Inter*net* Electronic Journal of Molecular Design

June 2004, Volume 3, Number 6, Pages 335–349

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday
Part 12

Guest Editor: Douglas J. Klein

The Branching Number of a Molecular Graph

Qian–Nan Hu and Yi–Zeng Liang

Institute of Chemometrics and Intelligent Analytical Instruments, College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, P. R. China

Received: July 28, 2003; Revised: November 24, 2003; Accepted: May 5, 2004; Published: June 30, 2004

Citation of the article:
Q.–N. Hu and Y.–Z. Liang, The Branching Number of a Molecular Graph, *Internet Electron. J. Mol. Des.* **2004**, *3*, 335–349, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

The Branching Number of a Molecular Graph[#]

Qian–Nan Hu and Yi–Zeng Liang*

Institute of Chemometrics and Intelligent Analytical Instruments, College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, P. R. China

Received: July 28, 2003; Revised: November 24, 2003; Accepted: May 5, 2004; Published: June 30, 2004

Internet Electron. J. Mol. Des. **2004**, *3* (6), 335–349

Abstract

Based on the structure information of the mathematical characteristics of the degree distribution of saturated hydrocarbons, a method to count the branching number of a saturated hydrocarbon is proposed and then extended to general molecules. In order to understand the structure information of the characteristics, some new concepts are introduced, and the edges in a molecule are partitioned into different cases. In the understanding, the edges that are essential to connecting the cycles are not considered as the branches. Subsequently, several formulas are tried to calculate the proposed branching number. Then, the formulas are generalized to other molecules with poly–cycles, heteroatoms and/or multiple bonds. Most of molecules are counted in agreement with intuitive view, but there also exits some molecules that are difficult to judge the branching numbers. The method offers an automatic and easy way to count the branching number of a molecular graph. The branching number is applied to be one of the classifiers to simplify the structure diversity, and the regression results are improved greatly. A new method, orthogonal block variables combining with canonical correlation analysis, is also used in the regression, which can avoid collinearity and includes, at the same time reducing the variable dimension significantly, almost all the information of original variables.

Keywords. Branching; branching number; cyclicity; cycle number; degree distribution; graph theory.

Abbreviations and notations

β, branching number	DD, degree distribution
E, expectation of degree distribution	MAD, the mean absolute deviation
SOM, the second order moment	TOM, the third order moment
TOAM, the third order absolute moment	

1 INTRODUCTION

In 1998, Prof. Randić reviewed that the quantitative characterizations of molecular structural features have been overlooked and neglected for too long time, and he analyzed the slow advance in quantitative characterization of molecular attributes is primarily due to lack of precise definition of such attributes [1]. The molecular attributes have been studied for many years, although most of the concepts have not been rigorously defined. The same dilemma extends to molecular shape, cyclicity, chirality, degree of folding, degree of planarity, molecular complexity, aromaticity,

[#] Dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday.

* Correspondence author; phone: 86–731–8830824; fax: 86–731–8825637; E–mail: yzliang@public.cs.hn.cn.

molecular similarity and molecular diversity, and even the branching in saturated hydrocarbons has not been rigorously defined [1].

Since the advent of the Wiener index [2], there are more than 400 topological indices emerged, and many topological indices claim that they code some specific structure information. It is the goal of topological index first to define the structural features of molecules mathematically and then to study the chemical consequences of the molecular features, just as the first general index of molecular complexity [3]. The basic idea is that topological indices can code molecular shape, size, branching, cyclicity, symmetry, centricity, compactness, diversity, and complexity as well.

The subject of branching has received considerable attention in the graph theory over the past decades years. In 1973, Lovasz and Pelikan suggested the leading or the first eigenvalue of the adjacency matrix as a molecular branching index [4]. Two years later, Randić [5] proposed the famous branching index, which is a useful descriptor of molecular branching. And later, Gutman and Randić [6] consider branching, to some degree, from the mode of a distribution, which allows a rigorous definition of the concept of branching and they suggest that structures having an identical distribution of valencies should not be discriminated [6]. In the same year, Bonchev and Trinajstić define a measure of branching from information theory, which is essentially based on a kind of distribution [7]. In 1988, Bertz [8] proposed a definition of branching in terms of the degrees of the central points in the star graphs. Kirby [9] discussed the limitations of some branching indices and offered some remedies improving the performance of the connectivity index for larger alkanes. Recently a novel branch index was proposed [10], which is based on the path matrix, a newly introduced matrix for graphs in which the matrix elements are expressed as the path subgraphs of a graph considered [11].

In the studies, many kinds of indices are proposed as branching index, however a dilemma is that even the branching number of molecular graph has not yet been generally accepted, especially for cyclic molecules, and a general method to compute the branching number has not been found.

In our former works [12], several mathematical characteristics, based on the degree distributions, are investigated in the total 530 saturated hydrocarbons [13] to mine out some structural features, and there are, mainly, two structural features, cyclicity and branching, coded by the characteristics.

In the present work, the former study is briefly introduced and some useful structure information is obtained. Based on the cyclicity and the branching information of the former research, some new concepts on partition of edges are suggested, and then a method to calculate the branching number is proposed for saturated hydrocarbons and then extended to general molecules with poly–cycles, heteroatoms and/or multiple bonds. The results are in agreement with intuitive view.

After getting the branching number of a molecular graph, it is applied to be one of the classifiers to simplify the structure diversity, and regression results are improved greatly. In order to enhance

the information contents of variables, more block variables [14,15] are introduced, and a newly proposed method, orthogonal block variables combining with canonical correlation analysis, is used in the regression, which can avoid collinearity and includes, at the same time reducing the variable dimension significantly, almost all the information of original variables.

2 METHODOLOGY AND DISCUSSION

2.1 The degree distribution (*DD*) and mathematical characteristics

As shown in studies [16, 17], the atomic vertex degrees constitute the degree distribution of a molecule. Once the distribution is given, the mathematical characteristics of the distribution can be easily calculated.

The expectation might be regarded as the center of the distribution. The expectation can be easily calculated by the average of vertex degrees in the molecule by $E(V_i)$, in which V_i is the vertex degree of atom i.

The kth central moment of DD is easily obtained by $E((V_i - E(V_i)))^k$, where the V_i is the vertex degree of atom i. The first central moment is zero. The second central moment is the variance using a divisor of n (number of carbon atom) instead of $n–1$ where n is the sample size. The structural feature of the variance by using n and $n–1$ is the same, only differing in the values.

The kth absolute central moment of DD can be calculated by $E(abs(V_i - E(V_i))^k)$, where the V_i is the vertex degree of atom i. The first order absolute central moment is larger than zero, and is called mean absolute deviation (*MAD*). The *MAD* of a distribution provides a measure of the spread or dispersion around its mean. A small value of the *MAD* indicates that the degree distribution is tightly concentrated around the mean; and a large value of the MAD typically indicates that the degree distribution has a wide spread around its mean. The second absolute central moment holds same structure information with the second central moments because their values are the same.

2.2 Cycle number of molecules and the expectation of degree distribution

The cycle number is helpful to understand the branching in the structures, and the proposed formula to count branching number is enlightened by the formula to count the cycle number. The expectations of the distributions hold information on the cyclicity of molecular graphs. Some saturated hydrocarbons are listed in the Table 1, in which the saturated hydrocarbons with same number of carbon atoms and same expectation values are grouped, and the branch seems have no influence on the expectation values. The labels of structures in Table 2 correspond to the labels of structures listed in Figure 1.

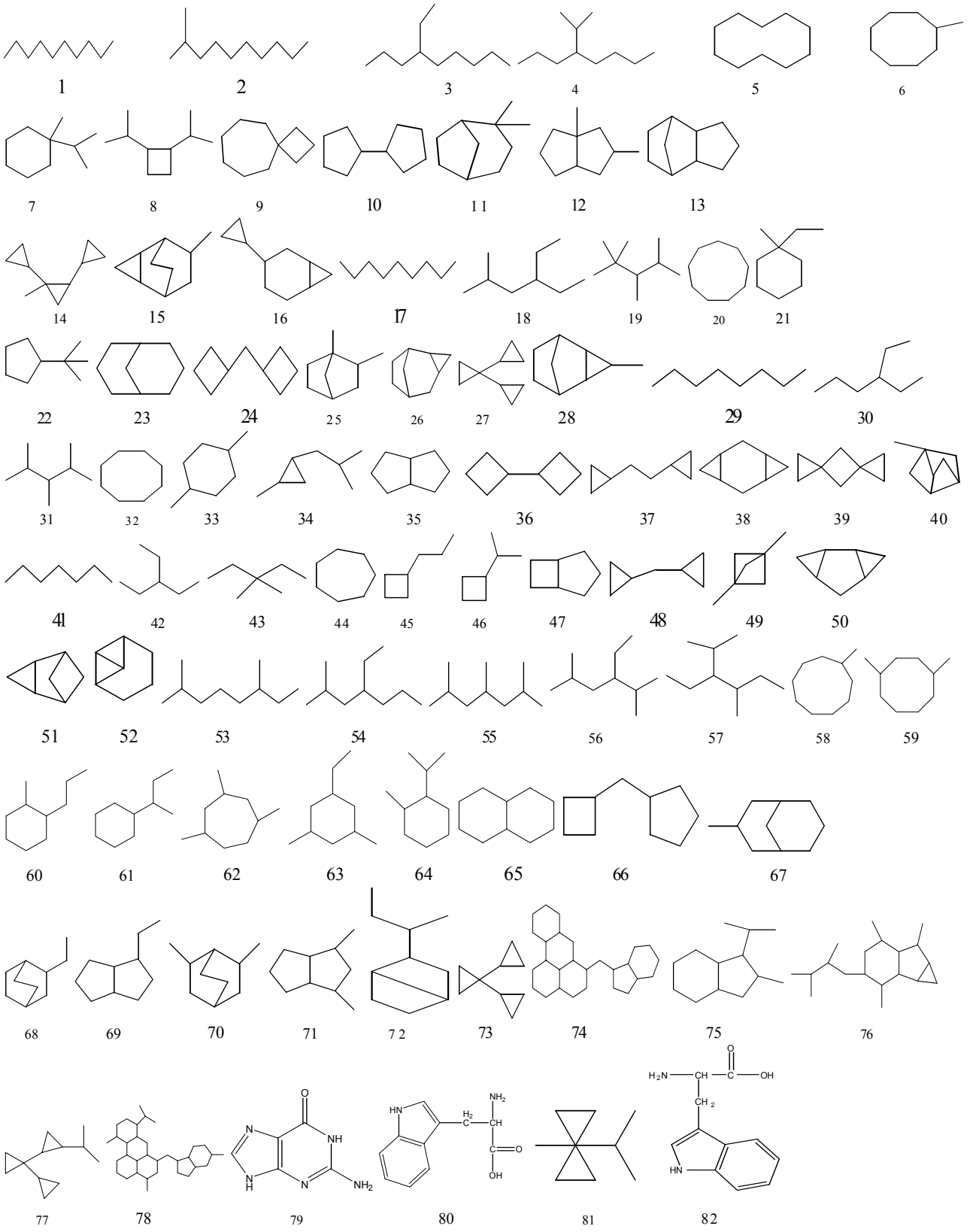

Figure 1. Eighty–two structures used in the present study.

Table 1. Structure examples of the cyclicity of saturated hydrocarbons and their related expectation of the degree distributions.

NC [a]	Cycle(s)	0	1	2	3	Difference
10	Expectations	1.8000	2.0000	2.2000	2.4000	0.2000
	Structures	1, 2, 3, 4	5, 6, 7, 8	9,10,11,12	13,14,15,16	
9	Expectations	1.7778	2.0000	2.2222	2.4444	0.2222
	Structures	17, 18, 19	20, 21, 22	23, 24, 25	26, 27, 28	
8	Expectations	1.7500	2.0000	2.2500	2.5000	0.2500
	Structures	29, 30, 31	32, 33, 34	35, 36, 37	38, 39, 40	
7	Expectations	1.7143	2.0000	2.2857	2.5714	0.2857
	Structures	41, 42, 43	44, 45, 46	47, 48, 49	50, 51, 52	

[a] NC indicates the number of carbon atoms

From Table 1, s36d, bcpe, 22mbc321o, and 13mbc330o are all with 10 carbon atoms and 2 cycles, which is calculated by the cyclicity used in defining Balaban *J* index [18] and is also equal to the so–called smallest set of rings [19]. The definition of cyclomatic number is as:

$$\mu = n_e - n + 1 \quad (1)$$

in which *n* denotes the number of carbon atoms, and n_e means the number of edges in the saturated hydrocarbons.

2.3 Branching information from the former study

The center (expectation) of a degree distribution corresponds to the cyclicity of a saturated hydrocarbon, and the dispersion around its center (or cyclicity) of a degree distribution is a measure of branching. The mean absolute deviation (*MAD*) of a distribution provides a measure of the spread or dispersion around its mean.

Table 2. Structural examples of decanes and the related moments of the degree distributions.

Structures		MAD	SOM	TOM	TOAM
1		0.3200	0.1600	–0.0960	0.1088
2, 3		0.4800	0.3600	0.0240	0.3312
4, 53, 54		0.6400	0.5600	0.1440	0.5536
55, 56, 57		0.8000	0.7600	0.2640	0.7760
	Difference	0.1600	0.2000	0.1200	0.2224
5		0	0	0	0
6, 58		0.2000	0.2000	0	0.2000
59, 60, 61		0.4000	0.4000	0	0.4000
62, 63, 64		0.6000	0.6000	0	0.6000
	Difference	0.2000	0.2000	0	0.2000
10, 65, 66		0.3200	0.1600	0.0960	0.1088
67, 68, 69		0.4800	0.3600	–0.0240	0.3312
70, 71, 72		0.6400	0.5600	–0.1440	0.5536
	Difference	0.1600	0.2000	–0.1200	0.2224

The molecular branching has a significant effect on the central moments such as the mean absolute deviation (*MAD*), the second order moment (*SOM*), the third order moment (*TOM*), and the third order absolute moment (*TOAM*) as well. The structures of some saturated hydrocarbons are listed in the Table 2, in which the saturated hydrocarbons with same number of carbon atoms and same moment values are classified. The moment values of saturated hydrocarbons, with arithmetic number of branch(es) and same number of carbon atoms, form an arithmetic series. The labels of structures in Table 2 correspond to the labels of structures listed in Figure 1.

For alkanes, the simplest cases are those only with methyl, in which the alkanes n10, 2mn9, 26mn8, and 246mn7 have 0, 1, 2, and 3 branches. The saturated hydrocarbons with ethyl or substituent with more carbon atoms are more complex. The saturated hydrocarbons 4en8, and 25m3en6, without much ambiguity, are regarded as with 1, 2, and 3 branches. The molecule 4ipn7, with isopropyl, is recognized as with the same moment values for saturated hydrocarbons with 2 branches. Intuitively, the isopropyl as a whole can be regarded as a branch of the main chain of heptane, and then partition the isopropyl from the main chain to be isobutane, which should be thought as a sub–chain on the propyl chain with another one branch. And so, the molecule 4ipn7 is with 2 branches. Other cases can be deduced by the above methods, for example 24m3ipn5 can be recognized as total four branches (three branches on the main chain of amyl, and one branch on the sub–chain of propyl (to be isobutane)).

For mono–cyclic saturated hydrocarbons, the molecules C10, 1mC9, 14mC8, and 135mC7 are with 0, 1, 2, and 3 branches; saturated hydrocarbons 1eC8, m2pC5, and 1e35mC6 are recognized as with 1, 2, and 3 branches; and the molecules 1SbC6, and 1m3ipC6 are found with 2, and 3 branches.

For bi–cyclic saturated hydrocarbons, the situations are relatively complicated due to the complexity as the different connection patterns between cycles and also the attachment of substituents. The three saturated hydrocarbons, 3mbc331n, 2ebc222o, and 2ebc330o, are found with one branch, and the saturated hydrocarbons 26mbc222o, and 24mbc330o are investigated as with two branches.

There are some confused cases, for instance, the saturated hydrocarbons bc440d, and bCpe, all with two cycles and different connecting edges, are investigated with the same *MAD*, *SOM*, *TOM*, and *TOAM* values as 0.3200, 0.1600, 0.0960, and 0.1088, and all are regarded as without branch. Someone will argue the branch situation in the three saturated hydrocarbons. In the understanding, the edges that are essential to connecting the cycles are not considered as the branches. From this view, the three saturated hydrocarbons hold the same moment values, and based on this point, the similar cases can be deduced, for example the molecules 65, 66, 73 and 74 from Figure 1 are thought to be without branch.

2.4 Cyclic edge, chain edge, and branch

In order to understand the structure information coded by the former studies, some concepts are, cautiously, proposed. The proposal is based on the idea of partitioning the molecules.

Cyclic edges are the edges that are essential to the construction of a cycle. Chain edges are the edges that construct the chain of a molecule, and the edges that are essential to connecting several cycles belong to this group.

Sub–chain edges are the edges on the chain of a functional group that is a substituent on the cycle or the chain, such as the propane of isobutene (having one vertex connecting to the chain) from 4ipn7. Similarly, sub–$(sub)_n$–chain edges can be deduced, for instance the methyl on the propyl (main–chain of the isobutene). The edges that are not essential to connecting the cycles in a molecule are recognized as sub–chain (or sub–$(sub)_n$–chain) edges. The sub–$(sub)_n$–chain edge can be one edge or a combination of several edges.

Thus, all kinds of edges can be partitioned into the cyclic, chain, sub–chain edges and sub–$(sub)_n$–chain edges. Next interest is how to consider the branch of a molecule. The chain edges and cyclic edges are not regarded as branch, while one sub–chain (or sub–$(sub)_n$–chain) is regarded as a branch. Some examples are listed below to give some deep understanding of the concepts. The alkane 2mn9 has a chain composed of 8 edges, and one sub–chain as methyl that is considered as one branch. The saturated hydrocarbons m2pC5 has a cycle formed by 6 cyclic edges, and two sub–chains (one is methyl and the other is propyl), and so the branching number of the structure is 2. The alkane 4ipn7 has one sub–chain as isopropyl, and a sub–sub–chain as methyl, and so the alkane has two branches. Logically, the saturated hydrocarbon 1m2ipC6 has two sub–chains and one sub–sub–chain, total three branches.

2.5 Branching number in saturated hydrocarbons

After the conceptual definition of the branch, enlightened by the formula (1) to count the cycle number, a method is proposed to calculate the branching number of saturated hydrocarbons.

(1) If n_e is equal to or bigger than n, that is, the molecule are cyclic saturated hydrocarbons, the branching number β is:

$$\beta=2*(n-n_e)+2*n_4+n_3 \quad (2)$$

(2) For alkanes, with their n_e being smaller than n, the branching number (β) is:

$$\beta=2*n_4+n_3 \quad (3)$$

in which n denotes the number of carbon atoms; n_e means the number of edges in the alkane; n_4 is the number of quaternary atoms; and n_3 represents the number of tertiary atoms.

The formula to count the branching number can be expressed by including cyclicity, and, for cyclic saturated hydrocarbons, the branching number is:

$$\beta=2*(n-n_e)+2*n_4+n_3=2*n_4+n_3-2\mu+2 \quad (4)$$

where μ is the cyclicity.

The branching numbers are obtained by using the number of non–hydrogen atoms, edges, and vertex with degree 3 and 4. Some examples are listed in Table 3 by applying the method to count the branching number of saturated hydrocarbons, in which molecule 73 from Figure 1, having 9 carbon atoms, 11 edges, 2 trinary atoms, and 1 quaternary atom, is recognized as with 0 branch; and similarly the molecule 74 from Figure 1 is regarded as with no branch. From the above discussion, the edges that are essential to connecting the cycles are not regarded as branch of the molecules, and so the two molecules are thought to be without branch. This case should be discussed to give a general definition of the branching number in similar situations, however the proposed branching number is, so far, a rational choice.

Table 3. The branching numbers of some saturated hydrocarbons

Structures	n	n_e	$Deg_{i=}3$	4	β
4	10	9	2	0	2
53	10	9	2	0	2
61	10	10	2	0	2
73	9	11	2	1	0
74	27	32	10	0	0
75	13	14	5	0	3
76	19	21	10	0	6
77	12	14	4	1	2
78	33	38	15	0	5

2.6 Branching number in the molecules with poly–cycles, heteroatoms and/or multiple bonds

For a general molecule, just for calculating the branching number, all non–hydrogen atoms are considered as the same, while all kinds of bonds are regarded as identical, that is, without distinguishing the atoms and bonds. Then, the algorithm to calculate the branching number of a general molecule is, carefully, generalized to be:

(1) for cyclic molecules:

$$\beta=2*(n-n_e)+\sum_{Deg_i\geq 3} n_{Deg_i}*(Deg_i-2) \quad (5)$$

(2) for acyclic molecules:

$$\beta=\sum_{Deg_i\geq 3} n_{Deg_i}*(Deg_i-2) \quad (6)$$

in which n denotes the number of carbon atoms; n_e is the number of edges in a molecule; Deg_i means the vertex degree of an atom; and n_{Deg_i} represents the number of atoms with specific Deg_i.

The branching numbers can be obtained by using the number of non–hydrogen atoms, edges, and vertex with degree equal to or higher than 3. The results of some real and virtual molecules are listed in Table 4 by the proposed method. The algorithm calculates easily the branching numbers, which are in agreement with the intuitive view.

Table 4. The branching numbers of some real and virtual molecules

Structures	n	n_e	Deg_i=3	4	5	6	7	8	β
79	11	12	4	0	0	0	0	0	2
80	15	16	5	0	0	0	0	0	3
81	9	10	1	0	0	1	0	0	3
82	14	15	5	0	0	0	0	0	3

2.7 Application of branching number as a classifier to simplify the structure diversity

In QSPR research, even for the simplest property, normal boiling point (bp), of saturated hydrocarbons, a perfect descriptor combination allowing to model accurately cycloalkane bps is not yet found [13]. The cited authors analyzed the reasons into four aspects as structural diversity, low precision of boiling points, stereochemistry, and the difficult extension of TI defined for acyclic molecules to cyclic compounds. The first reason is that the diversity in the structures of (poly)cyclic saturated hydrocarbons is overwhelming. The first question is weather it should simplify the structure diversity.

(1) The main indices, calculated by the in–house software, Heuristic Queue Notation (H.Q.N.) system [20], are: autocorrelation index [21], molecular connectivity index [22, 23], E–state index [24], Kappa index [25], MEDV–4 [26], MHDV [27], MEDV–13 [28], information indices [29], mpc [30], mpw [31], mwc [32], eigenvalue index [33], uvxy indices [34], uvxyoi indices [35], Balaban D index [36], J index [18], IB operator (using RD) [37], triplet [38], detour index [39], Harary index [40], Z [41], Idi [42], MTI [43], All Path version Wiener index [44], W [2], Zagreb [45], EFVCI [46], and centric index [47, 48] as well. However, even with so many indices, the regression results (s = 3.3553; however maximum absolute residual MAR = 25.0279) are still not satisfactory by multi–linear regression. Thus, it seems to be necessary to divide the total molecules into small groups (classes) based on the structure information of diversity, or to regress by non–linear methods.

(2) Consider the two sub–sets from the 530 saturated hydrocarbons: one is the set *A* of acyclic structures with one branch and another is the set *B* of acyclic structures with two branches. When the two data sets are regressed by seven chi indices, some interesting regression results are obtained. The regression results for data set A are R = 0.9992; s = 2.0044; and maximum absolute residual (MAR) equal to 4.9781, while those for data set B are R = 0.9981; s = 2.0988; and maximum absolute residual (MAR) equal to 6.8427. However, when the two data sets are applied

together, the regression results are: R = 0.9974; s = 2.8511; and especially the maximum absolute residual (MAR) changed so much to be 12.1857. From the analysis, the two data sets should belong to different classes. Further analysis is by using three block variables such as molecular connectivity index, Kappa index and E–State index to model the two sets. However even with three block variables, the regression results are still not good MAR = 10.9062. The same situation extends to two data sets of acyclic structures with both two and three branches, the regression is MAR = 11.9014. This phenomenon shows that the structure diversity should be classified further to obtain satisfactory results or there should introduce or design more descriptors. In the work, the regression by classification based on the three block variables is considered.

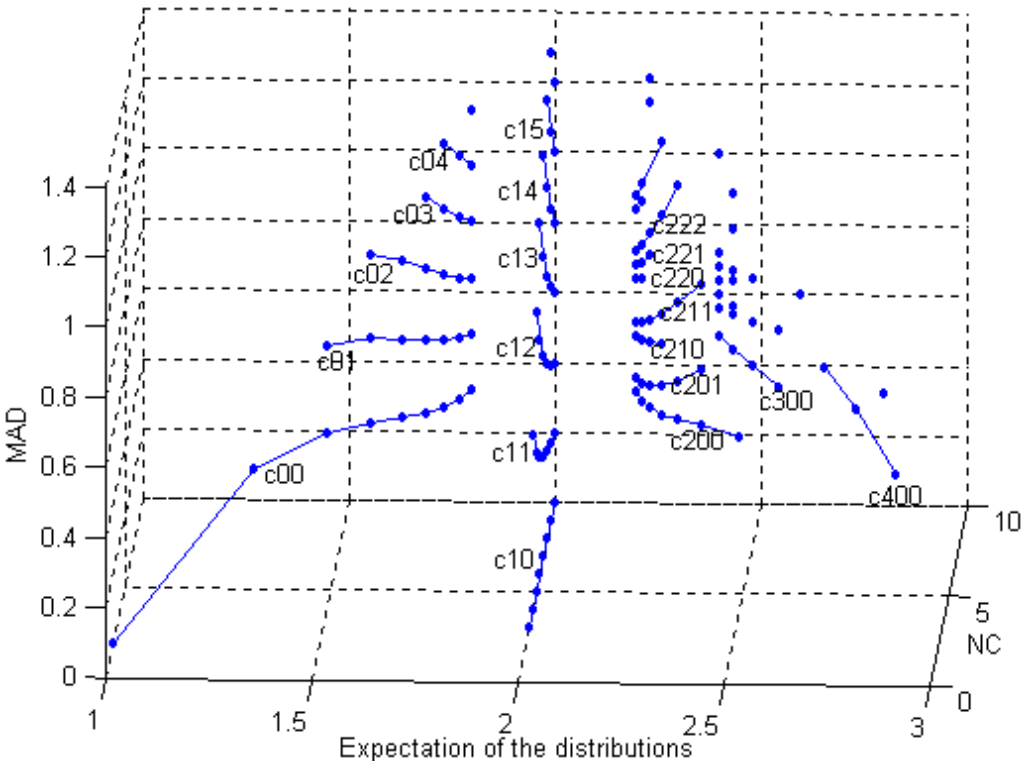

Figure 2. Structural diversity by the number of carbon atoms, the expectations and MADs of the ratio distributions. (c means classification; the first number following the c, such as 0, 1, 2, 3 *et al*., denotes the number of cycles in the alkanes; the second number following the c, such as 0, 1, 2, 3 *et al*., represents the number of branches in the alkanes; and the third number (if existing) following the c, such as 0, 1, 2, 3 *et al*., points to the number of quaternary atoms in the hydrocarbons.)

(3) Another example is that, from the size of structures, mathematical expectation and mean absolute deviation (MAD) of degree distribution of a molecular graph, the structure diversity shows some interesting results as shown in Figure 2, which can be easily repeated by the readers by using the formula given in the theory and methodology section. From the figure, the acyclic and monocyclic structures have not been further divided into sub–group, while multicyclic structures are grouped into different sub–groups by the number of quaternary atoms, which is completely from

the size, mathematical expectation, and MAD themselves with least human intervention. The work gives some interesting enlightenments of the structure diversity, and the results indicate that the diversity should be further classified to make regression satisfactory.

The key of classification and regression is how to classify the variables (X), which is essentially how to classify the molecular structures. A good classification should be the one holding chemical knowledge or information in order to interpret the models. The results in the research may bring a new way to model the properties of saturated hydrocarbons.

Table 5. Structure feature of the different classes

Classes	No. of Structures	No. of cycles	No. of branches	No. of quaternary atoms
C00	9	0	0	— [a]
C01	23	0	1	—
C02	52	0	2	—
C03	43	0	3	—
C04	20	0	4	—
C05	2	0	5	—
C10	8	1	0	—
C11	24	1	1	—
C12	75	1	2	—
C13	59	1	3	—
C14	35	1	4	—
C15	6	1	5	—
C16	2	1	6	—
C200	27	2	0	0
C201	11	2	0	1
C210	22	2	1	0
C211	18	2	1	1
C220	8	2	2	0
C221	14	2	2	1
C222	9	2	2	2
C231	8	2	3	1
C232	8	2	3	2
C242	1	2	4	2
C243	1	2	4	3
C300	15	3	0	0
C301	4	3	0	1
C302	5	3	0	2
C310	3	3	1	0
C311	3	3	1	1
C320	1	3	2	0
C321	2	3	2	1
C322	1	3	2	2
C332	3	3	3	2
C400	5	4	0	0
C401	2	4	0	1
C500	1	5	0	0

[a] no further partition of the class is needed

From the picture of structure diversity, the size and cyclicity can be easily calculated, while the branching information is intuitively related with branching number. However before the method in this work to count branching number, there still lack of a method to easily compute branching number, and so in the present work, the proposed branching number is applied as one of the classifiers to simplify the structure diversity. By following the structure information of the Figure 2, the total 530 alkanes are classified into 36 classes. The structural features of the classes are listed in Table 5. The first column of Table 5 is the names of the classes, the second column is number of molecules contained in the class, the third column is number of cycles in the molecule, the fourth column is number of branches in the molecule and the last column is number of quaternary atom. From the Table and Figure 2 of the structure diversity, the classes by the automatic classifiers reflect the same structure information hidden in the degree distributions.

After the classification of the structures, the problems in modeling different classes are: (*a*) number of variables are large, which indicates that a method to reduce the variable dimension should be introduced; (*b*) the structure information coded by individual block variable (descriptors from same resources) [14,15] is not enough to obtain satisfactory models, and so more block variables should be included; (*c*) the high collinearity between original variables will make the built model unstable, and orthogonal method to avoid collinearity had better be applied.. In the presnt work, the newly proposed method, orthogonal block variables, derived from subspace projection and canonical correlation analysis, is applied to model the different cases. 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 those by selecting only one or two of the original family variables. For each class, the molecular connectivity index, Kappa index and E–State index, representatives of three generations of topological index, which are all mainly extended or proposed by Kier and Hall, are used to model the classes that have more than 15 (3×5) structures. By our method [14,15], there will generate three new orthogonal variables to represent the former original variables, which also includes almost all the information contents of the original variables. The regression results are listed in the Table 6.

Table 6. Regression results of different classifications

Classes	R	s	RMSECV	max(abs(residual))
C01	0.9990	2.1292	2.4154	4.5650
C02	0.9985	1.9099	2.1618	5.3864
C03	0.9920	2.5732	2.9748	6.1702
C04	0.9938	1.8046	2.3745	3.2395
C11	0.9988	2.5943	2.8611	8.1256
C12	0.9975	2.8163	2.9401	8.1316
C13	0.9949	3.2775	3.6097	10.7345
C14	0.9925	3.3232	3.6590	7.4008
C200	0.9979	3.1480	3.8689	7.2981
C210	0.9947	2.8173	3.3471	7.2074
C211	0.9989	1.9949	2.4673	3.6546
C300	0.9972	2.5005	3.2316	4.4989

The first column of the Table 6 is the names of the classes, the second column is correlation coefficients, the third column is standard errors of regression, the fourth column is the cross–validated root mean square error of prediction (RMSECV) and the last column is maximum absolute residuals of regression.

From Table 6, the regression results of the classifications are satisfactory. Further analysis is by using three block variables such as molecular connectivity index, Kappa index and E–State index to model the two sets C01 and C02. However even with three block variables, the regression results are still not good MAR = 10.9062. The same situation extends to two data sets of acyclic structures with both two and three branches (C02 and C03), the regression is MAR = 11.9014. It should be expected that the simplification of the structural diversity contribute to the regression.

3 CONCLUSIONS

A method is proposed to count the branching number of saturated hydrocarbons and then extended to some general molecules. Some new concepts are introduced to interpret the structure information of the mathematical characteristics of degree distribution, and the edges in a molecule are partitioned into different cases. The method offers an automatic and easy way to count the branching number. The proposed branching number is applied as a classifier to simplify the evaluation of structural diversity in chemical libraries.

Acknowledgment

The authors appreciate the hospitality of HongKong Baptist University, when the authors attend "Workshop on Data Mining in Traditional Chinese Medicines".

4 REFERENCES

[1] M. Randić, On characterization of molecular attributes, *Acta Chim. Slov.* **1998**, *45*, 239–252.

[2] H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

[3] S. H. Bertz, The first general index of molecular complexity, *J. Am. Chem. Soc.* **1981**, *103*, 3599–3601.

[4] L Lovasz and J. Pelikan, On the eigenvalue of trees, *Period. Math. Hung.* **1973**, *3*, 175–182.

[5] M. Randić, On characterization of molecular branching, *J. Am. Chem. Soc.* **1975**, *97*, 6609–6013.

[6] I. Gutman and M. Randić, Algebraic characterization of skeletal branching, *Chem. Phys. Lett.* **1977**, *47*, 15–19.

[7] D. Bonchev and N. Trinajstić, Information theory, distance matrix, and molecular branching, *J. Chem. Phys.* **1977**, *67*, 4517–4533.

[8] S. Bertz, Branching in graphs and molecules, *Discrete Applied Math.* **1988**, *19*, 65–83.

[9] E. C. Kirby, 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.

[10] M. Randić, On molecular branching, *Acta Chim. Slov.* **1997**, *44*, 57–77.

[11] M. Randić, D. Plavsić, and M. Razinger, Double invariants, *MATCH* **1997**, *35*, 243–259.

[12] Q. N. Hu, Y. Z. Liang, Q. S. Xu, X. L. Peng and H. Yin, Mathematical characteristics of probability of vertex degree and structure features of molecular topological graph ; in: *Data Mining and Bioinformatics in Chemistry*

and Chinese Mediciens, Eds. K. T. Fang, Y. Z. Liang, and R. Q. Yu, Hong Kong Baptist University, Hong Kong, **2002**, P 242–256.

[13] G. Rücker, and C. Rücker, On the topological indices, boiling points, and cycloalkanes, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 788–802.

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

[15] Q. N. Hu, Y. Z. Liang, X. L. Peng, H. Yin and L. Zhu, Application of Orthogonal Block Variables and Canonical Correlation Analysis in Modeling Pharmacological Activity of Alkaloids from Plant Medicines, *J. Data Sci*. **2003**, *1*, 405–423.

[16] T. I. Bieber and M. D. Jackson, Applications of degree distribution. 1. (a) General discussion and computer generation of degree distributions. (b) Maximal degree distributions, *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 699–700.

[17] T. I. Bieber and M. D. Jackson, Applications of degree distribution. 2. Construction and enumeration of isomers in the alkane series, *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 701–708.

[18] A. T. Balaban, Highly discriminating distance–based topological index, *Chem. Phys. Lett*. **1982**, *89*, 399–404.

[19] B. T. Fan, A. Panaye, J. P. Doucet and A. Barbu, Ring perception. A new algorithm for directly finding the smallest set of smallest rings from a connection table, *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 657–662.

[20] Q. N. Hu, Y. Z. Liang, Y. L. Wang, F. Q. Guo and L. F. Huang, Heuristic queue notation: basic principles and applications in calculating matrices and topological indices, *Computer and Applied Chemistry* (*in Chinese*). **2003**, *4*, 386–390.

[21] G. Moreau and P. Broto, The autocorrelation of a topological structure: a new molecular descriptor, *Nouv. J. Chim*. **1980**, *4*, 359–360.

[22] M. Randić, On characterization of molecular branching, *J. Am. Chem. Soc*. **1975**, *97*, 6609–6615.

[23] L. B. Kier and L. H. Hall, *Molecular connectivity in chemistry and drug research*, Academic Press, New York, **1976**, p257.

[24] L. B. Kier and L. H. Hall, An electrotopological state index for atoms in molecules, *Pharm. Res*. **1990**, *7*, 801–807.

[25] L. B. Kier, A shape index from chemical graphs, *Quant. Struct. –Act. Relat*. **1985**, *4*, 109–116.

[26] S. S. Liu, C. Z. Cao and Z. L. Li, Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance–Edge (MDE) Vector, *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 387–394.

[27] S. S. Liu, *Structural Characterization and Application for Drug Molecule Based on Molecular Electronegativity–Distaqnce vector*(*MEDV*). Ph D Dissertation, Chongqing University, **2001**.

[28] S. S. Liu, C. S. Yin, Z. L. Li and S. X. Cao, QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV–13, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 321–329.

[29] D. Bonchev and N. Trinajstić, Information theory, distance matrix, and molecular branching, *J. Chem. Phys*. **1977**, *67*, 4517–4533.

[30] M. Randić, G. M. Brissey and R. E. Spencer, Search for all self–avoiding paths for molecular graphs, *Comput. Chem*. **1979**, *3*, 65–13.

[31] M. Randić, Novel Shape Descriptors for Molecular Graphs, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 607–613.

[32] G. Rücker and C. Rücker, Counts of all walks as atomic and molecular descriptors, *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 683–695.

[33] Y. Q. Yang, L. Xu and C. Y. Hu, Extended adjacency matrix indices and their applications, *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1140–1145.

[34] A. T. Balaban and T. S. Balaban, New vertex invariants and topological indices of chemical graphs based on information on distances, *J. Math. Chem*. **1991**, *8*, 383–397.

[35] O. Ivanciuc and A. T. Balaban, The graph description of chemical structures. Design on topological indices. Part 20. molecular structure descriptors computed with information on distances operators, *Rev. Roum. Chim*. **1999**, *44*, 220–228.

[36] A. T. Balaban, Topological indices based on topological distances in molecular graphs, *Pure Appl. Chem*. **1983**,

55, 199–206.

[37] O. Ivanciuc, T. Ivanciuc and A. T. Balaban, Design of Topological Indices. Part 10. Parameters Based on Electronegativity and Covalent Radius for the Computation of Molecular Graph Descriptors for Heteroatom–Containing Molecules, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395–401.

[38] P. A. Filip, T. S. Balaban and A. T. Balaban, A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlational ability, *J. Math. Chem.* **1987**, *1*, 61–83.

[39] I. Lukovits, The detour index, *Croat. Chem. Acta.* **1996**, *69*, 873–883.

[40] D. Plavsić, S. Nikolić, N. Trinajstić and Z. Mihalić, On the Harary index for the characterization of chemical graphs, *J. Math. Chem.* **1993**, *12*, 235–250.

[41] H. Hosoya, A newly proposed quantity characterizing topological nature of structural isomers of saturated hydrocarbons, *Bull. Chem. Soc. Japan.* **1971**, *44*, 2332–2339.

[42] M. Randić, On canonical numbering of atoms in a molecule and graph isomorphism, *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171–180.

[43] H. P. Schultz, Topoloical organic chemistry. 1. Graph theory and topological indices of alkanes, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227–228.

[44] I. Lukovits, An all–path version of the Wiener index, *J. Chem Inf. Comput. Sci.* **1998**, *38*, 125–129.

[45] I. Gutman, B. Ruscić, N. Trinajstić and C. F. Wilcox, Graph theory and molecular orbital. XII. Acyclic polyenes, *J. Chem. Phys.* **1975**, *62*, 3399–3409.

[46] Q. N. Hu, Y. Z. Liang, Y. L. Wang, C. J. Xu, Z. D. Zeng, K. T. Fang, X. L. Peng and H. Yin, External factor variable connectivity index, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 773–778.

[47] A. T. Balaban, Chemical Graphs. Part 35. Five New Topological Indices for the Branching of Tree–Like Graphs, *Theoret. Chim. Acta* **1979**, *5*, 239–261.

[48] Q. N. Hu, Y. Z. Liang and F. L. Ren, Molecular Graph Center, A Novel Approach to Locate the Center of a Molecule and a New Centric Index, *J. Mol. Struct.* (*Theochem*) **2003**, *635*, 105–113.

Biographies

Qian–Nan Hu is a PhD student of chemistry at the Central South University, P. R. China.
Yi–Zeng Liang is a Professor of chemistry at the Central South University, P. R. China.