Journal of Data Science **1**(2003), 405-423

Application of Orthogonal Block Variables and Canonical Correlation Analysis in Modeling Pharmacological Activity of Alkaloids from Plant Medicines

Qian-Nan Hu[1], Yi-Zeng Liang[1],
Xiao-Ling Peng[2], Yin Hong[2] and Lian Zhu[3]
[1,3]*Central South University and* [2]*Hong Kong Baptist University*

Abstract: A new kind of orthogonal block variables, derived from subspace projection and canonical correlation analysis, is applied to model pharmaological activity of alkaloids from plant drugs. The alkaloids are grouped into three cases by intravenous, intraperitoneal, and subcutaneous injections. Four block variables (family of variables) investigated in this work are valence molecular connectivity index, alpha kappa index, E-State index and element counts of molecules, respectively. The regression model embracing only few new orthogonal block variables against pharmaological activity shows significant improvement than those, say multiple linear regression (MLR) simply using original variables, principal component regression (PCR) and the ones selecting only one or two of the original family variables, both in fitting and prediction ability of the correlation model. The reason for this might be that the new orthogonal block variables in fact include almost all of the information of the original variables but without collinearity between them.

Key words: Alkaloids, canonical correlation analysis, orthogonal block variable, orthogonal variable, and plant drugs.

1. Introduction

Herbal medicine (HM) has a long therapeutic history over thousand years and is currently still serving many of the health needs of a large population in the world. However, currently existing approaches for quality assessment cannot fulfill the practical requirements of the safety and efficacy of HMs. One of these reasons might be that, unlike a chemically synthetic drug with much purity, a HM and/or a HM formula may consist of hundreds of complex phytochemicals. First to model the activity of individual composition from plant drugs, and then to study the synergistic action of its components might be useful for revealing the mystery of Chinese herb medicine. Thus, the technique developed in chemometrics, the so-called quantitative structure-activity relationships (QSAR), is used to fulfill the above-mentioned task.

The aim of QSAR is to relate the structure of a molecule to a biological activity by means of statistical tools, which can be expressed mathematically as follows (Devillers 1999):

$$A = f(\text{molecular structure}) = f(\text{molecular descriptors}), \qquad (1)$$

where A denotes the activity of chemical component, which is essentially a biological measurement value. In order to evaluate structural similarity and diversity of the molecules and/or to build QSAR model as shown in the above equation, one needs first to obtain the suitable numerical molecular descriptors associated with the molecular structure in QSAR researches. In fact, there are many molecular descriptors available, such as quantum chemical descriptors, physical chemical parameters, and topological indices, to describe the molecular structures. Only for topological indices, there emerged hundreds (Katritzky *et al.* 1994) of indices since 1947 (Wiener 1947). However, the multiplication of descriptors caused worry in some parts of the scientific community (Balaban and Ivanciuc 1999).

In QSAR research, to evaluate whether the information contents of descriptors are enough to describe the molecules and how much information is not "duplicated" by other descriptors are two very important aspects in building QSAR equations. Randić 1991 proposed orthogonal method to select variables. Xu and Zhang 2001 studied systematically some of the ingenious methods, such as forward selection, backward elimination, stepwise regression, leaps-and-bounds regression and genetic algorithm. In many

cases, the information of descriptors is not enough and, under this case, whether the variable selection method by deleting some variables for regression is the best choice is still a question. The methods in our former study (Du *et al.* 2002) and the present work might offer a new way to select variable by including almost all the information of original variables and, at the same time, reducing the number of variables.

In the research (Du *et al.* 2002), a subspace projection method is proposed to orthogonalize block variables in modeling the relationship between structure and retention index. The regression against retention index shows significant improvement both in fitting and predicting ability of the correlation model. Moreover, the quantitative intercorrelation between the different block variables of topological indices can also be evaluated by the proposed techniques. The basic idea is to first classify descriptors into different blocks (groups) and then, apply canonical correlation analysis to get new variables to represent the different blocks.

The alkaloids from natural sources are of importance in medical studies. Elbein and Molyneux 1999 reviewed alkaloids, isolated from natural resources, as inhibitor of glycoprotein processing. Wang and Xie 1999 reviewed the clinic effects of alkaloids of Chinese aconitum plants. To correlate the structure with activity of alkaloids is of use to predict the activity of other alkaloids, to deeply understand the changes of different chemical structures upon the activity and finally to make modification on the original structures to improve the activity.

Topological index has advantages of simplicity and quick speed of computation (important for large data) and so attracts attentions of scientist. What is important is that topological descriptors can explain most of the property modeled, as shown by some researchers (Basak *et al.* 1999, and Brown and Martin 1997). The research (Basak *et al.* 1999) indicates that the easily calculable topostructural and topochmical indices will be an effective first choice in QSAR studies. Brown and Martin 1997 concludes that 2D descriptors are better than 3D descriptors from information content. There are many kinds of topological descriptors in modeling pharmaological activity of drugs (Hu *et al.* 2003a). In this work, three most popular topological index families are first selected to build the statistical model. The first is valence molecular connectivity index (Kier and Hall 1976), sec-

ond is alpha kappa index (Kier 1986), and the third is e-state indices (Kier and Hall 1990). In order to describe the heteroatomic effect in investigated alkaloids, element counts are also included as the fourth block variables. The valence molecular connectivity index has wide applications (Kier and Hall 1986, and Hall and Kier 1991) in modeling activity of drugs. Kappa index codes information of cyclicity, spatial density, centrality of branching, and symmetry of molecules (Kier 1986, and Kier and Hall 1999) and it has been applied to many situations in QSAR researches (Kier 1985, 1997, and Shen 1967). The E-State index (Kier 1986, and Hall and Kier 1999a) is a very successful topological index for modeling activity of drugs, which is discussed in detail in a book (Hall and Kier 1999b). E-State indices have been used in molecular similarity and diversity research, and QSAR study (Hall *et al.* 1995, Kellogg *et al.* 1996, Hall and Vaughn 1997, and Hall and Story 1996). Furthermore, element counts combined with other topological indices have also been successfully used in QSAR studies (Balaban *et al.* 1992a, and Balaban *et al.* 1992b). It is worthy noting that none of any single family of the above mentioned variables could give satisfactory results if one tries to correlate them individually with pharmaological activity of alkaloids from plant drugs. Thus, in the present work, orthogonal block variables, derived from subspace projection and canonical correlation analysis, are applied to model pharmaceutical activity of alkaloids from plant drugs. The regression shows that the results by a few orthogonal block variables including almost all of the information of original descriptors are much better than by selecting one or two of the original family variables.

2. Methodology

In the former study (Du *et al.* 2002), orthogonal block variables that are from some families of topological indices or quantum chemical parameters were proposed by applying a subspace-projection method. The outline of the method is only briefly given in the following sections.

2.1 Orthogonalization of block variables by subspace projection

A series (or a family) of topological indices (not individual index) with

similar calculation strategy were often encountered, such as the molecular connectivity indices (${}^0\chi, {}^1\chi, {}^2\chi, {}^3\chi_p, {}^3\chi_c, {}^4\chi_p, \cdots$), Kappa indices (${}^0\kappa, {}^1\kappa, {}^2\kappa$, ${}^3\kappa$, $\cdots$). A series of descriptors were generally defined by accounting for more molecular structure information and less redundancy. Thus, a series of descriptors might be considered as an ensemble named block descriptor (variable), which includes all individual descriptors in this series. Being similar to the orthogonalization of individual descriptor, orthogonal block descriptors (variables) would also be obtained easily. The advantage of using block descriptor is that one may work with only a few block variables instead of many individual variables. The procedure of orthogolization of the block variables could be fulfilled in the following steps:

1. The procedure starts by selecting a block variable say X_1, as the first orthogonal matrix Ω_1. The second orthogonal matrix Ω_2 can be obtained through the orthogonal projection, that is

$$\Omega_2 = \mathbf{X}_{21} = (\mathbf{I} - \mathbf{X}_2(\mathbf{X}_2^t\mathbf{X}_2)^{-1}\mathbf{X}_2^t)\mathbf{X}_1 \tag{2}$$

2. Ω_3, which will be orthogonal with both Ω_1 and Ω_2, can be calculated easily by first defining $\mathbf{X}_j = [\Omega_1\Omega_2], \mathbf{X}_i = \mathbf{X}_3$ and then using the following equation, that is,

$$\mathbf{X}_{ij} = (\mathbf{I} - \mathbf{X}_j(\mathbf{X}_j^t\mathbf{X}_j)^{-1}\mathbf{X}_j^t)\mathbf{X}_i \tag{3}$$

Similarly, a series of orthogonal matrices of $\Omega_1, \Omega_2, \cdots, \Omega_n$ can be obtained.

2.2 Canonical correlation analysis (CCA)

Canonical correlation analysis (CCA) (Mardia *et al.* 1979) offers a way to establish the maximum correlation between variables. The original aim of CCA is to find linear combinations of **Xa** and **Yb**, which makes the correlation between **Xa** and **Yb** maximum. **Xa** and **Yb** are called canonical correlation variables. Only consider variance of v(**Xa**) and v(**Yb**) to be one and if there exits $\mathbf{a}_1$ and $\mathbf{b}_1$ making $R(\mathbf{Xa}_1, \mathbf{Yb}_1) = \max R(\mathbf{Xa}, \mathbf{Yb})$, then, $\mathbf{Xa}_1$ and $\mathbf{Yb}_1$ are called as the first pair canonical correlation variables.

After getting the first pair variables, second, third and so on pair variables can be found step by step. The canonical correlation variables reflect the linearity between $\mathbf{X}$ and $\mathbf{Y}$. The problem of obtaining the canonical correlation variables is how to calculate the eigenvalues and eigenvectors of the matrix $\mathbf{K}$=$(\mathbf{V}_{\mathbf{XX}}^{-1/2})\mathbf{V}_{\mathbf{XY}}(\mathbf{V}_{\mathbf{YY}}^{-1/2})$. Through singular value decomposition of the matrix $\mathbf{K}$, $\mathbf{u}_i$ and $\mathbf{v}_i$ can be obtained by $\mathbf{K}$=$[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r]\mathbf{S}[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r]$t. The canonical correlation variables can be calculated by the formula

$$\mathbf{b}_i = \mathbf{V}_{\mathbf{XX}}^{-1/2}\mathbf{u}_i \tag{4}$$

$$\mathbf{a}_i = \mathbf{V}_{\mathbf{YY}}^{-1/2}\mathbf{v}_i \qquad (i = 1, 2, \cdots, r) \tag{5}$$

And then, $\mathbf{Xa}_1$ and $\mathbf{Yb}_1$ are obtained as the ith pair of canonical correlation variables.

2.3 Outlines of the calculation procedure

1. Split all the given descriptors into a few subsets, say $\mathbf{X}_1$, $\mathbf{X}_2$, $\cdots$ $\mathbf{X}_n$, each of which comes from the same family of descriptors proposed by the same authors.

2. The block variables $\mathbf{X}_1$, $\mathbf{X}_2$, $\cdots$ $\mathbf{X}_n$ were then mean-centered.

3. Orthogonalize block variables by equation (3). Note that the order of variables strongly impacts on the orthogonalization result. Here we use "based on R_i" approach to orthogonalize variables. First pick up a block variable in the set of $\mathbf{X}_1$, $\mathbf{X}_2$, $\cdots$ $\mathbf{X}_n$ with maximum correlation coefficient R against the property $\mathbf{y}$ as the first orthogonal block variable Ω_1. Then for the remaining block variables, calculate their orthogonal block variables to Ω_1 by equation (2), and select the orthogonal block variables with maximum R in the left ones as the second orthogonal block variables Ω_2. The third orthogonal block variable Ω_3 is such orthogonal one to Ω_1 and Ω_2 that have maximum R in the remaining ones. Other orthogonal block variables have the same calculation procedure.

4. The canonical correlation variables can be calculated by using equations (4) and (5). Note that $\mathbf{Y}$ is actually the property vector $\mathbf{y}$ in the present work. Thus, $\mathbf{b}_1$ is a scalar and there is only one pair of canonical correlation variable for each orthogonal block variable with $\mathbf{y}$. The new orthogonal variables, $\omega_1, \omega_2, \cdots, \omega_n$, corresponding to the orthogonal block variables, say $\Omega_1, \Omega_2, \cdots, \Omega_n$ are then used to build the regression model.

5. Select a few variables with maximum correlation coefficient R_i to establish the descriptor-property correlation model if necessary.

3. Experimental Data

3.1 Drug data collection

The total 65 compounds, all of the alkaloids with LD50 for mice of the reference (Shakirov *et al.* 1996), are from plant drugs. The details of the compounds are listed in the Table 1, which is divided into three cases, according to different injections, intervenes, intraperitoneal, and subcutaneous injections, respectively. The column (NO.) of Table 1 corresponds to the names of the compounds. The data of activity values ($\mathbf{y}$) and all the numerical descriptors ($\mathbf{X}$) of the compounds are not given here for the sake of brevity of the paper. They are available from the corresponding author, if readers are interested in them.

3.2 Descriptor calculation

In the present work, four series of descriptors are selected. They are valence molecular connectivity index (${}^0\chi^v, {}^1\chi^v, {}^2\chi^v, {}^3\chi^v_p, {}^3\chi^v_C, {}^4\chi^v$) (Kier and Hall 1976), alpha kappa shape index (${}^1\kappa_\alpha, {}^2\kappa_\alpha, {}^3\kappa_\alpha, \Phi$) (Kier 1986), E-State index (Kier and Hall 1990), and element counts (N_C, N_O, N_N), respectively. The descriptors are calculated by the heuristic queue notation (H.Q.N.) system (Hu *et al.* 2003b). The descriptors used in the QSAR studies of the three cases are listed in Table 2. The indices from same sources, such as proposed by the same author or derived from the same invariants, should

Table 1: Active compounds from plant drugs and the biological activities

NO.	Name	LD50(I/v	I/P	S/C)mg/kg
1	14-DEHYDROBROWNINE	68		
2	DEOXYPEGANIDINE	143	254	380
3	DEOXYPEGANINE	24		54
4	DIPTERINE			550
5	DUBINIDINE		885	970
6	(+)-OTHSENINE		630	
7	PANCRATINE	280		
8	PACHYCARPINE(+)-	26		90
	SPARTEINE	26		
9	PEGANIDINE	143	254	380
10	(+-)-PEGANINE	78.7		220
11	PEGANOL	130		
12	PSEUDOKOPSININE		76	125
13	(+)-PSEUDOEPHEDRINE	100		
14	PUBERACONITINE	22.5		
15	RANACONITINE	6.2		
16	RESERPINE	28		
17	RESERPININE	148		
18	(-)-ROEMERINE	38.8		79.5
19	RETAMINE			1185
20	RINDERINE		562	
21	(+-)-SALSOLIDINE	170		
22	SARRACINE			1250
23	SENECIONINE	64.1		
24	SEPACONITINE	16.5		
25	SINAOACUTINE		115	
26	SKIMMIANINE		160	
27	SOPHORCARPINE	39.43		
28	STEPHARINE	245		
29	SUPININE	222.5		
30	AJMALINE		130	206
31	AKUAMMIDINE	391		550

Table 1 (continued)

NO.	Name	LD50(	I/v	I/P	S/C	)mg/kg
32	β-ALLOCRYPTOPINE				220	
33	ALSTONINE		8.8			
34	AMMODENDRINE				385	
35	ANABASAMINE			159		
36	(-)-ANABASINE			10.2	13.7	
37	ANONAINE				109	
38	ARMEPAVINE		22.2			
39	BERBERINE		9.5		13.3	
40	(+)-BICUCULLINE		0.3		1.48	
41	BREVICARINE				375	
42	BREVICOLLINE				146	
43	BUXTAUINE				220	
44	(-)-VASICINONE		152		1133	
45	VINERIDINE		125	485		
46	(-)-VINCADIFFORMINE		90	225		
47	(+)-VINCAMINE		57	411		
48	VINCANIDINE			85	85	
49	(-)-VINCANINE		5.6	13.6	14	
50	VINERVINE		24.5	100	102	
51	VINERVININE			115		
52	GALANTHAMINE		58.5	131.2	200	
53	HARMINE		75	124		
54	HELIOTRINE		274.4			
55	HAEMANTHAMINE				318	
56	GENTIANADINE			1210		
57	GENTIANAINE			1275		
58	GENTIANAMINE			770		
59	GENTIANINE			460	504	
60	HETERATISINE		192.5			
61	(+)-β-HYDRASTINE		0.102		0.97	
62	HIPPEASTRINE		195	670	800	
63	GLAUCINE		33		420	
64	HORDENINE		131			
65	GRAVEOLINE			363.5		

hold some common information and should be classified into the same group. Thus, the descriptors are divided into four block variables.

Table 2: The topological descriptors and their corresponding values for deoxypeganine

Indices	Values	Indices	Values
${}^0\chi^\nu$	9.7024	$S_{(-CH_3)}$	0
${}^1\chi^\nu$	8.3900	$S_{(-CH_{2-})}$	4.6834
${}^2\chi^\nu$	8.7493	$S_{(-CH<)}$	0
${}^3\chi^\nu$	8.9924	$S_{(>C<)}$	0
${}^3\chi^\nu$	1.2145	$S_{(=CH_2)}$	8.4389
${}^4\chi^\nu$	9.0905	$S_{(=C^-)}$	3.8378
${}^1\kappa$	4.3783	$S_{(=C<)}$	0
${}^2\kappa$	2.6717	$S_{(-NH-)}$	2.3911
${}^3\kappa$	1.2380	$S_{(-N<)}$	4.6498
	4.3783	$S_{(=N-)}$	0
Nc	11	$S_{(=0)}$	0
No	0	$S_{(-0-)}$	0
N_N	2	$S_{-OH)}$	0

In order to give the readers an intuitive impression of how to get the numerical quantifier of the molecular structures, an example (No. 3 deoxypeganine) from Table 1 is given to show the procedure. The chemical structure of deoxypeganine is shown in Figure 1. With the help of the structure, the topological indices can be calculated by the definitions listed in the proceeding paper (Hu *et al.* 2003a). What should be noted is that the original definitions for hydrocarbons are modified by introducing some chemical parameters for molecules with multiple bonds and/or hetero-atoms. An example of the indices of deoxypeganine is listed in Table 2.

4. Results and Discussion

The aim in QSAR is to use the equation (1) to build a model correlating

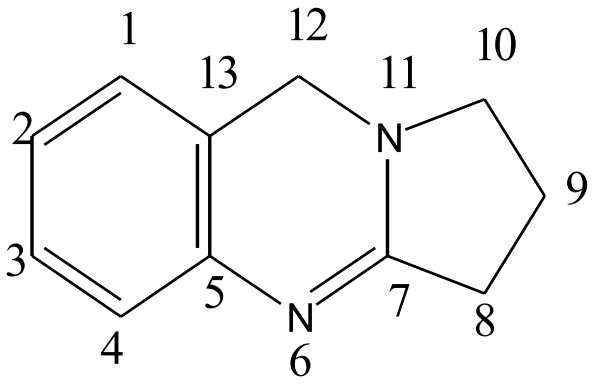

Figure 1. Molecular skeleton and numbering of atoms of deoxypeganine

the numerical molecular descriptors with their corresponding activities so as to further predict the activities of the similar molecules. In general, the linear model is the first choice, since the reason why the molecules have activities can be easily deduced with the linear model. From Table 1 and above discussion, one could easily see that the number of the samples is rather small, say 39, 26, and 32, respectively, in the present study. However, the number of variables included in the model is 26 (see Table 2), which hints that the overfitting might be the most serious problem to be faced in this work.

4.1 Correlation by different descriptors

First, we tried to use one family of molecular descriptors to build regression model. However, the regression results listed in the Table 3 are quite disappointed. The information contents of any individual group of variable are not enough to obtain satisfactory results. Then, the whole variables are used to model the activities, and the regression coefficients for the three cases are 0.8781, 0.9993, 0.9797, respectively. The fitting results seem to be quite good. In order to check the stability of the built models, leave-one-out cross-validation is applied for the three cases using cross-validated root

mean square error of prediction (RMSECV) criteria, that is

$$RMSECV = \sqrt{\frac{PRESS}{n}}, \tag{6}$$

in which n is the number of observation and PRESS is the predicted residual squared sum. The results are collected in Table 4. From the table, one can easily conclude that the prediction ability of the model is very bad, say 238.3407, 1.3676e+003, 1.3777e+003, respectively. This means the overfitting is clearly embedded in the MLR models. In order to cure such situations in QSAR researches, the chemists always resort to the principle component regression (PCR) and partial least squares (PLS) developed in chemometrics, since these techniques may reduce the dimension of variable space efficiently.

Table 3: Regression of alkaloids by multiple linear regression with different variables*

Variables and Methods	Regression Results	Case 1	Case 2	Case 3
X1	R=	0.5549	0.6861	0.4189
	s=	75.70	246.56	316.27
	F=	2.37	4.67	1.44
X2	R=	0.3614	0.5564	0.5354
	s=	84.84	281.62	294.18
	F=	1.28	2.35	2.71
X3	R=	0.7201	0.8697	0.7982
	s=	63.13	167.27	209.80
	F=	2.33	2.87	2.78
X4	R=	0.3962	0.7849	0.4833
	s=	83.55	210.01	304.92
	F=	2.17	11.77	2.84
all variables	R=	0.8781	0.9993	0.9797
	s=	43.53	12.26	69.82
	F=	1.75	31.81	5.73

*X1: valence molecular connectivity index; X2: Kappa index; X3: E-State index; X4: Element counts.

Table 4: RMSECV of CCA and MLR for the three cases

	iv39	ip26	Sc32
CCA	52.0379	15.6558	82.5655
MLR	238.3407	1.3777e+003	1.3676e+003

Table 5: Correlation coefficient and standard error of every principal component from PCR for the three cases

	sc32		iv39		ip26	
pc	R	s	R	s	R	s
1	0.2332	349.6210	0.2081	91.3389	0.4688	311.3538
2	0.0973	357.8284	0.0106	93.3780	0.5588	292.3163
3	0.3768	333.0342	0.4107	85.1436	0.0976	350.7967
4	0.2536	347.7805	0.0932	92.9769	0.3642	328.2708
5	0.4979	311.7970	0.0430	93.2967	0.0412	352.1793
6	0.1965	352.5266	0.0043	93.3824	0.2592	340.4349
7	0.2435	348.7104	0.0802	93.0824	0.2881	337.5363
8	0.0615	358.8534	0.3256	88.2959	0.1083	350.4053
9	0.1471	355.6232	0.3121	88.7173	0.1995	345.3909
10	0.0227	359.4420	0.2338	90.7953	0.1536	348.2951
11	0.1356	356.2144	0.2274	90.9375	0.2406	342.1270
12	0.1596	354.9270	0.0566	93.2334	0.0074	352.4697
13	0.0745	358.5364	0.1219	92.6863	0.0525	351.993
14	0.0782	358.4336	0.1438	92.4120	0.0799	351.3534
15	0.0339	359.3281	0.0122	93.3764	0.0436	352.1445
16	0.1026	357.6370	0.1224	92.6816	0.1026	350.6207
17	0.0296	359.3766	0.0656	93.1820	0.1395	349.0350
18	0.3634	334.9511	0.0119	93.3766	0.2817	338.2043
19	0.1032	357.6147	0.1587	92.1994	0.1180	350.0174
20	0.1014	357.6817	0.0468	93.2810	0.1461	348.6961
21	0.0737	358.5559	0.0866	93.0321	0.0278	352.3433
22	0.0883	358.1304	0.2409	90.6340	0.1282	349.5728
23	0.1488	355.5314	0.2519	90.3708	0.1339	349.3071
24	0.0935	357.9582	0.0111	93.3776	0.1245	349.7389
25	0.2380	349.2028	0.2298	90.8851	0.2252	343.4281

4.2 Limitation of PCR

The results obtained from PCR are shown in Table 5. From the table, the correlation coefficients for all the individual principal components from the first to the 25^{th} (Table 5) show that the order of the values of R has nothing to do with the order of the eigenvalues of the principal components. Thus, it is impossible to select reasonably the number of principal components to be included in PCR model.

Commonly in chemoemtrics, the leave-one-out cross-validation is adopted to choose the right number of the principal components, which are shown in Figure 2 for the three cases. From the plots, it can be seen that several minima in the curves are found, which makes the choice of right number of principal components very difficult. For instance, the R and s for the minima at five and seven principal components are 0.6809 and 263.3081; and 0.7682 and 230.1728 respectively, for Sc32 case, which definitely cannot be accepted by chemists.

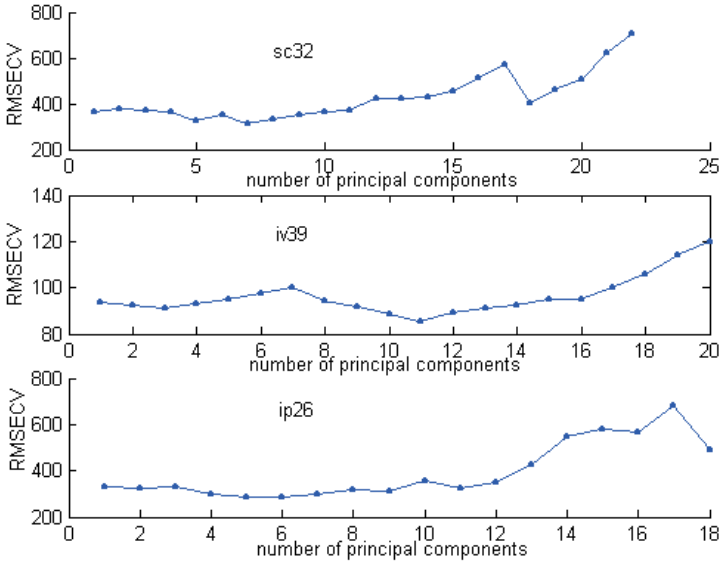

Figure 2: Relationships of cross-validation (leave-one-out) vs the number of principal components for the three cases.

Table 6: Regression results with Orthogonal Block Variables and RMSECV by Cross-validation

		iv39	ip26	Sc32
Regression with OBV	R=	0.8781	0.9993	0.9797
	s=	43.53	12.26	69.82
	F=	28.63	4007.8	161.19
RMSECV by Cross-validation		52.04	15.66	82.57

4.3 Improvement on correlation using block variables and CCA

Since the molecular descriptors are from four different families, they can be grouped into four blocks, and then all the blocks are replaced by new orthogonal block variables with the help of canonical correlation analysis. Then, the four orthogonal block variables are utilized to build the regression model through"based on R_i" approach described in methodology section. The regression results obtained by the method proposed in this work are shown in Table 6. One can see that the regression coefficients, standard errors and F test are quite satisfactory. In order to check the stability of the model, leave-one-out cross-validation is also adopted. The RMSECV are quite close to the size of the standard errors of the model, which indicates that there is no overfitting in the model and the prediction ability of the model is also quite good. All these show that the orthogonal block variables by subspace projection and canonical correlation analysis may offer a new method to reconstruct the variables and the method proposed in this paper might have a promising prospect in QSAR researches and data mining in chemistry.

Acknowledgement

This project is financially supported by National Nature Foundation Committee (NNFC) of P. R. China (No. 20235020, 20175036). And the authors also appreciate the hospitality of Hong Kong Baptist University, when the authors attend "Workshop on Data Mining in Traditional Chinese Medicines" from March to May in 2002. Finally, the valuable comments and suggestions for improving the paper from Prof. Chao, M. T. are highly

appreciated.

References

Balaban, A. T., and Ivanciuc, O. (1999). Historical development of topological indices. In *Topological Indices and Related Descriptors in QSAR and QSPR.* (Devillers, J. and Balaban, A. T. eds.) Gordon and Breach Science Publishers, 455-489.

Balaban, A. T., Kier, L. B., and Joshi, N. (1992a). Correlations between chemical structure and normal boiling points of acyclic ethers, peroxides, acetals, and their sulfur analogues. *J. Chem. Inf. Comput. Sci.,* **32**, 237-244.

Balaban, A. T., Joshi, N., Kier, L. B., and Hall, L. H. (1992b). Correlations between chemical structure and normal boiling points of halogenated alkanes C1-C4. *J. Chem. Inf. Comput. Sci.,* **32**, 233-237.

Basak, S. C., Gute, B. D., and Grunwald, G. D. (1999). A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters. In *Topological Indices and related descriptors in QSAR and QSPR.* (Devillers, J. and Balaban, A. T. eds.) Gordon and Breach Science Publishers, 675-696.

Brown, R. D., and Martin, Y. C. (1997). The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Compit. Sci.,* **37**, 1-9.

Devillers, J. (1999). No-free-lunch molecular descriptors in QSAR and QSPR. In. *Topological Indices and Related Descriptors in QSAR and QSPR.* (Devillers, J. and Balaban, A. T. eds.) Gordon and Breach Science Publishers, P1-17.

Du, Y. P., Liang, Y. Z., Li, B. Y., and Xu, C. J. (2002). Orthogonalization of block variables by subspace-projection for quantitative structure property relationship (QSPR) data. *J. Chem. Inf. Comput. Sci.,* **42**, 1128-1138.

Elbein, A. D., and Molyneux, R. J. (1999). Inhibitors of glycoprotein processing. *Iminosugars Glycosidase Inhib.,* 216-251.

Hall, L. H., and Kier, L. B. (1991). The molecular connectivity chi indexes and kappa shape indexes in structure-property relations. In. *Reviews of Computational Chemistry.* (Boyd, D. and Lipkowitz, K. eds) VCH Publishers, Inc. Chap. 9, pp 367-422.

Hall, L. H. and Kier, L. B. (1999a). The electrotopological state: structure modeling for QSAR and database analysis. In *Topological Indices and Related Descriptors in QSAR and QSPR.* (Devillers, J. and Balaban, A. T. eds.) Gordon and Breach Science Publishers.

Hall, L. H., and Kier, L. B. (1999b). *Molecular Structure Description: The Electrotopological State.* Academic Press.

Hall, L. H., Kier, L. B., and Brown, B. B. (1995). Molecular similarity based on novel atom-type E-State indices. *J. Chem. Inf. Comput. Sci.,* **35**, 1074-1080.

Hall, L. H., and Story, C. T. (1996). Boling point and critical temperature of a heterogeneous data set: QSAR with atom-type E-State indices using artificial neural networks. *J. Chem. Inf. Comput. Sci.,* **36**, 1004-1014.

Hall, L. H., and Vaughn, A. T. (1997). QSAR of phenol toxicity using E-state and kappa shape indices. *Med. Chem. Res.,* **7**, 407-416.

Hu, Q. N., Liang, Y. Z., and Fang, K. T. (2003a). The matrix sxpression, topological index and atomic attribute of molecular topological structure. *J. Data Science,* in press.

Hu, Q. N., Liang, Y. Z., Wang, Y., L. Guo, F. Q., and Huang, L. F. (2003b). The basic principles of heuristic queue notation and its applications in calculation of matrix and topological index for topological graphs. *Computers and Applied Chemistry,* **5**, in press.

Katritzky, A. R., Lobanov, B., and Karelson, M. (1994). *CODESSA, Comprehensive Descriptors for Structural and Statistical Analysis.* Reference Manual version 2.0, University of Florida, Gainesville, FL.

Kellogg, G. E., Kier, L. B., Gaillard, P., and Hall, L. H. (1996). E-state fields: application to 3-D QSAR. *J. Comp. -Aid. Molec. Des.,* **10**, 513-520.

Kier, L. B. (1985). A shape index from chemical graphs. *Quant. Strucut. Act. Relat.,* **4**, 109-116.

Kier, L. B. (1986). Shape indexes of orders one and three from molecular graphs. *Quant. Struct. -Act. Relat.,* **5**, 1-7.

Kier, L. B. (1997). Kappa shape indices for similarity analysis. *Med. Chem. Res.,* **7**, 8-12.

Kier, L. B., and Hall, L. H. (1976). *Molecular Connectivity in Chemistry and Drug Research.* Academic Press, New York.

Kier, L. B., and Hall, L. H. (1986). *Molecular Connectivity in Structure-Activity Analysis.* Research Studies Press Ltd, New York.

Kier, L. B., and Hall, L. H. (1990). An electrotopological state index for atoms in molecules. *Pharm. Res.,* **7**, 801-807.

Kier, L. B., and Hall, L. H. (1999). The kappa indices for modeling molecular shape and flexibility. In *Topological Indices and Related Descriptors in QSAR and QSPR.* (Devillers, J., and Balaban, A. T. eds.) Gordon and Breach Science Publishers. P455-490.

Mardia, K. V., Kent, J., and Bibby, J. (1979). *Multivariate Analysis.* Academic Press: New York.

Randić, M. (1991). Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.,* **31**, 311-317.

Shakirov, R., Telezhenetskaya, M. V., and Bessonova, I. A. (1996). Alkaloids, plants, structures, properties. *Chem. Nat. Comp.,* **32**, 216-858.

Shen, T. Y. (1967). Antiinflammatory agents. *Top. Med. Chem.,* **1**, 19-29.

Wang, X. W., and Xie, H. (1999). Alkaloids of Chinese aconitum plants. *Drugs Future.,* **24**, 877-882.

Wiener, H. (1947). Structural determination of parafin boiling points. *J. Am. Chem. Soc.,* **69**, 17-20.

Xu, L., and Zhang, W. J. (2001). Comparison of different methods for variable selection. *Anal. Chim. Acta.,* **446**, 477-483.

Received May 20, 2003; accepted August 2, 2003

Qian-Nan Hu
Research Center of Modernization of Chinese Herb Medicine
College of Chemistry and Chemical Engineering
Central South University
Changsha, P.R. China
qnhu@263.net

Yi-Zeng Liang
Research Center of Modernization of Chinese Herb Medicine
College of Chemistry and Chemical Engineering
Central South University
Changsha, P.R. China
yzliang@public.cs.hn.cn

Xiao-Ling Peng
Department of Mathematics
Hong Kong Baptist University
Hong Kong, P. R. China
01400908@hkbu.edu.hk

Yin Hong
Department of Mathematics
Hong Kong Baptist University
Hong Kong, P. R. China
01400924@hkbu.edu.hk

Lian Zhu
College of Information and Computer Science
Central South University
Changsha, P. R. China