

# Structural features hidden in the degree distributions of topological graphs

Qian-Nan Hu  
Yi-Zeng Liang\*

*Institute of Chemometrics and Intelligent Analytical Instruments, College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, Peoples Republic of China*

E-mail: yzliang@public.cs.hn.cn

Qing-Song Xu

*College of Mathematics and Econometrics, Hunan University, Changsha, 410082, P. R. China*

Kai-Tai Fang  
Xiao-Ling Peng  
Hong Yin

*Statistics Research and Consultancy Center, Hong Kong Baptist University, Hong Kong, China*

Received 23 November 2003; revised 29 June 2004

Based on a basic element, vertex degree, of topological graphs, insights are obtained on the structural features hidden in the degree distributions (DD) of saturated hydrocarbons. The investigation shows that the cyclicity and branching features are mainly coded by the different mathematical characteristics of the degree distributions. Surprisingly, the center (or mathematical expectation) of a degree distribution corresponds to the cyclicity of a saturated hydrocarbon, and the dispersion (mean absolute deviation or MAD) around its center of a distribution is a measure of branching. The structural feature such as number of quaternary atoms is also mined out as a special case of branching. The cyclicity and branching information in the present work is with least human intervention, and an interesting thing is that the two features can be unified into the mathematical characteristics of a degree distribution. By the strict mathematical characteristics of a distribution, the structure features within the degree distributions (DD isomer domains) are studied. The space spanned by the size (number of carbons), mathematical expectation, and MAD shows some enlightening results. The results also give a new idea on how to model the properties of diverse structures.

**KEY WORDS:** branching, cyclicity, degree distribution, mathematical characteristics, graph theory

\* Corresponding author.

## 1. Introduction

Any vertex of a connected graph has a certain degree. Vertex degree is one of the basic elements of a topological graph, which is defined as the number of edges issuing from a vertex [1]. The vertex degree is a main feature to describe the atomic attributes, and many indices [2–5] are based on it. The molecular connectivity index is interpreted by the bond additivity [6], which is, in nature, to split the bonds into different kinds of vertex degree combination. Atoms with different vertex degree constitute a distribution for each molecule. Some works [7,8] have been done before on degree distributions of alkanes to discuss the computer generation of degree distributions, maximal degree distributions, and construction and enumeration of isomers in the alkane series. The study [7] analyzed that all isomers that conform to a certain degree distribution can be considered to belong to the same domain (degree distribution), and no isomer can satisfy more than one domain. In the present study, the same and different aspects of structural features among the domains are discussed first.

In statistics, a distribution can be roughly described by the mathematical expectation and variance or mean absolute deviation (MAD) of the distribution. The mathematical expectation might be regarded as the center of the distribution. The variance or mean absolute deviation (MAD) of a distribution provides a measure of the spread or dispersion around its mean (mathematical expectation).

In image analysis, in which the image is described by a specific distribution, the zero- to third-order moment functions are used to describe the overall image characteristics, and the higher order moments contain better image details, but are sensitive to noise. The moment functions have been proven to be successful tools in extracting shape characteristics of images [9,10]. Moment functions represent global shape characteristics in an image, and have been used in many image analysis applications such as pattern recognition, pose estimation, and image classification.

Quantitative characterizations of molecular structural features have been overlooked and neglected for a long time [11]. A critique of topological index is always the lack of structural interpretation, although there are some prior researches [12,13] on this topic. It is the goal of topological index (TI) 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 [14]. To quantify structural features with least human intervention [11] is of pertinent interest in chemistry. The mathematical characteristics of degree distribution are from strict definitions, which are with least human intervention. A surprising thing is that the so simple invariants show some interesting results on structural features.

To compare the structural features that are similar or dissimilar among many molecules is an important aspect of quantitative structure activity/property relationship (QSAR/QSPR) studies. There are some studies [15–17] paying attention on the topics, in which most of them use molecular fingerprint (sub-structures) to evaluate the similarity or dissimilarity among different molecules. In those studies, emphasis is paid on similarity based on the fine (micro-) features. However, searching the major (macro-) features such as size, cyclicity, and branching, to provide information on structure similarity is also focused in the present work.

First, the simple definitions of mathematical expectation, MAD and moment functions are given briefly. Second, the structural features hidden in the degree distribution domains are discussed. Third, the structural features of 530 saturated hydrocarbons are compared. Then, the structure information spanned by the mathematical expectation and MAD is studied. Finally, the structure information by molecular size (number of carbons), mathematical expectation, and MAD is mined out.

## 2. Methodology

### 2.1. The degree distribution (DD)

As shown in the previous studies [7,8], the atomic vertex degrees constitute a degree distribution for a molecule, and the degree distribution has been applied to calculate maximal degree distributions, and construct isomers in the alkane series. The degree distribution (DD) for a certain skeleton is expressed as  $[a, b, c, d]$ . The degree distribution equation has the general form  $a + 2b + 3c + 4d = 2(n - 1 + r)$ , where  $n$  is the number of skeleton atoms (usually carbon atoms),  $r$  the number of rings, and  $a, b, c,$  and  $d$  denote the numbers of skeleton atoms of degrees 1, 2, 3, and 4, respectively. Once a degree distribution is given, the mathematical characteristics of the distribution can be obtained.

### 2.2. Mathematical expectation of the degree distribution (DD)

The mathematical 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  $E(V_i)$  can also be computed by

$$E(V_i) = \text{expectation} = (a + 2b + 3c + 4d)/(a + b + c + d) = 2m/n. \quad (1)$$

### 2.3. Mean absolute deviation (MAD)

The first order absolute central moment is larger than zero, and is called mean absolute deviation (MAD). The MAD can be computed by

$$\begin{aligned} \text{MAD} &= E(\text{abs}(V_i - E(V_i))) \\ &= [a * \text{abs}(1 - E) + b * \text{abs}(2 - E) + c * \text{abs}(3 - E) + d * \text{abs}(4 - E)]/n \end{aligned} \quad (2)$$

in which  $E$  means the expectation.

### 2.4. Central moment of the degree distribution (DD)

The  $k$ th 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 atoms) 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. A formula to calculate the central moments of the degree distribution is given below.

$$\begin{aligned} M^k &= E((V_i - E(V_i))^k) \\ &= [a * (1 - E)^k + b * (2 - E)^k + c * (3 - E)^k + d * (4 - E)^k]/n, \end{aligned} \quad (3)$$

where  $M^k$  denotes the  $k$ th central moments, and  $E$  means the expectation. The central moment functions in the study include second order moment (SOM), and third order moment (TOM) etc.

### 2.5. Absolute central moment of the degree distribution

The  $k$ th absolute central moment of DD can be calculated by  $E(\text{abs}(V_i - E(V_i))^k)$ , 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 absolute central moments can be computed by

$$\begin{aligned} \text{AM}^k &= E(\text{abs}(V_i - E(V_i))^k) \\ &= [a * \text{abs}(1 - E)^k + b * \text{abs}(2 - E)^k + c * \text{abs}(3 - E)^k + d * \text{abs}(4 - E)^k]/n, \end{aligned} \quad (4)$$

where  $\text{AM}^k$  denote the  $k$ th absolute central moments. The absolute central moment functions in the study include mean absolute deviation (MAD), second order absolute moment (SOAM), and third order absolute moment (TOAM) as well. The mathematical value of SOAM is the same to SOM, and the structure information of them is the same.

### 3. Data collection

There are two kinds of data selected. The first is the domains listed in table 5 of [7]. The second is the 530 saturated hydrocarbons (without methane) from [18]. The hydrocarbons are listed in the reference from methane to decanes and also from acyclic to penta-cyclic hydrocarbons. The nomenclature of the hydrocarbons is from [18].

### 4. Results and discussion

#### 4.1. Structure information of the degree distributions (DD isomer domains)

First, the structure information in the different degree distributions (DD isomer domains) is studied. By the equation (1) and (2), the expectation and MAD can be calculated for the degree distributions (DD isomer domains).

The expectations of degree distributions with 0, 1, and 2 ring number(s) are 1.8, 2.0, and 2.2, respectively. The structures with arithmetic number of rings show arithmetic change of expectation values.

For decanes, the MAD values and the possible structures of the different degree distributions (DD isomer domains) (from table 5 in [7]) are listed in table 1. From the table, the structures with arithmetic number of branches show arithmetic change of MAD values. For example, the MAD for n10 is 0.32; that of 2mn9 is 0.48; 0.64 for 22mn8; 0.80 for 223mn7, and another interesting result is that the change is 0.16 for acyclic decanes with one more branch, while the changes are 0.20 and 0.16 for monocyclic and bicyclic decanes with one more branch. However, the situation for bicyclic structures is complex due to the different number of quaternary atoms. For instance, the MAD for bcpe, which has not quaternary atom, is 0.32, while that of s45d, having one quaternary atom, is 0.36. Considering the bicyclic structures with four branches, the MADs are 0.96, 1.00, 1.04, and 1.08 for structures with 0, 1, 2, and 3 quaternary atoms.

From the table 1, the degree distributions (DD isomer domains), with the same mathematical expectation (center of a distribution) and MAD (dispersion around the center of a distribution), show same structural features such as number of rings and branches, although the degree distributions are different. Should the structures, based on the degree distributions, be further classified into same cluster? This is an interesting question that should be discussed deeply by the correlation with the molecular properties, which will appear in another paper.

#### 4.2. The mathematical expectation of the degree distributions of 530 hydrocarbons

The mathematical expectations of the degree distributions of the 530 saturated hydrocarbons are shown in the figure 1, and the mathematical expectations

Table 1  
The MAD values and possible structures within degree distributions (DD isomer domains).

r=0			r=1			r=2		
Domains	MAD	Structures	Domains	MAD	Structures	Domains	MAD (bn v4)	Structures
			[6 0 2 2]	1.2	112234mC4	[6 0 0 4]	1.44	6 4
[7 0 1 2]	1.12	22334mn5	[5 0 5 0]	1.0	12345mC5	[5 0 3 2]	1.20	5 2
[6 0 4 0]	0.96	2345mn6	[6 1 0 3]	1.2	112233mC4	[4 0 6 0]	0.96	4 0
[6 1 2 1]	0.96	2234mn6	[5 1 3 1]	1.0	11234mC5	[5 1 1 3]	1.24	5 3
[6 2 0 2]	0.96	2233mn6	[5 2 1 2]	1.0	11223mC5	[4 1 4 1]	1.00	4 1
[5 2 3 0]	0.80	234mn7	[4 2 4 0]	0.8	1234mC6	[4 2 2 2]	1.04	4 2
[5 3 1 1]	0.80	223mn7	[4 3 2 1]	0.8	1123mC6	[3 2 5 0]	0.80	3 0
[4 4 2 0]	0.64	24mn8	[4 4 0 2]	0.8	1122mC6	[4 3 0 3]	1.08	4 3
[4 5 0 1]	0.64	22mn8	[3 4 3 0]	0.6	123mC7	[3 3 3 1]	0.84	377mbc410h
[3 6 1 0]	0.48	2mn9	[3 5 1 1]	0.6	112mC7	[3 4 1 2]	0.88	lip5mbc310hx
[2 8 0 0]	0.32	n10	[2 6 2 0]	0.4	12mC8	[2 4 4 0]	0.64	23mbc321o
			[2 7 0 1]	0.4	11mC8	[2 5 2 1]	0.68	22mbc321o
			[1 8 1 0]	0.2	1mC9	[2 6 0 2]	0.72	15mbc321o
			[0 10 0 0]	0.0	C10	[1 6 3 0]	0.48	7mbc430n
						[1 7 1 1]	0.52	1mbc331n
						[0 8 2 0]	0.32	bcpe
						[0 9 0 1]	0.36	s45d

r means the number of rings, bn denotes the number of branches, and v4 represents the number of quaternary atoms.

of these hydrocarbons are listed in the table 2 with their structures given in table 3. From figure 1, tables 2 and 3, some conclusions might be drawn.

- (1.1) The mathematical expectation is 2 for any monocyclic hydrocarbons, and the value could be regarded as the threshold distinguishing the cyclic and acyclic hydrocarbons.
- (1.2) The mathematical expectation of hydrocarbons, with same number of carbon atoms and cycle(s), is identical, in which the branches seem have no influence on the mathematical expectation.
- (1.3) The mathematical expectation of alkane is approaching to 2 with the increase of the number of carbon atoms. For example, the mathematical expectation value is 1.9333 for alkanes with 30 carbon atoms, 1.9800 for 100, 1.9960 for 500, and 1.9980 for 1000.
- (1.4) The mathematical expectation of multicyclic hydrocarbons decreases and approaches to 2 with the increase of the number of carbon atoms. That is, for bicyclic hydrocarbons, the mathematical expectation is 2.0400 for hydrocarbons with 50 carbon atoms, 2.0200 for 100, 2.0100 for 200, and 2.0040 for 500. For tricyclic hydrocarbons, the mathematical expectation is 2.2000 for hydrocarbons with 20 carbon atoms, 2.0800 for 50, 2.0400 for 100, 2.0080 for 500, 2.0040 for 1000, and 2.0020 for 2000.

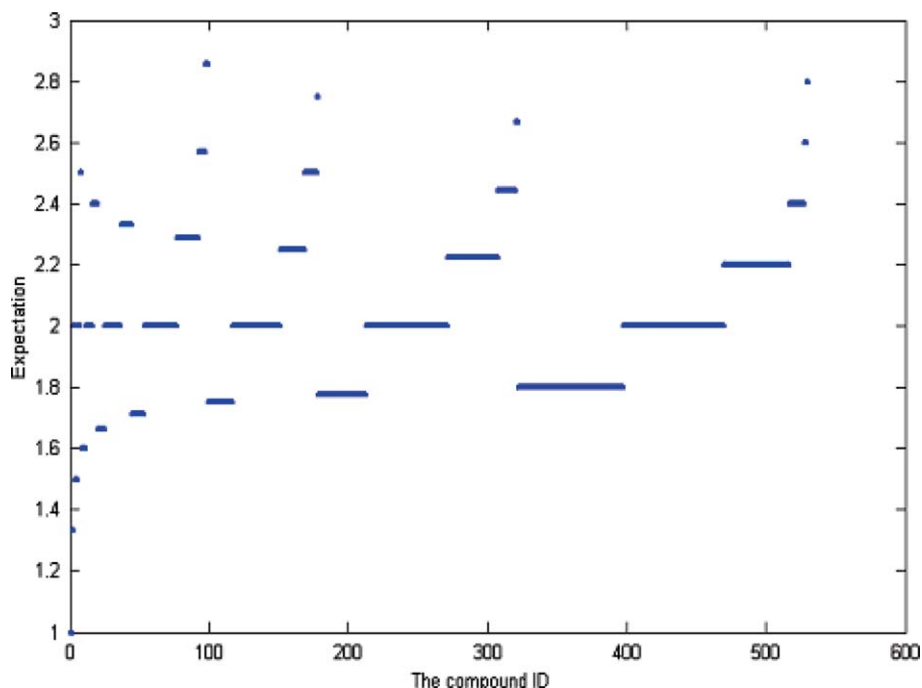


Figure 1. Structural features based on the expectations of the degree distributions of the 530 hydrocarbons. (The  $x$ -axis is the compound ID (without methane) in [18], and the  $y$ -axis is the expectation.)

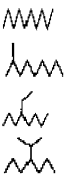
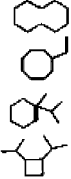
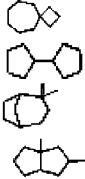
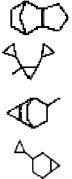
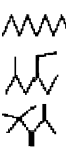
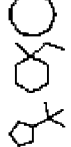
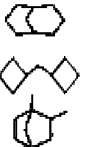
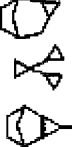
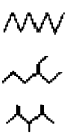
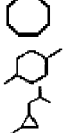
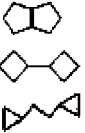





Table 2  
The expectations of the degree distributions of the 530 hydrocarbons.

NC*	Cycle(s)					
	0	1	2	3	4	5
2	1					
3	1.3333	2.0000				
4	1.5000	2.0000	2.5000			
5	1.6000	2.0000	2.4000			
6	1.6667	2.0000	2.3333			
7	1.7143	2.0000	2.2857	2.5714	2.8571	
8	1.7500	2.0000	2.2500	2.5000	2.7500	
9	1.7778	2.0000	2.2222	2.4444	2.6667	
10	1.8000	2.0000	2.2000	2.4000	2.6000	2.8000

\*NC indicates the number of carbon atoms.

- (1.5) The differences of mathematical expectation between two hydrocarbons, with the same number of carbon atoms and differing by one number of cycle (such as hydrocarbons with 0 cycle and 1 cycle; or 1 cycle and 2

Table 3  
Structure examples of some saturated hydrocarbons and the related expectations of the degree distributions.

NC*	Expectation:	Cycle(s)				Difference
		0	1	2	3	
10	Expectation:	1.8000	2.0000	2.2000	2.4000	0.2000
						
9	Expectation:	1.7778	2.0000	2.2222	2.4444	0.2222
						
8	Expectation:	1.7500	2.0000	2.2500	2.5000	0.2500
						
7	Expectation:	1.7143	2.0000	2.2857	2.5714	0.2857
						

\*NC indicates the number of carbon atoms.

cycles), are equivalent. And the differences are 0.5000, 0.4000, 0.3333, 0.2957, 0.2500, 0.2222, and 0.2000 for hydrocarbons with 4, 5, 6, 7, 8, 9, and 10 carbon atoms respectively. The differences are decreasing with the increase of carbon atoms. The mathematical expectations of hydrocarbons, with arithmetic number of cycle(s) and same number of carbon atoms, form an arithmetic series.



Next, the relationship among cycle number, cyclicity and the mathematical expectation will be discussed. Characterization of molecular cyclicity has received some attention. Bonchev, Mekenyan and Trinajstić [19] proposed the first cyclicity index for molecules. The cyclomatic number has been used in defining the Balaban  $J$  index [20] and is also equal to the so-called smallest set of smallest rings. Balaban et al. [21] proposed a program for finding all possible cycles in graphs. Bonchev et al. [22] have studied the cyclicity of polycyclic graphs. The definition of cycle number [23–27] is

$$\mu = n_e - n + 1 \quad (5)$$

in which  $n$  denotes the number of carbon atoms, and  $n_e$  means the number of edges in a saturated hydrocarbons. Recently a novel cyclicity index [28] was proposed, which has a high discrimination power and hopefully will resolve the questions of relative cyclicity among molecules with limited if any human intervention [28]. The index is based on the so-called D/DD quotient matrix constructed from the elements of the graph distance matrix  $D$ , and the graph detour matrix  $DD$ . Some other researchers [22–26] have also studied and applied the cyclicity, essentially the cycle number of a molecule, in different cases.

From the above analysis, the cycle number cannot completely explain the structure information of mathematical expectation, which also takes into consideration of number of atoms in the molecules. The cyclicity should be distinguished from the cycle number, and the cycle number and cyclicity have some common and uncommon aspects. The cycle number is the count of cycles in a molecule, however the cyclicity should be the parameter describing the whole molecule and, at the same time, taking into consideration of cycle number.

Consider two molecules tricyclic butane and tricyclic decane, with the same number of cycle but different carbon atoms, their cyclicity should be different. From the investigation and from figure 1, the mathematical expectation represents, to some extent, the cyclicity of a molecule, in which the increase of arithmetic number of cycle(s) and same number of carbon atoms, form an arithmetic series.

The molecular cyclicity has been studied for many years, although the concept in hydrocarbons has not been rigorously defined. The same dilemma extends to molecular shape, chirality, the degree of folding, degree of planarity, molecular complexity, aromaticity, molecular similarity and molecular diversity, and even the branching in alkanes has not been rigorously defined [11]. At here, the different definitions (about 10!) of cyclicity are listed to give a complete impression, however what is cyclicity and which is the cyclicity index are two problems that the scientific cycle should figure out.

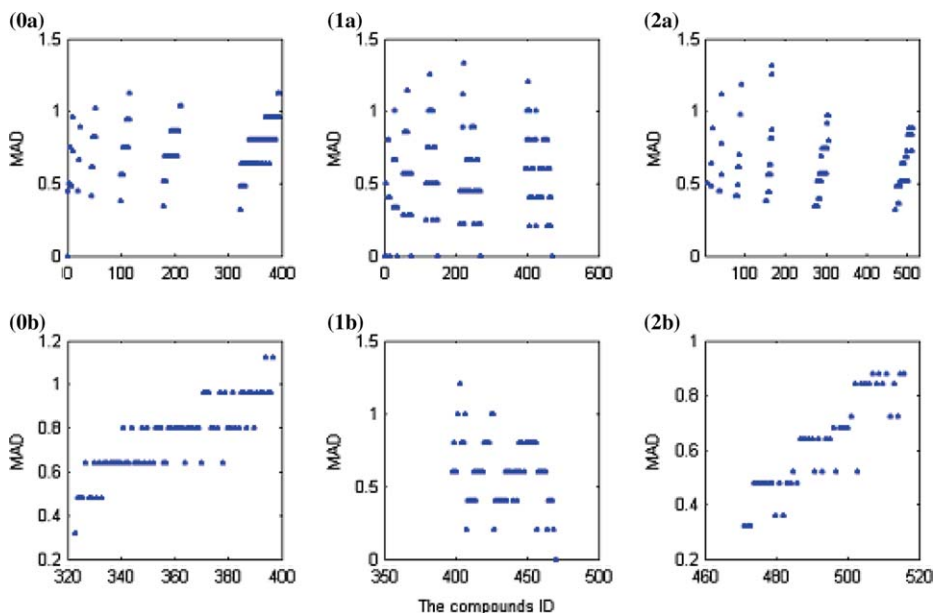


Figure 2. The MADs and structural features of hydrocarbons (The  $x$ -axis is the compound ID in [18], the  $y$ -axis is the MAD. The subplots as 0a and 0b, 1a and 1b, and 2a and 2b correspond to all acyclic alkanes and acyclic decanes, all monocyclic hydrocarbons and monocyclic decanes, and all bicyclic hydrocarbons and bicyclic decanes, respectively).

### 4.3. The MADs of the degree distributions of 530 hydrocarbons

In the above discussion (section 4.1), the MAD mine out branching to be an important global characterization, and the number of quaternary atoms is mined out as a special case of branching, in which two branches connect on a same atom. In this section, the structure information from MAD of 530 hydrocarbons is further investigated especially on the results among structures with different size. Figure 2 shows the MADs of the degree distributions of all hydrocarbons, and table 4 lists the MADs of these hydrocarbons. From figure 2, and table 4, some conclusions could be obtained.

- (2.1) For acyclic or monocyclic hydrocarbons, the hydrocarbons, having same number of carbon atom, cycle, and branch, hold the same MAD. For example, 2m3en7, 22mn8 and 45mn8, with ten carbon atoms, two branches and zero cycle, have the same MAD to be 0.6400.
- (2.2) The MAD for hydrocarbons with one more branch and same number of cycle will “jump” at one higher order with the same step for hydrocarbons with the same number of carbon atoms, for instance, the MAD of n10 is 0.3200; 2mn9 is 0.4800, 26mn8 is 0.6400, and 246mn7 is 0.8000.

Table 4

The MADs of the degree distributions of the hydrocarbons (r means the number of cycle(s), bn denotes the number of branch(es), and v4 represents the number of the quaternary atom(s)).

r	bn	Total carbon atoms:						
		5	6	7	8	9	10	
		v4 (only consider for bicyclic structures)						
0	0	0.4800	0.4444	0.4082	0.3750	0.3457	0.3200	
	1	0.7200	0.6667	0.6122	0.5625	0.5185	0.4800	
	2	0.9600	0.8889	0.8163	0.7500	0.6914	0.6400	
	3			1.0204	0.9375	0.8642	0.8000	
	4				1.1250	1.0370	0.9600	
	Difference for bn	0.2400	0.2222	0.2040	0.1875	0.1728	0.1600	
1	0	0	0	0	0	0	0	
	1	0.4000	0.3333	0.2857	0.2500	0.2222	0.2000	
	2	0.8000	0.6667	0.5714	0.5000	0.4444	0.4000	
	3		1.0000	0.8571	0.7500	0.6667	0.6000	
	4			1.1429	1.0000	0.8889	0.8000	
	Difference for bn	0.4000	0.3333	0.2857	0.2500	0.2222	0.2000	
2	0	0	0.4800	0.4444	0.4082	0.3750	0.3457	0.3200
		1	0.6400	0.5556	0.4898	0.4375	0.3951	0.3600
	1	0			0.6122	0.5625	0.5185	0.4800
		1	0.8800	0.7778	0.6939	0.6250	0.5679	0.5200
	2	0					0.6914	0.6400
		1				0.8125	0.7407	0.6800
		2		1.1111	0.9796	0.8750	0.7901	0.7200
3	0							
	1					0.9136	0.8400	
	2				1.1837	0.9630	0.8800	
	Difference for V4	0.1600	0.1112	0.0816	0.0625	0.0494	0.0400	
	Difference for bn	0.2400	0.2222	0.2041	0.1875	0.1728	0.1600	

\*Difference for bn means the difference of one more branch on MAD; Difference for V4 means the difference of one more V4 on MAD.

Investigate the hydrocarbons and their MADs: the MAD of C10 is 0, 1mC9 is 0.2000, 11mC8 is 0.4000, 12mC8 is 0.4000, 112mC7 is 0.6000, and 123mC7 is 0.6000. The changes differing by one branch on the MAD are 0.2500, 0.2400, 0.2222, 0.2041, 0.1875, 0.1728, and 0.1600 for alkanes with 4, 5, 6, 7, 8, 9, and 10 carbon atoms, respectively, and 0.5000, 0.4000, 0.3333, 0.2857, 0.2500, 0.2222, and 0.2000 for monocyclic hydrocarbons with 4, 5, 6, 7, 8, 9, and 10 carbon atoms, respectively.

(2.3) Bicyclic saturated hydrocarbons, having same number of carbon atom, cycle, branch, and quaternary atoms, hold the same MAD. For instance, the three molecules: 22mbc321o, 22mbc222o, and 14mbc321o, have the same MAD to be 0.6800. The MAD for hydrocarbons with one more




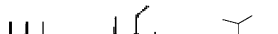


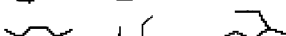




branch and same number of quaternary atoms will “jump” at one higher order with the same step for hydrocarbons with identical carbon atoms, that is, the MAD of bc440d is 0.3200, 3mbc331n is 0.4800, 6mbc322n is 0.4800, 23mbc321o is 0.6400, 26mbc222o is 0.6400, or the MAD of s45d is 0.3600, s36d is 0.3600, 1ms44n is 0.5200, 22mbc222o is 0.6800, 266mbc311h is 0.8400, and the MAD difference of bc440d and 3mbc331n is 0.16000, also 3mbc331n and 23mbc321o or 26mbc222o. The changes differing by one branch on the MAD are 0.2400, 0.2222, 0.2041, 0.1875, 0.1728, and 0.1600 for bicyclic hydrocarbons with same number of quaternary atom, and 5, 6, 7, 8, 9, and 10 carbons, respectively.

- (2.4) For acyclic and monocyclic molecules, the MADs for structures with same number of branch and atoms are the same, but for the bicyclic molecules, the MADs are grouped into different cases by the number of quaternary atoms. The changes differing by one quaternary atom on the MAD are 0.1600, 0.1112, 0.0816, 0.0625, 0.0494, and 0.0400 for bicyclic hydrocarbons with 5, 6, 7, 8, 9, and 10 carbons, respectively. The number of quaternary atoms seems have no effect on the MAD for acyclic and monocyclic hydrocarbons, while it should be distinguished for multicyclic hydrocarbons.

Next, the relationship and difference between branching and the MAD will be detailed. The subject of branching has received considerable attention in graph theory since the birth of Wiener index [29]. Prof. Randić’ reviewed the attributes of molecular characterization [11], which gave details of the development of topological index especially on branching and cyclicity. In 1973, Lovasz and Pelikan suggested the leading or the first eigenvalue of the adjacency matrix as a molecular branching index [30]. Two years later, Randić’ [2] proposed the branching index, which was proven to be a useful descriptor of molecular branching. And soon later, Gutman and Randić’ [31] considered branching, to some degree, from the mode of a distribution, which allows a rigorous definition of the concept of branching and they suggested that structures having an identical distribution of valencies should not be discriminated. In the same year, Bonchev and Trinajstić also defined a measure of branching from information theory, which is essentially based on a kind of distribution [32]. In 1988, Bertz [33] proposed a definition of branching in terms of the degrees of the central points in the star graphs. Kirby [34] 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 [35], which is based on the path matrix, a newly introduced matrix in which the matrix elements were expressed as the path subgraphs of a graph considered [36]. In 1998, Randić’ [37] studied the branching of acyclic saturated hydrocarbons.

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

Table 5  
Structure examples of decanes and the related moment functions of the degree distributions.

Structures	MAD	SOM	TOM	TOAM
	0.3200	0.1600	-0.0960	0.1088
	0.4800	0.3600	0.0240	0.3312
	0.6400	0.5600	0.1440	0.5536
	0.8000	0.7600	0.2640	0.7760
Difference	0.1600	0.2000	0.1200	0.2224
	0	0	0	0
	0.2000	0.2000	0	0.2000
	0.4000	0.4000	0	0.4000
	0.6000	0.6000	0	0.6000
Difference	0.2000	0.2000	0	0.2000
	0.3200	0.1600	0.0960	0.1088
	0.4800	0.3600	-0.0240	0.3312
	0.6400	0.5600	-0.1440	0.5536
Difference	0.1600	0.2000	-0.1200	0.2224

\*NC indicates the number of carbon atoms.

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. An interesting thing is that the structure information of MAD corresponds to the branching of molecules. The molecular branching in alkanes has not even been rigorously defined [11]. At here, the different definitions (about 8!) of branching are listed, however what is branching and which is the branching index should be solved.

#### 4.4. The second order moment (SOM) and structural features

The structures of some hydrocarbons and their corresponding SOMs are listed in the third column (SOM) in table 5. Since most of the structure information is the same (only different with the values) with that of the MAD, and so the results are given without details.

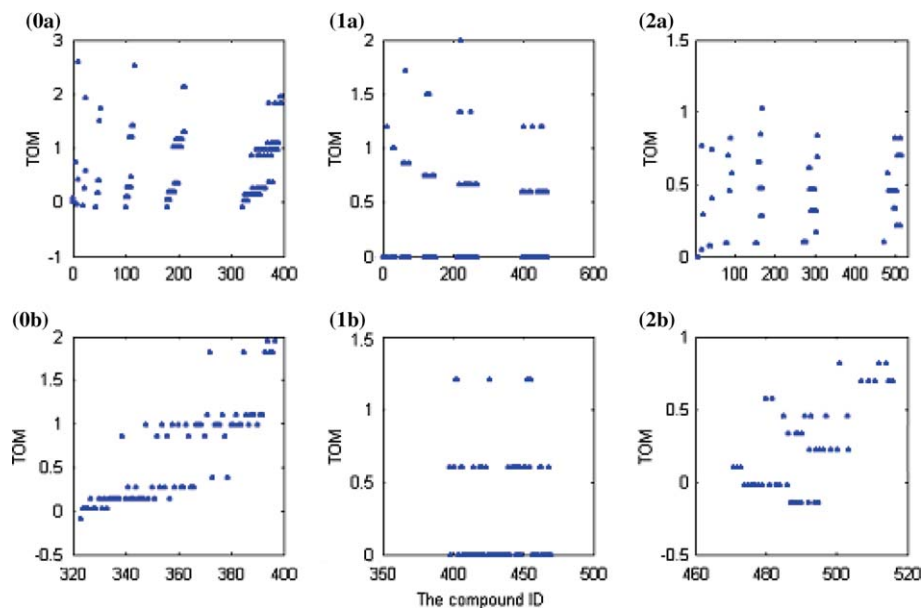


Figure 3. The third order moment and structural features of hydrocarbons (The  $x$ -axis is the compound ID in reference 18, the  $y$ -axis is the TOM. The subplots as 0a and 0b, 1a and 1b, and 2a and 2b correspond to all acyclic alkanes and acyclic decanes, all mono-cyclic hydrocarbons and mono-cyclic decanes, and all bi-cyclic hydrocarbons and bi-cyclic decanes, respectively).

The main character of this group is that the SOM of hydrocarbons with same number of carbon atoms, cycle(s) and one more number of quaternary atoms, will “jump” on the SOM of hydrocarbons with one more branch and one less quaternary atom, such as the SOM of 22mn10 is 0.7600 and 246mn7 is 0.7600; or 11mC8 is 0.6000 and 123mC7 is 0.6000 or 1mbc331n is 0.5600 and 23mbc321o is 0.5600.

#### 4.5. The third order moment (TOM) and structural features

From figure 3, where the TOMs are shown by the compounds order of [18], and the fourth column (TOM) in table 5 shows the TOMs of some hydrocarbons, the following information may be gotten for the TOM, without much duplicate information from those of the MAD and SOM.

The main characters of this group are

- (a) The hydrocarbons with same number of quaternary atom will be far away, as shown in figure 3, from the hydrocarbons differing by one number of quaternary atom. For example, for monocyclic hydrocarbons, the V4 effects (effects by quaternary atom) on the TOM are 1.2000, 1, 0.8571, 0.7500, 0.6667, and 0.6000 for hydrocarbons with 5, 6, 7, 8,

9, and 10 carbons, respectively. However, the Branch Effects (effects by branch) on the TOM are null for monocyclic hydrocarbons, that is, the monocyclic hydrocarbons, with same number of quaternary atom and whatever number of branch, have the same TOM.

- (b) The Branch Effects are different from case to case in the TOM. For alkanes, the Branch Effects are upward or positive, which means that the alkanes with more branches will hold higher TOM. For monocyclic hydrocarbons, the Branch Effects are null, that is, the hydrocarbons, with same number of quaternary atom and whatever number of branch, have the same TOM, while for bi-cyclic hydrocarbons, the Branch Effects are downward or negative.

#### 4.6. *The third order absolute moment (TOAM) and structural features*

The structures of some hydrocarbons and their corresponding TOAMs are listed in the fifth column of table 5. Some interesting results are obtained for TOAM.

The main aspects of this group are: (a) same to the information of TOM, the hydrocarbons with same number of quaternary atom will be far away from the hydrocarbons differing by one number of quaternary atom. (b) Different from TOM, the Branch Effects are all upward or positive, which means that the hydrocarbons with more branch will hold higher TOAM.

#### 4.7. *The fourth or higher order of moments and absolute moments*

The structure information from higher order of moments can also be explained by using the Branch and V4 Effects. The readers can easily deduce the information by themselves. An interesting thing is that the higher order (fourth or higher) moments offer almost no new information (only different in values from the structure information of the former moments).

#### 4.8. *Structure information from the center and dispersion of 530 degree distributions*

In statistics, a distribution can be roughly described by its mathematical expectation and dispersion around the mathematical expectation. After obtaining the several mathematical characteristics of the degree distributions, the next interest is to investigate the structure information in the space spanned by the characteristics, which is shown in the figure 4. From the figure, some interesting results, besides the conclusions discussed above, are concluded.

The line with mathematical expectation 2.0 is a set of mono-cyclic hydrocarbons and no matter how many branches in the molecules, in which, however, the hydrocarbons with the same number of branch(es) hold same MAD.

The left area of the line is for alkanes, in which alkanes with same number of atoms hold the same mathematical expectation. What is attractive to the chemist is that the alkanes, with same number of branch(es) but with different number of carbon atoms, show some regularity. For example, the alkanes without branch and different carbon atoms are belonged to line b0, and alkanes with three branches lie on the line b3.

For acyclic and bi-cyclic hydrocarbons, without quaternary atoms at the same time, if they have same number of carbon atom and branch, the MADs of them are identical, and their absolute mathematical expectation differences to line 2 (the set of mono-cyclic structures) are equal. For instance, the mathematical expectation and MAD of 2mn9 are 1.8000 and 0.4800, and those of 3mbc331n are 2.2000 and 0.4800.

The structure information on the right area of the line 2, due to the complexity of cycles and branches, is much more complicated and will be discussed below.

#### 4.9. Structural information by number of carbons (NC), expectation and MAD of the 530 degree distributions

In figure 4, hydrocarbons are plotted without the information of the change on both mathematical expectation and MAD with the variation of carbon

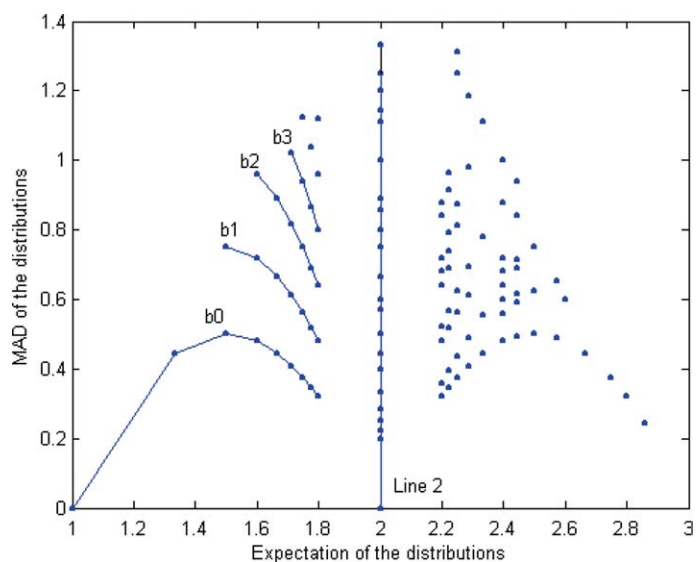


Figure 4. Structure information in the space spanned by the expectations and MADs of the degree distributions. (The lines b0, b1, b2, b3 represent alkanes with none, one, two, three branches respectively.)



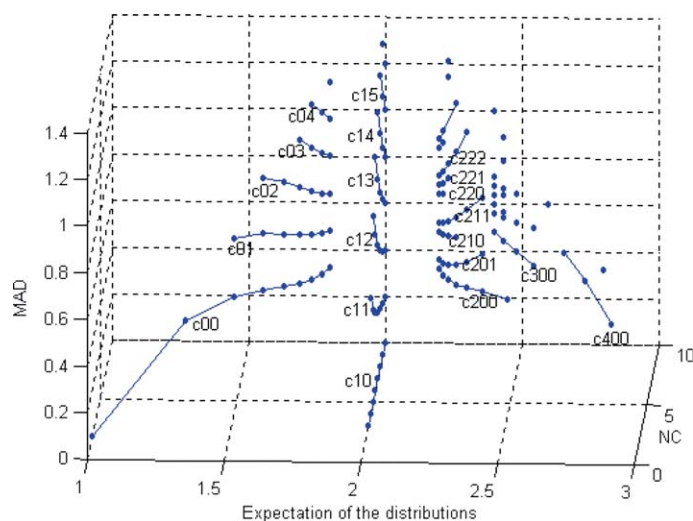


Figure 5. Structural diversity by the NC (number of carbon atoms), the expectations and MADs of the degree distributions. (c – means classification, the first number following the c, such as 0, 1, 2, 3 et al., denotes the number of cycles in the alkanes, the second number following the c, such as 0, 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.)

atoms. The structure information by the molecular size (number of carbons), and two mathematical characteristics of the degree distributions is shown in figure 5, in which the hydrocarbons with same number of carbons, cycles, branches and quaternary atoms (for multi-cyclic hydrocarbons) hold the same expectation and MAD value, and the hydrocarbons with same number of cycles, branches and quaternary atoms (for multi-cyclic hydrocarbons) show some regularity. The groups such as c00, c01, c02, c03, and c04 are alkanes with zero, one, two, three, and four branches, respectively. The c10 group is a set of monocyclic hydrocarbons without branch, c11 for monocyclic hydrocarbons with one branch, and c12 for monocyclic hydrocarbons with two branches. When considering the multi-cyclic hydrocarbons, the situation is much more complex. The structural feature such as number of quaternary atom is mined out as an important aspect. The c200 group is for unbranched bi-cyclic hydrocarbons without quaternary atoms, and c201 is for unbranched bi-cyclic hydrocarbons with one quaternary atom. The groups such as c220, c221, and c222 are for bi-cyclic hydrocarbons with two branches, and zero, one, and two quaternary atoms, respectively.

From figure 4 if the hydrocarbons are with the same number of carbon atoms, cycles, branches and quaternary atoms (for multi-cyclic hydrocarbons), they hold the same mathematical expectation and MAD. And if the hydrocarbons are with different carbon atoms, but with same number of cycles, branches

and quaternary atoms (for multi-cyclic hydrocarbons), they seem to show regularity and are in the same group. Another interesting thing is that the structures in different groups (different in number of cycles, branches, and/or quaternary atoms) differ from other groups with variation, which is changed regularly as mentioned in above sections, on expectation and/or MAD values. The different groups in the hydrocarbons form an overview of the structural similarity or 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 [18]. 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 study of structural diversity attracts much attention of chemists in evaluating databases [38,39], combinatorial chemistry [40]. Some methods [41,42] were proposed, largely based on similarity between two molecules, to quantify structural diversity. In intuitive view of point, the discussed structural diversity in many references is a comprehensive and qualitative concept, which should be, mainly, composed of the major structural features such as size, cyclicity, branching, and also other finer ones. To quantify the size, cyclicity, branching, and other finer ones should be helpful to obtain a numerical form of structural diversity. The present 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 this research may bring a new way to model the properties of saturated hydrocarbons.

## **5. Concluding remarks**

Based on the strict mathematical characteristics of a distribution, the structure features hidden in the degree distributions (DD isomer domains) are studied. Some interesting results on cyclicity, branching, similarity and diversity are obtained. The cyclicity and branching information in the present work is with least human intervention, and an interesting thing is that the two features can be unified into two mathematical characteristics of the degree distribution, in which the center (expectation) of the distribution corresponds to the cyclicity of saturated hydrocarbons, and the dispersion around its center (or cyclicity) of the distribution is a measure of branching. The space spanned by the size (number of carbons), mathematical expectation, and MAD shows some interesting informa-

tion on structure similarity. The results also give some enlightening idea on how to model the properties of diverse structures.

## Acknowledgments

This project is financially supported by National Nature Foundation Committee (NNFC) of P. R. China (No.20235020, 20175036). And the authors also appreciate the hospitality of HongKong Baptist University, when the authors attend “Workshop on Data Mining in Chemistry and Traditional Chinese Medicines” in Oct. 2003.

## References

- [1] N. Trinajstić, *Chemical Graph Theory*, 2nd edition (CRC Press, LLC, 1992).
- [2] M. Randić, *J. Am. Chem. Soc.* 97 (1975) 6609.
- [3] G. Moreau and P. Broto, *Nouv. J. Chim.* 4 (1980) 359.
- [4] Q.N. Hu, Y.Z. Liang, Y.L. Wang, C.J. Xu, Z.D. Zeng, K.T. Fang, X.L. Peng and H. Yin, *J. Chem. Inf. Comput. Sci.* 34 (2003) 773.
- [5] Q.N. Hu, Y.Z. Liang and F.L. Ren, *THEOCHEM.* 635 (2003) 105.
- [6] M. Randić and J. Zupan, *J. Chem. Inf. Comput. Sci.* 41 (2001) 550.
- [7] T.I. Bieber and M.D. Jackson, *J. Chem. Inf. Comput. Sci.* 33 (1993) 699.
- [8] T.I. Bieber and M.D. Jackson, *J. Chem. Inf. Comput. Sci.* 33 (1993) 701.
- [9] R. Mukundan, *Moments Functions in Image Analysis* (World Scientific Publishing, Co. Pte. Ltd., 1998).
- [10] T.J. Wang and T.W. Sze, *Pattern Recognit.* 34 (2001) 2145.
- [11] M. Randić, *Acta Chim. Slov.* 45 (1998) 239.
- [12] M. Randić, A.T. Balaban and S.C. Basak, *J. Chem. Inf. Comput. Sci.* 41 (2001) 593.
- [13] M. Randić and J. Zupan, *J. Chem. Inf. Comput. Sci.* 41 (2001) 550.
- [14] S.H. Bertz, *J. Am. Chem. Soc.* 103 (1981) 3599.
- [15] J.W. Raymond, E.J. Gardiner and P. Willett, *J. Chem. Inf. Comput. Sci.* 42 (2002) 305.
- [16] J.D. Holliday, N. Salim, M. Whittle and P. Willett, *J. Chem. Inf. Comput. Sci.* 43 (2003) 819.
- [17] J.W. Raymond and P. Willett, *J. Chem. Inf. Comput. Sci.* 43 (2003) 908.
- [18] G. Rücker and C. Rücker, *J. Chem. Inf. Comput. Sci.* 39 (1999) 788.
- [19] D. Bonchev, O. Mekenyan and N. Trinajstić, *Int. J. Quant. Chem.* 17 (1980) 845.
- [20] A.T. Balaban, *Chem. Phys. Lett.* 89 (1982) 399.
- [21] A.T. Balaban, P. Filip and T.S. Balaban, *J. Comput. Chem.* 6 (1985) 316.
- [22] D. Bonchev, A.T. Balaban, X. Liu and D.J. Klein, *Int. J. Quantum Chem.* 50 (1994) 1.
- [23] D.J. Klein and O. Ivanciuc, *J. Math. Chem.* 30 (2001) 271.
- [24] G. Rücker and C. Rücker, *Chimia.* 44 (1990) 116.
- [25] J.B. Hendrickson and C.A. Parks, *J. Chem. Inf. Comput. Sci.* 31 (1991) 101.
- [26] A. Srikrishna, *J. Chem. Educ.* 73 (1996) 428.
- [27] M. Petitjean, B.T. Fan, A. Panaye and J.P. Doucet, *J. Chem. Inf. Comput. Sci.* 40 (2000) 1015.
- [28] M. Randić, *J. Chem. Inf. Comput. Sci.* 37 (1997) 1063.
- [29] H. Wiener, *J. Am. Chem. Soc.* 69 (1947) 17.
- [30] L. Lovász and J. Pelikan, *Period. Math. Hung.* 3 (1973) 175.
- [31] I. Gutman and M. Randić, *Chem. Phys. Lett.* 47 (1977) 15.
- [32] D. Bonchev and N. Trinajstić, *J. Chem. Phys.* 67 (1977) 4517.

- [33] S. Bertz, *Discrete Appl. Math.* 19 (1988) 65.
- [34] E.C. Kirby, *J. Chem. Inf. Comput. Sci.* 34 (1994) 1030.
- [35] M. Randić, *Acta Chim. Slov.* 44 (1997) 57.
- [36] M. Randić, D. Plavšić and M. Razinger, *MATCH.* 35 (1997) 243.
- [37] M. Randić, *J. Math. Chem.* 24 (1998) 345.
- [38] A.M. Munk and J.T. Pedersen, *J. Chem. Inf. Comput. Sci.* 41 (2001) 338.
- [39] D.M. Bayada, H. Hamersma and V.J. Geerestein, *J. Chem. Inf. Comput. Sci.* 39 (1999) 1.
- [40] W.A. Warr, *J. Chem. Inf. Comput. Sci.* 37 (1997) 134.
- [41] D.K. Agrafiotis, *J. Chem. Inf. Comput. Sci.* 37 (1997) 576.
- [42] D.B. Turner, S.M. Tyrrell and P. Willett, *J. Chem. Inf. Comput. Sci.* 37 (1997) 18.