Internet Electronic Journal of

Molecular Design

October 2008, Volume 7, Number 10, Pages 216–224

Editor: Ovidiu Ivanciuc

On the Degeneracy of Molecular Identification Number MID06

Damir Vukičević,[1] Tanja Vojković [1]

[1] Department of Mathematics, Faculty of Natural Sciences and Mathematics, University of Split, HR–21000 Split, Croatia

Received: April 3, 2009; Accepted: August 14, 2009; Published: November 14, 2009

Citation of the article:

D. Vukičević and T. Vojković, On the Degeneracy of Molecular Identification Number MID06, *Internet Electron. J. Mol. Des.* **2008**, *7*, 216–224, http://www.biochempress.com.

Inter*net* **Electronic** Journal
of Molecular Design

BioChem Press
http://www.biochempress.com

On the Degeneracy of Molecular Identification Number MID06

Damir Vukičević,[1,*] Tanja Vojković [1]

[1] Department of Mathematics, Faculty of Natural Sciences and Mathematics, University of Split, HR–21000 Split, Croatia

Received: April 3, 2009; Accepted: August 14, 2009; Published: November 14, 2009

Internet Electron. J. Mol. Des. **2008**, 7 (10), 216–224

Abstract
The topological index MID06 (Molecular Identification Number 06) has been proposed by Chang–Yu Hu and Lu Xu in their paper "Developing Molecular Identification Numbers by an All–Paths Method". It has been tested successfully on a large number of alkane isomers and molecules containing heteroatoms. Here we analyze the degeneracy of MID06 for alkanes, and we demonstrate that there are many more alkanes then possible MID06 identifiers. Although MID06 has very good discriminative properties, it has degenerate values for certain alkane pairs.

Keywords. Topological index; molecular descriptor; molecular graph; index degeneracy.

Abbreviations and notations

AID, atom identification number	NIST, Chemistry WebBook
IU–PAC, International Union of Pure and Applied Chemistry	NP, non polynomial
MID, molecular identification number	PI, path identifier hips

1 INTRODUCTION

NIST Chemistry WebBook [1] is one of the internet databases of chemical compounds. In order to search through those databases every molecule must be uniquely identified. International Union of Pure and Applied Chemistry (IUPAC) [2] has developed a standard for naming chemical elements and compounds. To search the databases more successfully we need a quick algorithm for identifying and discriminating substances, and the most effective way to do that is by using their structural formula. In mathematics that problem comes down to determining a canonical graph form, which is extremely difficult problem and falls into class of NP–difficult problems [3]. Concept of isomorphism, finding if two graphs are isomorphic, is also very important in this problem. It also falls into NP–problems class, although there are very good algorithms for this, like Nauty [4]. Molecular descriptors are therefore very important and scientists continuously try to

[*] Correspondence author; E–mail: vukicevi@pmfst.hr.

develop better identifiers. However, it has been shown that even the most discriminative descriptors, such as the Balaban topological index *J* [5] cannot discriminate every set of graphs, moreover, when the number of vertices in a graph is large enough, every graph has its indiscriminative counterpart [6].

Chang–Yu Hu and Lu Xu in their article "Developing Molecular Identification Numbers by an All–Paths Method" [7] have suggested new molecular descriptors, based on all–paths method. The best one was Molecular Identification Number 06 (MID06). It was tested on the family of alkane trees with p to 22 atoms, that is 3 807 434 alkane isomers, and on 430 472 structures containing heteroatoms having 8 carbon atoms and up to one oxygen or/and nitrogen heteroatoms. No duplicate with an identical MID06 number were found for observed substances.

It is interesting to explore if MID06 descriptor works in general. We will show that that is not the case, that two graphs with identical MID06 exist. We will show that even the family of trees is not discriminative, although we have certificates for that in mathematics [8].

2 PRELIMINARIES

Let G be a graph. We denote number of vertices of G by $v(G)$, and by $e(G)$ we denote number of edges of G. We say that graph is d–regular if d is a degree of every vertex $v \in V(G)$. Two graphs G and H are isomorphic if bijections $f : V(G) \to V(H)$ and $g : E(G) \to E(H)$ exist, so that vertex v in G is incident with edge e in G if and only if vertex $f(v)$ in H is incident with edge $g(e)$ in H. We denote an array of vertices $v_0 v_1 ... v_k$ in a graph G as a walk, and if all the vertices $v_0 v_1 ... v_k$ are different, we call that walk a path in G. Distance $d(u,v)$ of vertices u and v is the length of the shortest path between u and v. Diameter of a graph G is the maximum distance between two vertices in G. Tree is connected acyclic graph. A root of the tree is one fixed vertex. Vertices that have a degree of 1 are called leaves (or pendant vertices).

Algorithm for determining MID06 of given molecule, described in paper [7], is as follows. Let $b_{k,k-1}$ be the code for the bond between nodes k and $k-1$, (its values are 1,2,3 for single, double and triple bond, and 1.5 for aromatic bond), $\delta'(k)$ for atom k is $\delta'(k) = \delta(k)\sqrt{Z}$, where $\delta = Z - h$, is connectivity for atom k (the number of non–hydrogen atoms attached to it), Z is the atomic number, and h is the number of attached hydrogen atoms. For every two atoms i and j in a molecule we determine Path Identifier 06:

$$\text{PI06} = \prod_{k=2}^{n_{ij}} \sqrt{\frac{b_{k,k-1}}{k} \cdot \frac{1}{\delta'(k) \cdot \delta'(k-1)}} \qquad (1)$$

where k is the sequence number of the nodes along the path between nodes i and j, n_{ij} is the total number of nodes in the path. For an atom (vertex) i in the graph, AID06 (Atom Identification

Number) is obtained by adding the *PI*s of all paths started at that vertex:

$$AID_i = \sum_j PI_{ji} \tag{2}$$

where *j* is the sequence number of every other atom in the molecule. We define MID06 as the sum of AID^2:

$$MID = \sum_i AID_i^{\,2} \tag{3}$$

where *i* is the sequence number of every atom in the molecule.

We will observe molecular structures as chemical graphs. There are two kinds of chemical graphs, plerograph and kenograph [9]. A plerograph is a graph where every atom is represented by a vertex, and two vertices are connected if responding atoms are chemically connected. A kenograph is a graph where every non–hydrogen atom is represented by a vertex.

We will use kenographs and observe only those with following properties:

– Alkanes with *n* carbon atoms (trees with *n* vertices);

– All vertices must have degree of 1 or 4, i.e. every carbon atom is connected with 1 or 4 carbon atoms (these graphs will be called 1,4–regular graphs);

– Diameter of tree with *n* vertices must be less or equal to $3\log_3 2\text{n}$;

The family of such trees will be denoted by $\tau(n)$.

We will show that there are more described graphs than possible different MIDs, and then conclude our claim about degeneracy for all molecular structures.

3 RESULTS AND DISCUSSION

Because of 1,4–regularity for an *n* vertices graph we have $n = 3k + 2, k \in N_0$. So the first few members of our observed alkane family $\tau(n)$ have *n*=2,5,8,11,14,... First we will calculate how many possible MIDs are for the observed graphs. Let us look at the Path Identifier formula for vertices *i* and *j* in a molecule:

$$\text{PI06} = \prod_{k=2}^{n_{ij}} \sqrt{\frac{b_{k,k-1}}{k} \cdot \frac{1}{\delta'(k) \cdot \delta'(k-1)}} \tag{4}$$

It holds $Z = 4$, $b = 1$ and $\delta = 1$ or 4 for all such graphs. So we have:

$$\text{PI06} = \prod_{k=2}^{n_{ij}} \sqrt{\frac{1}{k \cdot Z}} \cdot \sqrt{\frac{1}{\delta(k) \cdot \delta(k-1)}} \tag{5}$$

First part of the formula is a constant, while the other part depends on the path. Graph is 1,4–regular, hence all paths are between 1–1, 1–4, or 4–4 vertices. Therefore, there are three different kinds of paths. This means that if we choose one vertex as first, in that graph there are $9\log_3 2\mathrm{n}$ paths at most, hence there are at most $9\log_3 2\mathrm{n}$ different Path Identifiers.

For an atom (vertex) *i* in a graph, we calculate AID by:

$$AID_i = \sum_j PI_{ji} \tag{6}$$

and then we get MID for the molecule with the formula:

$$MID = \sum_i AID_i^{\,2} \tag{7}$$

Now we have:

$$\mathrm{MID} = \sum_i (\sum_{\mathrm{j}} \mathrm{PI}_{\mathrm{ji}})^2 = \sum_i \sum_{j'} \sum_{j''} \mathrm{PI}_{\mathrm{j'i}} \mathrm{PI}_{\mathrm{j''i}} \tag{8}$$

For fixed *j′*, there are $\le 9\log_3 2\mathrm{n}$ different $PI_{j'i}$ s, and also the same $PI_{j''i}$ s for fixed *j″*, so in the MID formula there are $\le\ 81\cdot\log_3{}^2 2\mathrm{n}$ different summand. Altogether there are $n(n-1)^2 \le n^3$ summands (not necessarily different), because the sum by *j′* and *j″* has $n-1$ summands, and by *i* there are exactly *n* summands.

Weak decomposition of number N into R summands is sum $N = x_1 + \ldots + x_R$ where $x_1, x_2, \ldots, x_R \ge 0$. The number of such sums [10] is $\begin{pmatrix} N+R-1 \\ R-1 \end{pmatrix}$. Note that:

$$\begin{aligned} \begin{pmatrix} N+R-1 \\ R-1 \end{pmatrix} &= \frac{(N+R-1)\cdot(N+R-2)\cdot\ldots\cdot(N+1)}{(R-1)\cdot(R-2)\cdot\ldots\cdot 1} = \\ &= \left(1+\frac{N}{R-1}\right)\cdot\left(1+\frac{N}{R-2}\right)\ldots\left(1+\frac{N}{2}\right)\cdot(1+N) \le N^R \end{aligned} \tag{9}$$

Since, in our case $R \le 81\cdot\log_3{}^2 2\mathrm{n}$ and $N \le n^3$, it follows that there are at most $\left(n^3\right)^{81\cdot\log_3{}^2 2n}$ different MIDs. We have:

$$\left(n^3\right)^{81\cdot\log_3{}^2 2n} = n^{243\cdot\log_3{}^2 2n} = \left(2^{\log_2 n}\right)^{243\cdot\log_3{}^2 2n} \le 2^{400\cdot\log_3{}^3 2n} \tag{10}$$

Theorem 1. The number of trees in $\tau(n)$ is at least $2^{\frac{1}{2}\sqrt{n}}$.

Proof. We will construct a tree from family $\tau(n)$ with *n* vertices in the several steps. Denote by k the largest number such that $1+1+3+3^2+\ldots+3^{k-1} = \dfrac{3^k+1}{2} \le \dfrac{n}{4}$. Let us take one vertex and add to it 3 leaves.

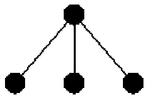

Figure 1. Basic block of the graph.

Let us denote that as basic block. To make this block into a tree of our family we have to add one more vertex to that first one and denote it as the root.

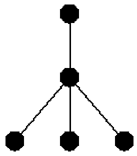

Figure 2. Basic block with its root.

Now let us construct a tree with height increased by 1. Let us take 3 basic blocks and identify their roots (Figure 3, left hand–side) and add one pendant vertex which becomes our new root (Figure 3, right hand–side).

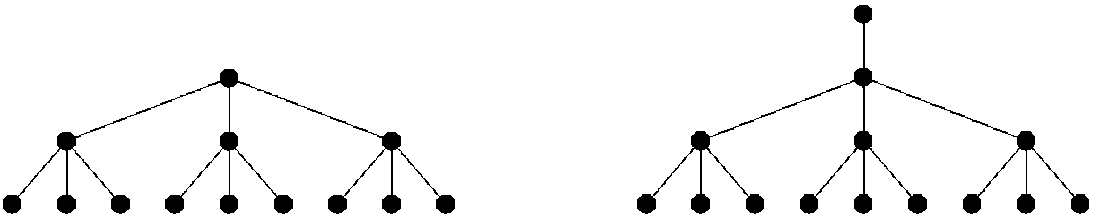

Figure 3. Building larger trees from basic blocks.

Now we continue to expand the tree, each time increasing the height by 1 and expanding all the vertices on the previous level. We obtain the tree of height $k+1$. In this case there are basic blocks on k levels. The example for $k=3$ is given on the following Figure:

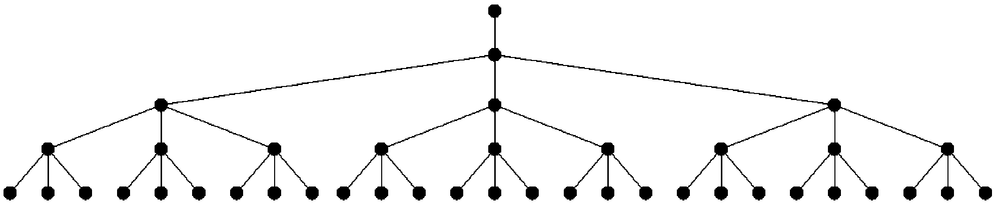

Figure 4. Branched tree with height 4.

Once we have built such tree with $\frac{3^k+1}{2}$ vertices, we are going to branch vertices on the bottom level. We can branch either 0, 1, 2 or 3 vertices. However, in order to avoid technical difficulties, we shall branch at least one vertex in each of the basic blocks on the bottom. So we have 3 possibilities for each basic block. We will represent those possibilities with strings of 0s and 1s. We assign 0 to the vertex that we haven't branched and 1 if we did. For a basic block strings have the length of 3. Since it is the same which vertex we branch, because their positions aren't fixed, strings 001 and 010 represent the same possibility. Hence, we have the following possible strings for a basic block:

$$\begin{gathered}100, 010, 001 \text{ – if we branched 1 vertex,}\\ 110, 011, 101 \text{ – if we branched 2 vertices,}\\ 111 \text{ – if we branched 3 vertices.}\end{gathered}$$

Note that we have enough remaining vertices for all these branchings, since at most $3\cdot 3^{k-1}\le 3\cdot\frac{n}{4}$ vertices can be added at this stage, and still we have some vertices left. We will add that "leftover" vertices on any vertex at bottom most level, branching uniformly. The way those vertices are arranged can only increase total number of different trees, so we can ignore it for the purpose of our calculation (and just say that the observed vertex is branched). It can be easily seen that diameter of this graph is $<3\log_3 2n$.

The tree we get with this construction looks like uniformly branched tree, and on the bottom most level some vertices are branched and some are not. So down to that level trees all look the same and we conclude that our tree of $\tau(n)$ family is uniquely described by the string assigned to the bottom most level of the tree. For a *k*–height tree that string has the length of 3^{k-1}. Hence, the number of different trees of given family is equal to the number of different strings described above. Let us denote with a_{p-1} number of different strings of length 3^{p-1} (taking into the account identification under isomorphism). It can be easily seen that $a_1=3$ (since 000 is not allowed) and we got a_2 as number of combinations with repetitions of arranging 3 elements into a_1 places. We now conclude that 3^{p-1}–length strings, that is the number a_{p-1} will be the number of combinations with repetitions [10] of a_{p-2}:

$$a_{p-1}=\left(\!\!\binom{a_{p-2}}{3}\!\!\right)=\binom{a_{p-2}+3-1}{3}=\binom{a_{p-2}+2}{3}=\frac{(a_{p-2}+2)(a_{p-2}+1)a_{p-2}}{3!}\ge\frac{\left(a_{p-2}\right)^3}{6} \qquad (11)$$

Now we have:

$$\begin{aligned}&a_1\ge 3\\ &a_2\ge\binom{a_1+2}{3}=10\\ &a_3\ge\frac{{a_2}^3}{6}\end{aligned} \qquad (12)$$

$$a_4 \ge \frac{a_3^{\;3}}{6} \ge \frac{\left(\frac{a_2^{\;3}}{6}\right)^3}{6} = \frac{a_2^{\left(3^2\right)}}{6^{3+1}} \ge \frac{a_2^{\left(3^2\right)}}{6^{\left(3^2\right)}}$$

$$a_5 \ge \frac{a_4^{\;3}}{6} \ge \frac{\left(\frac{a_2^{\left(3^2\right)}}{6^{3+1}}\right)^3}{6} = \frac{a_2^{\left(3^3\right)}}{6^{9+3+1}} \ge \frac{a_2^{\left(3^3\right)}}{6^{\left(3^3\right)}}$$

...

$$a_{k-1} \ge \frac{a_2^{\left(3^{k\text{-}3}\right)}}{6^{\left(3^{k\text{-}3}\right)}} = \left(\frac{10}{6}\right)^{3^{k-3}} .$$

Hence, $a_{k-1} \ge 2^{2^{k-1}}$ for k big enough. So, for the tree of height k we have $a_{k-1} \ge 2^{2^{k-1}}$, number of different strings, and hence number of different trees. Let us express that by number of vertices n.

Since k is the largest number such that $\frac{3^k+1}{2} \le \frac{n}{4}$, i.e. the largest number such that $3^k \le \frac{n}{2} - 1$, it follows that $k > \log_3 \frac{n-2}{6} > \log_3 \frac{n}{4} > \log_4 \frac{n}{4}$.

When we put that in our claim we get:

$$a_{k-1} \ge 2^{2^{k-1}} \ge 2^{2^{\log_4 \frac{n}{4}}} = 2^{2^{\frac{1}{2}\log_2 \frac{n}{4}}} = 2^{\left(\frac{n}{4}\right)^{\frac{1}{2}}} = 2^{\frac{1}{2}\sqrt{n}} \qquad (13)$$

So the number of trees in the family $\tau(n)$ is at least $2^{\frac{1}{2}\sqrt{n}}$, which is what we wanted to prove. ■

Theorem 2. Number of different MIDs is smaller than the number of alkane graphs they are associated with.

Proof. Let us look at the ratio of different MIDs and observed graphs when number of vertices n is increasing. We have:

$$\lim_{n\to\infty} \frac{2^{400\log_3^{\,3} 2n}}{2^{\frac{1}{2}\sqrt{n}}} = \lim_{n\to\infty} 2^{400\log_3^{\,3} 2n - \frac{1}{2}\sqrt{n}} \qquad (14)$$

This is equivalent to $2^{\lim\limits_{n\to\infty}\left(400\log_3^{\,3} 2n - \frac{1}{2}\sqrt{n}\right)} = 2^{\lim\limits_{n\to\infty}\sqrt{n}\cdot\left(\frac{400\log_3^{\,3} 2n}{\sqrt{n}} - \frac{1}{2}\right)} = 2^{\lim\limits_{n\to\infty}\sqrt{n}\cdot\lim\limits_{n\to\infty}\left(\frac{400\log_3^{\,3} 2n}{\sqrt{n}} - \frac{1}{2}\right)}$

And then:

$$\lim_{n\to\infty}\left(\frac{400\log_3^{\,3} 2n}{\sqrt{n}}\right) = 400\cdot\lim_{n\to\infty}\left(\frac{\log_3^{\,3} 2n}{n^{1/2}}\right) = 400\cdot\lim_{n\to\infty}\left(\frac{\log_3 2n}{n^{1/6}}\right)^3 = c\cdot\lim_{n\to\infty}\left(\frac{\ln 2n}{n^{1/6}}\right)^3 \qquad (15)$$

where $c=\frac{400}{(\ln 3)^3}$. Let us apply L'Hospital rule to $\lim\limits_{n\to\infty}\frac{\ln 2n}{n^{\frac{1}{6}}}$:

$$\lim_{n\to\infty}\frac{\ln 2n}{n^{\frac{1}{6}}}\overset{L'H}{=}\lim_{n\to\infty}\frac{\frac{1}{n}}{\frac{1}{6}n^{\frac{-5}{6}}}=\lim_{n\to\infty}\frac{6}{n^{1/6}}=0. \tag{16}$$

And since we have:

$$\lim_{n\to\infty}\sqrt{n}=\infty \text{ and } \lim_{n\to\infty}\frac{1}{2}=\frac{1}{2} \tag{17}$$

in the starting expression we get:

$$2^{\lim\limits_{n\to\infty}\sqrt{n}\cdot\lim\limits_{n\to\infty}\left(\frac{400\log_3^3 2n}{\sqrt{n}}-\frac{1}{2}\right)}=2^{\infty\left(0-\frac{1}{2}\right)}=2^{-\infty}=0 \tag{18}$$

Limit of the observed ratio goes into 0 when number of carbon atoms in the molecule increases. So for large enough *n* we have many more alkanes than MIDs and there will be two different alkanes with the same MID06. ■

Remark 1. From previous theorem we conclude that for almost every graph *G* of the given family, exists a graph *G'* so that MID06(G)=MID06(G').

4 CONCLUSIONS

In this paper we have analyzed MID06 index. Although it has been shown that it has perfect discriminative properties for the given family it can be shown that this cannot be extended to general graphs. Here we construct large family of graphs such that almost every member of this family has an indiscriminative counterpart. This shows that MID06 might not be useful in some cases for the identification of the compounds in the large data–bases, because there is always danger that in practice two different compounds with the same value of MID06 may occur. We show, that when we observe theoretical models, there is a very large number of such indiscriminative pairs.

However, this is still highly discriminative index and useful tool to show that two molecular graphs are not isomorphic (*i.e.*, that they do not model the same compound). Namely, if two compounds have different MID06 values, they are not isomorphic; and otherwise additional research is necessary.

Acknowledgment

Partial support of the Ministry of Science, Education and Sports of the Republic Croatia (Grants No. 037–0000000–2779 and No. 177–0000000–0884) is gratefully acknowledged.

5 REFERENCES

[1] NIST Chemistry WebBook **–** http://webbook.nist.gov/chemistry/**.**

[2] IUPAC – International Union of Pure and Applied Chemistry – http://www.iupac.org/.

[3] Wolfram MathWorld, "*NP–Problem*" – http://mathworld.wolfram.com/NP–Problem.html.

[4] B. D. McKay and R. G. Stanton, Some graph isomorphism computations, *Ars Comb*. **1980**, *9*, 307–313.

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

[6] D. Vukičević and A. T. Balaban, On the Degeneracy of Topological Index J, *Internet Electron*. *J*. *Mol*. *Des*. **2005**, *4*, 491–500.

[7] C–Y. Hu and L. Xu, Developing Molecular Identification Numbers by an All–Paths Method, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1997**, *37*, 311–315.

[8] D. L. Kreher and D. R. Stinson, *Combinatorial Algorihtms*, CRC Press, Boca Raton, 1999.

[9] D. Vukičević, A. Miličević, S. Nikolić, J. Sedlar, and N. Trinajstić, Paths and Walks in Acyclic Structures: Kenographs vs. Plerographs, *ARKIVOC*, **2005**, *10*, 33–44.

[10] P. J. Cameron, *Combinatorics: Topics*, *Techniques*, *Algorithms*, Cambridge University Press, Cambridge, 1994.

Biographies

Damir Vukičević is associate professor at Department of Mathematics, Faculty of Natural Sciences and Mathematics, University of Split. He has collaborated with more than 50 scientists including Nobel Prize laureate and four members of National Academies Science and Arts from three countries. He is winner of International Academy of Mathematical Chemistry Award for Young Scientists. His research interests include discrete mathematics (especially graph theory), mathematical chemistry, algorithms and their optimization and applications of computer science.

Tanja Vojković is junior research assistant at Department of Mathematics, Faculty of Natural Sciences and Mathematics, University of Split. She got her BSc in Mathematics and Computer Science, 2008, University of Split, and since then is working on a project “Discrete Mathematical Models in Chemistry.” Her research interests include discrete mathematics and mathematical chemistry.