MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **69** (2013) 785-794

ISSN 0340 - 6253

The Discrimination Power of Molecular Identification Numbers Revisited

Matthias Dehmer, Martin Grabner

Institute for Bioinformatics and Translational Research, UMIT, Eduard Wallnoefer Zentrum 1, 6060 Hall in Tyrol, Austria

matthias.dehmer@umit.at, martin.grabner@umit.at

(Received June 27, 2012)

Abstract

In this paper, we explore the discrimination power of existing molecular ID numbers as the potential of these measures has not yet been investigated on a large scale. First, we find that many ID numbers are computationally insufficient and, hence, can not be calculated on large sets of graphs. Second, we also determine the discrimination power of recently developed eigenvalue-based indices which possess polynomial time complexity. Particularly we find that some of these measures outperform specific molecular ID numbers in terms of their discrimination power.

1 Introduction

Network-based methods have been proven useful in various disciplines. Hence, there is a strong need to understand the mathematical apparatus in-depth. A standard method to characterize the structure of complex networks are structural graph descriptors [13, 10, 11, 26]. A challenging problem when designing such structural descriptors for networks is to convert structural information into numbers or hashes uniquely [2, 8, 15, 14, 19, 21]. Also, this relates to investigate the discrimination power or uniqueness of a structural graph measure by using specific sets of graphs, see [3, 4, 8, 12, 14]. In the eighties, Randić and Balaban made an attempt in this direction by introducing several molecular identification numbers representing topological indices [17, 19, 21, 20]. Note that the discrimination power of those ID numbers has been first explored by Szymanski et al. [17]. Apart from investigating the uniqueness of these measures, applications thereof in QSAR and drug design have also been reported, see [5, 6].

To define the Randić molecular ID, he summed certain property functions involving vertex degrees of all edges along all possible paths in a graph [19]. This measure turned out to be highly discriminating for alkane trees. Afterwards, Szymanski et al. [18] found counterexamples that the Randić molecular ID does not represent a complete graph invariant for alkane trees, i.e., it is not fully unique on this graph class. Further, the Balaban ID has been defined by using distance sums rather than vertex degrees in the path weighting scheme [2]. However, the discrimination power of this measure has only been little investigated [2].

When investigating the computational complexity of these ID numbers, it turned out that these measures are not suitable to characterize either large graphs or large sets of graphs efficiently as the number of all paths in a graph is approximately $n!$. In particular, the time complexity of the ID's increases tremendously with cyclic graphs. In order to reduce the time complexity, Ivanciuc et al. [24] developed similar quantities by summing edge properties along the shortest paths (not all paths). Also, Szymanski et al. [18] defined new quantities by considering weighted walks rather than paths and improved the time complexity of the resulting quantities. However, we prove in Section 3 that these measures are less unique than the ID's by using chemical and exhaustively generated graphs. Finally, Randić [21] explored so-called ring ID's which assign numerical values to ring structures with high discrimination power. Hence, a large cyclic molecular structure could be decomposed into several ring-like subgraphs which can be characterized uniquely by using the ring ID's. But note that the ring ID's do not allow deriving scores for characterizing cyclic graphs globally. Related indices namely the so-called prime ID's also developed by Randić [19] turned out to be highly unique for chemical alkane trees.

The contributions of this paper is twofold: First, we compare the uniqueness of the molecular ID numbers on a large scale by using chemical alkane trees and exhaustively generated graphs. Here, exhaustively generated graphs are graphs without any structural constraints, e.g., bounded degrees, hierarchy etc. We find that the molecular ID numbers are not suitable to tackle this problem for exhaustively generated graphs due to their insufficient computational complexity. Second, we also evaluate the uniqueness of recently developed eigenvalue-based entropies [9] and prove that these indices outperform some of the ID numbers. As a strong point, the eigenvalue-based entropies due to Dehmer et al. [9] have a much better time complexity than the ID numbers. Note that in this paper, we

will not develop any new graph measures. Instead, the main contribution is to perform a numerical study to demonstrate that classical molecular ID numbers are not sufficient to evaluate the uniqueness compared to recently developed eigenvalue-based indices [9].

2 Materials and Methods

In this section, we briefly recall the definitions of the ID numbers we are going to apply in this paper. Also, we state the eigenvalue-based entropies due to Dehmer et al. [9].

2.1 Molecular ID Numbers

Let $G = (V, E)$ be a connected graph, $|V| := N$. We start by stating the general formula of the Randić and Balaban ID number [2, 19], namely

$$ID_{C,B}(G) := N + \sum_{m_{p_{ij}}} w_{ij}. \tag{1}$$

$m_{p_{ij}}$ are all paths of length $m > 0$ and w_{ij} is a distinctively defined path weight. In case of ID_C, we define

$$w_{ij} = \prod_{b=1}^{m} (k_{b(1)} k_{b(2)})_b^{-1/2}. \tag{2}$$

Note that the sum runs over all edges in the graph and $k_{b(1)}$, $k_{b(2)}$ are the vertex degrees of the two vertices incident to the b-th edge [26].

In order to define ID_B, the path weight has been defined by [2, 26]

$$w_{ij} = \prod_{b=1}^{m} (\sigma_{b(1)} \cdot \sigma_{b(2)})_b^{-1/2}, \tag{3}$$

where σ_k is the vertex distance degree [2, 26] and $b(1)$, $b(2)$ are the vertices adjacent to the edge b.

Instead of summing the path weights over all paths $m_{p_{ij}}$ between two vertices v_i and v_j, Ivanciuc and Balaban [24] developed related measures

$$ID^{min}(G) := N + \sum_{min_{p_{ij}}} w_{ij}, \tag{4}$$

by only considering the shortest paths $min_{p_{ij}}$. By using the path weights w_{ij} represented by Equation 2, 3, the corresponding formulas for ID_C^{min} and ID_B^{min} follow straightforwardly.

Finally, Szymanski et al. [18] further developed this general approach and defined the so-called Weighted ID number by [18, 26]:

$$WID(G) := N - \frac{1}{N} + \frac{ID^*}{N^2}, \tag{5}$$

where

$$ID^* := \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}^*, \tag{6}$$

$$W^* := \sum_{k=0}^{N-1} {}^{\sigma}\chi^k. \tag{7}$$

${}^{\sigma}\chi^k$ stands for the k-th power of distance-sum-connectivity matrix [18, 26].

2.2 Eigenvalue-based Entropies

Recently, a family of eigenvalue-based entropies H_{M_s} has been proposed by Dehmer et al. [9]:

$$H_M^s(G) := \sum_{i=1}^{k} \frac{|\lambda_i|^{\frac{1}{s}}}{\sum_{j=1}^{k} |\lambda_j|^{\frac{1}{s}}} \log \frac{|\lambda_i|^{\frac{1}{s}}}{\sum_{j=1}^{k} |\lambda_j|^{\frac{1}{s}}}. \tag{8}$$

$\lambda_1, \lambda_2, \ldots, \lambda_k$ are the non-zero eigenvalues of a molecular matrix M [9].

Index	Symbol
Randić Connectivity ID Number [19]	ID_C
min Randić Connectivity ID Number [24]	ID_C^{min}
Balaban ID Number [2]	ID_B
min Balaban ID Number [24]	ID_B^{min}
Weighted ID Number [18]	WID
H_M^1 by using the distance path matrix [9, 26]	H_{DP}
H_M^1 by using the extended adjacency matrix [9, 26]	H_{EA}
H_M^1 by using the augmented vertex degree matrix [9, 26]	H_{AV}
H_M^1 by using the first weighted structure function matrix [9, 26]	H_{IM_1}
H_M^1 by using the second weighted structure function matrix [9, 26]	H_{IM_2}

Table 1. The descriptors and their symbols.

See Table 1 for the explanation of the used symbols H_{DP}, H_{EA}, H_{AV}, H_{IM_1} and H_{IM_2}. It is known that simple algorithms to calculate the eigenvalues of a $n \times n$ matrix require cubic time complexity, see [7]. Hence, the time complexity for this family of eigenvalue-based entropies (Equation 8) is $O(n^3)$.

2.3 Datasets and Software

The datasets C_i, $19 \leq i \leq 22$, are alkane trees with i vertices (carbon atoms) generated by using `Molgen` [1].

Finally, the sets N_i, $9 \leq i \leq 10$, are exhaustively generated non-isomorphic unlabeled graphs having i vertices each generated by using the `Nauty` package [22].

To calculate the descriptors shown in Section 2, we used the R package `QuACN` [23] which is public available via the CRAN-archive. `QuACN` contains already over 150 quantitative network measures which have been used in disciplines such as social network analysis, mathematical chemistry and network physics.

3 Results

We start interpreting the numerical results with Table 2.

	C_{19}		C_{20}		C_{21}		C_{22}	
Index	ndv	S	ndv	S	ndv	S	ndv	S
ID_C	46	0,999690	174	0,999525	454	0,999501	1327	0,999418
ID_B	0	1,000000	4	0,999989	4	0,999996	132	0,999942
ID_C^{min}	46	0,999690	176	0,999520	454	0,999501	1327	0,999418
ID_B^{min}	0	1,000000	4	0,999989	4	0,999996	132	0,999942
WID	4	0,99997	144	0,999607	308	0,999662	2674	0,998827
H^1_{DP}	0	1,000000	0	1,000000	0	1,000000	0	1,000000
H^1_{EA}	0	1,000000	0	1,000000	2	0,999998	0	1,000000
H^1_{AV}	48	0,999676	98	0,999732	82	0,999910	236	0,999896
$H^1_{IM_1}$	0	1,000000	0	1,000000	0	1,000000	0	1,000000
$H^1_{IM_2}$	0	1,000000	0	1,000000	0	1,000000	0	1,000000

Table 2. Chemical alkane trees with $|V| = 19, 20, 21, 22$. $|C_{19}| = 148284$, $|C_{20}| = 366319$, $|C_{21}| = 910726$, $|C_{22}| = 2278658$.

In general, ndv are the number of non-distinguishable values and S is the well-known sensitivity measure due to Konstantinova [14]. First, we see that the uniqueness of several indices such as ID_C, ID_C^{min}, ID_B, ID_B^{min} and WID slightly decreases with an increasing size of the underlying set of graphs. This is in accordance with an earlier finding [8] that many topological graph measures induce a dependency between their uniqueness and the size of the graph set in question. But in any way, the uniqueness of all measures evaluated on alkane trees is very high. Interestingly, all eigenvalue-based entropies, except H^1_{AV}, possess a better discrimination power than the molecular ID numbers for the alkane trees. This proves that these measures capture structural information significantly. In

particular, H^1_{DP}, H^1_{EA}, $H^1_{IM_1}$, $H^1_{IM_2}$ are fully unique for all sets of chemical alkane trees. Based on the fact that the entropies have cubic time complexity (in n), the calculation of the uniqueness is much more efficient than by using the ID numbers. Particularly ID_C and ID_B turned out computationally insufficient for all large sets of graphs we have used in this study.

Table 3, 4 show the results by using exhaustively generated graphs without any structural constraints. Note that these graphs contain cycles. We see that ID_C and ID_B are fully unique for $N_6, \ldots, N_9$. Due to the insufficient time complexity of these measures, it turned out to be impossible to generate all values for N_{10} (see Table 4).

	N_6		N_7		N_8	
Index	ndv	S	ndv	S	ndv	S
ID_C	0	1,000000	0	1,000000	0	1,000000
ID_B	0	1,000000	0	1,000000	0	1,000000
ID_C^{min}	0	1,000000	4	0,995311	263	0,976343
ID_B^{min}	0	1,000000	0	1,000000	240	0,978411
WID	4	0,964286	24	0,971864	284	0,974454
H^1_{DP}	0	1,000000	20	0,976553	512	0,953944
H^1_{EA}	2	0,982143	8	0,990621	46	0,995862
H^1_{AV}	0	1,000000	0	1,000000	0	1,000000
$H^1_{IM_1}$	5	0,955357	16	0,981243	140	0,987407
$H^1_{IM_2}$	5	0,955357	10	0,988277	99	0,991095

Table 3. Exhaustively generated sets of non-isomorphic and generated graphs. $|N_6| = 112$, $|N_7| = 853$ and $|N_8| = 11117$.

	N_9		N_{10}	
Index	ndv	S	ndv	S
ID_C	0	1,000000		
ID_B	0	1,000000		
ID_C^{min}	19842	0,924000	1912752	0,836748
ID_B^{min}	18341	0,929750	1782776	0,847841
WID	3343	0,987195	73073	0,993763
H^1_{DP}	19982	0,923464	1141560	0,902569
H^1_{EA}	479	0,998165	16394	0,998601
H^1_{AV}	0	1,000000	6940	0,999408
$H^1_{IM_1}$	4402	0,983139	350726	0,970066
$H^1_{IM_2}$	3693	0,985855	302916	0,974146

Table 4. Exhaustively generated sets of non-isomorphic and generated graphs. $|N_9| = 261080$ and $|N_{10}| = 11716571$.

The second best measure is H_{AV}^1 (see Table 3, 4) whose computational complexity is only cubic. Also, this measure does not have a strong dependency between its uniqueness and the size of the graph set. This makes it a useful index for discriminating graphs on a large scale.

All other eigenvalue-based indices possess lower discrimination power than H_{AV}^1 but they are able to distinguish $\geq 97\%$ of the graphs of N_{10}, except H_{DP}^1. We emphasize that N_{10} contains almost 12 million graphs. An alternative to ID_C and ID_B for discriminating graphs on a large scale (e.g., see N_{10}) is WID as it can distinguish $\geq 99\%$ of the graphs of N_{10} and has a better computational complexity. This can be understood by the fact that Szymanski et al. [18] used only shortest paths and not all paths involved.

Further, we observe that the discrimination power of ID_C^{min} and ID_B^{min} is worse than by calculating all other indices based on the set N_{10}. Again, this proves the intuitive conjecture that determining the shortest paths instead of all paths may lead to an aggravation of the discrimination power. Because of their still high computational complexity, they turned out to be not feasible for discriminating graphs uniquely among all used molecular ID numbers.

4 Discussion

In this paper, we determined the discrimination power of existing molecular ID numbers on a large scale. Not surprisingly, most of the molecular ID numbers, particularly the one due to Randić and Balaban ID [2, 19] possess insufficient time complexity. As a result, the discrimination power of these measures by using N_{10} could not be determined. Interestingly, we found that recently developed eigenvalue-based entropies clearly outperform some ID numbers in terms of their uniqueness by using exhaustively generated graphs and chemical alkane trees. Particularly for exhaustively generated alkane trees, these indices did not produce any degeneracies. Also, all ID numbers turned out to be less unique than by using the eigenvalue-based entropies.

Again, this shows that the discrimination power of a structural graph measure strongly depends the underlying graph class, see [8]. Crucially, we see that the potential of many existing measures has not yet been investigated properly for tackling this problem. In particular, it is surprising that even eigenvalue-based measures outperform the ID numbers in terms of their uniqueness as many examples can be found in the literature, where

eigenvalue-based quantities failed to characterize graphs structurally, see [16, 25]. As a conclusive remark, this calls for exploring existing measures more deeply on a large scale rather than developing new measures.

5 Acknowledgements

Matthias Dehmer and Martin Grabner thank the Austrian Science Funds and the Standortagentur Tirol for supporting this work (project P22029-N13). We also thank the 'Zentraler Informatikdienst' of the Technical University of Vienna for providing computing resources to perform large scale computations on the Phoenix Cluster.

References

[1] Molgen isomer generator software, www.molgen.de (2000), Institute of Mathematics II, Univ. Bayreuth, Germany.

[2] A. T. Balaban, Numerical modelling of chemical structures: Local graph invariants and topological indices, in: R. King, D. Rouvray (Eds.), *Graph Theory and Topology in Chemistry*, Elsevier, Amsterdam, 1987, pp. 159–176.

[3] D. Bonchev, O. Mekenyan, N. Trinajstić, Isomer discrimination by topological information approach, *J. Comp. Chem.* **2** (1981) 127–148.

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

[5] S. Carter, N. Trinajstić, S. Nikolić, A note on the use of ID numbers in QSAR studies, *Acta Pharm. Jugosl.* **37** (1987) 37–42.

[6] S. Carter, N. Trinajstić, S. Nikolić, On the use of ID numbers in drug research: A QSAR of neuroleptic pharmacophores, *Med. Sci. Res.* **16** (1988) 185–186.

[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, *Introduction to Algorithms*, MIT Press, Cambridge, 1990.

[8] M. Dehmer, M. Grabner, K. Varmuza, Information indices with high discrimination power for arbitrary graphs, *PLoS ONE* **7** (2012) e31214.

[9] M. Dehmer, L. Sivakumar, K. Varmuza, Uniquely discriminating molecular structures using novel eigenvalue-based descriptors, *MATCH Commun. Math. Comput. Chem.* **67** (2012) 147–172.

[10] M. Dehmer, K. Varmuza, S. Borgert, F. Emmert-Streib, On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures, *J. Chem. Inf. Model.* **49** (2009) 1655–1663.

[11] M. V. Diudea, I. Gutman, L. Jäntschi, *Molecular Topology*, Nova, New York, 2001.

[12] M. V. Diudea, A. Ilić, K. Varmuza, M. Dehmer, Network analysis using a novel highly discriminating topological index, *Complexity* **16** (2011) 32–39.

[13] F. Emmert-Streib, M. Dehmer, Information theoretic measures of UHG graphs with low computational complexity, *Appl. Math. Comput.* **190** (2007) 1783–1794.

[14] E. V. Konstantinova, The discrimination ability of some topological and information distance indices for graphs of unbranched hexagonal systems, *J. Chem. Inf. Comput. Sci.* **36** (1996) 54–57.

[15] W. D. Ihlenfeldt, J. Gasteiger, Hash codes for the indentification and classification of molecular structure elements, *J. Comput. Chem.* **15** (1994) 793–813.

[16] O. Ivanciuc, T. Ivanciuc, M. V. Diudea, Polynomials and spectra of molecular graphs, *Roman. Chem. Quart. Rev.* **7** (1999) 41–67.

[17] K. Szymanski, W. Müller, J. Knop, N. Trinajstić, On Randić's molecular identification numbers, *Int. J. Quantum Chem.* **25** (1985) 413–415.

[18] K. Szymanski, W. Müller, J. Knop, N. Trinajstić, On the identification numbers for chemical structures, *J. Chem. Inf. Comput. Sci.* **30** (1986) 173–183.

[19] M. Randić, On molecular identification numbers, *J. Chem. Inf. Comput. Sci.* **24** (1984) 164–175.

[20] M. Randić, Molecular ID numbers: By design, *J. Chem. Inf. Comput. Sci.* **26** (1986) 134–136.

[21] M. Randić, Ring ID numbers, *J. Chem. Inf. Comput. Sci.* **28** (1988) 142–147.

[22] B. D. McKay, Nauty, http://cs.anu.edu.au/∼bdm/nauty/ (2010).

[23] L. A. J. Müller, K. G. Kugler, A. Dander, A. Graber, M. Dehmer, QuACN – An R package for analyzing complex biological networks quantitatively, *Bioinformatics* **27** (2011) 140–141.

[24] O. Ivanciuc, A. Balaban, Design of topological indices. Part 3. New identification numbers for chemical structures: MINID and MINSID, *Croat. Chem. Acta.* **69** (1996) 9–16.

[25] M. Randić, M. Vračko, M. Novič, Eigenvalues as molecular descriptors, in: M. V. Diudea (Ed.), *QSPR/QSAR Studies by Molecular Descriptors*, Nova, New York, 2001, pp. 93–120.

[26] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2002.