Front. Chem. China 2007, 2(1): 55–63
DOI 10.1007/s11458-007-0012-x

RESEARCH ARTICLE

ZHOU Peng, MEI Hu, TIAN Feifei, WANG Jiaona, WU Shirong, LI Zhiliang

A new two-dimensional approach to quantitative prediction for collision cross-section of more than 110 singly protonated peptides by a novel moecular electronegativity-interaction vector through quantitative structure-spectrometry relationship studies

Abstract Based on two-dimensional topological characters, a novel method called molecular electronegativity-interaction vector (MEIV) is proposed to parameterize molecular structures. Applying MEIV into quantitative structure-spectrometry relationship studies on ion mobility spectrometry collision cross-sections of 113 singly protonated peptides, three models were strictly obtained, with correlative coefficient r and leave-one-out cross-validation q of 0.983, 0.979, 0.981, 0.979 and 0.980, 0.978, respectively. Thus, the MEIV is confirmed to be potent to structural characterizations and property predictions for organic and biologic molecules.

Keywords molecular electronegativity-interaction vector, quantitative structure–spectrometry relationship, ion mobility spectrometry, collision cross-section, peptide

1 Introduction

Ion mobility spectrometry (IMS) [1,2], an analytical technique with high sensitivities at normal pressure, has been widely used in trace analysis of aviation, customhouse, battlefield and crime scenes. Recently, advents of some novel ionization techniques such as the matrix-assisted laser desorption and ionization [3] and electrospray ionization [4] fulfill the IMS test of macromolecules. In this context, the IMS, in combination with other analytical equipment such as the GC-IMS, IMS-TOFMS, etc., becomes more efficacious in applications into biomolecules [5,6]. The collision cross-section (Ω), an important IMS parameter indicating ionic structural characters, is calculated by Eq. (1) by Mason et al. [7].

$$\Omega = \frac{(18\pi)^{1/2}}{16} \cdot \frac{ze}{(k_b T)^{1/2}} \cdot \left(\frac{1}{m} + \frac{1}{M}\right)^{1/2} \cdot \frac{t_D E}{L} \cdot \frac{760}{P} \cdot \frac{T}{273.3} \cdot \frac{1}{N} \quad (1)$$

where z denotes ionic charges; e is an elementary charge of $1.6021892 \times 10^{-19}$C; k_b is the Boltzmann constant; m and M represent the masses of ion and mobile gas molecules, respectively; P and T are the experimental pressure and temperature; E is the intensity of the electric field; t_D is the mobile time; N is the density of the mobile gas; and L is the length of the mobile tube. Thus, Ω can be accurately calculated by measuring E, L, P, T, and t_D via Eq. (1). However, large-scale measurements of collision cross-section are not feasible due to experimental and technical limits. Here, the quantitative structure–spectrometry relationship (QSSR) [8,9] has been employed to predictions on collision cross-section, anticipating to obtain a practically valuable quantitative model. Considering that collision cross-section closely relates to analyte structures, a novel molecular representation method called the molecular electronegativity-interaction vector (MEIV) is proposed and successfully employed for the 113 singly protonated peptides.

2 Principle and methodology

2.1 Several concepts about MEIV

Atomic type: in organic molecules, common atoms such as H, C, N, O, etc., are mostly located in groups IA, IVA, VA, VIA,

Translated from Chinese Journal of Analytical Chemistry, 2006, 34(6) (in Chinese)

ZHOU Peng, MEI Hu, TIAN Feifei, WANG Jiaona, WU Shirong, LI Zhiliang (✉)
College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, China;
State Key Laboratory of Chemo/Biosensing and Chemometrics, Changsha 410082, China
E-mail: zlli-cqu@163.com, zlli2662@163.com

and VIIA in the periodic table of elements. Considering the definitions of group in the periodic table of elements (i.e., atoms of the same family possess similar properties), typing atoms according to their families is deemed to be reasonable and meaningful. Thus, the common atoms are divided into five types (Table 1). While what should be elucidated here is that MEIV does not exclude any other new atoms, i.e., in case a certain new atom (e.g., unfamiliar atoms as Se and Si etc.) is present, the MEIV could be extended out of the above-mentioned five types just by the same rule.

Table 1 Five atomic types and fifteen atomic interactions among them

No.	Atomic type	1	2	3	4	5
1	H	1–1	1–2	1–3	1–4	1–5
2	C	—	2–2	2–3	2–4	2–5
3	N, P	—	—	3–3	3–4	3–5
4	O, S	—	—	—	4–4	4–5
5	F, Cl, Br, I	—	—	—	—	5–5

Atomic relative electronegativity (ARE): it is often regarded that atoms in a molecule are of neither volume nor weight and inter-atomic interactions occur via chemical bonds transmitting electricity. Besides, taking into account that the definition of electronegativity as the ability of an atom in a molecule to attract electrons to itself indicates the internal charge distributions to some extent, Pauling's electronegativity [10] here is introduced as the electricity criteria. While in practical calculations, Pauling's electronegativity needs to be further regulated by adopting ARE, i.e., electronegativity ratio of a certain molecule to that of atom C. For example, the ARE value of O is given as $ARE_O = ENP_O/ENP_C = 1.3490$. Pauling's electronegativity and the ARE value for each atom are listed in Table 2.

Table 2 Pauling's electronegativity (ENP) and atomic relative electronegativity (ARE) for common atoms in organic compounds

No.	Atom	ENP [11]	ARE
1	H	2.20	0.8627
2	C	2.55	1.0000
3	N	3.04	1.1922
4	P	2.19	0.8588
5	O	3.44	1.3490
6	S	2.58	1.0118
7	F	3.98	1.5608
8	Cl	3.16	1.2392
9	Br	2.96	1.1608
10	I	2.66	1.0431

Atomic relative bond distance (ARB): atoms in a molecule are connected together via different kinds of chemical bonds which are thus deemed to be the most direct conduction media for atomic interactions. Interatomic interaction dramatically decreases with the increase of their direct connecting distance, so atomic direct connecting distance is regarded as the interacting distance. To fulfill the unification of data, relative bond length (RBL) is defined here in the same way as ARB as the ratio of a certain bond length to the C—C bond length which is taken as the standard MEIV bond length. For example, the RBL of the C—O bond is shown as: $RBL_{C-O} = BL_{C-O}/BL_{C-C} = 0.9286$. Furthermore, the ARB value for a molecule is defined as the sum of all RBL values along the shortest distance connecting two atoms. The common chemical bonds and the ARB value are presented in Table 3.

Table 3 The practical bond length and relative bond length (RBL) for common chemical bonds in organic compounds

No.	Bond type	Bond length [12]	RBL
1	C—C	0.154	1.0000
2	C=C	0.134	0.8701
3	C≡C	0.120	0.7792
4	C≈C(butadiene)[a]	0.144	0.9351
5	C≈C(benzene)[a]	0.139	0.9026
6	C—O	0.143	0.9286
7	C=O	0.122	0.7922
8	C≈O[a, b]	0.137	0.8896
9	C—S	0.182	1.1818
10	C=S	0.161	1.0455
11	C≈S[a, c]	0.171	1.1104
12	C—N	0.147	0.9545
13	C=N	0.130	0.8442
14	C≡N	0.116	0.7532
15	C≈N[a, d]	0.134	0.8701
16	C—P	0.181	1.1753
17	C—F	0.142	0.9221
18	C—Cl	0.178	1.1558
19	C—Br	0.191	1.2403
20	C—I	0.213	1.3831
21	N≈O[a, e]	0.122	0.7922
22	N≈N[a, f]	0.130	0.8442
23	N—N[g]	0.137	0.8896
24	P—O	0.156	1.0130
25	P=O	0.149	0.9675
26	N=N	0.124	0.8052
27	C—H	0.110	0.7143
28	N—H	0.103	0.6688
29	O—H	0.097	0.6299
30	S—H	0.134	0.8702

[a]≈: conjugation bond; [b]in furan; [c]in thiophene; [d]in pyridine; [e]in nitryl; [f]in pyridazine; [g]in pyrazole.

2.2 Calculations for MEIV

The charged atoms in a molecule exist in the form of microcosmic particles, and the interactions among these charges determine the internal structural characteristics of the molecule, while with external reflections as physicochemical properties in the macroscopic state. As described by the basic formula as Coulomb's law, internal interactions in molecules are given out. According to families in the periodic table of elements, common atoms are divided into five types, and a further division is implemented due to atom of different chemical properties lending itself to distinct chemical effects. Ultimately, the 15 MEIV descriptors are generated, expressed as 1–1, 1–2, 1–3, 1–4, 1–5, 2–2, 2–3, 2–4, 2–5, 3–3, 3–4, 3–5,

4–4, 4–5 and 5–5, respectively (Table 1) with calculations as follows

$$v_{kl} = \sum_{i\in k}\sum_{j\in l}\frac{ARE_i \cdot ARE_j}{ARB_{ij}^{\ 2}} \quad (i\neq j, 1\leqslant k\leqslant l\leqslant 5) \tag{2}$$

Where k and l represent the atomic type; ARE is the relative atomic electronegativity; ARB_{ij} represents the sum of each chemical bond along the shortest path connecting atom i with atom j.

2.3 Calculation example for MEIV

Alanine is exemplified with its molecular skeleton and atomic codes given in Fig. 1. Since halogen atoms are absent from amino acids, 5 out of the 15 descriptors are zero. Removing all these zero items, 10 interaction items remain with calculations as follows. As for C—C interactions in alanine, interactions among atoms 1–2, 1–3 and 2–3 are included with the calculation expressed as Eq. (3)

$$\begin{aligned} v_{CC} &= \frac{ARE_1\times ARE_2}{ARB_{12}^{\ 2}} + \frac{ARE_1\times ARE_3}{ARB_{13}^{\ 2}} + \frac{ARE_2\times ARE_3}{ARB_{23}^{\ 2}} \\ &= \frac{1.0000\times 1.0000}{1.0000^2} + \frac{1.0000\times 1.0000}{2.0000^2} + \frac{1.0000\times 1.0000}{1.0000^2} \\ &= 2.2500 \end{aligned} \tag{3}$$

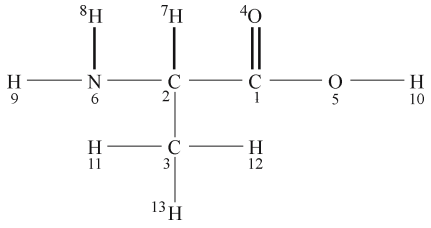

Fig. 1 Structure and molecular skeleton of alanine

The remaining 9 MEIV descriptors are calculated in the same way: $v_{HH} = 2.8389$, $v_{HC} = 10.2937$, $v_{HN} = 5.4847$, $v_{HO} = 4.4266$, $v_{CN} = 1.9327$, $v_{CO} = 4.8269$, $v_{NN} = 0.0000$ (it maybe non-zero value in other amino acids or peptides), $v_{NO} = 0.4067$, $v_{OO} = 0.6146$.

3 Results and discussion

3.1 Multiple linear regression model (M1)

All Primary structures of 113 singly protonated peptides and their observed collision cross-sections taken from results reported by Mosier et al. [13], are utilized as the overall sample set, with peptide lengths ranging from 5 to 10 to be 6 ranks in sum and lysine served as the C-terminator, which exerts effects to fix the positive charges at this site. Being all the tryptic digestion products, the peptides have their collision cross-section measured by IMS-TOFMS (Table 4). First, M1 is constructed to relate MEIV descriptors with the collision cross-sections by multiple linear regression (MLR) [14]

$$\begin{aligned} \Omega = {} & (73.994\pm 23.577) - (4.812\pm 3.697)\cdot v_{HH} \\ & + (2.412\pm 1.488)\cdot v_{HC} + (1.743\pm 1.555)\cdot v_{HN} \\ & + (1.444\pm 1.082)\cdot v_{HO} - (1.342\pm 1.339)\cdot v_{CC} \\ & - (1.324\pm 1.239)\cdot v_{CN} - (0.260\pm 1.503)\cdot v_{CO} \\ & + (22.934\pm 37.967)\cdot v_{NN} - (4.754\pm 7.932)\cdot v_{NO} \\ & - (1.228\pm 6.726)\cdot v_{OO} \end{aligned} \tag{4}$$

Simultaneously, leave-one-out cross-validation (LOO CV) [15] is employed to validate the M1, with relative statistics presented in Table 5, indicating a favorable relation between MEIV and the collision cross-section. However MLR equation may usually include problems of multicolinearity and insignificant variables, a deep discussion is required here. Implementing regressive diagnosis against M1 by the software SPSS 13.0, statistical value t and each of the variance inflation factor (VIF) for all the 10 MEIV descriptors are calculated, indicating that M1 is partially multicolinear (VIF >20) and that not all the variables are significant ($|t|<2$). Such a phenomenon can be ascribed to the unification of the MEIV calculating method, which may lend itself to some information overlap among different descriptors. To avoid this, two approaches available are adopted: (a) stepwise multiple regressions (SMR) are employed to select the variables; (b) Partial least square (PLS) regression is used to construct the model.

3.2 Stepwise multiple regression model (M2)

Stepwise multiple regression (SMR) [16], as a classical variable selection method for linear models, introduce variables in turn according to significance test of variables to ultimately determine their introduction or removal. Here, in combination with correlative coefficient q in LOO CV, which is taken as criteria-determining variable number, the plot of q and fitting correlative coefficient r of the model versus introduced variables is shown in Fig. 2 wherein the changing trend for the two ascending curves are found to be in an approximate agreement, with q achieving the maximum in case of seven variables. Besides, it was also found that in case variable number is more that 4, both r and q are not advanced obviously while when q (0.979) has been closely accessed to the maximum (0.980), thus the model with 4 variables is ultimately generated to decrease complexities of the model (M2)

$$\begin{aligned} \Omega = {} & (96.581\pm 4.812) + (0.528\pm 0.104)\cdot v_{HC} \\ & + (0.394\pm 0.229)\cdot v_{CC} - (0.708\pm 0.616)\cdot v_{CN} \\ & + (7.011\pm 1.728)\cdot v_{NO} \end{aligned} \tag{5}$$

Table 4 Various sequences and MEIV descriptors of 113 singly protonated peptides with observed and calculated collision cross-section Ω/Å² with various models

No.	Peptide[a]	v_{HH}	v_{HC}	v_{HN}	v_{HO}[a]	v_{CC}	v_{CN}	v_{CO}[a]	v_{NN}	v_{NO}[a]	v_{OO}[a]	Obsd	C(M1)	C(M2)	C(M3)
1	AAWGK	16.1735	73.5394	27.7205	9.5822	34.1121	23.2212	16.8955	0.8819	3.5556	1.1236	157.36	156.715	157.366	160.858
2	Ac-GDVEK	21.0839	85.5669	22.7978	21.7529	27.9712	20.5943	31.8859	0.7232	4.8112	3.1759	163.18	166.777	171.962	170.346
3	ADLAK	20.8646	78.3344	24.6010	15.0760	24.0883	18.4976	22.4732	0.7232	3.8812	2.0480	159.31	160.134	161.575	161.406
4	AFDEK	17.6460	81.9831	24.3236	20.0159	35.8834	19.2627	29.2319	0.7232	4.2754	3.1988	168.36	169.186	170.373	168.969
5	AIAEK*	22.4211	84.0285	24.8847	15.3878	25.7239	18.7327	22.8109	0.7232	3.7625	2.0812	160.73	165.253	164.229	164.612
6	APNAK	17.7109	70.2102	28.1519	11.0986	22.3672	21.8972	20.0202	0.9410	4.3889	1.3402	147.31	158.059	157.759	157.931
7	AWGGK	14.3573	67.4237	27.3820	9.2562	31.8061	22.4677	16.2750	0.8662	3.5201	1.1236	152.16	153.706	153.511	156.642
8	DIAAK	21.5926	78.4200	24.7390	10.0886	22.2888	18.2760	16.6996	0.7232	3.4170	1.1236	155.37	157.429	157.814	157.578
9	DLLFK	28.5200	113.5033	25.4897	10.8560	41.8637	19.8280	18.1139	0.7232	3.4170	1.1236	183.11	182.440	182.962	179.593
10	FFSDK	16.2432	74.7424	24.0775	19.2491	32.3635	18.8003	25.6429	0.7232	4.3219	2.5755	172.73	163.748	165.814	163.711
11	GGNMK	17.3500	65.7358	29.5377	12.8846	19.6445	19.8606	21.8160	0.9410	4.5694	1.5163	147.44	158.807	157.029	156.459
12	GITWK	24.4484	101.7695	28.7660	15.2871	43.2488	24.8010	21.1650	0.8790	3.9952	1.4269	169.34	177.051	177.843	180.673
13	GTFAK	18.8967	81.3054	24.4801	14.5977	32.7077	18.8779	20.1150	0.7232	3.7841	1.3427	153.97	165.739	165.591	163.782
14	IFVQK	23.0445	95.4740	30.4994	11.9621	37.9397	21.7303	20.9922	0.8711	4.2683	1.2985	181.96	176.759	176.514	174.434
15	IIAEK	22.4211	84.0285	24.8847	15.3878	25.7239	18.7327	22.8109	0.7232	3.7625	2.0812	172.54	165.253	164.229	164.612
16	LDALK	20.9396	78.4554	24.5904	15.0799	24.1425	18.4851	22.4792	0.7232	3.8812	2.0480	172.36	159.994	161.669	161.504
17	NLNEK	22.3530	84.2461	30.2257	16.7790	27.3947	21.1103	26.2738	0.9410	4.7690	2.3383	167.97	171.027	170.376	170.353
18	NTYEK	19.5066	85.0926	24.6126	23.7698	36.2733	19.3346	29.5388	0.7232	4.2346	2.4672	175.90	174.050	171.832	172.595
19	TAWEK	18.8053	85.1891	28.1556	15.1807	39.6230	24.2734	23.5448	0.8819	3.9304	2.0812	170.03	167.515	167.575	171.826
20	TGQIK	21.8352	79.6557	30.2284	11.7464	24.5427	20.3877	20.2008	0.8738	4.2566	1.2807	157.62	163.742	163.746	163.651
21	TLTGK*	20.8878	76.6570	24.5747	14.5744	22.0795	17.9172	19.3933	0.7232	3.8189	1.4269	157.34	160.531	159.871	159.085
22	TPGSK	15.3927	62.6068	22.4725	13.9729	18.8558	18.9108	18.7940	0.7232	3.8111	1.4604	145.50	151.715	150.421	150.363
23	TVGGK	16.3307	60.9524	24.0765	9.3808	17.0552	17.0123	15.8403	0.7232	3.4170	1.1236	140.18	147.363	147.418	146.982
24	YYPLK	24.4705	102.6382	23.6199	14.4361	40.5296	21.1418	21.2248	0.7232	3.5115	1.1989	187.31	176.328	176.429	175.983
25	AAAAEK	20.4804	78.6068	28.5957	16.6326	24.3825	22.7078	26.1383	0.9828	4.8170	2.2746	160.38	166.155	165.416	166.773
26	ANIDVK	27.8523	103.1204	34.7888	19.0383	33.6085	26.2621	30.6584	1.1864	5.9365	2.5737	176.78	184.790	187.335	187.295
27	ASEDLK	24.3938	94.6968	28.9479	26.8909	31.5497	23.6792	35.3513	0.9828	5.7501	3.6996	175.16	182.069	182.595	183.028
28	EAMAPK	23.8211	92.5075	27.4816	13.6744	27.5081	24.6673	23.1313	0.9828	4.6305	1.4377	176.19	173.301	171.297	171.779
29	EMPFPK	26.8193	118.6433	26.3963	14.0316	46.1897	27.7352	24.6269	0.9828	4.6110	1.3999	193.26	191.155	190.157	188.688
30	IEEIFK	28.3237	119.5129	29.6657	22.9311	47.6493	25.0030	33.8534	0.9828	5.1554	3.1401	197.00	199.096	196.942	196.291
31	IVAPGK	22.8287	87.9994	27.3441	11.5434	26.7508	24.2506	20.6154	0.9828	4.4533	1.2923	173.33	167.130	167.670	168.171
32	LIFAGK	24.3979	101.3286	29.0078	11.7111	38.2398	23.4561	20.9009	0.9828	4.4533	1.2923	186.10	180.432	179.800	177.390
33	LVEDLK*	28.5177	108.1166	29.4792	22.8606	35.0490	24.2148	33.3461	0.9828	5.3743	3.3939	192.17	186.979	188.049	188.256
34	MQIFVK	32.0847	126.8376	35.5097	14.3540	47.9736	27.2573	25.5237	1.1241	5.2512	1.4128	203.90	200.118	200.016	198.285
35	NDIAAK	24.1593	90.5932	28.9697	16.8774	27.9701	23.2844	26.3219	0.9828	4.9326	2.2243	173.78	172.253	173.565	173.910
36	NLDNLK	29.3672	108.6631	34.8894	19.0184	34.9388	26.3472	30.7023	1.2162	6.0467	2.6325	192.40	189.194	191.500	190.354
37	NVPLYK	31.5764	127.5943	28.4488	16.6665	48.5781	26.5189	25.7559	0.9828	4.5639	1.4105	195.27	195.556	196.358	196.219
38	NYQEAK	23.2183	99.2438	34.4851	22.4072	41.4382	26.3731	33.6780	1.1422	5.8266	2.5581	191.16	190.164	187.523	188.325

(*Continued*)

No.	Peptide[a]	v_{HH}	v_{HC}	v_{HN}	v_{HO}[a]	v_{CC}	v_{CN}	v_{CO}[a]	v_{NN}	v_{NO}[a]	v_{OO}[a]	Obsd	C(M1)	C(M2)	C(M3)
39	TEAEMK*	23.8587	94.3140	29.0503	24.0227	29.8984	23.5697	34.4859	0.9828	5.3376	3.3576	182.54	184.729	178.928	180.200
40	TPVSEK	24.7788	98.1784	27.6580	21.9183	31.5661	25.1903	29.7097	0.9828	5.2410	2.6607	175.98	182.327	179.802	180.651
41	YLTTLK	32.8262	117.4330	29.9005	22.4384	34.9388	24.0942	27.6103	0.9828	5.2938	1.9775	197.94	192.761	192.449	191.725
42	Ac-SIPETQK	34.8254	136.5724	36.3773	27.5885	45.1630	34.8942	41.6720	1.4096	7.8637	3.2134	205.42	212.402	216.963	215.366
43	APVDAFK	28.3842	122.7711	32.0013	19.0714	47.9413	30.8022	31.8816	1.2483	6.0739	2.5638	189.05	200.033	201.114	199.460
44	ATDEQLK*	31.3103	119.3937	39.3153	30.8702	40.6234	31.4120	43.5636	1.4196	7.7861	4.1465	205.89	207.408	208.019	208.571
45	ATEEQLK	33.0840	125.6017	39.6667	31.2926	42.3697	31.6131	43.6892	1.4196	7.6159	3.9993	206.40	213.417	210.651	212.038
46	DGADFAK	21.7638	95.7600	32.5893	18.1371	39.0914	27.8324	30.7284	1.2483	6.0739	2.5638	185.83	182.524	185.459	183.971
47	DSAIMLK	36.3129	131.0093	34.3762	21.3362	38.2105	28.8442	30.0532	1.2483	6.0712	1.8699	203.77	206.150	202.999	202.769
48	ELTEFAK	31.6853	129.8020	33.9014	24.2639	49.4093	29.4811	34.9002	1.2483	6.3274	2.7986	209.40	209.416	208.119	206.716
49	EVTEFAK*	29.9493	123.7941	33.7745	24.1228	47.7637	29.4104	34.8041	1.2483	6.3274	2.7986	202.91	205.180	204.347	203.410
50	FNDLGEK	26.3120	102.6881	38.4386	25.0888	34.9850	30.1662	39.5243	1.4564	7.4001	3.7765	206.73	194.693	195.148	195.322
51	GDVAFVK	30.9306	126.1042	33.8554	19.3082	48.2059	29.5194	31.5569	1.2483	5.9904	2.4034	200.45	201.413	203.302	202.035
52	GGVVGIK	32.1383	116.0117	33.9559	14.2198	34.1872	27.8023	24.7440	1.2483	5.5006	1.4648	175.87	190.427	190.227	190.260
53	IATAIEK	30.5598	113.5604	33.8005	24.1753	35.5633	28.5161	33.9442	1.2483	6.3021	2.7929	202.92	195.587	194.589	195.849
54	ILLSSAK	32.3171	117.6732	33.7598	23.4539	35.0738	28.2101	30.1100	1.2483	6.3758	2.2012	202.97	198.373	197.302	196.567
55	IVTDLAK	32.2147	119.0085	33.8510	24.2252	37.1263	28.7772	34.2166	1.2483	6.4992	2.9382	204.75	197.293	199.280	199.172
56	IVTDLTK	33.5189	123.3994	33.9970	29.2938	38.9600	29.0731	37.6129	1.2483	6.9138	3.3488	207.02	202.969	205.020	204.915
57	LVTDLTK	33.5189	123.3994	33.9970	29.2938	38.9600	29.0731	37.6129	1.2483	6.9138	3.3488	205.76	202.969	205.020	204.915
58	MIFAGIK	33.1966	131.9623	34.0089	14.1968	47.6312	28.9208	25.3926	1.2483	5.5006	1.4648	207.13	204.177	203.159	200.746
59	MLTAEEK	31.0034	118.4442	33.8098	29.3375	38.4864	28.8193	39.9943	1.2483	6.7016	3.8760	209.77	203.576	200.907	202.140
60	NPDPWAK	26.5689	122.5113	33.9706	18.7768	53.8553	37.6266	32.3241	1.4283	6.2429	2.5186	198.09	197.438	199.662	204.167
61	VAAALTK	29.6202	107.7257	33.5742	18.7429	31.9056	27.9120	27.7031	1.2483	5.9152	1.8300	190.32	188.151	187.780	188.412
62	VADALTK	28.7532	107.2812	33.4039	23.7387	33.6403	28.2156	33.7156	1.2483	6.4677	2.9045	194.69	189.926	191.888	192.192
63	VDPVNFK	28.7946	123.4389	37.5620	20.5144	50.5493	33.5576	35.4914	1.4910	7.0428	2.6800	208.72	204.173	207.338	206.334
64	VLAAVYK	33.0971	130.1223	34.0745	18.2417	47.8450	29.2233	28.9205	1.2483	5.6162	1.5896	206.91	203.863	202.868	202.874
65	VLPVPQK	37.0680	140.8640	37.0957	16.1981	46.0564	34.9351	29.7423	1.4096	6.3853	1.6605	206.94	207.555	209.188	210.263
66	VLSAADK	26.5048	100.5628	33.0470	23.1221	31.4124	27.8152	32.8943	1.2483	6.4645	2.9309	193.03	186.746	187.721	187.626
67	VLSPADK	28.1497	109.9991	31.7927	23.4059	35.8031	29.8747	33.7457	1.2483	6.4645	2.9309	196.51	190.972	192.979	193.300
68	VLTSAAK	28.6080	105.2642	33.3612	23.0706	31.4008	27.8229	30.0504	1.2483	6.3646	2.1581	194.37	190.606	189.495	189.407
69	VSEALTK	29.8718	111.7811	33.6446	28.6270	35.3388	28.4092	36.2642	1.2483	6.6821	3.0698	198.59	198.454	196.300	197.323
70	VVTDLTK	33.5189	123.3994	33.9970	29.2938	38.9600	29.0731	37.6129	1.2483	6.9138	3.3488	202.35	202.969	205.020	204.915
71	WNMQNGK	25.2784	96.2959	49.2761	19.3844	31.9283	34.3352	36.2654	1.9022	8.6325	2.3037	206.26	200.508	196.256	195.758
72	AADALLLK	42.4312	152.8104	38.9877	22.0129	46.9233	34.0546	35.5163	1.5180	7.1189	2.7009	223.86	218.477	221.607	221.531
73	ADFAEISK	32.6653	136.3553	38.0195	31.0367	53.4800	34.3705	44.2487	1.5180	7.8934	4.0247	218.06	220.328	220.704	220.023
74	ADFAEVSK*	30.7489	130.0182	37.8038	30.7421	51.5811	34.1654	44.0023	1.5180	7.8934	4.0247	214.17	216.348	216.753	216.258
75	ADFTDVTK	32.1149	134.1803	37.8631	35.8186	53.5260	34.5763	48.1062	1.5180	8.5157	4.6189	214.56	219.334	223.791	222.410
76	ADFTEISK	33.9893	140.8607	38.1877	36.1507	55.4050	34.6902	47.7916	1.5180	8.3409	4.4644	219.61	225.904	226.754	225.936

(*Continued*)

No.	Peptide[a]	v_{HH}	v_{HC}	v_{HN}	v_{HO}[a]	v_{CC}	v_{CN}	v_{CO}[a]	v_{NN}	v_{NO}[a]	v_{OO}[a]	Obsd	C(M1)	C(M2)	C(M3)
77	ALQASALK	37.9998	138.0861	43.9722	22.4217	42.3802	35.5761	34.9339	1.6872	7.8900	2.1813	218.92	218.653	216.368	216.378
78	DIVGAVLK	42.7034	152.0990	39.1343	16.6777	44.7631	33.6573	29.2104	1.5180	6.5560	1.6407	206.07	217.049	216.715	216.671
79	DLGEENFK	32.6717	137.7757	43.4679	27.7292	54.9087	36.4609	45.0961	1.7668	8.4979	4.0173	223.98	226.384	224.776	223.343
80	DLGEQYFK*	34.9706	153.7283	43.8826	26.8620	67.9369	37.2794	43.2317	1.6981	8.0442	3.0066	232.22	237.008	234.577	232.976
81	DSADGFLK	29.7789	124.5971	37.4155	25.0060	48.1332	33.1715	37.4104	1.5180	7.5304	3.0103	209.30	209.610	210.689	208.653
82	EYEATLEK	35.9156	146.1916	38.4973	35.6206	56.5110	34.6406	48.0961	1.5180	7.8935	4.1509	229.28	230.283	226.904	228.398
83	FGVNGSEK	26.3235	101.9327	42.5007	26.1082	33.8349	34.0817	39.7332	1.7685	8.4542	3.3498	201.76	200.343	198.915	199.401
84	GASIVEDK	35.0641	132.0957	38.3169	31.4562	43.3661	33.9077	44.0923	1.5180	7.9477	4.2205	205.22	213.361	215.177	216.429
85	IDALNENK	31.6806	121.6739	48.7422	28.8840	42.6924	38.3060	47.4848	2.0153	9.6731	4.2710	225.09	216.304	218.390	218.683
86	IGDYAGIK	28.4092	118.3393	37.3239	24.1126	46.4402	32.7491	37.8608	1.5180	7.2476	2.8494	210.60	203.914	205.032	205.021
87	LIVTQTMK	42.4813	153.3085	44.7585	30.1286	46.9969	36.7687	41.5951	1.6981	8.5865	2.8344	243.91	232.930	230.268	230.216
88	TYETTLEK	37.2632	150.7003	38.6655	40.7748	58.4145	34.9603	51.6263	1.5180	8.3409	4.6143	239.55	235.815	232.947	234.330
89	VLTPDLYK	40.4217	162.4376	37.3340	30.7968	61.7399	36.7500	43.4966	1.5180	7.7012	3.2705	230.73	232.170	234.706	234.642
90	YLGEEYVK	36.3490	148.2863	38.5106	30.8197	57.1706	34.4673	45.2767	1.5180	7.5184	3.8536	238.79	228.569	225.764	226.649
91	AAVTAFWGK	35.7477	156.6716	46.1603	22.3887	70.0220	44.4644	37.1572	1.9802	8.2599	2.1677	237.76	233.619	233.381	236.115
92	AAVTGFWGK*	33.7175	150.3015	45.8234	22.0010	67.9114	43.8451	36.7054	1.9802	8.2599	2.1677	233.00	230.648	229.622	231.933
93	ANELLINVK	51.8622	185.1885	55.4195	27.9714	60.2741	44.8919	47.2264	2.2615	10.0633	3.2700	249.69	255.477	256.946	258.199
94	EAVLGLWGK	42.3120	166.2833	46.8755	18.0717	62.7212	43.7766	33.4913	1.9802	7.7985	1.8192	235.37	234.513	232.834	236.586
95	FMMFESQNK	38.3328	158.3445	53.7152	34.8647	60.1851	44.2378	53.9926	2.2231	10.9402	4.1564	259.37	255.908	249.341	247.201
96	FQPLVDEPK	44.0392	171.5177	45.9303	30.8477	59.3811	45.2447	50.3698	1.9415	9.4793	4.2113	255.89	241.954	245.028	247.365
97	MFLGFPTTK	40.1335	173.8031	41.4034	27.7193	72.2294	41.3680	40.8587	1.7907	8.4920	2.6103	250.02	247.423	247.120	243.865
98	MFLSFPTTK	41.6001	178.5870	41.6086	32.8758	74.3774	41.9702	44.2385	1.7907	8.9453	2.9788	255.16	252.538	253.245	250.116
99	QSALAELVK	44.9970	163.5035	43.4905	28.7938	50.8039	38.9306	42.0172	1.7907	8.4389	3.0778	234.39	235.722	234.589	235.968
100	QTALVELLK	52.8945	188.4878	44.4512	29.8958	58.2656	39.9532	43.2967	1.7907	8.4389	3.0778	245.01	249.545	250.004	251.183
101	QTALVELVK*	50.9830	182.2124	44.3010	29.7090	56.5218	39.8660	43.1681	1.7907	8.4389	3.0778	242.43	245.564	246.063	247.620
102	SAVTALWGK	39.9247	158.2713	46.6421	22.8101	60.9664	43.9648	36.6943	1.9802	8.2599	2.1677	231.09	231.760	231.011	235.355
103	SLVSGLWGK	41.3177	163.8469	46.6234	22.5998	62.2487	43.6544	36.0716	1.9802	8.2599	2.1677	236.29	237.022	234.682	237.731
104	TFQSFPTTK	37.7258	166.4048	46.6930	33.9530	72.3095	44.0286	47.3841	1.9629	9.8524	3.1851	245.22	250.831	250.898	248.365
105	AQSDFGVDTK	36.5346	153.0207	51.6782	40.4550	60.9642	45.8198	58.6147	2.2185	11.6278	5.3457	241.43	247.058	250.537	248.890
106	DGAGDVAFVK	37.4693	151.9007	46.4425	24.4096	56.6248	42.6953	42.9117	2.0658	9.3043	3.1427	229.25	231.863	234.156	233.171
107	LVNELTEFAK	46.8958	186.4085	53.3823	37.7969	70.6528	47.5770	57.4260	2.3113	11.0959	4.7909	267.51	267.184	267.020	266.904
108	LVNEVTEFAK*	44.9875	180.0877	53.2004	37.5514	68.8756	47.4669	57.2556	2.3113	11.0959	4.7909	262.66	263.025	263.058	263.273
109	SEEEYPDLSK	42.5751	180.7810	45.8302	50.0485	73.3444	46.9673	69.2626	2.0658	11.0635	6.8620	257.78	265.023	265.312	268.085
110	TAAYVNAIEK	44.6813	174.6657	53.1759	31.1636	65.5606	46.9574	50.7945	2.3318	10.3176	3.4583	246.52	254.766	253.791	255.925
111	VLDSFSNGMK*	38.0993	157.4955	51.7849	37.0693	59.4735	45.6573	54.2829	2.3296	11.5072	4.4454	252.25	253.192	251.583	248.343
112	VLNSFSDGLK*	40.9463	165.4873	52.1063	35.5951	62.6258	45.9007	52.4776	2.3113	11.2937	4.2769	255.16	253.918	255.378	252.338
113	VLQSFSDGLK	42.7607	171.7457	52.4433	35.9352	64.4028	46.0617	52.6281	2.2404	11.1481	4.2068	255.95	257.875	258.249	255.560

[a] "Ac-" indicates an acetylated N-terminus, "*" superscript indicates that the peptide was chosen to be a member of the test set.

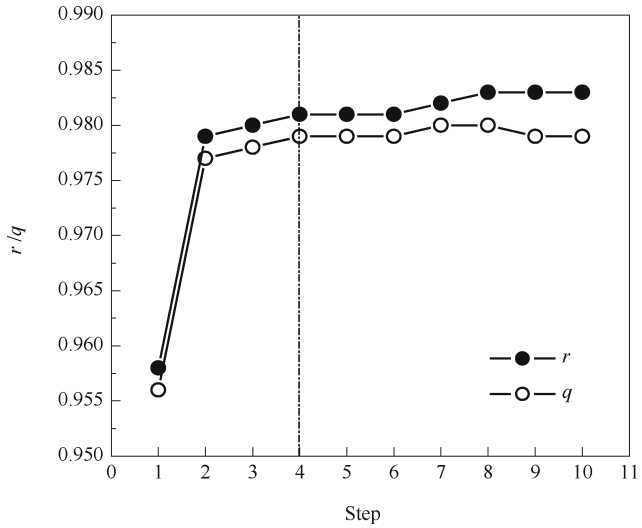

Fig. 2 Plot of r/q versus the number of SMR-introduced variables

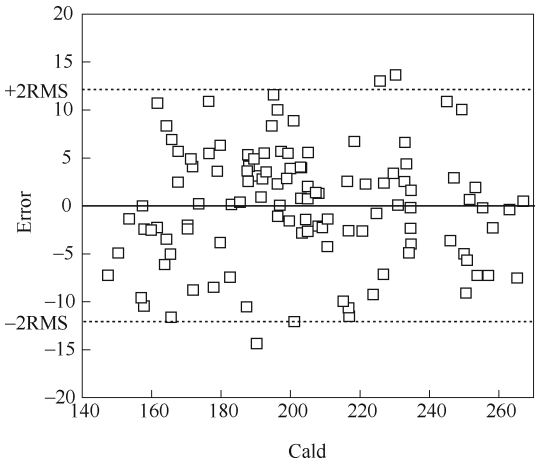

Fig. 4 Scatter plot of residue distribution by M2 (The dashed is double root mean square error)

Statistics related to M2 are listed in Table 5, and in contrast with M1, complexity of M2 has been found to be dramatically decreased but not for r and q, and the prominence F of the overall variables remarkably increased. Then by regressive diagnosis on this model, it is demonstrated that for the four variables participating in modeling, the $|t|$ values are all above 2 (the smallest is 2.278) while VIF notably decreases (the maximum is 21.162), indicating that the model is robust. Here, the plot of the observed versus calculated values for the collision cross-section of 113 peptides is delineated in Fig. 3, the scatter plot of residual errors has been given out in Fig. 4, demonstrating that approximately no outliers are presented and only three residue values slightly exceed the positive-negative double root mean squares ($\pm$2RMS) but dramatically smaller than $\pm$3RMS. Besides, as is well employed to test outliers, Cook's distance is deemed to be well reflecting the influence of samples on modeling and their deviations from normal states, and thus the plot of Cook's distance to centered leverage is delineated in Fig. 5. The fact that all sample points are distributed in a small scope in the lower left corner of Fig. 5 (Cook's distance <0.1, centered leverage <0.13) indicates no prominent outliers here. However, as shown by Tropsha et al. [17–19], modeling stabilities are underdetermined only by q value, so 13 out of the 113 samples are selected out into the test set. This is further employed to validate the modeling predictabilities to the external samples, and the remaining 100 samples are served as the training set, constructing a regressive model. The selection rule is as follows: (a) to meet the demand of randomness and uniformity, two peptides at each length rank are randomly introduced into the test set; (b) each amino acid at each position in the test set should also be present at that position in the training set; (c) the collision cross-section range of the test set should not exceed that of the training set; (d) more peptide is added into the test set from the rank of longest peptide length, making up 13 test samples because it may be

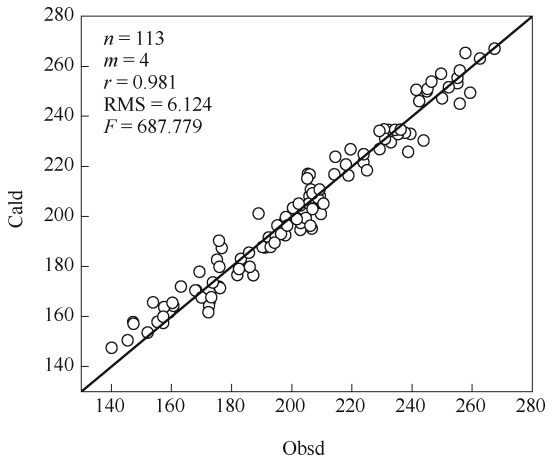

Fig. 3 Plot of calculated versus observed collision cross-sections of 113 singly protonated peptides by M2

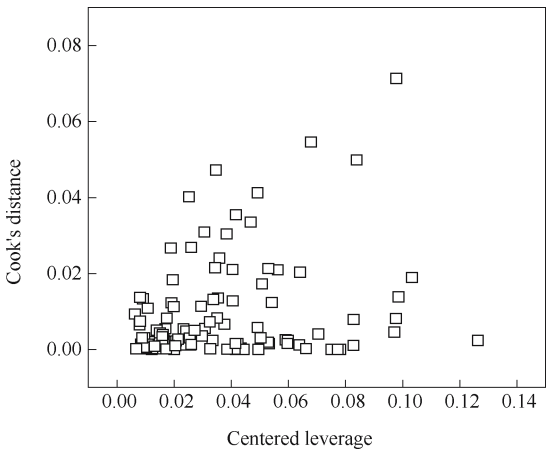

Fig. 5 Plot of the Cook's distance vs. the centered leverage values

difficult to express and predict long peptides. A 4-variable regressive model on the 100 training samples is ultimately given out with relative statistics as $r = 0.978$, RMS = 6.420, $F = 520.026$, $q = 0.976$, $\mathrm{RMS_{CV}} = 6.749$ and $F_{CV} = 468.141$, and the four variables used are the same as the four MEIV descriptors in M2. Using such a model to predict 13 test samples, satisfying results are generated with $r_{ext} = 0.997$ and $\mathrm{RMS_{ext}} = 2.761$, indicating the MEIV model is well preformed to be stable and generalized on peptide collision cross-section.

3.3 Partial least square regression model (M3)

Partial least square (PLS) [20,21] regression is a widely used data analyzing method in fields of analytical, physical and medicinal chemistry. Using the software Simca-p 10.0, a PLS model (M3), is constructed to relate each peptide MEIV descriptor (10 descriptors) with corresponding collision cross-section. The number of principal components (PCs) is determined unless the cross-validation leave-one-out q reaches the maximum, and the resulting three significant PCs comprise 97.0% information content of the original variable matrix X, explaining 96.0% variance of Y and 95.7% by cross-validation (Table 5). For a further exploration, the scoring scatter for 113 peptides at the top two PCs is contoured (Fig. 6), demonstrating that most samples fall into the Hotelling T^2 confidence ellipse with a 95% confidence level except for #105 and #109, which are structurally intricate and dramatically different from the others at the same rank (both belong to the rank of longest peptide length of 10). In addition, it is also found that according to length, the peptide sequences are obviously distributed from left to right as shown in Fig. 6, indicating a good reflection of peptide structural characteristics by the PLS PC space. Loading distributions of the 10 variables are presented in Fig. 7 wherein all the MEIV descriptors are positively related to the dependent variables Y at the first PC, suggesting a consistent changing trend between the collision cross-section and the peptide sequence length. Amongst, the variables v_{HH}, v_{HC} and v_{CC} contribute considerably to the top two PCs (loadings >0.3), representing the dominant C—H interactions as H—H, C—H, and C—C to the peptide structures. By a further external validation (100 samples as the training set and 13 as the test set), the resulting 3-PCs PLS model ($r = 0.977$, RMS = 6.569, $q = 0.974$, $\mathrm{RMS_{CV}} = 6.798$) obtains a series of good results on the 13 test samples with its $r_{ext} = 0.996$ and $\mathrm{RMS_{ext}} = 2.919$.

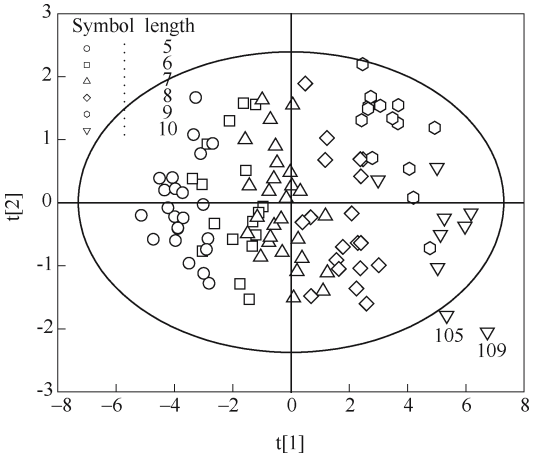

Fig. 6 Scoring plot for 113 samples scattered at the first and second principal components in M3 (Sequences of different lengths are highlighted by different symbols)

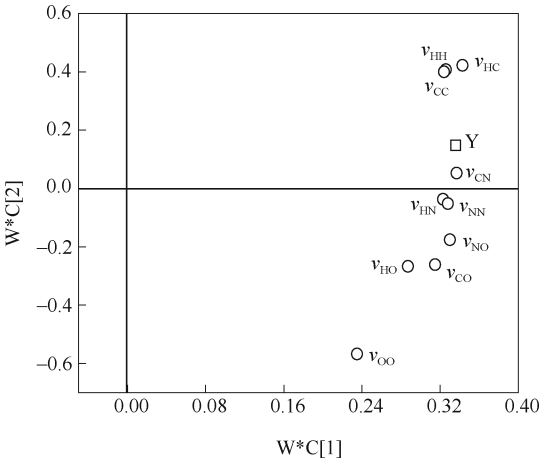

Fig. 7 Loading plot for distributions of 10 MEIV descriptors at the first and second principal components of M3

4 Conclusions

Excellent molecular structure descriptors should be potent not only to a strong characteristic extraction and property correlativity, but also to simple operation and easy calculation. Seeing that, MEIV, expressing the electrical topological properties for organic compounds, is developed from two dimensional invariants of molecular sketches. Three robust linear QSSR models for the 113 peptides are derived, obtaining satisfactory results. Dividing atoms in terms of groups in the periodic table of elements, the MEIV is proved to be

Table 5 Statistical data in M1, M2 and M3

Model	Method	n	m	r	RMS	F	q	$\mathrm{RMS_{CV}}$	F_{CV}
M1	MLR	113	10	0.983	5.944	293.256	0.979	6.575	293.256
M2	SMR	113	4	0.981	6.124	687.779	0.979	6.366	628.481
M3	PLS	113	3	0.980	6.245	/	0.978	6.417	/

efficacious to structural representation and activity predictions for drug and biomolecules. However, its idea is waited for a further improvement, including pattern recognition [22], multivariate analysis, variable combination with our proposed vectors [23–28] like molecular electronegative distance (MED) vector.

Acknowledgements The authors thank the State Key Laboratory of Chemo/Biosensing and Chemometrics Foundation (KCBCF 0501201), State Chuihui Project Fund (SCPF 990404+03-07), Fok-Yingtung Educational Foundation (Grant No. 98-7-6) and Chongqing University ZYX Juche Innovation Fund (CUIF 030506) for partly financial supports.

References

1. Shao S Y, Kan R F, Hou K Y, Li H Y. The development of research in ion mobility spectrometer. Modern Scientific Instruments, 2004, 4: 9–12 (in Chinese)
2. Hill H H, Siems W F, Louis R H, Mcminn D G. Ion mobility spectrometry. Anal Chem, 1990, 62: 1201A–1209A
3. Karas M, Hillenkamp P. Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons. Anal Chem, 1988, 60: 2299–2301
4. Fenn J B, Mann M, Meng C K, Wong S F, Whitehouse C M. Electrospray ionization for mass spectrometry of large biomolecules. Science, 1989, 246: 64–71
5. Beegle L W, Kanik I, Matz L, Hill H H. Electrospray ionization high-resolution ion mobility spectrometry for the detection of organic compounds 1. Amino acids Anal Chem, 2001, 73: 3028–3034
6. Myung S, Lee Y J, Moon M H, Taraszka J, Sowell R, Koeniger S, Hilderbrand A E, Valentine S J, Cherbas L, Cherbas P, Kaufmann T C, Miller D F, Mechref Y, Novotny M V, Ewing M A, Sporleder C R, Clemmer D E. Development of high-sensitivity ion trap ion mobility spectrometry time-of-flight techniques: a high-throughput Nano-LC-IMS-TOF separation of peptides arising from a drosophila protein extract. Anal Chem, 2003, 75: 5137–5145
7. Revercomb H E, Mason E A. Theory of plasma chromatography/gaseous electrophoresis. Anal Chem, 1975, 47: 970–983
8. Liao C, Chen Z, Yin Z, Li S Z. Preliminary approach to estimation and prediction of infrared spectroscopy for Mannich bases by atomic electronegativity distance vector (VAED). Comput Biol Chem, 2003, 27: 229–239
9. Zhou P, Zhou Y, Mei H, Tian F F, Li Z L. Atomic electronegativity interaction vector and atomic hybridation state index for spectroscopic simulation of ^{13}C nuclear magnetic resonance of amino acids. Chinese Journal of Analytcal Chemistry, 2006, 34(2): 200–204 (in Chinese)
10. Pauling L. The nature of chemical bond IV. Energy of single bonds and the relative electronegativity of atoms. J Am Chem Soc, 1932, 54: 3570–3582
11. Aylward G H, Findlay T J. SI Chemical data. 2nd ed. Beijing: Higher Education Press, 1985, 94–95
12. ZHANG Haiqi, CHEN Zheng, LIN Yingjie, MA Xiouli, ZHANG Yihua, SONG Lizhu, YANG Hua, WANG Dejun. Handbook of Chemical Data. Beijing: Science Press, 2001: 682–723
13. Mosier P D, Counterman A E, Jurs P C, Clemmer D E. Prediction of peptide ion collision cross-sections from topological molecular structure and amino acid parameters. Anal Chem, 2002, 74: 1360–1370
14. Eriksson L, Johansson E. Multivariate design and modeling in QSAR. Chemom Intell Lab Syst, 1996, 34: 1
15. Wold S. Cross-validation estimation of the number of components in factor and principal components models. Technometrics, 1978, 20: 897–903
16. Xu L, Shao X G. Methods of Chemometircs. 2nd ed. Beijing: Science Press, 2004: 52–57
17. Golbraikh A, Tporsha A. Beware of q2! J Mol Graphics Mod, 2002, 20: 269–276
18. Tpopsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and inerpretation of QSPR models. QSAR & Comb Sci, 2003, 22: 69–77
19. Gramatica P, Pilutti P, Papa E. Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. J Chem Inf Comput Sci, 2004, 44: 1794–1802
20. Wold S, Ruhe A, Wold H, Dunn W J. The collinearity problem in linear regression—the partial least squares (PLS) approach to generalized inverses. Siam J Sci Statist Comput, 1984, 5: 735–743
21. Wold S, Sjöström M, Eriksson L. PLS regression: a basic tool of chemometrics. Chemom Intell Lab Syst, 2001, 58: 109–130
22. Miyashita Y, Li Z, Sasaki S. Chemical pattern recognition and multivariate analysis for QSAR Studies, Trend Anal Chem (TrAC), 1993, 12(2): 50–60
23. Liu S S, Cai C Z, Li Z. Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector, λ. J Chem Inf Comput Sci, 1998, 38(3): 387–394
24. Liu, S S, Liu H L, Xia Z N, Cai C Z, Li Z S. Molecular Distance-Edge (MDE) Vector μ: An Extension from Alkanes to Alcohols. J Chem Inf Comput Sci, 1999, 39(6): 951–957
25. Li Z S, Fu B H, Wang Y Q, Liu S S. On Structural Parameterization and Molecular Modeling of Peptide Analogues by Molecular Electronegativity Edge Vector (VMEE): Estimation and Prediction for Biological Activity of Dipeptides. J Chin Chem Soc, 2001, 48(5): 937–944
26. Deng H, Huang P, Hu Y, Ye N, Li Z. A novel molecular distance edge vector as applied to chemical modeling of quantitative structure-retention relationships: Various gas chromatographic retention behaviors of polychlorinated dibenzofurans on different polarity-varying stationary phases, Chin Sci Bull, 2005, 50, 16: 1683–1687
27. Liu S S, Yin C S, Cai S X, Li Z. QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV-13. J Chem Inf Comput Sci, 2001, 41(2): 321–329
28. Liu S S, Cai S X, Cai C Z, Li Z L. Molecular Electronegative Distance Vector (MEDV) Relating to 15 Properties of Alkanes. J Chem Inf Comput Sci, 2000, 40(6): 1337–1348