West Lafayette, IN, 1985; Abstr. 62.
(14) Speiser, B. "Multiparameter Estimation: Extraction of Information from Cyclic Voltammograms". *Anal. Chem.* **1985**, *57*, 1390-1397.
(15) Borman, S. A. "New Electroanalytical Pulse Techniques". *Anal. Chem.* **1982**, *54*, 698A-705A.
(16) Eklund, J. A.; Faulkner, L. R. "Pursuing Major Conclusions: The Intelligent Director of an Electrochemical Repertoire". *1985 Electroanalytical Symposium*; BAS Press: West Lafayette, IN, 1985; Abstr. 64.
(17) Georgeff, M. P. "Strategies in Heuristic Search". *Artificial Intelligence* **1983**, *20*, 313-425.
(18) Barr, A.; Feigenbaum, E. *Handbook of Artificial Intelligence*; William Kaufman: Los Altos, CA, 1982; Vol. 1.
(19) H. Gunasingham, K. P. Ang, and C. C. Ngo, in preparation.
(20) Nau, D. S. "Expert Computer Systems". *Computer* **1983**, *15*, 63-84.
(21) Fuchi, K. "The Direction the FGCS Project Will Take". *New Generation Comput.* **1983**, *1*, 3-9.
(22) Gunasingham, H.; Srinivasan, B.; Ananda, A. L. "Design of an Expert System for Planning HPLC Separations". *Anal. Chim. Acta*, in press.
(23) Dahl, V. "Logic Programming as a Representation of Knowledge". *Computer* **1983**, *16*, 106-111.
(24) Rich, E. *Artificial Intelligence*; McGraw Hill: New York, 1983.
(25) Adeloju, S. B.; Bond, A. M.; Briggs, M. H. "Multielement Determination in Biological Materials by Differential Pulse Voltammetry". *Anal. Chem.* **1985**, *57*, 1386-1390.
(26) A. L. Ananda and H. Gunasingham, in preparation.
(27) Bard, A. J.; Faulkner, L. R. *Electrochemical Methods*; Wiley: New York, 1980.

Molecular ID Numbers: By Design†

MILAN RANDIĆ*

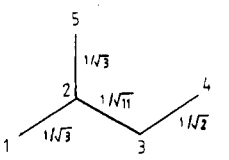
Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311, and Ames Laboratory, Iowa State University, Ames, Iowa 50011

Received November 27, 1985

The paper improves the discrimination ability of ID (identification) numbers by making use of a new set of weights for bonds, based on prime numbers.

Recently, I proposed a novel structural index—called molecular ID number—as a potentially useful label for molecular skeletons.[1] The index was a result of developing characterizations of complex molecules with weighted paths.[2] It proved valuable in a nonempirical approach to structure–activity. For example, use of the index has made it possible to classify a dozen anticholinergic compounds among some 40 therapeutically useful drugs exhibiting other activities.[3] When ID values were derived for all alkanes up to, and including, undecanes (in all, over 300 acyclic structures), it was observed that in no case did a duplicate numerical value occur. This suggested that this parameter, which is the total number of all suitably weighted paths (vide infra), is a *potential* structure discriminator. One can contrast this index to other graph theoretical (frequently referred to somewhat imprecisely as topological) indices of limiting discriminatory power: Hosoya's Z index,[4] the connectivity index,[5] and Balaban's J index,[6] which have been recently examined in a comparative study.[7] The uniqueness of ID was neither claimed nor was easy to prove (as is true for many other schemes related to graph isomorphism problem). Usually, the lack of uniqueness can be established by finding a counterexample. The report on molecular ID numbers[1] ended with an invitation to search for a counterexamples. A systematic search can provide insight into the number of counterexamples over a certain field size, an important information in evaluating practial use of such discriminators. While in mathematics interest in a conjecture immediately collapses as a single counterexample is found, in *applied* fields this is not necessarily the case. For example, many cluster analyses frequently do not suggest a single compound as candidate.[8] Thus, despite the unresolved issue of uniqueness, molecular ID numbers remain of considerable practical potential. Figueras[9] upgraded the existing ALL PATH program, which enumerates paths of different length in a graph, written in BASIC[10] by offering a much faster (turbo) Pascal version and at the same time extended the considerations to incorporate heteroatoms as distinctive items. It appeared that the ID numbers—arrived at by accident—have desirable features and promise in differentiating a large body of chemical compounds.

Table I. Bond Weights Based on the First Nine Prime Numbers and Enumeration of All Paths in 2-Methylbutane Using Prime Number Weights $1/[P_i]^{1/2}$ [a]



bond type	weight
(1, 2)	$1/2^{1/2}$
(1, 3)	$1/3^{1/2}$
(1, 4)	$1/5^{1/2}$
(2, 2)	$1/7^{1/2}$
(2, 3)	$1/11^{1/2}$
(2, 4)	$1/13^{1/2}$
(3, 3)	$1/17^{1/2}$
(3, 4)	$1/19^{1/2}$
(4, 4)	$1/23^{1/2}$

atom	p_1	p_2	p_3
1, 5	$1/3^{1/2}$	$1/(3^{1/2}\cdot3^{1/2}) + 1/(3^{1/2}\cdot11^{1/2})$	$1/(3^{1/2}\cdot11^{1/2}\cdot2^{1/2})$
2	$2(1/3^{1/2}) + 1/11^{1/2}$	$1/(11^{1/2}\cdot2^{1/2})$	
3	$1/2^{1/2} + 1/11^{1/2}$	$2[1/(11^{1/2}\cdot3^{1/2})]$	
4	$1/2^{1/2}$	$1/(2^{1/2}\cdot11^{1/2})$	$2[1/(2^{1/2}\cdot11^{1/2}\cdot3^{1/2})]$

$P_1 = 1/2^{1/2} + 2/3^{1/2} + 1/11^{1/2} = 2.163\,318$
$P_2 = 1/(2^{1/2}\cdot11^{1/2}) + 2/(3^{1/2}\cdot11^{1/2}) + 1/3^{1/2} = 0.894\,689$
$P_3 = 2/(2^{1/2}\cdot3^{1/2}\cdot11^{1/2}) = 0.246\,182$
$ID = P_1 + P_2 + P_3 + N = 8.304\,191$

[a] Bond type case (1, 1) is unimportant because it concerns only ethane.

However, Szymanski, Müller, Knop, and Trinajstić[11] undertook to systematically examine all acyclic structures beyond undecanes up to $n = 20$ carbon atoms using their very efficient program of generating trees.[12] In the field of 618 050 (i.e., well over half a million) structures, they found 124 pairs of nonisomorphic alkanes and a triple having exactly the same ID number (some of these are shown in Figure 1). The smallest pair has 15 vertices, two graphs among 13 476 different possible alkanes with 15 carbon atoms![13] There is one pair among alkanes having 16 carbon atoms (i.e., among 10 359); there are four pairs of alkanes having 17 carbon atoms (i.e., among 60 529); etc. All the computations were performed

† Ames Laboratory is operated for the U.S. Department of Energy by Iowa State University under Contract W-7405-Eng-82. This work was supported by the office of R. S. Hansen, Director.

* Address correspondence to the author at Drake University.

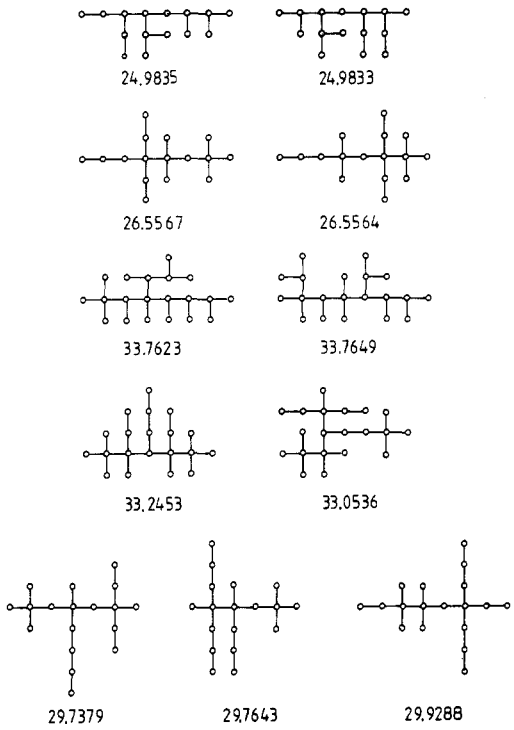


Figure 1. A selection of alkane trees with equal connectivity ID indices on each line (15, 16, 20, 20, and 18-vertex graphs, respectively; cf. footnote 11) found to have different prime ID indices (shown under each structure).

by double-precision arithmetic as well as by use of only integer arithmetic (vide infra) in order to avoid accidental overlap in a large number of integer digits. The verdict is clear: the molecular ID numbers as developed (accidentally) are not unique.

In Table I we illustrate the origin of molecular ID numbers on a simple skeleton of 2-methylbutane. Individual bonds are first classified in (m, n) bond types, where m and n are the number of neighbors for end atoms of any bond. Each bond is given weight $1/(mn)^{1/2}$, and each path is taken into account, but only once as seen in Table I. Paths of longer length are weighted by the product of factors of the bonds involved. The sum of all paths is the molecular ID number. Because in the earlier ID numbers only $1/2^{1/2}$, $1/3^{1/2}$, and $1/6^{1/2}$ can appear as irrational coefficients, one can perform the enumeration *exactly* over the field of three irrational numbers. One can therefore write the result of the count of weighted paths as: $a + b/2^{1/2} + c/3^{1/2} + d/6^{1/2}$. This form immediately suggests that coincidental ID numbers may arise due to *restricted* selection of weights for the *nine* (m, n) bond types of interest, which leads to only *three* irrational weight coefficients. Hence, we here propose new weights for the nine (m, n) bond types, so that no collapse of contributions from different bond types can result. A way to achieve this is to select the first nine *prime numbers* as the basis for weights (as shown in Table I). Now the molecular ID numbers will be expressed as: $a + b/2^{1/2} + c/3^{1/2} + d/5^{1/2} + e/7^{1/2} + f/11^{1/2} + g/13^{1/2} + h/17^{1/2} + i/19^{1/2} + j/23^{1/2}$ + all possible product terms. The number of distinct terms now increased from 4 (based on two prime numbers) to 2^9 or 512 (based on nine prime numbers).

We tested these new ID numbers—arrived at *by design* rather than by accident—on the graphs of Figure 1 and found that all previously not distinguished graphs are now differentiated. Moreover, we examined all graphs having $n = 18$ vertices and less among the counterexamples of Szymanski et al. In all cases we found different "prime" ID values. Because graphs having already different "connectivity" ID values will necessarily have different new "prime" ID, it follows that in the field of over 100 000 acyclic structures there is not a single duplicate ID number. This is not so surprising, because the graphs found previously to have identical (connectivity) ID numbers have different path counts,[14] sometimes even different bond types. For example, one graph of the triple of nonisomorphic graphs has no (4, 4) bond type, necessarily; hence, its ID will be missing the contribution from the $1/23^{1/2}$ term and cannot equal others. All this indicates that the present version of ID numbers, referred to as "prime ID numbers", in contrast to earlier "connectivity ID numbers" (in view of the nature of bond weights based on prime numbers or the connectivity weights, respectively), may *by virtue of design*, be less prone to occurrence of duplicates. We neither claim nor conjecture as to the uniqueness of the newly proposed ID numbers. On the contrary, we invite further searching for potential counterexamples. Clearly, however, the pool of alkanes for which this eminently practical approach will also be mathematically impeccable may be further dramatically increased.

CONCLUDING REMARKS

The pattern of design, i.e., use of prime numbers in choosing the calculation method as compared with accidental choice techniques, is a standard mathematical procedure in design of codes. It ensures, as long as all possible products of nine numbers are different, that the resulting sum will be unique. Choice of the first nine prime numbers ensures the above condition, but any nine relative primes could have been selected. The present area of application of prime ID numbers is somewhat limited by the fact that the index calculation is carried out essentially on alkane "black" graphs (nonintervention of multiplicity of bonds and vague introduction of heteroatoms). Because the concept of ID numbers is new and is evolving, it ought to be tested first on alkane graphs, to find out if its performance is satisfactory there. This paper appears encouraging in that respect and further developments may follow. Finally, the question can be raised as to the advantages of this new ID number in studying properties (structure–activity correltions). A likely answer here is to be negative. Graph theoretical (topological) indices have two distinct roles that are not necessarily compatible: (1) to provide code for structure–property studies and (2) to provide code for chemical information needs. The emphasis in the two roles is different: Molecules may have (and many have) very similar selected properties, and it is appropriate to require that "similar" molecules have "similar" indices, not barring the occurrence of a same index value when differences for a particular pair of structures are small. For chemical documentation, different molecules ought to have different codes, and there is no requirement that similar molecules should have similar codes, although if that happens it is a bonus.

In conclusion, this calculation of ID values, which are of graph theoretical (topological) origin, is interesting since the code arrived at is not too unwiedly, computationally and conceptually.

ACKNOWLEDGMENT

Correspondence with J. Figueras concerning the Pascal version of the ALL PATH program and its extension to heteroatms and with J. V. Knop and N. Trinajstić, who sent the list of all trees having the same connectivity ID numbers, is appreciated. Thanks are expressed also to J. E. Dubois and A. T. Balaban for helpful comments that improved the presentation of the material.

REFERENCES AND NOTES

(1) Randić, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164.
(2) Randić, M. In *Molecular Basis of Cancer*; Rein, R., Ed.; Liss: New York, 1985.
(3) Randić, M. *Int. J. Quant. Chem., Quant. Biol. Symp.* **1984**, *11*, 137.

(4) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(5) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(6) Balaban, A. T. *Theor. Chim. Acta* **1979**, *53*, 355.
(7) Razinger, M.; Chretien, J. R.; Dubois, J. E. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23.
(8) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(9) Program runs on Apple IIe home computer. Revised version involving prime number weights is available upon request to noncommercial users.
(10) Figueras, J., a preprint, private communication.
(11) Szymanski, K., Müller, W.R., Knop, J. V., Trinajstić, N., a preprint, private communication.
(12) Knop, J. V.; Müller, W. R. Jericević, Z.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1983**, *21*, 91.
(13) Triansjtić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vol. II, Table 4, 153.
(14) Randić, M., submitted for publication in *Croat. Chem. Acta*.

Compact Molecular Codes†

MILAN RANDIĆ*

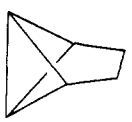
Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311, and Ames Laboratory, Iowa State University, Ames, Iowa 50011

Received May 22, 1985

In this paper we introduce structural codes that are easy to derive for most molecular skeletal forms of chemical interest yet satisfy numerous desirable properties, being linear, unique, reconstructable, derivable and decodable by hand, brief, based on familiar symbols, easily comprehensible, and efficient. Codes in general imply a resolution of the following problems: (1) canonical numbering of atoms; (2) graph isomorphism; (3) discernment of the symmetry of the structure (graph). Our approach resolves these problems in a remarkably simple way, at least for the examples selected. The approach is based on an extension of the *N*-tuple codes of Knop and co-workers, which apply only to trees (acyclic graphs). By excising selected vertices in a polycyclic graph, one arrives at subspanning trees for the polycyclic graph for which *N*-tuple codes of Knop et al. are adopted. Subsequently, such an incomplete code is augmented by the listing of adjacencies for the vertices, which represent ring closures. This paper presents numerous illustrations of the compact codes and discusses the rules that govern construction of the compact codes and the relative ease of the search for the codes. In order to more clearly show the relative simplicity of the new codes, we end with a comparison of the compact codes with a selection of alternative codes currently in use.

INTRODUCTION

The history of chemical nomenclature and the search for codes with desirable qualities is old and continuing. As early as 1881, Friedrich Konrad Beilstein[1] initiated a nomenclature system that is still of interest and serves as a basis for the naming of numerous structures. In 1900, Adolph von Bäyer[2] suggested the nomenclature for bridged bicyclic molecules, which is still the basis for the systematic naming of compounds like norbornane etc. Already, the extension of the nomenclature to tricyclic systems pointed to some difficulties. Besides *digits* used to indicate the number of carbon atoms in individual bridges, one needs *labels* to indicate the particular bridges in polycyclic structures. For example



is named (by IUPAC rules) tricyclo[3.1.0.0^{2,6}]hexane. Observe two *kinds* of uses of digits: 3.1.0.0 indicates *structural* data, the number of carbon atoms in the four branches of the structure, and 2,6 is a *label* referring to selected carbon atoms.

Much progress followed the early interest in chemical nomenclature. Coding is important not only for chemical documentation but also for enumeration of isomers and the construction of graphs. Finally, structural codes are of interest in structure–property and structure–activity studies.[3] Recently, Read[4] reviewed desirable qualities for codes for chemical structures. These are listed in Table I. Various codes proposed in the past satisfy to some degree several of the suggested desirable features, but no code has been found that would satisfy all the requirements satisfactorily. Not all the requirements are, however, equally important, nor can they be resolved with similar efforts. Of the attributes required of codes, according to Goodson,[5] the ones most difficult to comply with are that names be based on linear character strings to permit lexicographic ordering and that names be brief. This means that codes should be short and that standard symbols (e.g., digits, letters, and other common mathematical or typographical symbols, such as brackets, slashes, asterisks, etc.)

Table I. List of Requirements of Codes as Proposed by Read[4]

(1) codes should be a linear string of symbols
(2) coding algorithm should produce a unique code
(3) structure should be recoverable by a clearly defined process
(4) coding should be simple; preferably, it should be possible to code a compound by hand (without the use of a computer)
(5) decoding process should be simple, preferably one that can be carried out by hand
(6) coding process should not depend on chemical intuition or properties of chemicals
(7) coding should not depend on any list of names or other nonsystematic items
(8) codes should be brief
(9) codes should be pronounceable
(10) symbols used should be familiar (available on standard typewriter or computer keyboard)
(11) codes should be easily comprehensible
(12) coding and decoding algorithms should be efficient

† Dedicated to Professor Vladimir Prelog on the occasion of his 80th birthday. Ames Laboratory is operated by the Iowa State University for the U.S. Department of Energy, under Contract W-7405-Eng-82, Division of Basic Sciences. This work is supported in part by the Office of the Director.

* Address correspondence to the author at Drake University.

0095-2338/86/1626-0136$01.50/0