Structural Interpretation of a Topological Index. 1. External Factor Variable Connectivity Index (EFVCI)

Qian-Nan Hu,[†] Yi-Zeng Liang,*[,†] Xiao-Ling Peng,[‡] Hong Yin,[‡] and Kai-Tai Fang[‡]

Institute of Chemometrics and Intelligent Analytical Instruments, College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, P.R. China, and Statistics Research and Consultancy Centre, Hong Kong Baptist University, Hong Kong, China

Received October 19, 2003

The external factor variable connectivity index (EFVCI) is interpreted by mining out the structural features hidden in the space spanned by the EFVCI indices through projection pursuit combining with number-theory net (NT-net) on the unit sphere U(Us). Projection pursuit is concerned with "interesting" projections of high-dimensional data sets to machine-pick "interesting" low-dimensional projections of a high-dimensional point cloud by numerically maximizing a certain objective function or projection index. At first, the optimal EFVCI index reaches to −0.80 in the correlation with a retention index of 207 hydrocarbons produced by insects. The EFVCI indices, with regression results of $R = 0.99998$, $s = 3.49$, RMSECV = 3.90, and $F = 7.9560e{+}005$, obtain high regression quality. The model is proven valid by leave-one-out cross validation. Second, the EFVCI index is interpreted by the structure information, that is, size, branch number, graph center, and branching position of topological structures, which is searched out on the unit sphere U(U^s) by projection pursuit. Finally, the interpretation information is used to discover some chemical knowledge concerning the variation of the retention index with the change in chemical structures.

INTRODUCTION

Since the advent of the Wiener index,[1] there have emerged more than 400 indices. Their advantage over the "traditional" molecular descriptors used in the so-called Hansch Analysis, or the quantum chemical descriptors derived, is that topological indices (TIs) are easily available and can be quickly computed for existing or virtual structures. However, even though the mathematical invariants are well defined, often they are accompanied with an important drawback: a lack of interpretation in terms of simple structural and physicochemical concepts. An apparent answer is that interpretation of a topological index is difficult.[2,3]

The chemists recognize that the problem of interpretation of topological indices is very important and as such should not be avoided. The topic of interpretation of topological indices has attracted some attention from different groups by using various methods in their work. Recently, Randić et al.[2,3] studied systematically the interpretation of topological indices. They have used bond additive and bond contribution to interpret the molecular connectivity index and several other topological indices. With the basic idea that topological indices can code molecular shape, size, branching, cyclicity, symmetry, centricity, compactness, diversity, and complexity as well, there are some authors[4−7] who consider the interpretation of a topological index by using various structural features. In the former studies, the structural interpretation is mainly carried out on the individual index.

In practice, the tackled problems are always involved with a multidimensional description of structures. Thus, to interpret the topological index by mining out the structural features hidden in the space spanned by several indices might provide some insights about the nature of the topological index.

In the same spirit our basic position when considering the interpretation of TIs is that we insist that topological indices have an interpretation within structural chemistry, which is also the position of Randić et al.[2,3] It is the goal of the topological index first to define the structural features of molecules mathematically and then to study the chemical consequences of the molecular features. The mystery of the topological index should be the structural features described by it, and so to mine out the features will be helpful to interpret the topological index and also the built model.

The first topological index[1] was triggered by the idea to describe the branching of structures. There are also some other topological indices claimed that they coded specific structural features. The molecular connectivity index[8] described mainly five general categories of molecular structure information, which include the degree of branching, the variable branching pattern, the position and influence of heteroatoms, the patterns of adjacency, and the degree of cyclicity. The Kappa shape indices[9] encoded, in combination, significant information on the degree of cyclicity and the degree of centralization/separation in branching. The third famous topological index created by Kier and Hall, the representative of the third generation of the topological index, is the E-State index[10,11] combining both electronic and topological factors together, which is interpreted by the combination with element content, electronic organization,

* Corresponding author phone: 86-731-8822841; fax: 86-731-8825637; e-mail: yzliang@public.cs.hn.cn.
[†] Central South University.
[‡] Hong Kong Baptist University.

10.1021/ci034225f CCC: $27.50
Published on Web 02/17/2004

and local topological state of atom or group. And there are also many indices[12−22] describing diversity, complexity, chirality, cyclicity, graph center, and branching et al.

The next interest is how to search the structural features in the space spanned by descriptors, which needs to reduce the high dimension to lower dimension. Projection pursuit[23−25] is concerned with "interesting" projections of high-dimensional data sets to machine-pick "interesting" low-dimensional projections of a high-dimensional point cloud by numerically maximizing a certain objective function or projection index. The projection pursuit has been applied successfully to tackle several chemical problems.[26−29] Two keys in the projection pursuit are as follows: (1) how to rotate the projection directions in a high-dimensional space to make the procedure searching efficiently in the whole space within an acceptable running time and (2) how to define the projection index.

In the present work, the TFWW (**T**ashiro, Y.; **F**ang, K. T.; **W**ang, Y.; and **W**ong, K. L.)[30,31] method, coproposed by one of the authors, of generating the number-theory net (NT-net) on the unit sphere U(U^s) is introduced, which is based on the good lattice point (GLP)[32,33] on cubic sphere. The TFWW method can generate uniform points on the unit sphere, and the calculation time is quick, which makes the method suitable in projection pursuit (PP).

Typically, PP uses a *projection index*, an objective function computed on a projected density (or data set), to measure the "interestingness" of the current projection, and then uses a numerical optimizer to move the projection direction to a more interesting position. In the projection pursuit, the interesting structure clusters or projection spaces are obtained, which is calculated by minimizing the "entropy" of the projections. Entropy is an index of "disorder" that measures the degree of the state of chaos. The lower the entropy, the more the data tend to segregate into clusters, corresponding to more ordered with some structure embedded in data set. Thus, if the entropy of state is minimized, which might provide some interesting structure clusters, when the high-dimensional data is projected on one dimension. Each projection direction has an entropy value. With the rotation of projection directions on Unit sphere, there will generate many entropy values. Finally, several minimum entropy are selected to find some interesting structure clusters hidden in the projection directions.

In 2000, Katritzky[34] et al. studied the QSPR correlation and prediction of GC retention indexes for methyl-branched hydrocarbons produced by insects, in which the studied structures are mainly mono-, di-, tri-, and tetramethylalkanes. The main differences between the compounds are the length of the chain, the position of the methyl groups, and the number of the methyl groups connected to the backbone. The cited authors analyzed that the descriptors chosen should reflect the relative positions and the number of the multiple methyl groups attached to the carbon chain, the conformation of the compounds, and the length of the carbon backbone, and four descriptors, the average information content, average structural information content, maximum electron−electron repulsion, and the Balaban J index, are selected to build the QSPR model with the regression coefficient equal to 0.9585 and $s = 4.6$. The cited authors also concluded that the topological factors govern the chromatographic retention behavior of methylalkanes.

To interpret quantitative structure−property relationships[35] seems to be more valuable than the model itself. Interpretation of the correlation of the property by molecular structures is the so-called knowledge discovery, which can give some insights on the changes of structures on the property and so can predict the property of other structures with some prior knowledge. However, in many cases, the built models are difficult to interpret due to the descriptors, which hide the described structural features in numerical form. The interpretation of the built model depends on the descriptors used, and another interesting thing is that the interpretation of the topological index itself is also a very difficult topic, although there are several prior works[2,3] on it.

Enlightened by former studies,[2,3,34,35] the authors try to apply the newly proposed external factor variable connectivity index (EFVCI)[36] to correlate the retention indexes produced by insects and interpret the EFVCI indices by the structure information hidden in EFVCI. The EFVCI is one of the latest developments of the molecular connectivity index, which introduced some flexible factors[37−40] to improve regression results, with the idea that a variable parameter undergoes change during the regression analysis. In the former study,[36] the structure information of EFVCI included intuitively size, branch number, graph center, and position of branching. In the present work the structure information is quantified by projection pursuit to interpret the EFVCI indices.

THEORY AND METHODOLOGY

Projection Pursuit.[23−25] Given the ($k \times n$) data matrix X, where k is the number of observed variables and n is the number of units, and an orthonormal matrix A($m \times k$), the ($m \times n$) matrix $Y = AX$ represents the coordinates of the projected data onto the m-dimensional ($m \ll k$) space spanned by the rows of A. Considering only one-dimensional projections of the form $Y = \alpha' X$, with α a ($k \times 1$) unit vector, projection pursuit consists of searching, with respect to α, the local maxima and global maximum of a function of the projected data H(Y), called the projection index, which measures the departure of Y from the normal distribution.

Method of Generating GLP Set[32,33] on Cubic Sphere C^s. The set based on the good lattice point by mode n is the so-called good lattice point (GLP) set.

Definition 1: Let ($n; h_1,.....h_s$) be a integral vector, in which $1 \le h_i < n$, $h_i \ne h_j$ (i≠j), $s < n$, and the greatest common division (n, h_i) = 1, $i = 1,.....s$. Denote

$$q_{ki} \equiv kh_i \pmod{n} \tag{1}$$

$$x_{ki} = (2q_{ki} - 1)/2n, \quad k = 1,...,n; i = 1, ...,s \tag{2}$$

in which $1 \le q_{ki} < n$. Then, the set $P_n = \{x_k=(x_{k1},.....,x_{ks}), k=1,.....,n\}$ is called the lattice point set of the spanning vector ($n; h_1,.....h_s$). If it holds the minimum discrepancy in all possible vectors, the P_n, is the GLP set. From eqs 1 and 2, the x_{ki} can be calculated by

$$x_{ki} = \{(2kh_i - 1)/2\}, \quad k = 1,...,n; i = 1,...,s \tag{3}$$

TFWW Algorithm.[30,31] After getting the GLP set, the NT-net on the unit sphere U(U^s) can be calculated by the TFWW

algorithm. If the dimension (s) is even, the outlines of the algorithm are mainly composed of the following steps:

(1.1) Generate NT-net on C^{s-1} $\{c_k=(c_{k1},.....,c_{k,\ s-1}), k=1,.....,n\}$.

(1.2) Set $g_{km} = 1$ and $g_{k0} = 0$, $k = 1,..., n$.

(1.3) When $k = 1, ..., n$, iterate to compute

$$g_{kj} = g_{k,j+1}c_{kj}^{1/j}\ j = m - 1, m - 2, ..., 1 \quad (4)$$

(1.4) Compute $d_{kl} = \text{sqrt}(g_{kl}-g_{k,l-1})$ and

$$x_{k,2l-1} = d_{k,l}cos(2\pi c_{k,m+l-1}),\ l = 1,..., m \quad (5)$$

$$x_{k,2l} = d_{k,l}sin(2\pi c_{k,m+l-1}),\ k = 1,...,n \quad (6)$$

Then, $\{x_k=(x_{k1},....., x_{ks},\ k=1,...,n)$ is the NT-net on U^s.

If the dimension (s) is odd, the steps are as follows:

(2.1) Generate NT-net on C^{s-1} $\{c_k=(c_{k1},.....,c_{k,s-1}), k=1,.....,n\}$.

(2.2) Set $m = (s-1)/2$, $g_{km} = 1$, and $g_{k0} = 0$, $k = 1,..., n$.

(2.3) When $k = 1, ..., n$, iterate to compute

$$g_{kj}=g_{k,j+1}c_{kj}^{2/(2j+1)}\ j = m - 1, m - 2, ..., 1 \quad (7)$$

(2.4) Compute $d_{kj} = sqrt(g_{kj} - g_{k,j-1}),\ j = 1, ...m,\ j = 1,...,n$ (8)

(2.5) When $k = 1,..., n$, compute

$$x_{k,l}=d_{k,1}(1\text{-}2c_{k,m})$$

$$x_{k,2} = d_{k,l}sqrt(c_{km}(1 - c_{km}))cos(2\pi c_{k,m+1})$$

$$x_{k,3} = d_{k,l}\ sqrt(c_{km}(1 - c_{km}))sin(2\pi c_{k,m+1})$$

$$x_{k,2l} = d_{k,l}cos(2\pi c_{k,2l})$$

$$x_{k,2l+1} = d_{k,l}sin(2\pi c_{k,2l})\ l = 2,...,m \quad (9)$$

Then, $\{x_k=(x_{k1},.....,x_{ks},k=1,...,n)\}$ is the NT-net on U^s.

Projection Index.[26,27,29] The projection function H(·) can be chosen in a wide set of functions, all minimized by the Gaussian density: Shannon entropy, Fisher information, orthonormal polynomial expansions of the distance between the empirical density and the normal one, and the classical tests of normality based on the Kolmogorov-Smirnov and the χ^2 statistics. Each of these indices picks up different aspects of nonnormality (skewness, kurtosis or multimodality), but this is irrelevant for our goal. i.e. the detection of the variables that contribute most to the definition of the structure appearing in a projection, independently from the index employed to obtain it.

In the study, the high-dimensional data are projected on one dimension, and for the one-dimensional data, the entropy is defined as

$$\text{entropy} = -\int_{-\infty}^{+\infty} f(x)logf(x)dx \quad (10)$$

where $f(x)$ is a kernel density estimate. When the data are discrete, an applicable form of entropy in eq 10 is changed to be

$$\xi = -p_i\log(p_i) \quad (11)$$

in which $p_i = m_i/m$, where m is the number of all samples, and m_i is the number of samples falling into the ith interval. Entropy is an index of "interestingness" that measures the nonnormality of the data. The lower the entropy, the more the data tend to segregate into clusters.

DATA COLLECTION AND DESCRIPTORS

The retention indexes of 207 (without 22-0822, for its structure seems incorrect) hydrocarbons are taken from ref 34. The indices used are as follows: ${}^0\chi^{efvci}$, ${}^1\chi^{efvci}$, ${}^2\chi^{efvci}$, ${}^3\chi^{efvci}$, ${}^3\chi^{efvci}_{cluster}$, ${}^4\chi^{efvci}$, ${}^4\chi^{efvci}_{path-cluster}$. The descriptors are calculated by the in-house software, Heuristic Queue Notation system (H.Q.N.s),[41] which can compute about 320 topological indices. To compare the regression performance of ref 34, the transformed, by subtracting the number of carbons in the main chain*100, retention index is applied. To follow the common methods in modeling retention behavior, the original retention index data are also studied, which are listed in Table 1. The table also contains the number of atoms, the branch number, and the first branching position of all 207 structures.

RESULTS AND DISCUSSION

Brief Introduction of the Former Study[34] on Insect Alkanes. Four descriptors, the average information content (AIC), the average structural information content (ASIC), the maximum electron−electron repulsion ($E^{max}_{el-el.repuls}$), and the Balaban J index, are selected to build the QSPR model with $R^2 = 0.9585$ and $s = 4.6$, which is validated by the cross-validated correlation. The first-order AIC is dependent on the number of atoms involved in the molecules, and it arranges the molecules in the order of rising chain length and the number of substituents of aliphatic alkanes. The ASIC shows how branched the molecule is and how complex the neighborhood of the various carbon atoms is. The $E^{max}_{el-el.repuls}$ behaves as an indicator showing a different rotation and/or inversion behavior of 3- and 4-substituted compounds in comparison with other compounds. And the J index reflects the changes of increase in branching and the number of atoms in the molecule. Katritzky et al. further analyzed[34] the retention index, and it depends on (i) the length of carbon backbone, (ii) the positions of the methyl groups connected to the backbone, and (iii) the number of branches on the backbone. In the structure−activity correlations, the final aim is to obtain chemical knowledge on how the structure influences the activity and then to improve the structures to design new molecules. The study paying attention on the structure factors that affect the retention behavior gives the readers some chemical knowledge, and the authors try to follow the knowledge by introducing EFVCI to model the retention index.

Correlation with the Transformed Retention Index by EFVCI. Katritzky et al.[34] selected four from about 129 descriptors to build a model with $s = 4.6$. To test the information contents of EFVCI, the transformed retention index is regressed by seven indices (${}^0\chi^{efvci}$, ${}^1\chi^{efvci}$, ${}^2\chi^{efvci}$, ${}^3\chi^{efvci}$, ${}^3\chi^{efvci}_{cluster}$, ${}^4\chi^{efvci}$, ${}^4\chi^{efvci}_{path-cluster}$) with $R^2 = 0.98$, $s = 3.55$ at optimal $x = 0.8$ (those in ref 34 are as follows: $R^2 = 0.9585$ and $s = 4.6$). The standard error(s) obtained in the present work is lower than the experimental error (4).[34]

Table 1. Retention Index of 207 Alkanes[a]

no.	structure	cn	bn	bp	RTI(exp)	RTI(cal)	Res	no.	structure	cn	bn	bp	RTI(exp)	RTI(cal)	Res
1	2mC9	10	1	2	966.5	965.5	1	76	2m10mC30	32	2	2	3099	3098	1
2	3mC9	10	1	3	973	973.6	−0.6	77	2m12mC30	32	2	2	3095	3097.4	−2.4
3	2mC11	12	1	2	1166.5	1164.4	2.1	78	3m7mC30	32	2	3	3108	3109.1	−1.1
4	3mC11	12	1	3	1172.5	1172.6	−0.1	79	4m10mC30	32	2	4	3094	3087.8	6.2
5	2mC13	14	1	2	1366.5	1364.1	2.4	80	6m10mC30	32	2	6	3075	3074.7	0.3
6	3mC13	14	1	3	1373	1372.4	0.6	81	3m7mC31	33	2	3	3209	3208.7	0.3
7	2mC15	16	1	2	1566.5	1564.1	2.4	82	3m13mC31	33	2	3	3203.5	3206.1	−2.6
8	3mC15	16	1	3	1573.7	1572.6	1.1	83	3m15mC31	33	2	3	3209	3206.1	2.9
9	2mC17	18	1	2	1765.8	1764.2	1.6	84	5m13mC31	33	2	5	3180.5	3179.8	0.7
10	3mC17	18	1	3	1774	1772.9	1.1	85	5m17mC31	33	2	5	3182	3181.1	0.9
11	2mC19	20	1	2	1966	1964.3	1.7	86	7m11mC31	33	2	7	3170.2	3169.7	0.5
12	3mC19	20	1	3	1974.3	1973.2	1.1	87	11m21mC31	33	2	11	3162.9	3166.5	−3.6
13	10mC19	20	1	10	1943	1938.8	4.2	88	2m8mC32	34	2	2	3297	3298	−1
14	2mC21	22	1	2	2166	2164.4	1.6	89	4m8mC32	34	2	4	3292	3287.7	4.3
15	3mC21	22	1	3	2174.5	2173.4	1.1	90	6m10mC32	34	2	6	3273.5	3273.6	−0.1
16	11mC21	22	1	11	2141	2136.8	4.2	91	8m12mC32	34	2	8	3266	3265.8	0.2
17	2mC23	24	1	2	2364	2364.5	−0.5	92	9m21mC32	34	2	9	3262	3268.5	−6.5
18	3mC23	24	1	3	2374.5	2373.6	0.9	93	14m18mC32	34	2	14	3257.5	3257.8	−0.3
19	12mC23	24	1	12	2337	2335.1	1.9	94	3m9mC33	35	2	3	3403	3406.3	−3.3
20	2mC25	26	1	2	2563	2564.5	−1.5	95	3m15mC33	35	2	3	3409	3404.7	4.3
21	3mC25	26	1	3	2574.4	2573.7	0.7	96	5m17mC33	35	2	5	3380	3379.3	0.7
22	13mC25	26	1	13	2534.5	2533.6	0.9	97	5m19mC33	35	2	5	3382	3380.1	1.9
23	2mC27	28	1	2	2763	2764.4	−1.4	98	7m17mC33	35	2	7	3370	3370.1	−0.1
24	3mC27	28	1	3	2774.4	2773.7	0.7	99	11m23mC33	35	2	11	3362.4	3365.9	−3.5
25	14mC27	28	1	14	2733	2732.2	0.8	100	2m10mC34	36	2	2	3494	3496	−2
26	2mC29	30	1	2	2962.2	2964.2	−2	101	4m16mC34	36	2	4	3489	3485.5	3.5
27	3mC29	30	1	3	2974	2973.7	0.3	102	6m10mC34	36	2	6	3473.8	3472.7	1.1
28	15mC29	30	1	15	2931.5	2931	0.5	103	8m12mC34	36	2	8	3465	3464.7	0.3
29	2mC31	32	1	2	3161.5	3164	−2.5	104	12m22mC34	36	2	12	3461.4	3461.7	−0.3
30	3mC31	32	1	3	3174.1	3173.6	0.5	105	13m17mC34	36	2	13	3455	3456.4	−1.4
31	4mC31	32	1	4	3157.5	3155.4	2.1	106	3m7mC35	37	2	3	3609.5	3607.2	2.3
32	5mC31	32	1	5	3150	3148.6	1.4	107	3m15mC35	37	2	3	3601	3603.5	−2.5
33	6mC31	32	1	6	3143.2	3143.7	−0.5	108	5m9mC35	37	2	5	3580	3578.2	1.8
34	7mC31	32	1	7	3140	3140.1	−0.1	109	5m19mC35	37	2	5	3580.5	3578.3	2.2
35	13mC31	32	1	13	3130.8	3130.6	0.2	110	7m17mC35	37	2	7	3569.7	3568.5	1.2
36	16mC31	32	1	16	3129.8	3129.8	0	111	9m21mC35	37	2	9	3561	3564.7	−3.7
37	2mC33	34	1	2	3362	3363.8	−1.8	112	2m12mC36	38	2	2	3695	3694	1
38	3mC33	34	1	3	3374.5	3373.4	1.1	113	5m17mC36	38	2	5	3680	3677.1	2.9
39	4mC33	34	1	4	3357.5	3355.1	2.4	114	13m23mC36	38	2	13	3661	3658.9	2.1
40	5mC33	34	1	5	3350	3348.3	1.7	115	3m15mC37	39	2	3	3801	3802.4	−1.4
41	6mC33	34	1	6	3343.7	3343.4	0.3	116	5m9mC37	39	2	5	3779	3777.4	1.6
42	13mC33	34	1	13	3328.5	3329.9	−1.4	117	5m17mC37	39	2	5	3780	3776.5	3.5
43	17mC33	34	1	17	3328.5	3328.7	−0.2	118	13m23mC37	39	2	13	3759	3757.7	1.3
44	2mC35	36	1	2	3562	3563.5	−1.5	119	5m17mC38	40	2	5	3878	3875.9	2.1
45	3mC35	36	1	3	3574.3	3573.1	1.2	120	4m8m12mC24	27	3	4	2520	2522.5	−2.5
46	18mC35	36	1	18	3527.3	3527.6	−0.3	121	5m9m13mC25	28	3	5	2610	2613.3	−3.3
47	3m9mC23	25	2	3	2410	2411.9	−1.9	122	4m8m12mC26	29	3	4	2719	2719.5	−0.5
48	5m9mC24	26	2	5	2485	2484.1	0.9	123	3m7m11mC27	30	3	3	2838	2838.9	−0.9
49	3m11mC25	27	2	3	2609	2610.4	−1.4	124	4m8m12mC28	31	3	4	2918	2917.2	0.8
50	3m15mC25	27	2	3	2605	2612.4	−7.4	125	3m7m11mC29	32	3	3	3037	3037	0
51	5m11mC25	27	2	5	2582	2583.7	−1.7	126	5m13m17mC29	32	3	5	3007	3009.7	−2.7
52	5m17mC25	27	2	5	2585	2589.9	−4.9	127	6m14m18mC30	33	3	6	3100	3103.3	−3.3
53	7m11mC25	27	2	7	2577	2573.9	3.1	128	3m7m11mC31	34	3	3	3236.5	3235.4	1.1
54	2m6mC26	28	2	2	2704	2701.8	2.2	129	5m13m17mC31	34	3	5	3205.4	3206.6	−1.2
55	4m8mC26	28	2	4	2695	2690.6	4.4	130	7m13m17mC31	34	3	7	3191.3	3196	−4.7
56	5m11mC26	28	2	5	2682	2682.8	−0.8	131	11m15m19mC31	34	3	11	3181	3187.8	−6.8
57	6m10mC26	28	2	6	2678	2677.1	0.9	132	2m10m16mC32	35	3	2	3324	3323.4	0.6
58	7m11mC26	28	2	7	2675	2673.1	1.9	133	4m12m16mC32	35	3	4	3316	3312.7	3.3
59	3m7mC27	29	2	3	2809	2810.4	−1.4	134	6m14m18mC32	35	3	6	3299	3300.3	−1.3
60	3m15mC27	29	2	3	2805	2809.7	−4.7	135	12m16m20mC32	35	3	12	3281	3285.7	−4.7
61	5m11mC27	29	2	5	2782	2782	0	136	3m7m15mC33	36	3	3	3436.5	3435.5	1
62	5m17mC27	29	2	5	2786	2786	0	137	5m13m17mC33	36	3	5	3405	3404.2	0.8
63	7m23mC27	29	2	7	2774	2772.8	1.2	138	7m11m15mC33	36	3	7	3389	3393.7	−4.7
64	9m19mC27	29	2	9	2765	2774.9	−9.9	139	11m15m19mC33	36	3	11	3379	3384.8	−5.8
65	2m6m28C	30	2	2	2905	2900.8	4.2	140	2m10m16mC34	37	3	2	3524	3521.3	2.7
66	2m10mC28	30	2	2	2899	2899.2	−0.2	141	4m8m12mC34	37	3	4	3515.5	3512	3.5
67	4m10mC28	30	2	4	2895	2889	6	142	6m14m18mC34	37	3	6	3497	3497.8	−0.8
68	5m15mC28	30	2	5	2882	2883	−1	143	8m12m16mC34	37	3	8	3486.4	3489.4	−3
69	7m13mC28	30	2	7	2873	2871.9	1.1	144	12m16m20mC34	37	3	12	3478	3482.6	−4.6
70	3m7mC29	31	2	3	3008	3009.5	−1.5	145	3m7m15mC35	38	3	3	3636.3	3633.8	2.5
71	3m13mC29	31	2	3	3004	3007.4	−3.4	146	5m9m13mC35	38	3	5	3605	3602.8	2.2
72	5m13mC29	31	2	5	2982	2981.3	0.7	147	7m11m15mC35	38	3	7	3588.3	3592	−3.7
73	5m19mC29	31	2	5	2983	2985.1	−2.1	148	13m17m21mC35	38	3	13	3577	3580.8	−3.8
74	7m17mC29	31	2	7	2973	2974.1	−1.1	149	13m17m23mC35	38	3	13	3583	3581.8	1.2
75	2m6mC30	32	2	2	3105	3099.9	5.1	150	4m8m16mC36	39	3	4	3715	3712	3

Table 1 (Continued)

no.	structure	cn	bn	bp	RTI(exp)	RTI(cal)	Res	no.	structure	cn	bn	bp	RTI(exp)	RTI(cal)	Res
151	8m12m16mC36	39	3	8	3685	3687.6	−2.6	180	7m11mC21	23	2	7	2172	2179	−7
152	14m18m22mC36	39	3	14	3676	3679.1	−3.1	181	3m11mC23	25	2	3	2405	2412.3	−7.3
153	3m7m15mC37	40	3	3	3835	3832.3	2.7	182	3m7mC25	27	2	3	2608.5	2611.4	−2.9
154	5m13m17mC37	40	3	5	3803	3800.3	2.7	183	5m9mC25	27	2	5	2586	2583.4	2.6
155	7m13m19mC37	40	3	7	3784	3789.7	−5.7	184	4m10mC26	28	2	4	2692.5	2690.3	2.2
156	15m19m23mC37	40	3	15	3775	3777.6	−2.6	185	6m13mC26	28	2	6	2681	2678.4	2.6
157	16m20m24mC38	41	3	16	3873.5	3876.3	−2.8	186	5m15mC27	29	2	5	2783.2	2784.2	−1
158	5m13m17mC39	42	3	5	4001	3998.7	2.3	187	7m11mC27	29	2	7	2767.2	2772.3	−5.1
159	15m19m23mC39	42	3	15	3972.4	3975.1	−2.7	188	9m11mC27	29	2	9	2765	2757.2	7.8
160	14m18m22mC40	43	3	14	4071	4074.4	−3.4	189	4m8mC28	30	2	4	2895	2889.5	5.5
161	3m7m11m15mC29	33	4	3	3062	3062.8	−0.8	190	5m9mC29	31	2	5	2982	2981	1
162	3m7m11m15mC31	35	4	3	3261	3259.9	1.1	191	7m19mC31	33	2	7	3166	3173.3	−7.3
163	4m8m12m16mC31	35	4	4	3249	3239	10	192	9m19mC31	33	2	9	3165	3167.8	−2.8
164	3m7m11m15mC33	37	4	3	3459	3457.6	1.4	193	2m10mC32	34	2	2	3291	3297	−6
165	4m8m12m16mC33	37	4	4	3448	3436.4	11.6	194	2m12mC34	36	2	2	3494	3495.1	−1.1
166	3m7m11m15mC35	39	4	3	3658	3655.5	2.5	195	6m14mC34	36	2	6	3475	3472.6	2.4
167	7m11m15m19mC35	39	4	7	3628	3615.5	12.5	196	3m7m13mC27	30	3	3	2840	2840	0
168	9m13m17m21mC35	39	4	9	3617	3610	7	197	2m10m18mC28	31	3	2	2918	2932.5	−14.5
169	11m15m19m24mC35	39	4	11	3605	3607.2	−2.2	198	9m13m17mC29	32	3	9	2995	2993.2	1.8
170	6m10m12m16mC36	40	4	6	3723	3731	−8	199	5m9m13mC31	34	3	5	3200	3206	−6
171	8m12m16m20mC36	40	4	8	3713	3710.9	2.1	200	7m11m15mC31	34	3	7	3191.3	3195.7	−4.4
172	10m14m18m22mC36	40	4	10	3703.5	3706.8	−3.3	201	9m13m17mC31	34	3	9	3192.2	3190.2	2
173	3m7m11m15mC37	41	4	3	3855	3853.7	1.3	202	5m9m23mC33	36	3	5	3409	3412.7	−3.7
174	7m11m15m19mC37	41	4	7	3823	3813.1	9.9	203	7m13m17mC33	36	3	7	3395	3393.5	1.5
175	9m13m17m21mC37	41	4	9	3813	3807.1	5.9	204	9m13m17mC33	36	3	9	3391.9	3387.8	4.1
176	11m15m19m24mC37	41	4	11	3803	3803.2	−0.2	205	6m10m14mC34	37	3	6	3496	3497.4	−1.4
177	10m14m18m22mC38	42	4	10	3900	3904	−4	206	6m12m16mC34	37	3	6	3500	3496.7	3.3
178	5mC27	28	1	5	2750.3	2749	1.3	207	10m14m18mC34	37	3	10	3489	3484.8	4.2
179	7mC29	30	1	7	2939.8	2940.4	−0.6								

[a] cn means the number of atoms; bn denotes the branch number; bp indicates the first branching position of branch(es); RTI(exp) is the experimental retention index; RTI(cal) represents the calculated retention index; and Res is the abbreviation of regression residuals.

Table 2. Intercorrelations among the EFVCI Descriptors

	${}^{0}\chi^{efvci}$	${}^{1}\chi^{efvci}$	${}^{2}\chi^{efvci}$	${}^{3}\chi^{efvci}$	${}^{3}\chi^{efvci}_{cluster}$	${}^{4}\chi^{efvcief vci}$	${}^{4}\chi^{efvci}_{path-cluster}$
${}^{0}\chi^{efvci}$	1.0000	0.9913	0.9584	0.9367	0.7093	0.9097	0.7071
${}^{1}\chi^{efvci}$		1.0000	0.9877	0.9744	0.7957	0.9556	0.7936
${}^{2}\chi^{efvci}$			1.0000	0.9972	0.8803	0.9886	0.8787
${}^{3}\chi^{efvci}$				1.0000	0.9079	0.9968	0.9090
${}^{3}\chi^{efvci}_{cluster}$					1.0000	0.9295	0.9963
${}^{4}\chi^{efvci}$						1.0000	0.9326
${}^{4}\chi^{efvci}_{path-cluster}$							1.0000

Correlation with Untransformed Retention Index by χ and EFVCI. To follow the common methods in modeling the retention index, the untransformed data are obtained by addition of the KI (Kovats retention index) with the number of carbons in the main chain*100. Then, the regressions by χ and EFVCI are studied.

The regression results of the retention index by seven χ indices are as follows: $s = 6.06$ and $F = 2.68e{+}005$. To improve the information content of ${}^{efvci}\chi$, the variable connectivity index is extended to more orders (0−4). By the same searching process in refs 36 and 42, the optimal EFVCI index will reach at $x = -0.8$, and at the optimal point the regression results are as follows: $R = 0.99998$, $s = 3.49$, and $F = 7.96e{+}005$, which is a very good regression by indices from the same series. The regression results (calculated retention index and regression residuals) are given in Table 1, in which there are six structures with an absolute regression residual exceeding 8, that is, no. 64 (9m19mC27), 163 (4m8m12m16mC31), 165 (4m8m12m16mC33), 167 (7m11m15m19mC35), 174 (7m11m15m19mC37), and 197 (2m10m18mC28), and the biggest is 14.4 for no. 197 (2m10m18mC28).

The regression equation is

$$\mathrm{RTI} = -607.6 + 1799.8*{}^{0}\chi^{efvci} - 3973.8*{}^{1}\chi^{efvci} + 2733.4*{}^{2}\chi^{efvci} - 488.6*{}^{3}\chi^{efvci} - 1802.7*{}^{3}\chi^{efvci}_{cluster} + 41.4*{}^{4}\chi^{efvci} - 47.6*{}^{4}\chi^{efvci}_{path-cluster}$$

$$n = 207,\ s = 3.49,\ F = 7.95e + 005,\ R = 0.99998, \text{ and RMSECV} = 3.94 \quad (12)$$

The leave-one-out cross-validation is applied to validate the built model by EFVCI with the cross-validated root-mean-square error of prediction (RMSECV) equal to 3.94, which shows the stability of the built model.

Moreover, the intercorrelations among the seven EFVCI indices are studied with their correlation coefficients listed in Table 2, in which some coefficients are very high, such as $R({}^{0}\chi^{efvci}, {}^{1}\chi^{efvci}) = 0.9913$, $R({}^{2}\chi^{efvci}, {}^{3}\chi^{efvci}) = 0.9972$, and $R({}^{3}\chi^{efvci}, {}^{4}\chi^{efvci}) = 0.9968$ et al. Should the descriptor be deserted due to the high collinearity with other indices? Can the commonly used orthogonal method work significantly in this case?

At first, we investigate the situation when a descriptor is deserted. If the ${}^{0}\chi^{efvci}$ is not included in the regression due to its high collinearity with ${}^{1}\chi^{efvci}$ (0.9913), the standard error

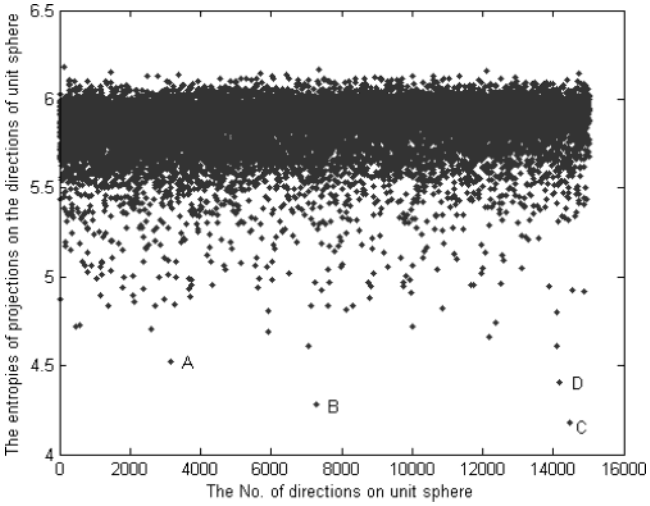

Figure 1. The entropies of the projections on the projection directions of unit sphere.

will increase to 6.1885. Similarly, the standard error will be 6.7691 without $^{3}\chi^{efvci}$ (0.9972 with $^{2}\chi^{efvci}$). Second, by using the orthogonal method, the RMSECV is also about 3.92, that is, the RMSECV is only improved a little, say from 3.94 to 3.92.

Searching Optimal Values for χ^{f} and Optimal EFVCI. The idea of variable connectivity and EFVCI is that a variable parameter undergoes change during the regression analysis, in which the multilinear regression and varying the variable parameter show the change of standard error *s* and the variable *x* for a different combination of several χ^{efvci}. For each combination of different EFVCI, the optimal variable value is obtained when the standard error reaches the minimum (as shown in Figure 4 of ref 36). When searching for the optimal value for two to seven indices of EFVCI by observing the variation of standard error as *x* ranging from −0.9 to 0.9 with a step length of 0.01, the optimal values for two to seven EFVCI indices are as follows: 0.81 for $^{0}\chi^{efvci}$ and $^{1}\chi^{efvci}$; −0.78 for $^{0}\chi^{efvci}$, $^{1}\chi^{efvci}$ and $^{2}\chi^{efvci}$; −0.79 for $^{0}\chi^{efvci}$, $^{1}\chi^{efvci}$, $^{2}\chi^{efvci}$ and $^{3}\chi^{efvci}$; −0.81 for $^{0}\chi^{efvci}$, $^{1}\chi^{efvci}$, $^{2}\chi^{efvci}$, $^{3}\chi^{efvci}$, and $^{3}\chi^{efvci}_{cluster}$; −0.80 for $^{0}\chi^{efvci}$, $^{1}\chi^{efvci}$, $^{2}\chi^{efvci}$, $^{3}\chi^{efvci}$, $^{3}\chi^{efvci}_{cluster}$, and $^{4}\chi^{efvci}$; −0.81 for $^{0}\chi^{efvci}$, $^{1}\chi^{efvci}$, $^{2}\chi^{efvci}$, $^{3}\chi^{efvci}$, $^{3}\chi^{efvci}_{cluster}$, $^{4}\chi^{efvci}$, and $^{4}\chi^{efvci}_{path-cluster}$. From the analysis, the optimal variable parameter used in the EFVCI index varies in a narrow range and will approach a constant at about −0.81. The former five EFVCI indices give a regression standard error about 5.36, while the seven indices give a standard error about 3.49. The information contained in the last two indices ($^{4}\chi^{efvci}$, and $^{4}\chi^{efvci}_{path-cluster}$) is, to some extent, rich, which should not be ignored.

Structural Features Hidden in the Molecular Connectivity Indexes. The good performance of EFVCI drives us to ponder why the EFVCI can improve the regression significantly. In the former study,[36] the reasons are analyzed to be the following: (1) bond additivity of the EFVCI. (2) The atomic attributes in EFVCI reflect, to some extend, the center sequence, which is located by the balance function (BF) method[12] proposed by the authors or the IVEC method proposed by Bonchev.[43] The atomic attributes with different kinds of center sequences can be distinguished by EFVCI, which, in the long run, helps make improvements on regression. (3) Innate and external division of atomic attributes in EFVCI.

The intuitive structure information hidden in EFVCI indices is as follows: (i) size of molecules; (ii) branch number; and (iii) branch position. The study[34] also regards the three features domains the retention behavior of 208 alkanes produced by insects. In the present work, the structure information coded by EFVCI is further studied by the projection pursuit (PP).[23−25] By using the projection pursuit combined with NT-net, generated by TFWW, on the unit sphere, the entropies of the projections are shown in Figure 1, in which there are several minima in the entropies. The projection points such as A, B, C, and D are selected to investigate the structure information in the directions.

In the projection direction (corresponding to point A in Figure 1) *a* = [0.8452, −0.1341, 0.0071, −0.3664, 0.0193, 0.1507, 0.2370] for $^{0}\chi^{efvci}$, $^{1}\chi^{efvci}$, $^{2}\chi^{efvci}$, $^{3}\chi^{efvci}$, $^{3}\chi^{efvci}_{cluster}$,

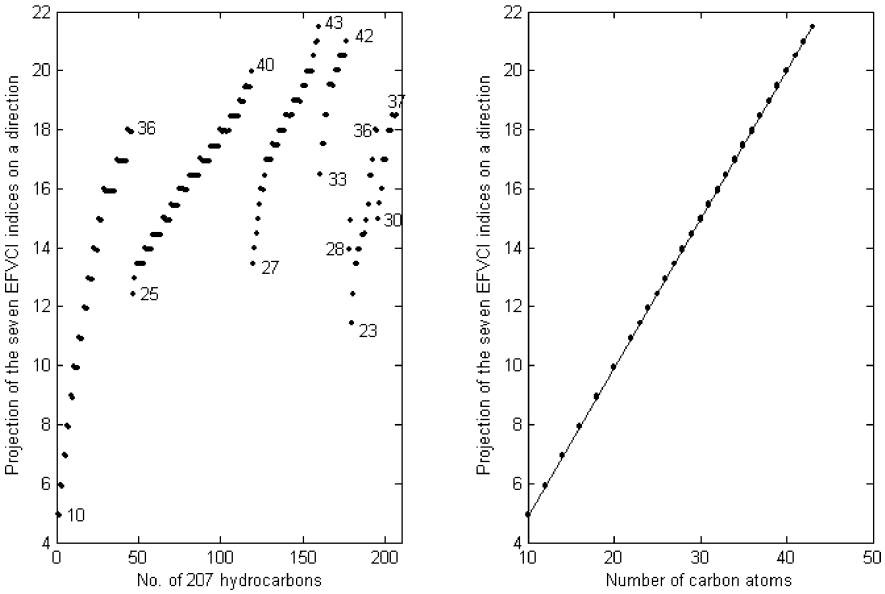

Figure 2. Size feature in a projection direction for the 207 alkanes (the *x*-axis of the left subplot is the no. of the 207 alkanes listed in Table 1; the numbers in the left subplot are the carbon atoms in the structures; the *x*-axis of the right subplot is the number of carbon atoms; and the *y*-axis of both plots is the projection of the seven EFVCI indices on a given projection direction).

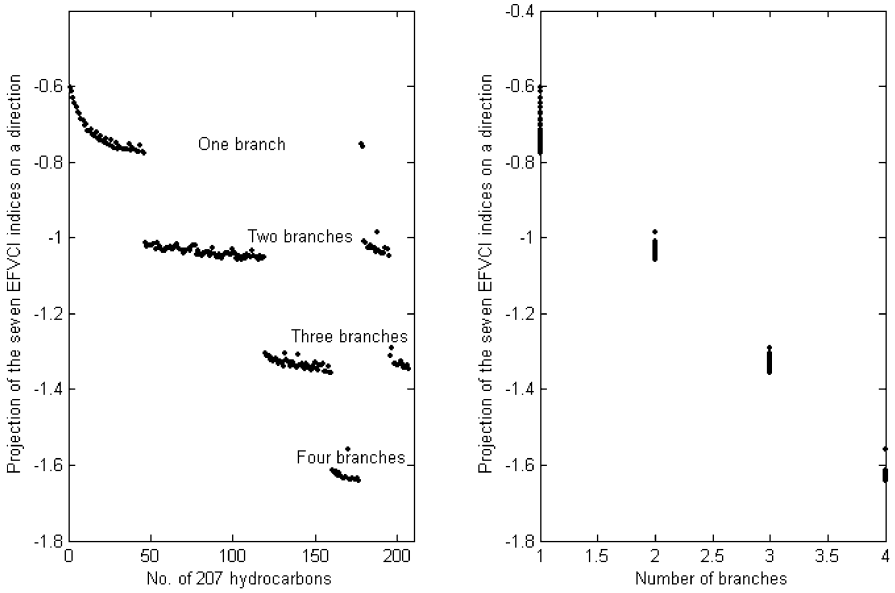

Figure 3. Branch number information in a projection direction for the 207 alkanes (the *x*-axis of the left subplot is the no. of the 207 alkanes listed in Table 1; the *x*-axis of the right subplot is the number of branches; and the *y*-axis of both plots is the projection of the seven EFVCI indices on a given projection direction).

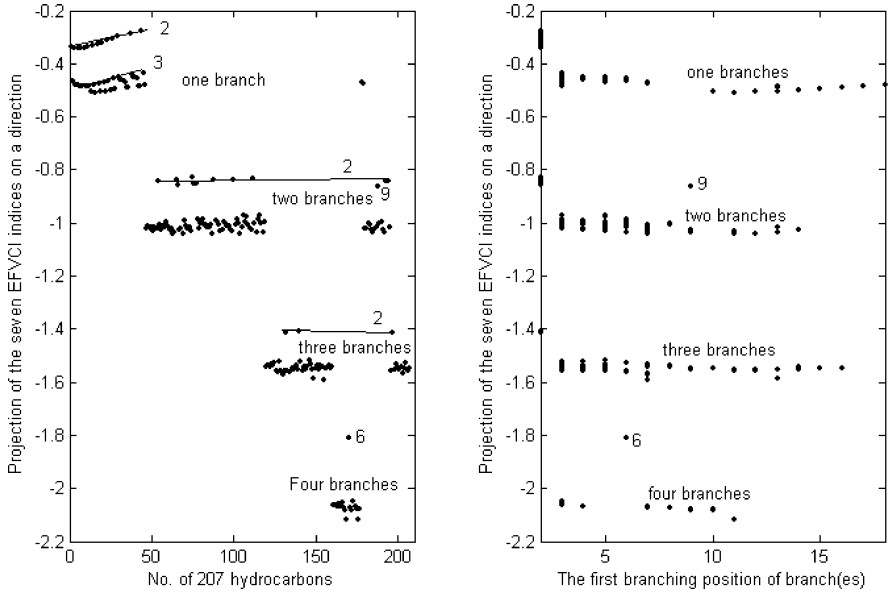

Figure 4. Branch position information in a projection direction for the 207 alkanes (the *x*-axis of the left subplot is the no. of the 207 alkanes listed in Table 1; the *x*-axis of the right subplot is the first branching position of branch(es); and the *y*-axis of both plots is the projection of the seven EFVCI indices on a given projection direction).

$^{4}\chi^{efvci}$, $^{4}\chi^{efvci}_{path-cluster}$, the size (carbon number of a molecule) can be described very well, which is shown in Figure 2. From the figure, structures with same number of carbons hold the same value, and structures with arithmetic carbon atoms differ from others with arithmetic values. The size structure information can be coded by EFVCI.

In the projection direction (corresponding to point B in Figure 1) a = [−0.0803, −0.2925, −0.1293, −0.6927, −0.3063, 0.0764, −0.0712] for $^{0}\chi^{efvci}$, $^{1}\chi^{efvci}$, $^{2}\chi^{efvci}$, $^{3}\chi^{efvci}$, $^{3}\chi^{efvci}_{cluster}$, $^{4}\chi^{efvci}$, $^{4}\chi^{efvci}_{path-cluster}$, the branch number can be described relatively well, which is shown in Figure 3. From the figure, the structures with the same branch number can be intuitively grouped into one class, and the structures with arithmetic branches differ from others with arithmetic values. However, the branch number information cannot be completely coded by the EFVCI, and say the structures with the same branch number cannot be projected on the EFVCI to be on the same line.

In the projection direction (corresponding to point C, similar results obtained from point D in Figure 1) a = [0.0527, −0.0070, −0.1618, −0.0401, −0.9223, 0.1931, 0.0520] for the seven indices, the branch position can also be, to some degree, coded by EFVCI as shown in Figure 4. From the figure, the position 2, relating to the terminal branch, is distinguished from other positions, while other kinds of branch positions are mixed. While two structures (6m10m12m16mC31 and 9m11mC27) exhibit different behavior, the reason might be that there is a branch pattern (the two atoms with branch bridged by another atom) such as 10m and 12m or 9m and 11m, which is an interesting phenomenon. To check the results, dozens of outside structures are generated to hold that kind of branch pattern

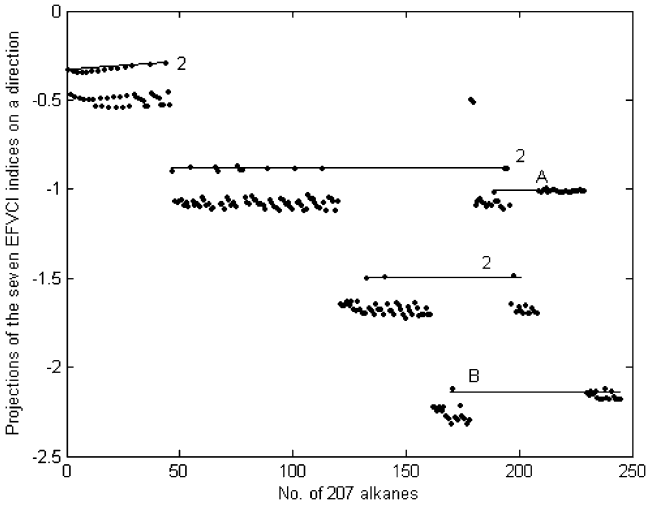

Figure 5. The variation of the average retention index by the change of number of carbon atoms.

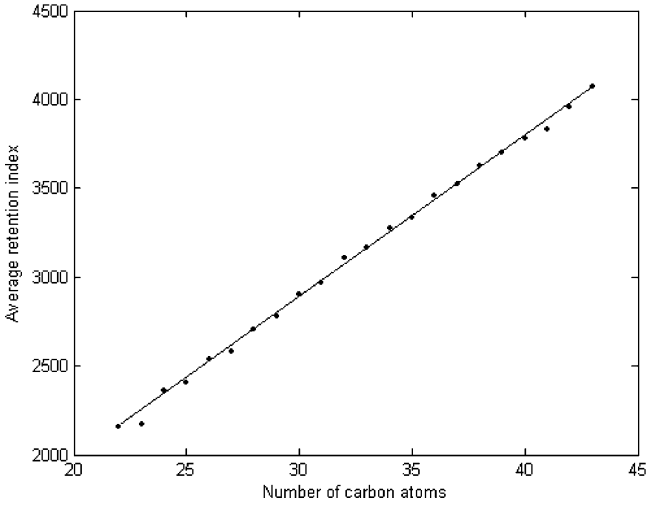

Figure 6. The projection results of the branch pattern (the two atoms with a branch bridged by another atom: lines A and B).

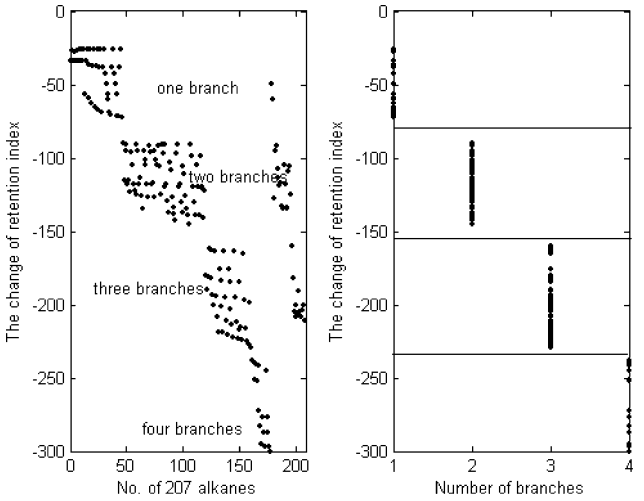

Figure 7. The variation of the average retention index by the change of the branch number.

and included in the data set, and then the EFVCI of them are calculated by using the optimal value at −0.8. After getting the EFVCI of both 207 and the new dozens of structures, the projection pursuit method is used to search interesting structure clusters by selecting several projections with minimum entropy. And in the same projection direction (a = [0.0527, −0.0070, −0.1618, −0.0401, −0.9223, 0.1931, 0.0520]), the projection results are shown in Figure 5. From the figure, the branch pattern (the two atoms with branch bridged by another atom: line A and B) is really different from other kinds of patterns! In the direction, the branch number information is also mined out. What should be pointed out is that the branch position information is not so obvious. How to define quantitatively branch position information still needs much endeavor in graph theory.

The graph center and related concepts have been applied to analyze intuitively the branch position information hidden in the EFVCI.[36] By considering the atomic attributes in EFVCI and graph center sequence of the molecules, the EFVCI index describes, to some extent, the center sequence and also the branch position information. That is, the more central[43] the atom is, the bigger the influence the atom will have. The center sequence information codes,[12] to some extent, the branch position information, which is one of the hidden structural features in EFVCI.

Application of the Interpretation Information. After mining out the three structural features, we tried to use them to discover some chemical knowledge to understand the changing tendency of the retention index with the change of structures. In the study, it is very hard to distinguish the finer structure features such as the different positions of methyl on the chain. Thus, in the present work, we use statistical results to illustrate roughly the changing tendency of the retention index by different structure information.

For the variation of the retention index by the change of size, at first, we sum up the retention indexes of structures with the same number of carbon atoms and then average the sum to get the average retention index for a specific number of carbon atoms. Next, the average retention index is plotted by the size (the number of carbon atoms) as shown in the Figure 6 (it can be regarded as a linear relationship between the average retention index and the number of carbon atoms), and finally, we use the number of carbon atoms to regress the average retention index, and the regression equation is

$$\mathrm{RTI(size)} = 144.5 + 91.2*\mathrm{CN} \qquad (13)$$

Here CN is the number of carbon atoms. From the regression, the effect of one carbon atom on the average retention index is about 91.2. Of course, the result is a just statistic phenomenon. However, it can be roughly predicted that with the addition of the number of atoms in a structure, the retention index will, in general, increase.

For the variation of the retention index by the change of the branch number, we use the nonbranched structures to be the references to evaluate the effects of the branch number. For a structure with a specific branch number, its change of the retention index by a branch is calculated by its own retention index subtracted from the retention index of a reference (the nonbranched structure with the same carbon atoms, that is, NC*100), which is shown in the left plot (the x-axis is the no. of 207 alkanes listed in Table 1) of Figure 7. At first, we sum up the changes of the retention index (after subtracted the retention index of a reference) of structures with the same branch number and then average

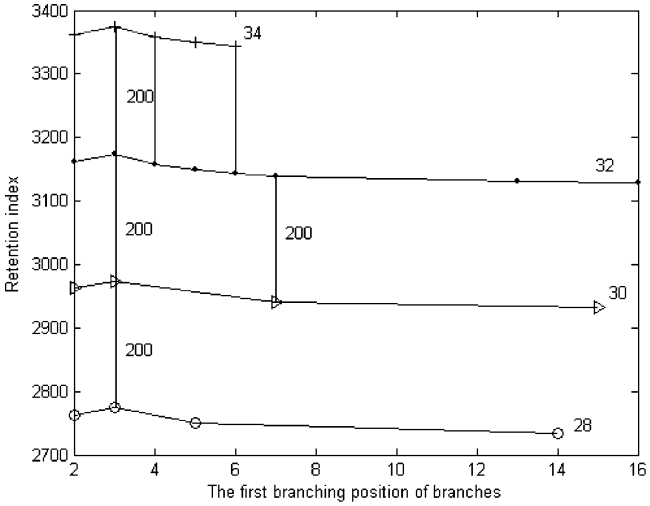

Figure 8. The branching position information and the change of retention index (the numbers 28, 30, 32, and 34 are the number of atoms; 200 is the difference of the retention index of two structures with same branch position and differing by two atoms).

the sum to obtain the average change retention index for structures with a specific branch number. Next, the average retention index is plotted by the branch number in the right plot of Figure 7. It is intuitively observed that with the addition of the branch number in a structure, the retention index will be, in general, decreased, which is also a statistic result.

For the variation of the retention index by the change of the branch position, due to the inseparable of finer structural features, that is when several kinds of features (such as several branch number combining with several kinds of branching positions) are mixed, it is not easy to distinguish which play the role influencing the change of the retention index. And so, we only consider the simplest case—structures with one branch, and the retention index is plotted by the branch position in Figure 8, in which only four kinds of carbon atoms are selected because there are few structures with one branch in other carbon atoms. From the figure, (1) the change is 200 for structures with the same branching position and differing by two atoms, about 100/per atom. (2) The retention index of different structures shows a regular relationship with the different branching position except for position 2 (relating with terminal branch). The fact tells that the position 2 should have a significantly different influence on the retention index, which meets roughly with the structure information mined out in Figures 4 and 5.

These phenomena indicate that the structural features hidden in EFVCI such as size, branch number, and branch position should mainly contribute to the high quality regression with the retention index. By projection pursuit, the structural features hidden in the space spanned by several EFVCI indices are mined out, which provide some insights about the nature of the topological index.

CONCLUSION

To interpret build models in QSAR/QSPR depends on the descriptors used by the authors. Topological indices have been applied in different kinds of fields. But one of the criticisms is that of the difficult interpretation of the topological index. Based on the spirit that the interpretation of TIs should have an interpretation within structural chemistry, the structural features coded by EFVCI indices are used to interpret the indices themselves, and the interpretation information is applied to understand the chemical knowledge of the variation of the retention index by the change of structures.

ACKNOWLEDGMENT

This project is financially supported by the National Nature Foundation Committee (NNFC) of P. R. China (no. 20235020 and 20175036). The authors appreciated the hospitality of Hong Kong Baptist University, when the authors attended "Workshop on Data Mining in Chemistry and Traditional Chinese Medicines" in Oct 2003. The valuable comments and suggestions for improving the paper from the referee are highly appreciated.

REFERENCES AND NOTES

(1) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(2) Randic', M.; Zupan, J. On Interpretation of well-known Topological Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550−560.
(3) Randic', M.; Balaban, A. T.; Basak, S. C. On structural interpretation of several distance related topological indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 593−601.
(4) Motoc, I.; Balaban, A. T. Topological indices: intercorrelations, physical meaning, correlation ability. *Rev. Roum. Chim.* **1981**, *265*, 593−600.
(5) Kier, L. B.; Hall, L. H. Derivation and significance of valence molecular connectivity. *J. Pharm. Sci.* **1981**, *70*, 583−589.
(6) Labanowski, J. K.; Motoc, I.; Dammkoehler, R. A. The physical meaning of topological indices. *Comput. Chem.* **1991**, *15*, 47−53.
(7) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological indices: their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891−898.
(8) (a) Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic Press, Inc.: New York, 1976. (b) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure−activity analysis*; Research Studies Press Ltd.: New York, 1986.
(9) (a) Kier, L. B. A shape index from chemical graphs. *Quantum Struct.−Act. Relat.* **1985**, *4*, 109−116. (b) Kier, L. B. Shape indexes of orders one and three from molecular graphs. *Quant. Struct.−Act. Relat.* **1986**, *5*, 1−7. (c) Kier, L. B. Kappa shape indices for similarity analysis. *Med. Chem. Res.* **1997**, *7*, 8−12. (d) Kier, L. B.; Hall, L. H. The kappa indices for modeling molecular shape and flexibility. In *Topological indices and related descriptors in QSAR and QSPR;* Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: 1999; pp P455−490.
(10) (a) Kier, L. B.; Hall, L. H. An electrotopological state index for atoms in molecules. *Pharm. Res.* **1990**, *7*, 801−807. (b) Hall, L. H.; Kier, L. B. *Molecular structure description: the electrotopological state;* Academic Press: 1999. (c) Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular similarity based on novel atom-type E-State indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074−1080. (d) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-State fields: Application to 3-D QSAR. *J. Comput. Aid. Mol. Des.* **1996**, *10*, 513−520. (e) Hall, L. H.; Vaughn, A. T. QSAR of phenol toxicity using E-State and Kappa shape indices. *Med. Chem. Res.* **1997**, *7*, 407−416. (f) Hall, L. H.; Story, C. T. 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.* **1996**, *36*, 1004−1014. (g) Hall, L. H.; Kier, L. B. The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 784−791.
(11) (a) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493. (b) Huuskonen, J. J.; Livingstone, D. J.; Tetko, I. V. Neural network modeling for estimation of partition coefficient based on atom-type electrotopological state indices. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 947−55.
(12) Hu, Q. N.; Liang, Y. Z.; Ren, F. L. Molecular Graph Center, A Novel Approach to Locate the Center of a Molecule and a New Centric Index. *THEOCHEM.* **2003**, *635*, 105−113.
(13) Bonchev, D.; Mekenyan, O.; Trinajstic, N. Topological characterization of cyclic structures. *Int. J. Quantum Chem.* **1980**, *17*, 845−893.

(14) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(15) Balaban, A. T.; Filip, P.; Balaban, T. S. Computer program for finding all possible cycles in graphs. *J. Comput. Chem.* **1985**, *6*, 316−329.
(16) Bonchev, D.; Balaban, A. T.; Liu, X.; Klein, D. J. Molecular cyclicity and centricity of polycyclic graphs. Part 1. Cyclicity based on resistance distances and reciprocal distances. *Int. J. Quantum Chem.* **1994**, *50*, 1−20.
(17) Lovasz, L.; Pelikan, J. On the eigenvalue of trees. *Period. Math. Hung.* **1973**, *3*, 175−182.
(18) Gutman, I.; Randić, M. Algebraic characterization of skeletal branching. *Chem. Phys. Lett.* **1977**, *47*, 15−19.
(19) Bonchev, D.; Trinajstic, N. Information theory, distance matrix, and molecular branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.
(20) Bertz, S. Branching in graphs and molecules. *Discrete Applied Math.* **1988**, *19* 65−83.
(21) Kirby, E. C. Sensitivity of topological indices to methyl group branching in octanes and azulenes, or what does a topological index index? *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1030−1035.
(22) Randić, M. On molecular branching. *Acta Chim. Slov.* **1997**, *44*, 57−77.
(23) Friedman, J. H.; Tukey, J. W. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **1974**, *C-23*, 881−889.
(24) Huber, P. J. Projection Pursuit (with discussion). *Ann. Statistics* **1985**, *13*, 435−475.
(25) Friedman, J. H. Exploratory projection pursuit. *J. Am. Stat. Assoc.* **1987**, *82*, 249−266.
(26) Du, Y. P.; Liang, Y. Z.; Yun, D. Data Mining for Seeking an Accurate Quantitative Relationship between Molecular Structure and GC Retention Indices of Alkenes by Projection Pursuit. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1283−1292.
(27) Du, Y. P.; Liang, Y. Z. Data Mining for Seeking Accurate Quantitative Relationship between Molecular Structure and GC Retention Indices: Regression Model for Alkanes. *Comput. Biol. Chem.* **2003**, in press.
(28) Kvalheim, O. V.; Liang, Y. Z. Heuristic evolving latent projections: resolving two-way multicomponent data. 1. Selectivity, latent-projective graph, datascope, local rank, and unique resolution. *Anal. Chem.* **1992**, *64,* 936−946.
(29) Guo, Q.; Wu, W.; Questier, F.; Massart, D. L. Sequential Projection Pursuit Using Genetic Algorithms for Data Mining of Analytical Data. *Anal. Chem.* **2000**, *72*, 2846−2855.
(30) Tashiro, Y. On methods for generating uniform points on the surface of a sphere. *Ann. Institute Statistical Mathematics* **1977**, *29*, 295−300.
(31) Fang, K. T.; Wang, Y.; Wong, K. L. *A new method of generating an NT-net on the unit sphere*; Technical Report; MATH-002; Hong Kong Baptist College: 1992.
(32) Niederreiter, H. Pseudo-random numbers and optimal coefficients. *Adv. Math* **1977**, *26*, 99−181.
(33) Hua, L. K.; Wang, Y. *Applications of number theory to numerical analysis;* Spring-Verlag and Science Press: Berlin and Beijing, 1981.
(34) Kartrizky, A. R.; Chen, K.; Maran, U.; Carlson, D. A. QSPR correlation and predictions of GC retention indexes for methyl-branched hydrocarbons produced by insects. *Anal. Chem.* **2000**, *72*, 101−109.
(35) Katrizky, A. R.; Petrukhin, R.; Tatham, D.; Basak, S. C.; Benfenati, E.; Karelson, M.; Maran, U. Interpretation of quantitative structure−property and -activity relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 679−685.
(36) Hu, Q. N.; Liang, Y. Z.; Wang, Y. L.; Xu, C. J.; Zeng, Z. D.; Fang, K. T.; Peng, X. L.; Yin, H. External factor variable connectivity index. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 773−778.
(37) Randić, M. Novel graph theroretical approach to heteroatoms in QSAR. *Chemom. Intel. Labl. Syst.* **1991**, *10*, 213−223.
(38) Randić, M. On computation of optimal parameters for multivariate analysis of structure−property relationship. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 970−980.
(39) Randić, M.; Pompe, M. The variable connectivity index $^1\chi^f$ versus the traditional molecular descriptors: A comparative study of $^1\chi^f$ against descriptors of CODESSA. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 631−638.
(40) Randić, M.; Basak, S. C. On use of the variable connectivity index $^1\chi^f$ in QSAR: Toxicity of Aliphatic Ethers. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 614−618.
(41) Hu, Q. N.; Liang, Y. Z.; Wang, Y. L.; Guo, F. Q.; Huang, L. F. Heuristic queue notation: basic principles and applications in calculating matrices and topological indices. *Comput. Appl. Chem. (in Chinese)* **2003**, *20*, 386-390.
(42) Randic', M.; Pompe, M. The variable connectivity index $^1\chi^f$ versus the traditional molecular descriptors: A comparative study of $^1\chi^f$ against descriptors of CODESSA. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 631−638.
(43) Bonchev, D.; Mekenyan, O.; Balaban, A. T. Iterative procedure for the generalized graph center in polycyclic graphs. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 91−97.

CI034225F