Expansion of Matula Numbers to Heteroatoms and to Ring Compounds

Seymour B. Elk

Elk Technical Associates, 321 Harris Place, New Milford, New Jersey 07646

Received November 14, 1994®

Matula numbers, an application from number theory that had been of very limited value for canonically nomenclating compounds (inasmuch as, in its original form, it requires a one parameter system, such as alkanes) is expanded to include heteroatoms and polycyclic compounds. The resultant code can now be manipulated to produce a canonical ordering of all chemical compounds that is readily handled by a computer.

1. INTRODUCTION

Although David Matula created the set of numbers that now bears his name for the purpose of organizing rooted trees in graph theory and as an exercise in number theory,[1] he immediately recognized them as the basis of a potential nomenclature system for that very limited class of chemical compounds called alkanes. However, because nomenclature systems for a very small class of compounds are NOT in great demand, especially when combined with the unfamiliarity of prime factorization techniques to the practicing chemists, they remained buried in the mathematics literature until we re-examined the underlying foundations of these numbers[2] in terms of a one-parameter system and extended their usage to polycyclic aromatic compounds of ring size 6 and to polymantanes.[3] Nevertheless, this dependence on one parameter systems was sufficiently damning to make the study of such a system of nomenclature of only minor interest to a small clique of mathematical and computational chemists.[4] In this report, we present a method of circumventing the need for a single parameter system by using information already coded into the individual branches of the rooted tree and then affixing a vector listing heteroatoms and, by analogy, positions of ring scission to polycyclic compounds. The concomitant result is the formulation of a system that, although still "exotic" in its use of the prime factorization of integers, is (1) much more logical in its usage than the I.U.P.A.C.[5] system and (2) gives a canonical ordering of all chemical compounds that is readily handled by a computer.

2. DESCRIPTION OF SYSTEM FOR A GENERAL ACYCLIC MOLECULE

The first step in the algorithm is to write the "associated alkane"; i.e., the molecule that would be formed by replacing all heteroatoms (except hydrogen) with carbon atoms and all multiple bonds by single bonds. In the case of an acyclic compound, the Matula number may now be computed by the method described in ref 1 and 2. For example, consider the amino acid arginine, $C_6N_4O_2H_{14}$ (Figure 1), a molecule containing 12 non-hydrogen atoms. Consequently, we first give the Matula name to the corresponding dodecane (Figure 2): 11269. Since each of the atoms as well as each of the bonds has been assigned an integer value in formulating the Matula number, we now wish to number the hetero atoms

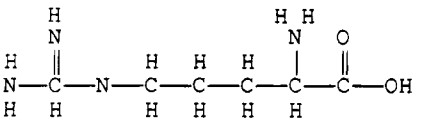
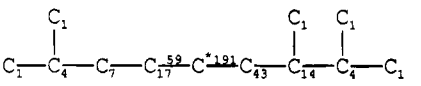


Figure 1. Arginine, $C_6H_4O_2H_{14}$.



11269 (191:-N(1-14),=O(1-4),-O(1-4); 59:N(0-7),=N(1-4),-N(1-4)

Figure 2. Matula name of associated alkane of arginine.

and bond types in each branch. Consequently, let us begin with the highest numbered branch first (191 for arginine). In this branch we have hetero atoms (−O, =O) attached as leaves (i.e., in position no. 1) to a carbon atom at position no. 4 and a single bonded nitrogen atom leaf attached to carbon no. 14. In the other branch, we have a nitrogen atom in the primary chain at position no. 7 and −N and =N attached as leaves at position no. 4. All of this information could now be coded as

$$11\ 269\ (191{:}{-}N_{1-14}, {=}O_{1-4}, {-}O_{1-4};\ 59{:}N_{0-7}, {=}N_{1-4}, {-}N_{1-4})$$

The following items should be noted at this point:

(1) Once the initial computation of the Matula number has been made, the remaining branch designations are available, and do NOT have to be redetermined. Furthermore, there is an ordered sequencing of which branches to follow (largest first), etc.

(2) Multiple occurrences of the same heteroatom (say oxygen), but with different bond attachment schemes are considered separately. Certain combinations, which represent familiar functional groups such as $=O_4$, $-O_4$ for carboxylic acids, etc. will have a characteristic "signature" number that will become familiar with usage. Furthermore, the presence of identical atoms with different bonding will be recognizable as resonance forms, and the appropriate, say, "one and one-half" bond, etc. will be an important part of that signature.

(3) The location of each heteroatom is given a subscript with the atom number being given first and the atom to which it is attached next. However, because the presence of heteroatoms in leaf position is so much more common than in the interior of a chain, we may simplify this name by ignoring the "1" part of the subscripts and include the

® Abstract published in *Advance ACS Abstracts*, February 1, 1995.

0095-2338/95/1635-0233$09.00/0

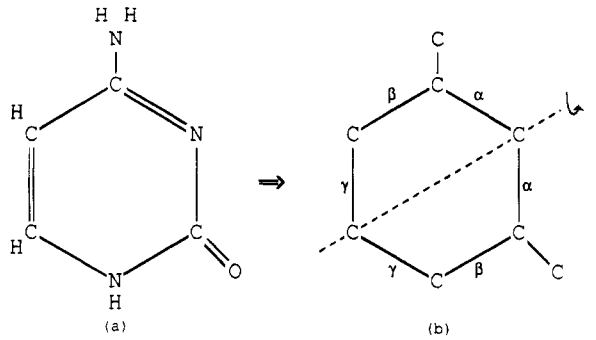


Figure 3. (a) cytosine, $C_4N_3OH_5$ and (b) matula name of associated alkane of cytosine.

position of the atom only when it is at some other location. In other words, the name for arginine would be

$$11\ 269\ (191{:}{-}N_{14}, {=}O_4, {-}O_4;\ 59{:}N_{0-7}, {=}N_4, {-}N_4)$$

3. DESCRIPTION OF SYSTEM FOR A GENERAL MONOCYCLIC MOLECULE

The main difference between an acyclic and a monocyclic system is the fact that (with the trivial exception of $M = 3$ or 4) in an acyclic system, the logical choice of a node that will produce the smallest Matula number for a compound is one of the "center" nodes of the graph. For a monocyclic system, on the other hand, any one of the bonds in the ring may be selected as the "cut point" and then a "convenient" node chosen, so that the entire system may then be viewed as an acyclic system. Consequently, for an *n*-cycloalkane, there are *n* different potential cut points and possibly more than one logical choice for the root—even though, in general, the cut point will be as far removed from the root as possible. All of these must be examined, and the one creating the smallest Matula number of the resultant tree is then selected as the desired canonical name. For example, consider the cytosine molecule, $C_4N_3OH_5$, shown in Figure 3 with its associated alkane. Since there exists an axis of symmetry in the associated alkane, only three, rather than all six potential cut points (named α, β, and γ, respectively) need to be examined. The three resultant trees are shown in Figure 4, along with the Matula number formed for that tree. Note that scission at β gives the smallest Matula number and thus is the one to use in forming the nomenclating tree. Furthermore, the choice of which node to consider as the root was made easy in each case, by starting from the two ends and working toward the middle: Whichever edge had the smallest name as we approached a new node was the one to use in naming the next node. For example, in part c of Figure 4, the clockwise progressing edge was labeled 13, while the product of the two counterclockwise edges was only 10, thus we shall obtain a smaller value for the name of the next edge by extending this counterclockwise branch to the next node, rather than the "13" branch.

We must now return to the given molecule in Figure 3b and consider the two possible locations of scission β with respect to the heteroatoms. For scission in the upper left part of the picture (Figure 5a), the molecule is named 374 ($17{:}{-}N_{0-7}$, $={C}_{0-4}$, $-N_1$; $11{:}{-}N_{0-5}$; $2{:}{=}O_1$), while for scission in the lower right part of the picture (Figure 5b), the name is 374 ($17{:}{=}N_{0-7}$, $={O}_{0-1}$; $11{:}{=}C_3$, $-N_2$; $2{:}{-}N_1$). Now, our algorithm directs us to select the smallest numbers;

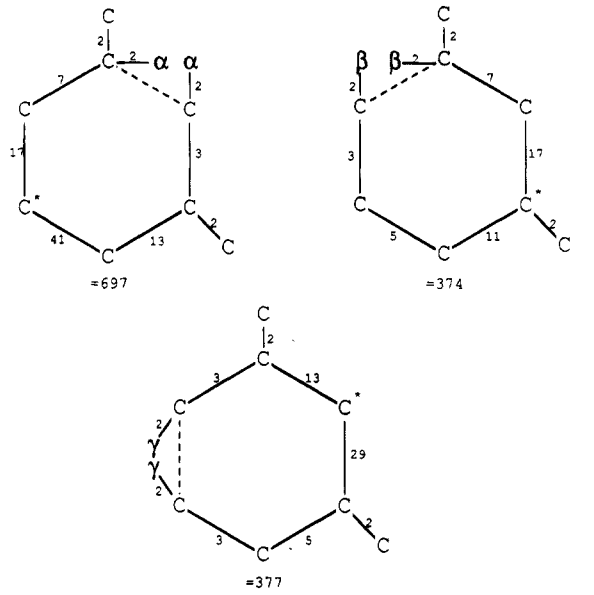


Figure 4. Potential locations of scission of cytosine.

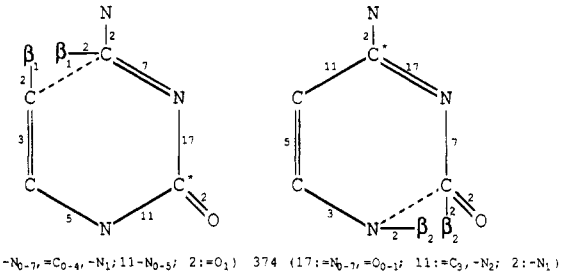


Figure 5. Potential matula names of cytosine.

however, we have equality for each of the branches emanating from the root. Consequently, we must expand the algorithm to next compare some other parameter. We chose the lowest locant numbers in the highest numbered chain; i.e., in chain 17 the 0−7's are the same, but the 0−1 for =O is lower than the 0−4 for =C, thus Figure 5a is the one to use in nomenclating cytosine.

A couple of observations at this point are as follows:

(1) The number of colons in the name is the degree of the root.

(2) For monocyclic compounds, despite that the scission point is an integral part of determining which atom to designate as the root, it appears that they need not be included in the name. However, without their inclusion, one might have difficulty distinguishing whether the compound being named is acyclic vs monocyclic. Consequently, we indicate scission points by unused symbols, such as the letters of the Greek alphabet. For example, cytosine would be named

$$374\ (17{:}{-}N_{0-7}, {=}C_{0-4}, {-}N_1, {-}\alpha;\ 11{:}{-}N_{0-5}, {-}\alpha;\ 2{:}{=}O_1)$$

4. DESCRIPTION OF SYSTEM FOR A GENERAL POLYCYCLIC MOLECULE

Just as the introduction of a scission point in a monocyclic compound created a spanning tree; similarly, we want to introduce a minimum number of scission points in a polycyclic compound. The most efficient way to do this is to start with the Schlegel projection[6] of the molecule and to "cut" edges as near to the perimeter as possible. Two

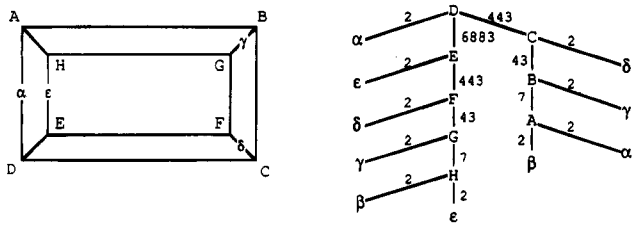


Figure 6. $2 \times 6883 \times 443 = 6\,098\,338$.

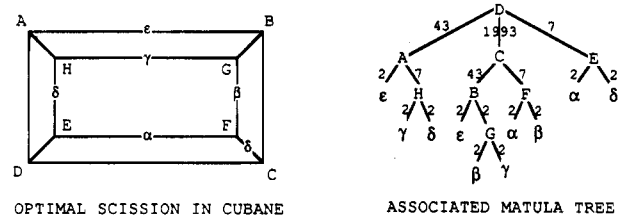


Figure 7. $43 \times 1993 \times 7 = 599\,893$.

comparable size (equal number of nodes and edges) molecules are now examined as part of our study of the various possible "cut schemes": In Figures 6 and 7 we examine applications of the system with respect to a highly symmetric molecule, cubane. For Figure 6, we have chosen the cut set so that the resultant tree is a Hamiltonian path,[7] with the concomitant result that the maximum degree of a node is 2; while for Figure 7, we have maximized the number of higher degree value nodes. Comparison of the best Matula numbers (with the root chosen as near to center of graph as possible) that we can get for these two trees suggests that these are the pessimal ($M = 6\,098\,338$) and optimal ($M = 599\,893$) ways of cutting the graph. Additionally, we note that the Matula number part of the name would be found by selecting the root of the resultant tree as far removed as possible from the various scission points. The rest of the name requires the inclusion of the location of the various selected scission points: For Figure 6, this becomes

6 098 338 (6883/443/43/7/2:$-\beta$;
6883/443/43/7/2:$-\epsilon$; 6883/443/43/2:$-\gamma$;
6883/443/2:$-\delta$; 6883/2:$-\epsilon$, 443/43/7/2:$-\alpha$;
443/43/7/2:$-\beta$; 443/43/2:$-\gamma$; 443/2:$-\delta$; 2:$-\alpha$)

Similarly, for Figure 7, the desired name is

599 893 (1993/43/7/2:$-\beta$; 1993/43/7/2:$-\gamma$;
1993/43/2:$-\epsilon$; 1993/7/2:$-\alpha$; 1993/7/2:$-\beta$; 43/7/2:$-\gamma$;
43/7/2:$-\delta$; 43/2:$-\epsilon$; 7/2:$-\alpha$; 7/2:$-\delta$)

Observe that because of the high symmetry of cubane, we did not have to actually examine all of the $\binom{12}{5} = 792$ combinations of cuts that could be made. Furthermore, this number of potential candidates for the structure to be named may be immediately greatly reduced by noting that the desired structure must be connected; i.e., no combination in which we have a single node or segment is viable. In practice, for small symmetric molecules, the number of cases to examine is few. Even when the molecule chosen to be named is less symmetric, and, consequently, we have to examine more combinations, our intuition which cases to examine will *usually* suffice for "small" molecules. In Figures 8 and 9, the pessimal and optimal Matula names

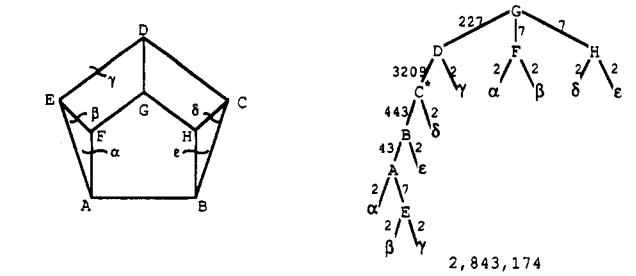


Figure 8.

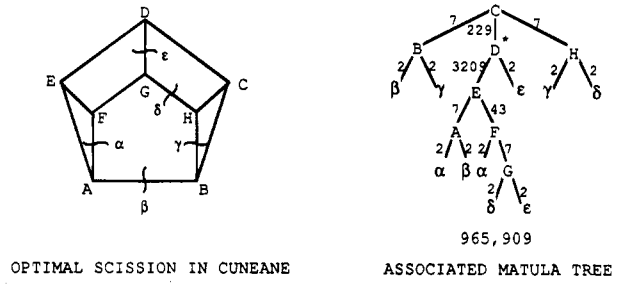


Figure 9.

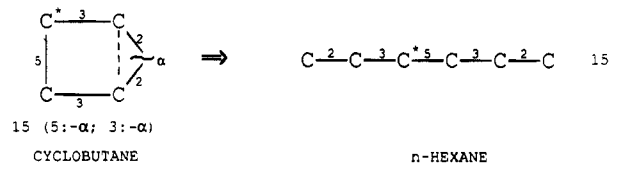


Figure 10.

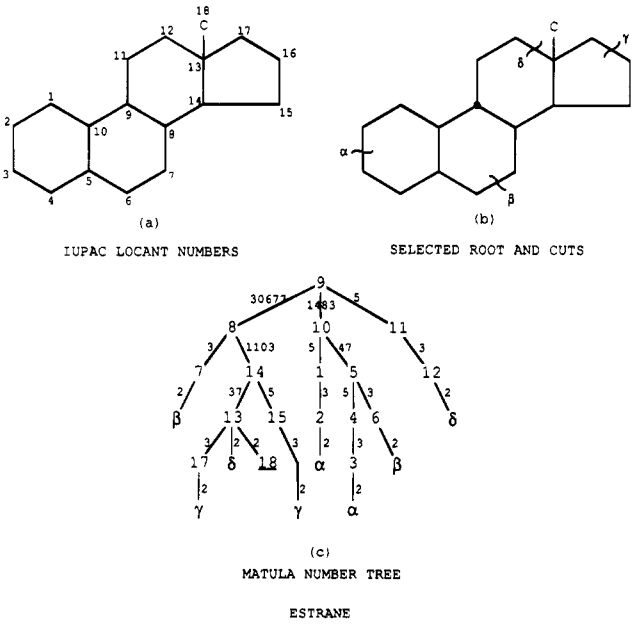


Figure 11.

(for the root at the center of the graph) are derived for cuneane. The intuitive derivation begins by first introducing the desired number of cuts so that we minimize for the pessimal and maximize for the optimal the number of higher degree vertices, then we select any vertex and, working in both directions from that node, form the tree. The desired root of this tree is now either the single one or one of the two central nodes. Figure 8 shows the optimal ($M = 965\,909$) and Figure 9 the pessimal ($M = 2\,843\,174$) names that we found for cuneane. Note that cubane has both a lower optimal and a higher pessimal name than cuneane.

At this point, we note that the relative magnitude of the Matula number of an alkane is a function of both the number of atoms used in the name and how compact the molecule is. Similarly, the presence of heteroatoms does not affect either of these conditions; however, the presence of rings does. In fact, each ring scission functions as though there were exactly two more atoms in the name. For example, the Matula name for cyclobutane is 15 (5:-α; 3:-α), while that of *n*-hexane is 15 (Figure 10).

Note that this procedure is applicable even for a much larger polycyclic system, such as a steroid. The smallest such steroid, if we include estrane (Figure 11), has a minimum of 18 carbon atoms in the fused four ring structure. Thus, with the four scission points, there are effectively 26 nodes to be considered. By examining the graph theoretical distances from each of the nodes to all of the other nodes, we find that either vertex numbered 8 or 9 in the standardization by IUPAC[5] will be the center (Figure 11a). Next, our guess as to where to make the scissions in each of the four rings is predicated on what choices will minimize the number of levels in the resultant tree (Figure 11b). Note that every letter of scission (α thru δ) must appear as a leaf in this tree twice—on either the same or on adjacent levels. The minimum diameter cut scheme we found is given in Figure 11c, and the resultant Matula number for this structure is 227 469 955. Since there were no heteroatoms in this molecule, we complete the Matula name by affixing the scission locations, thereby distinguishing it form the acyclic alkane with eight more carbon atoms; i.e.

227 469 955 (30677/1103/37/3:−γ; 30677/1103/37/2:−δ; 30677/1103/5:−γ; 30677/3:−β; 1483/47/5:−α; 1483/47/3:−β; 1483/5:−α; 5:−δ)

REFERENCES AND NOTES

(1) Matula, D. W. A Natural Rooted Tree Enumeration by Prime Factorization, *SIAM Rev.* **1968**, *10*, 273.

(2) Elk, S. B. A Problem with the Application of Matula's Method of Prime Numbers and Rooted Trees for Canonical Nomenclature of Alkanes. *Graph Theory Notes of New York* **1989**, *XVIII*, 40−43.

(3) Elk, S. B. A Canonical Ordering of Polybenzenes and Polymantanes Using a Prime Number Factorization Technique *J. Math. Chem.* **1990**, 55−68.

(4) (a) Gutman, I.; Ivic, A.; Elk, S. B.Matula Numbers of Coding Chemical Structures and Some of their Properties. *J. Serb. Chem. Soc.* **1993**, *58*, 193−201. (b) Gutman, I.; Ivic, A. Graphs with Maximal and Minimal Matula Numbers. *Bull. Acad. Serb. Sci.* **1994**, *19*, 65−74. (c) Elk, S. B.; Gutman, I. Further Properties Derivable from the Matula Number of an Alkane. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 331−334. (d) Müller, W. R.; Szymanski, K.; Knop. J. V.; Trinajstic, N. A Comparison Between the Matula Numbers and Bit-tuple Notation for Rooted Trees. publication pending.

(5) *I.U.P.A.C. Nomenclature of Organic Chemistry, Sect. A*; Pergamon Press: Oxford, 1979.

(6) (a) Loeb, A. Space Structure—Their Harmony and Counter-point. Addison-Wesley: Reading, MA, 1976. (b) Coxeter, H. S. M. *Regular Polytopes*, 2nd ed.; Macmillan Company: New York, 1963; p 10. (c) Schlegel, V. Theorie der homogen zusammengesetzten Raum gebilde. *Verhandlungen der Kaiserlichen Leopoldinisch- Carolinischen Deutschen Akademie Naturforscher*, **1883**, *44*, 343−459.

(7) By a Hamiltonian path is meant that it is possible to draw one continuous line that passes through each vertex exactly once.

CI940125L