenic compounds from MDL Toxicity Database, 3,789 and 497 non-mutagens from CMC and agricultural chemicals database (in-house) respectively.

The performance of the two models for the three sets were listed in Table 5 in detailed. The correct predictions of the two models were greater than 80% for all the three sets, and the SVM model was more robust than RP model. Their prediction performances were also

compared in Table 5. For most of the tests, the results predicted by them are the same. However, a little part of the tests was inconsistently predicted by two models. Although SVM model was better than RP model, there were still 3.0% (for MDL ToxDB), 2.0% (for MDL CMC) and 3.2% (for in-house DB) of the three test sets that were correctly predicted by RP but not by SVM. The information indicated that two models complemented each other. Table 6, 7 and 8 showed some examples of consistent and inconsistent prediction of the two models.

Conclusion

In this paper, we presented the work about prediction of mutagenic toxicity by Recursive Partitioning (RP), Support Vector Machines (SVM) and substructural descriptors published [13, 14]. Two models, RP model and SVM model, were constructed and compared. From these computational experiments, we observed:

- (1) Performances of the two models are greater than 80%.
- (2) Performance of SVM model based on the descriptors selected by RP method and SVM is better than that of RP model.

The test result indicates that this SVM model has satisfied performance in prediction of mutagenic toxicity.

Acknowledgments The authors thank Dr. R. Bursi, Dr. S. S. Young and Dr. C. Helma for supplying the data sets. This work was supported in part by the National Basic Research Program (also called 973 Program) of China, through Grants 2003CB114400; by the National High-Tech. Program (also called 863 Program), through Grants 2006AA02Z39; by National Natural Science Foundation of China through Grants 20473112 and 20572120; by Chinese Academy of Sciences, through Grants KG CX2-SW-213-05 and KG CX2-SW-213-01.

References

1. Benigni R. (2005) Structure–activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem Rev* 105:1767–1800
2. World Health Organization (WHO) (1985) Guide to short-term tests for detecting mutagenic and carcinogenic chemicals. *Environmental Health Criteria* 51:100–114
3. Ashby J, Tennant RW (1991) Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat Res* 257:229–306
4. Klopman G, Rosenkranz HS (1992) Testing by artificial intelligence: Computational alternatives to the determination of mutagenicity. *Mutat Res* 272:59–71
5. Ridings JE, Barratt MD, Cary R, Earnshaw GG, Egginton E, Ellis MK, Judson PN, Langowski JJ, Marchant CA, Payne MP, Watson WP, Yih TD (1996) Computer prediction of possible toxic action from chemical structure; an update on the DEREK system. *Toxicology* 106:267–279
6. Klopman G (1992) MULTICASE 1. A hierarchical computer automated structure evaluation program. *Quant Struct Act Relat* 11:176–184
7. Enslein K, Gombar VK, Blake BW (1994) Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program. *Mutat Res* 305:47–61
8. Young SS, Gombar VK, Emptage MR, Cariello NF, Lambert C (2002) Mixture-deconvolution and analysis of Ames mutagenicity data. *Chem Intel Lab Sys* 60:5–11
9. Bacha PA, Gruver HS, Den Hartog BK, Tamura SY, Nutt RF (2002) Rule extraction from a mutagenicity data set using adaptively grown phylogenetic-like trees. *J Chem Inf Comput Sci* 42:1104–1111
10. (a) Kazius J, McGuire R, Bursi R Derivation and validation of toxicophores for mutagenicity prediction, *J Med Chem* 48 312–320 (b) Data from <http://www.cheminformatics.org/>
11. (a) Helma C, Cramer T, Kramer S, Raedt L (2004) Data mining and machine learning techniques for the identification of mutagenicity: inducing substructures and structure activity relationships of noncongeneric compounds, *J Chem Inf Comput Sci* 44 1402–1411, (b) Data from http://www.predictive-toxicology.org/data/cpdb_mutagens/
12. (a) Feng J, Lurati L, Ouyang H, Robinson T, Wang Y, Yuan S, Young SS (2003) Predictive toxicology: benchmarking molecular descriptors and statistical methods. *J Chem Inf Comput Sci* 43 1463–1470, (b) Data from <http://www.niss.org/publications.html>
13. Liao Q, Yao JH, Li F, Yuan SG, Doucet JP, Panaye A, Fan BT (2004) CISOC-PSCT: a predictive system for carcinogenic toxicity. *SAR QSAR Environ Res* 15:217–235
14. Liao Q, Yao JH, Yuan SG (2006) SVM approach for predicting LogP. *Mol Divers* 10:301–309
15. Breiman L, Friedman JH, Olshen RA, Stone CG (1984) Classification and regression trees. Wadsworth International Group, Belmont, CA
16. Myles AJ, Brown SD (2003) Induction of decision trees using fuzzy partitions. *J Chemomet* 17:531–536
17. Vapnik VN (ed) (1998) Statistical learning theory. John Wiley & Sons, New York
18. Cristianini N, Shawe-Taylor J (eds) (2000) An introduction to support vector machines. Cambridge University Press, Cambridge, UK
19. Burges CJC (1998) A tutorial on support vector machine for pattern recognition. *Data Min. Knowl. Disc* 2:121–167
20. <http://www.mdli.com/products/predictive/toxicity/>
21. http://www.mdli.com/products/knowledge/medicinal_chem/
22. http://www.nature.com/nrg/journal/v5/n4/glossary/nrg1317_glossary.html
23. Rusinko A, Farnen MW, Lambert CG, Brown PL, Young SS (1999) Analysis of a large structure/biological activity data set using Recursive Partitioning. *J Chem Inf Comput Sci* 39:1017–1026
24. Blower P, Fligner M, Verducci J, Bjoraker J (2002) On combining Recursive Partitioning and Simulated Annealing to detect groups of biologically active compounds. *J Chem Inf Comput Sci* 42:393–404

25. Tong W, Hong H, Fang H, Xie Q, Perkins R (2003) Decision forest: combining the predictions of multiple independent decision tree models. *J Chem Inf Comput Sci* 43:525–531
26. Daszykowski M, Walczak B, Xu QS, Daeyaert F, de Jonge MR, Heeres J, Koymans LM, Lewi PJ, Vinkers HM, Janssen PA, Massart DL (2004) Classification and Regression Trees-studies of HIV reverse transcriptase inhibitors. *J Chem Inf Comput Sci* 44:716–726
27. DeLisle RK, Dixon SL (2004) Induction of Decision Trees via Evolutionary Programming. *J Chem Inf Comput Sci* 44:862–870
28. Bai JPF, Utis A, Crippen G, He HD, Fischer V, Tullman R, Yin HQ, Hsu CP, Jiang L, Hwang KK (2004) Use of classification regression tree in predicting oral absorption in humans. *J Chem Inf Comput Sci* 44:2061–2069
29. Furnkranz J (1997) Pruning algorithms for rule learning. *Mach Learn* 27:139–172
30. Burbidge R, Trotter M, Buxton B, Holden S (2001) Drug design by machine learning: Support Vector Machines for pharmaceutical data analysis. *Comput Chem* 26:5–14
31. Song M, Breneman CM, Bi J, Sukumar N, Bennett KP, Cramer S, Tugcu N (2002) Prediction of protein retention times in anion-exchange chromatography systems using Support Vector Regression. *J Chem Inf Comput Sci* 42:1347–1357
32. Kramer S, Frank E, Helma C (2002) Fragment generation and Support Vector Machines for inducing SARs. *SAR QSAR Environ Res* 13:509–523
33. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV (2003) Drug discovery using Support Vector Machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci* 43:2048–2056
34. Luan F, Zhang RS, Zhao CY, Yao XJ, Liu MC, Hu ZD, Fan BT (2005) Classification of the carcinogenicity of N-Nitroso compounds based on Support Vector Machines and Linear Discriminant Analysis. *Chem Res Toxicol* 18:198–203
35. Byvatov E, Fechner U, Sadowski J, Schneider G (2003) Comparison of Support Vector Machine and Artificial Neural Network systems for drug/nondrug classification. *J Chem Inf Comput Sci* 43:1882–1889
36. Chang CC, Lin CJ, LIBSVM – A library for Support Vector Machines, <http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>
37. Hsu CW, Chang CC, Lin CJ, A practical guide to Support Vector Classification, <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>