Exposition, Carlos Cuadra, President of Cuadra Associates, Inc., said that "many publishers have come to accept the idea that there are legitimate reasons for the temporary retention of data obtained from an online search, and they have given [such] permission.... More and more users want to retain records permanently, to make them part of their own local electronic library." He urged publishers "to give immediate attention to...user needs and to develope pricing policies...that serve those needs, while protecting their own rights in the data".[7] Although I am not in a position to give any legal advice, I suggest you look at the contract you have with the copyright owner or data base producer. If you do not already have a contract and you are reusing information gleaned from an on-line system, perhaps you should request a copy. I suggest you ensure that what you want to do with data base information is permissible and spelled out in your agreement. Also, do not hesitate to seek legal advice. These steps might save you a lot of headaches (and possible lawsuits) in the long run. Keep in mind that the copyright owner has the exclusive right to determine the conditions under which data is made available to others.

(2) What happens to information after users receive it? This is an old and a present concern that we cannot really do anything about, except perhaps through negotiated contracts, licenses, and agreements. I still believe that most people are honest and would be willing to abide by the copyright law and by reasonable terms set by publishers.

(3) Does copyright infringement of works in a computer system occur at the point of input or output? The Subcommittee on Databases of the National Commission on New Technological Uses of Copyrighted Works, better known as CONTU, came to the same conclusion as the UNESCO/WIPO group: the act of storing a computerized data base in the memory of a computer is the exclusive right of the copyright owner.[8]

(4) Does "fair use" apply to computer-generated works? CONTU determined that fair use of machine-readable data bases follows "the same guidelines as are applicable to print materials. This means that the user could retrieve and use information derived from the data base, such as a citation or in fact, just as he would with any other copyrighted work. He may not, however, use a substantial portion of the data base without violating the owner's copyright."[8]

I am sure there are countless other concerns and questions that one may have with the copyrightability of computer-generated works or works accessed from computerized systems. It is well-known that users want and need information that is timely and cost effective. It is almost certain that "downloading will soon become a way of life",[9] and we will have to pay for the right to use and reuse that valuable resource known as information.

Until our copyright law and other laws catch up, or at least come close, to the rapidly advancing and ever-changing technology, I urge you to use good judgement concerning the use and reuse of copyrighted material. I leave you with a word of advice, which I have adapted from the late Joseph McDonald: Take not from others to the extent that you would be resentful if they took that from you.

REFERENCES AND NOTES

(1) UNESCO/WIPO/CEGO/II/7, Annex I, 1982: Recommendations for Settlement of Copyright Problems Arising from the Use of Computer Systems for Access to or the Creation of Works.

(2) UNESCO/WIPO/CEGO/II/7, Aug 13, 1982: Report of the Second Committee of Governmental Experts on Copyright Problems Arising from the Use of Computers for Access to or the Creation of Works, Paris, June 7-11, 1982.

(3) UNESCO/WIPO/CEGO/I/7, Feb 20, 1981: Report of the Committee of Governmental Experts on Copyright Problems Arising from the Use of Computers for Access to or the Creation of Works, Paris, Dec 15-19, 1980.

(4) Peters, M. "General Guide to the Copyright Act of 1976"; Copyright Office, Library of Congress: Washington, DC, 1977, p 1.1.

(5) Title 17, USC, Copyrights, Oct 19, 1976.

(6) Boorstyn, N., "Copyrights, Computers, and Confusion" *J. Patent Office Soc.* **1981**, *63* (5), 277.

(7) Cuadra Associates, Inc. *NFAIS Newslett.* **1982**, *24* (6), 9.

(8) Wolfe, M. "Copyright and Machine Readable Databases" *Online (Weston, Conn.)* **1982**, July, 54.

(9) Keenan, S. "The End User's Point of View" *ASIDIC Newslett.* **1982**, *No. 43*, p 6.

Unique Numbering and Cataloguing of Molecular Structures[1]

JAMES B. HENDRICKSON* and A. GLENN TOCZKO

Department of Chemistry, Brandeis University, Waltham, Massachusetts 02254

Received February 28, 1983

A simple procedure is offered for creating a unique canonical numbering of a molecular skeleton, or general graph, based on the maximization of the linear binary number corresponding to its adjacency matrix. This numbering allows ready comparison of two molecules for identity, i.e., of two graphs for isomorphism, and a catalog of these identity numbers may be assembled in numerical order for quick searching. The method also identifies equivalence classes for automorphism and is error free and rapid both by hand and computer. Comparisons with other such systems are made to show its superiority in computer speed and work space.

For our program in synthesis design[1] we required a catalog of available starting material molecules arranged for rapid search and comparison with the starting material molecules generated by the program. We required a notation system for the computer which possessed the following features: (1) unique canonical numbering of the atoms; (2) separation of the skeleton of the molecule from its functionality; (3) rapid comparison of two molecules for identity; (4) rapid search of the catalog; (5) minimal storage requirements for each molecule in the catalog. The skeleton of the molecule is simply a graph of points (atoms) and lines (bonds) and may be represented most easily in the computer by its adjacency matrix, an $n \times n$ matrix of the n skeletal atoms in which the elements are simply 1 or 0, indicating atoms bonded or nonbonded,

0095-2338/83/1623-0171$01.50/0 © 1983 American Chemical Society

respectively. Such a matrix fully represents the connectivity of the molecular skeleton (which may be reconstructed from it).[2] There are, however, $n!$ ways of numbering the atoms in any skeleton and therefore $n!$ equivalent matrices, all interconvertible by row/column interchanges in the matrix. Hence there is a need for a clear definition of a single numbering so that any two molecules may be so numbered and then compared for identity. This is the requirement (1 above) for a unique canonical numbering of the atoms.

In the computer the adjacency matrix is more easily handled as a linear string of 1/0 bits; thus it is spun out with the successive rows of the matrix following each other in a long string of bits (n^2 bits). Since the matrix is symmetrical, only those bits in each row to the right of the diagonal need to be included, and so the binary string is only half the size. Such a binary string fully conserves the connectivity information in the matrix and so in the molecular skeleton. Each of the $n!$ different possible numberings of the molecular skeleton will have a different matrix and so a different binary string. However, if these binary strings are simply read as long binary numbers, only one can be an absolute maximum in the set, and so this defines a single, particular adjacency matrix and in turn describes a single, unique canonical numbering of the molecular skeleton. This binary string is therefore a unique binary identification number for the skeleton.

GENERATION OF CANONICAL NUMBERING

The generation of this unique numbering is quite simple. In principle the computer could start with any arbitrary numbering and make all possible row/column interchanges until the maximum binary string is found. However, for certainty this must require $n!$ operations and is far too time consuming. A much more powerful approach is to seek the maximum for each successive matrix row, assigning numbers to the skeleton successively as each row is maximized.

The maximum value for the first row is obtained by placing the atom with the greatest connectivity, or valence, in the first row. If there are several such atoms, several first rows are separately initiated and their growing matrices followed until one shows a maximum. If atom 1 in the first row is quaternary, its four attached atoms must be numbered (2345) in whatever combination, in order for the first row to be the maximum binary number (i.e., 11110000...). One of those four will then be atom 2, and they are compared to find the largest connectivity, assigning the maximum binary number to the second row and so identifying atom 2. If two (or more) cannot be distinguished, a separate matrix is again initiated for each. Unassigned atoms attached to atom 2 must now be the next numbered atoms (e.g., 6, 7, ...) in whatever order. In this fashion one proceeds, identifying each successive atom as that which yields the maximum binary number for its row and then assigning the next unused numbers to its attached atoms.

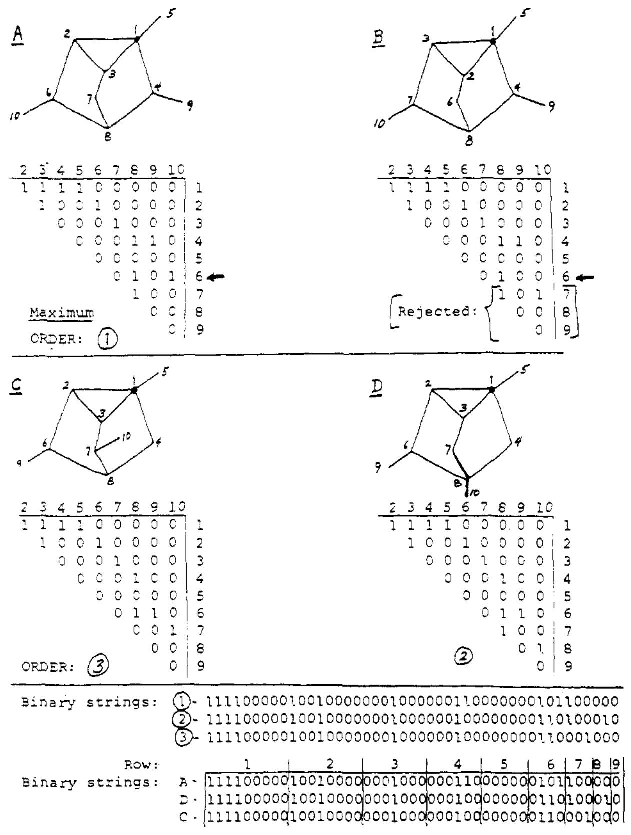


Figure 1. Unique maximal numbering and adjacency matrices.

The 10-atom skeleton (A and B) in Figure 1 can serve to illustrate the number generation procedure: (1) Atom 1 is quaternary; row 1 (above diagonal) is 111100000. (2) Attached to atom 1 are four atoms numbered (2345) in any permutation. (3) Of the four, three are tertiary, but the two in the cyclopropane have maximal rows (10010000). (4) Each is assigned as atom 2 and the other necessarily as 3, in two separate matrices (A and B). (5) Attached to atom 2 is an atom assigned as 6 and to 3 one assigned as 7. (6) Atom 4 is now the last tertiary atom on atom 1, and it in turn assigns 8 and 9 in either order to its attached atoms. (7) Atom 5 is trivial (all primary atoms will appear with an all-zero row). (8) Atom 6 now assigns atom 8 of the pair attached to atom 4 in order to maximize its row. (9) The matrix A with atom 6 tertiary is now preferred (row 6 = 0101) for the first time, and the other (B) with row 6 = 0100 is discarded. (10) The matrix A is now completed: atom 9 is the remaining atom assigned to atom 4, and atom 10 is assigned by default to the last atom.

The maximal numbering derived in A now fully identifies the skeleton, and its maximal binary string or number, shown below, is a unique identification of the skeleton. The skeleton may also be uniquely reconstructed from it by reconstituting the matrix and building the skeleton visually from it as a graph.

Whenever the developing maximization procedure encounters two or more equivalent atoms, it initiates as many separate matrices to develop further. Some of these may later be dropped as less than maximal as this proceeds, but if they are not, they will end up as equivalent maximal matrices, implying a structure with some finally indistinguishable atoms. The number of such identical and redundant matrices so obtained equals the number of possible equivalent maximal numberings of the skeleton and is a measure of its symmetry. These identical matrices, of course, all exhibit the same binary string which is the identification number for that skeleton. Thus a simple cycloalkane of n atoms has $2n$ equivalent numbers (the label 1 can be placed in any atom and label 2 placed either to the right or left of it), the maximal binary string (number) identifying cyclopentane being 1100010011. The third tricyclic structure (C) in Figure 1 actually has two sets of equivalent atoms, 2–6–9 and 3–7–10, and two equivalent matrices would be generated from row 2 down. These are the matrix shown, with the same binary identification number, shown below.

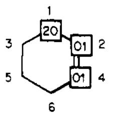
Our catalog of starting materials is now constructed by skeleton, ordering all the skeletons first by size (number of carbons) and then in numerical order of their binary identification numbers. Three separate tricyclic structures (A, C, D) are maximally numbered in Figure 1. The binary strings identifying them are listed below and show an ordering of A > D > C. In order to establish whether any given molecular skeleton is represented in the catalog, it is first maximized as above and the maximal binary string so obtained compared with those in the catalog. The search of the catalog is a simple binary search, requiring s steps for a 2^s-size catalog; our starting material catalogue presently contains about 3000 entries and so requires only $s = 12$ steps to search.

In our system we clearly separate skeleton from functionality.[1] If the skeleton is considered to be composed of carbon atoms only (see below), we designate the number of attachments of any carbon (i) to other carbons (σ bonds) as σ_i, i.e., the skeletal connectivity denoted by its row sum in the adjacency matrix. The functionality at each carbon i is designated by z_i, the number of bonds (π or σ) to electronegative heteroatoms, and by π_i, the number of π bonds to carbon. The number of hydrogens (and electropositive attachments) is then designated by $h = 4 - (\sigma + z + \pi)$. Hence the functionality at any carbon is the "$z\pi$ value". This number can be entered into the adjacency matrix on the diagonal,[3] i.e., as element a_{ii} in the matrix for carbon i. Alternatively, the functionality list, or $z_i\pi_i$ list, is regarded as row $n + 1$ of the matrix and as such is appended to the end of the binary string of connectivity, derived above. Thus the functionality list is a list of $z\pi$ values for the carbons of the skeleton, canonically numbered from the matrix maximization. When this $z\pi$ list is appended to the binary identification number for the skeleton, the catalog may be searched in the same fashion either for skeletal identity only or for the whole structure (=skeleton + functionality). It should be emphasized here that multiple bonds in the skeleton are treated strictly as *functionality*, annotated to the atoms so joined as functional groups. This convention keeps the *skeleton* quite separate conceptually as the σ framework and also avoids the complexities inherent in trying to incorporate multiple bonds into the adjacency matrix, i.e., using entries larger than 1.

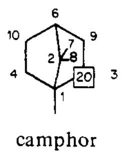
Since the values of $z = 0$–3 and $\pi = 0$–2, only 2 bits are needed to express each one, or 4 bits per $z\pi$ value. Hence the $n + 1$ row of functionalities is $4n$ bits in length. These $z\pi$ values indicating functionality are "atom attributes", i.e., the numbers of heteroatom attachments (z) and of C–C π bonds (π) attached are expressed as attributes of the individual carbon atoms of the skeleton. There are 16 possible 4-bit values for $z\pi$, but they are interdependent so that actually only nine of them, those above the line in the following table, are structurally viable. Since $\sigma \geq 1$, $z + \pi$ can be no more than 3. The remaining seven $z\pi$ values (below the line) therefore remain available for expressing other atom attributes such as distinctions among z-atom attachments or identifying atoms other than carbon in the skeleton (see below).

$z\pi$ = 00	10	20	30
01	11	21	31
02	12	22	32
03	13	23	33

Redundancy in the skeleton is usually reduced by the presence of functionality. When any two or more carbons are skeletally equivalent, the $z\pi$ list is ordered numerically to afford the maximum number, and this in turn fixes specific numbering for the previously equivalent skeletal carbons. For example, there are 12 equivalent numberings for cyclohexane but only two for cyclohexanone, in which the $z\pi$ value for the ketone carbon is 20, and it becomes carbon 1. For 2-cyclohexenone the functionalized carbons have $z\pi$ lists of 20, 01, and 01 (boxed on the structure below), but the maximized skeletal numbering will place them at positions 1, 2, and 4, giving a $z\pi$ list for the molecule of 200100010000 in canonical order. Similarly, there are two identical bridges in the camphor skeleton, numbered as shown, but the redundancy is removed by the presence of a ketone ($z\pi = 20$ in the box at position 3) so that its bridge takes on the lower pair (3–9) of position numbers, leaving 4–10 to number the skeletally equivalent other bridge.



2-cyclohexenone



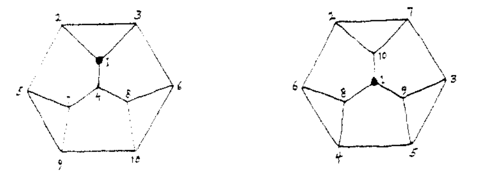
camphor

COMPARISON WITH OTHER SYSTEMS

We developed this procedure first just as a simple, straightforward logical way to solve our own need for a fast accurate computer search of our starting materials catalog, without much attention to the literature save for an acquaintance with the Wiswesser notation[5] and some variants, which did not suit our purpose. A more comprehensive examination of the literature revealed this to be a much-studied graph theory problem,[6,7] that of isomorphism identification, i.e., comparing graphs for identity. There are indeed several specific schemes focussed on the canonical numbering of molecules.[8-20] There appear to have been two conceptual approaches to the problem: visual and mathematical. The visual approach[5,8,11,12,15,16] implies an observer of the graph tracing out the extended connectivities of each atom and combining these and other atom attributes in enough detail to supply each atom with a separate identity. This approach, called vertex partitioning in graph theory, is a device for initial priority ordering of atoms to reduce the otherwise $n!$ problem of comparing numberings. These algorithms, based on ad hoc rules for assigning priorities, have been critized as not containing rigorous proofs of the uniqueness of the final numberings: "verifying that an algorithm is never *known* to fail...does not constitute proof that the algorithm is valid beyong the specific cases which are tested."[13]

The mathematical approach is "blind" and consists only of manipulating numbers in a numerical presentation of the graph, e.g., the adjacency matrix, and this approach is, of course, much more amenable to computer manipulation. The first mathematical approach was that of Randic,[9] who treated the adjacency matrix as we have but sought a *minimum* binary string from it.[17] His first suggestion of seeking this by row/column interchanges from an arbitrary numbering was found to be faulty in that it can lock onto incorrect local minima;[10] this is, of course, a pitfall of any search for maxima or minima which follows a direction of change from a single starting point on a surface. Randic has since corrected this[9e] by developing the minimum matrix stepwise, minimizing row by row, a procedure inverse to that given above. The mathematical literature[18] provides another approach similar to ours in that row-wise or column-wise maximization is applied to the *incidence* matrix of the graph, an unsymmetrical atom-bond matrix.[19] This procedure is actually slower since the "combinatorial explosion" problem must also be first reduced by an initial priority ordering of atoms, and the computer storage of the unsymmetrical incidence matrix is necessarily more than twice that of the symmetrical adjacency matrix. Finally, there are two other mathematical approaches which are much more complex to apply.[9c,20]

Although at first sight the minimum[9e] and maximum approaches appear equivalent, as inversions of each other, there are several considerable advantages to the maximization in practical use. First, the number of computer operations is much reduced. For the first row the computer must examine all n atoms to find the ones of maximum (or minimum) valence (σ) to initiate matrices as atom 1, in either approach. For the second row, to assign atom 2, in the maximization the computer needs only to examine the direct neighbors of atom 1, which are σ_1 in number (never more than four), while the



MAXIMUM

Row	Trials	Retained	Assigned Atom No.
1	10	10	1, (234)
2	30	18	2,3,4.5
3	18	18	6
4	18	18	(78)
5	30	12	7,8,9
6	12	6	10
	Σ =106	Max=18	

MINIMUM

Row	Trials	Retained	Assigned Atom No.
1	10	10	1, (8910)
2	60	42	2. (67), (89), 10
3	126	48	3,5,6, 7,8,9
4	48	48	4
5	48	12	-
6	12	6	-
	Σ =304	Max=48	

Figure 2. Maximal and minimal numbering.

minimization assigns the atoms adjacent atoms to atom 1 as the highest numbers (n, $n - 1$, etc.) and so must assess all the remaining atoms, numbering $n - \sigma_1 - 1$, in order to label atom 2. With a molecule of any size there will continue to be a fair number of unassigned atoms which must be examined at each of the next several rows in this minimization protocol and thus a substantially larger number of necessary trials at each row. The number of retained matrices in the early rows also reduces much faster in the maximization method.

This may be exemplified with a graph cited[13] as difficult since all atoms are trivalent but not equivalent. There are three distinguishable sets of equivalent atoms. Shown in Figure 2 are the minimal and maximal numberings of this 10-atom graph. Below each is enumerated the number of searches the computer must make at each row, and, since many atoms are of equivalent sets, also the number of growing identical matrices which must be kept at each step. The new atom numbers assigned at each row are also shown, and those in parentheses are assigned as a group but are not yet ordered. In the maximization process there are three atoms which may be assigned as 2, as neighbors of the ten possible carbons 1; hence there are 10 × 3 or 30 necessary trials to establish precedence. Of these only 18 afford a maximal row for atom 2, those with atom 1 at the center being rejected since atoms 2 and 3 must both be in a cyclopropane to afford a maximal second row. Also some combinations with atom 1 at the other two possible distinguishable locations are rejected for the same reason. In the minimization procedure there are $10 - 4 = 6$ possible locations for atom 2, implying 60 necessary trials, of which 42 combinations must be retained. In the course of establishing the priority of row 2, the maximization has necessarily assigned positions to atoms 2–5, while the minimization has assigned atoms 2 and 10 and the unordered pairs of (6,7) and (8,9). In the maximization, row 3 is simply already fixed and assigns atom 6 as its neighbor, rejecting none of the 18 equivalent matrices grown so far. In the minimization, atom 3 can now be in three places in each of the retained 42 growing labelings, hence 126 necessary trials. Assignment of atom 3 also implies the larger number for its neighbors when they may be part of an unordered pair from before, in this case (6,7) and (8,9). Hence, in this row all atoms are defined, and only 48 equivalences remain at this row (no. 3). Atom 4 has already been assigned now in each case, and all the retained matrices have been tried; all give the same row 4 in each case and are retained. These are further reduced in each approach to a final number of six fully equivalent numberings by continuing to test each subsequent row for maximum (minimum) values, the tests being complete in each case by row 6.

The results are tabulated in Figure 2. The maximization requires 106 trials and retains at any one time no more than 18 parallel and equivalent growing matrices, while the other requires 304 trials, retaining as many as 48 growing matrices. This is a common result in all examples, owing to the larger number of necessary atom examinations ($n - \sigma_1 - 1$) in the minimum search in the early rows. Hence the minimization method requires substantially more time to run and at the same time requires the use of much more memory space in the computer to retain currently equivalent matrices.[19]

It is also very clear that for the same reason the maximization procedure is far easier to follow by hand, as a number of simple trials make clear. A set of various sample molecules is offered in Figure 3, numbered maximally and each quickly and easily reproduced by hand. When number choices are made by hand, preference goes either to higher valent atoms in the largest number of smallest rings or to numbering out into rings toward the lowest previously assigned numbers, with the second rule taking priority in a conflict. While these simple guides assist assignment by hand, the real test in case of doubt is still writing out the relevant rows of competing numberings and comparing them for a maximum.

The regular dodecahedrane skeleton (A, Figure 3) has all faces and atoms equivalent. It therefore is a kind of worst case example since high symmetry increases the number of equivalent matrices tried and also retained throughout. In the regular solids of n trivalent ($\sigma = 3$) points (atoms), as in tetrahedrane, cubane, and dodecahedrane (A), there are $n\sigma!$ equivalent numberings, i.e., 24, 48, and 120, respectively. In this instance (A) the maximization procedure must make a total of 2120 trials, retaining 120 matrices, but never more than 120 en route; the minimization procedure must make 6100 trials, retaining as many as 240 matrices en route to reach 120 finally. However, the steroid skeleton (B), though larger, is typically much easier owing to asymmetry, requiring only 85 trials to establish the full numbering shown. There are only two quaternary centers to label 1 at first, and one drops out as less than maximal by row 10. From then on only one matrix is required to develop, becoming two equivalent matrices only at the very end with the placement of equivalent atoms 26 and 27.

Examples C–E have been offered[13,20] as counter examples on the breakdown of other numbering schemes. They required 60, 84, and 402 trials, respectively, to establish the unique numberings shown in Figure 3, and none exceeded en route the final number of retained equivalent matrices, e.g., 4, 8, and 32, respectively. Example D illustrates a case in which the choice for atom 2 is dictated by closure of the small ring (2367) in preference to highest valency.[22] Two equivalent numberings are shown for the left ring in E. The final example (F) is the 17-carbon molecule recently synthesized[23] as a molecule illustrating the K_5 graph of five points (dots in F) all connected to each other. The molecule is shown both in perspective and in projection down the three-fold axis (the 1–2 bond). The symmetry is apparent in the cycling of the numbers and its combinatorics required 321 trials, with no more than the 24 final equivalent matrices retained at any time en route.

A final example in Figure 4 has been offered as a graph with no equivalent vertices and is so regarded as a test of vertex-classification schemes.[15] The graph is redrawn here to emphasize its apparent symmetry. The circled atoms, distinguished as different atom types before,[15] were here taken as CH–CH_3 groups to retain that distinction. The maximization procedure affords the single numbering shown, with no atoms equivalent.

The computing time necessary for establishing a canonical numbering has been quoted for a few examples in several systems and is surprisingly large. Thus the Jochum–Gasteiger[11b] times for naphthalene, anthracene, and ethyl acetoacetate are 410, 950, and 190 ms, respectively, as obtained on a high-speed IBM 360/91 computer, while for this system

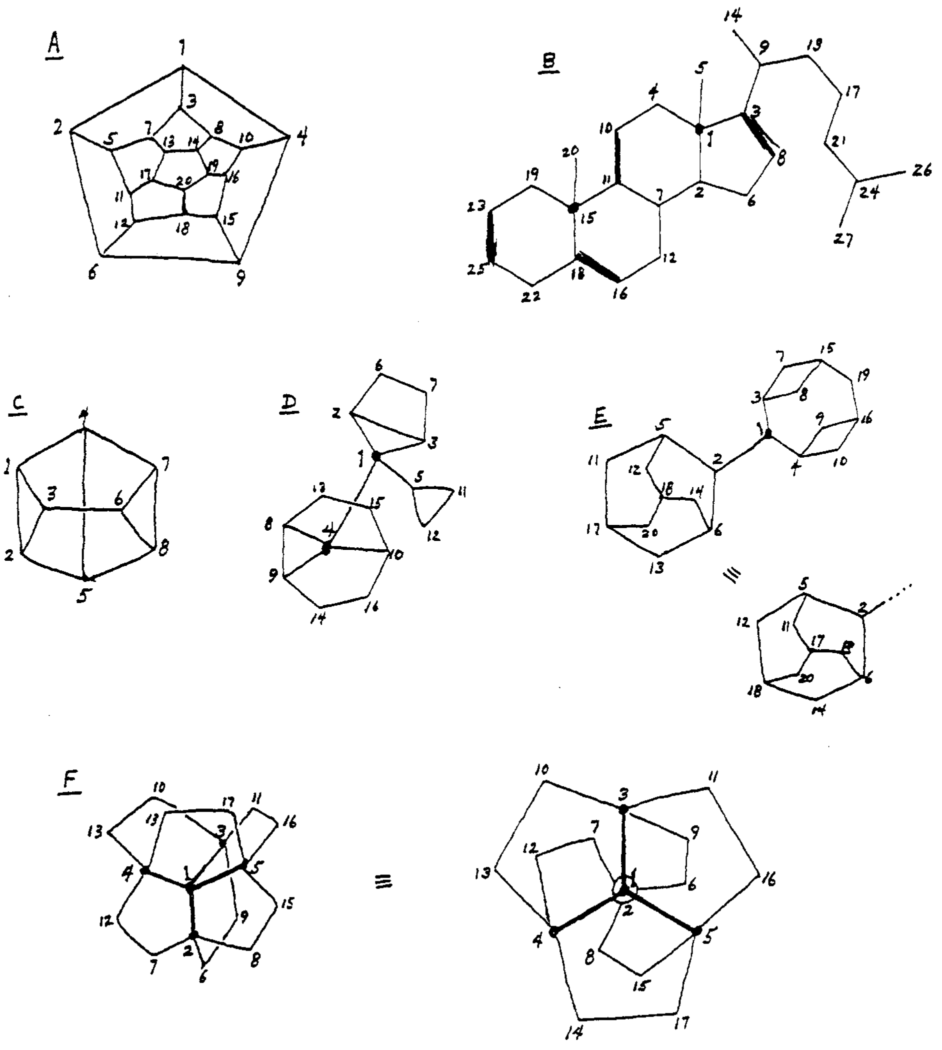


Figure 3. Selected examples of maximal numbering.

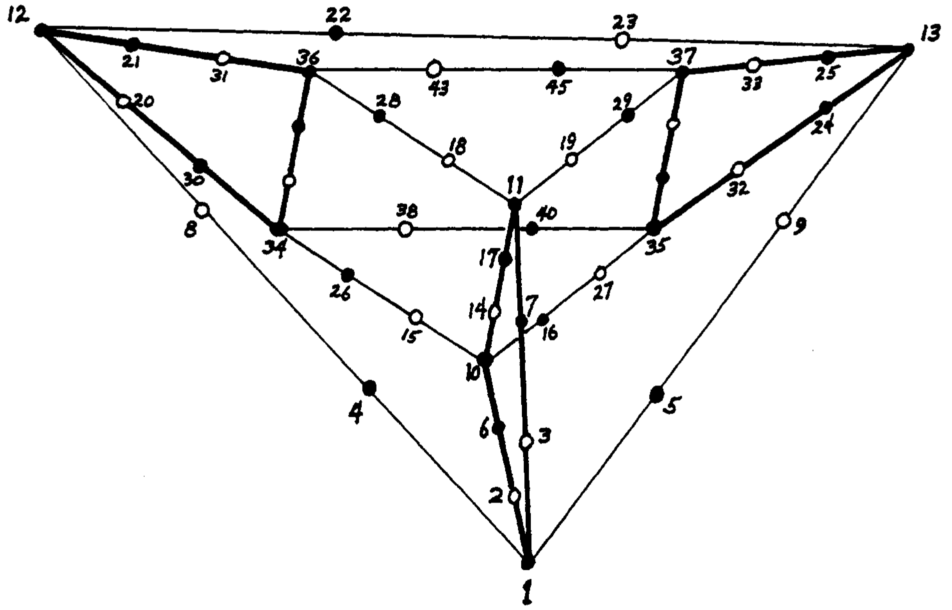


Figure 4. Unique numbering of a graph with no equivalent atoms.

the times are, respectively, only 75, 163, and 33 ms; for comparison the oxygens of the acetoacetate are counted (and prioritized over carbon in the $n + 1$ row) as skeletal, as described below. Furthermore, these times are obtained on the much slower (30–40 times) DEC 11/23 minicomputer. Skelley and Munk[15] quote an average time of 100 ms for 184 compounds of average size $n = 18$ on a UNIVAC 1100/42, as well as an average of 253 ms each for some 59 $n = 10$ graphs and 1980 ms for the dodecahedron. Randic[9f] quotes a variety of times for his minimization procedure, with 10-carbon polycyclics ranging from 470 to 750 ms on a high-speed IBM 370/158 computer. The same examples on our minicomputer run at 70–90 ms, and the dodecahedron required 300 ms. The fully trivalent C_{10} structures and some C_{14} polycyclics required 1–7 s in the Randic method and only 50–500 ms on ours. However, when we operated the program on several such samples using a high-speed DEC PDP-10 computer, our times never exceeded 8 ms (the structure in Figure 2 required 5250 ms/Randic minimization against our 170 ms/minicomputer and 6 ms/PDP-10).

A final, more trivial, basis for preference of the maximization over the minimization numbering method can be seen in the distribution of the numbers on the skeleton. The minimization approach scatters numbers seemingly at random over the skeleton, while the maximal numbering grows outward from a single highest valency starting atom like a seed crystal. In fact, this numbering appears very close to (but no identical with) that of the Morgan algorithm.[5]

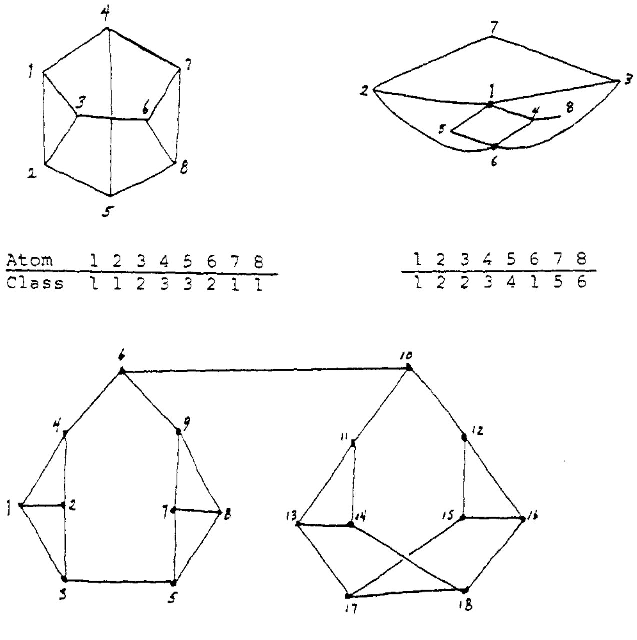
BROADER UTILITY OF THE METHOD

For the purposes of our synthesis-design system we initially defined the skeleton as composed only of linked carbon atoms, with all heteroatoms subsumed as functionality and indicated only as z values on the carbon to which they are attached. We have now, however, incorporated nitrogen atoms with carbons as skeletal atoms. Here the skeleton is defined as all connected carbon and nitrogen atoms for purposes of skeletal numbering via maximization. Nitrogen is treated as a special kind of carbon and can be given one of the seven unused $z\pi$ values (above) as an atom attribute to identify it. The same treatment could be extended to incorporate other skeletal atom types such as oxygen or sulfur in heterocycles.

In a broader vein, however, the whole procedure can be applied to identify any molecular structure, assuming the skeleton to be maximized as containing all nonhydrogen atoms of any kind. In this treatment the whole (nonhydrogen) skeleton is maximized without regard to atom type, and either the $(n + 1)$ row, the $(n + 1)$ column, or the diagonal is utilized for individual atom types. The extra entries can also be expanded to incorporate any other atom attributes as well, such as stereochemical annotations, isotopes, etc., as desired for any specialized use. In our particular requirements, the extra row is reserved for functionality as a $z\pi$ list and appended at the end of the binary string in order to allow separate catalog searches for skeleton only as well as full searches for whole structure. The more general usage here is made more efficient if each atom attribute is appended to its own *row*, i.e., the list of attributes is the *n + 1 column* rather than the $n + 1$ row. In this way the maximization will directly incorporate the value of the atom in maximizing the matrix, and that atom value will then serve to reduce structural redundancy and so speed up the maximization process.

For graph theory applications beyond molecular structures the procedure is also applicable. The major change in such applications is simply the removal of the molecular limitation on the valency, or degree, of any point being no more than four. This allows much denser adjacency matrices but does not change the maximization procedure. A major need in graph theory is a protocol for detecting isomorphism or identity of graphs.[6,7] Given a unique canonical numbering procedure, it is then trivial to establish the identity of two graphs, and indeed this is exactly what is done in our comparison of molecules with the starting material catalog. The central requirement here is then a proof of the uniqueness of the numbering.

Although apparently self-evident, an outline of a proof for maximal numbering follows. The adjacency matrix represents fully the connectivity information in the graph itself, which can be reconstructed from it. This 1:1 correlation is also fully preserved in the binary string representation of the upper right half of the symmetrical matrix. Considered as binary numbers, the $n!$ possible binary strings must include one and only one maximum number, and so this represents a unique numbering. If this number appears more than once, this indicates more than one equivalent numbering of the graph and hence local identities of points, i.e., local isomorphism or automorphism. The procedure for establishing the one maximal number is without ambiguity, being developed and tested point by point by assessing each successive matrix row for maxima.



Atom	1	2	3	4	5	6	7	8
Class	1	1	2	3	3	2	1	1

1	2	3	4	5	6	7	8
1	2	2	3	4	1	5	6

Atom	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Class	1	1	2	3	2	4	1	1	3	5	6	6	7	7	7	7	8	8

Figure 5. Equivalence classes.

The concept of equivalence classes looks to establishing the identity of several points (atoms) in the graph (molecular structure) as local isomorphs, or automorphs. In our procedure for maximal numbering, one finishes with several identical matrices when there are automorphic atoms. The initial input of the graph necessarily provides an initial arbitrary numbering, in our case the order in which the points are drawn onto the CRT screen. The finished maximal numberings are all mapped with a 1:1 correspondence to the initial input numbering. Hence a reading across any given row of all maximal matrices yields the numbers of all atoms that may equivalently occupy that row, i.e., all automorphs for that position. We may assign each of these an equivalence class label and so identify and enumerate all automorphic points in the graph. Thus, in the graph of Figure 2, maximally numbered, we may chart these results in the following convenient partition[20] (others are illustrated in Figure 5):

atom numbers =	1	2	3	4	5	6	7	8	9	10
equivalence class:	1	2	2	3	2	2	1	1	2	2

The graph in Figure 4 was designed as a test of vertex-classification methods, which failed to establish its automorphism partitioning.[15] It has only the single numbering shown by our procedure with no equivalent atoms.

CONCLUSION

The procedure outlined here affords a unique canonical numbering for molecular structures and graphs by maximizing the binary string corresponding to the adjacency matrix. The method is fast, efficient, infallible either by hand or computer, and, indeed, very simple to apply by hand. It allows easy ordering of a catalog of structures for rapid search and retrieval, as well as fast comparison of two structures for identity.

The method also provides for enumeration and identification of automorphic (locally isomorphic) atoms. The procedure is simple and fast enough for adaptation to microcomputer use.

ACKNOWLEDGMENT

We are grateful to the National Science Foundation for a grant (CHE-9102972) which has generously supported this work.

REFERENCES AND NOTES

(1) Systematic Synthesis Design. 10. Paper 9: Hendrickson, J. B.; Braun-Keller, E.; Toczko, A. G. *Tetrahedron Suppl.* **1981**, *No. 1*, 359.

(2) The adjacency matrix is described more fully in: Harary, F. "Graph Theory"; Addison-Wesley: Reading, MA, 1969. It is discussed for molecules in paper 3.[3]

(3) Hendrickson, J. B. *J. Am. Chem. Soc.* **1975**, *97*, 5763.

(4) Smith, E. G.; Baker, P. A., "The Wiswesser Line-Formula Chemical Notation (WLN)", 3rd ed.; Chemical Information Management: Cherry Hill, NJ, 1975.

(5) Morgan, H. L. *J. Chem. Doc.* **1965**, *6*, 107.

(6) Read, R. C.; Corneil, D. G. "The Graph Isomorphism Disease" *J. Graph Theory* **1977**, *1*, 339.

(7) Colbourn, C. J. "Bibliography of The Graph Isomorphism Problem". Technical Report 123/78; Computer Science Department, University of Toronto: Toronto, Canada, 1978.

(8) Penny, R. H. *J. Chem. Doc.* **1964**, *5*, 113.

(9) (a) Randic, M. *J. Chem. Phys.* **1974**, *60*, 3920. (b) *Ibid.* **1975**, *62*, 309. (c) *J. Chem. Inf. Comp. Sci.* **1975**, *15*, 105. (d) *Chem. Phys. Lett.* **1976**, *42*, 283. (e) *J. Chem. Inf. Comp. Sci.* **1977**, *17*, 171. (f) Randic, M.; Brissey, G. M.; Wilkins, C. L. *Ibid.* **1981**, *21*, 52.

(10) Mackay, A. L. *J. Theor. Biol.* **1975**, *54*, 399; *J. Chem. Phys.* **1975**, *62*, 308.

(11) (a) Blair, J.; Gasteiger, J.; Gillespie, C.; Gillespie, P. D.; Ugi, I. *Tetrahedron* **1974**, *30*, 1845. (b) Jochum, C.; Gasteiger, J. *J. Chem. Inf. Comp. Sci.* **1977**, *17*, 113. (c) *Ibid.* **1979**, *19*, 49.

(12) Masinter, L. M.; Sridharan, N. S.; Carhart, R. E.; Smith, D. H. *J. Am. Chem. Soc.* **1974**, *96*, 7714.

(13) Carhart, R. E. *J. Chem. Inf. Comp. Sci.* **1978**, *18*, 108.

(14) Dyott, T. M.; Howe, W. J. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 187.

(15) Shelley, C. A.; Munk, M. *J. Chem. Inf. Comp. Sci.* **1979**, *19*, 247.

(16) Bersohn, M. *Comput. Chem.* **1978**, *2*, 113.

(17) The idea of minimal number may have arisen from IUPAC numbering practice which placed atom 1 at a terminus, i.e., a *lowest* valency ($\sigma = 1$) atom.

(18) The 197 references cited in the bibliography[7] are divided into groups by subject focus, and the group of chemical references are all quoted here. However, in another section of mathematical references was found the incidence matrix approach of Proskurowski,[19a] which was not cited in any of the chemical group of papers. We have not located another paper which bears on our method.

(19) (a) Proskurowski, A. *BIT* **1974**, *14*, 209. (b) Overton, M. L.; Proskurowski, A. *Ibid.* **1979**, *19*, 271.

(20) Uchino, M. *J. Chem. Inf. Comp. Sci.* **1980**, *20*, 116, 121.

(21) The maximization is also more efficient in that organic molecules usually exhibit fewer highest valency atoms (especially quaternary) than primary or terminal ones, and hence there are fewer false-start generations of separate matrices at the beginning.

(22) Example D in Figure 3 has been discussed[13,17] as a molecule containing several double bonds and exemplifies the extra complexity inherent in a system which incorporates multiple bonds in the numbering protocol. As already noted, multiple bonds in our system are indicated as atom functionality with π values in the $z\pi$ list appended after the skeletal connectivity in the binary string.

(23) Simmons, H. E., III; Maggio, J. E. *Tetrahedron Lett.* **1981**, *22*, 287. Paquette, L. A.; Vazeux, M. *Ibid.* **1981**, *22*, 291.

A Reasonable Triamantane Rearrangement Path Searched by the Selective Disource Propagation Algorithm[†]

NOBUHIDE TANAKA

Gakushuin University Computer Center, Mejiro, Tokyo, Japan 171

TADAYOSHI KAN

Department of Physics, Faculty of Science, Gakushuin University, Mejiro, Tokyo, Japan 171

TAKESHI IIZUKA*

Department of Chemistry, Faculty of Education, Gunma University, Maebashi, Gunma, Japan 371

Received July 5, 1983

Triamantane rearrangement reactions reported by McKervey et al. are studied theoretically. A great many isomers must be considered to find rearrangement paths. A searching procedure by a graph-theoretical method and computational techniques along with force field calculations, "selective disource propagation algorithm", is developed to find a path efficiently. Several reasonable rearrangement paths with about 15 steps are found and are shown together with the structures of the isomers in these paths. These isomers and paths are expected to be experimentally verified.

INTRODUCTION

Some rearrangement reactions of the first member of the diamondoid hydrocarbons, adamantane, were studied graph theoretically by Whitlock and Siefken.[1] They studied relationships obtained by the 1,2 alkyl shift between 16 adamantane isomers. In addition to a graph-theoretical study of adamantane rearrangement reactions, Schleyer et al.[2] adopted empirical force field calculations to estimate rearrangement paths. They applied their method to rearrangement reactions of diamantane, the second member of the diamondaoid hydrocarbons, and searched for rearrangement paths by computer.[3] They generated more than 20 000 isomers of diamantane and found reasonable paths by calculating their strain energies. Ōsawa et al.[4] searched rearrangement paths of tricycloundecane, a homologue of adamantane, by using the same method mentioned above. We were also interested in enumerating isomers[5] of adamantane analogues and studying rearrangement relationships between them.[6] McKervey et al.[7] succeeded in synthesizing triamantane, the third member of the diamondoid hydrocarbon series,[8] from the two elaborated pentacyclic hydrocarbons **1** and **2** through rearrangement reactions catalyzed by $AlCl_3$ (Figure 1). The rearrangement path in the synthesis had not yet been found experimentally. We were tempted by this report to use our path-finding algorithm, the disource propagation algorithm (DSPA),[9] by which we succeeded in finding diamantane rearrangement paths efficiently. Soon after we started, we found that it is

[†] This work composes a part of Dr. Tanaka's thesis.