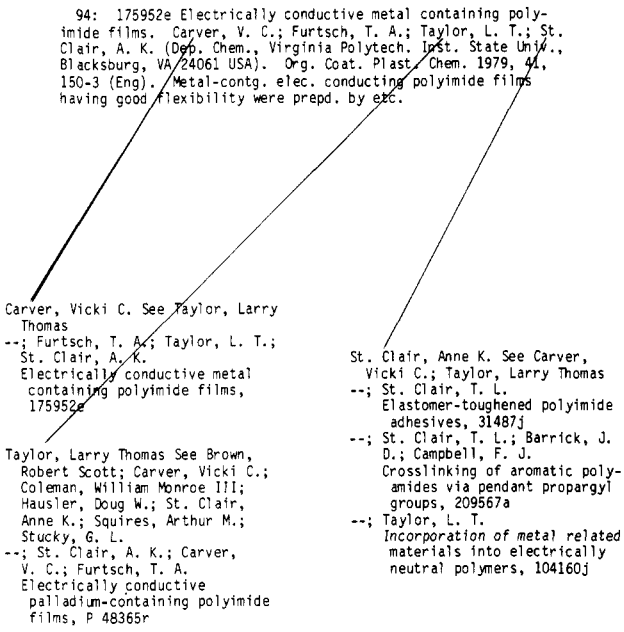


Figure 2. Relationship between author index entries and abstract.

adequate, job without overloading the students with outside work. Even with the restricted nature of the coverage, the work time required of the student is still relatively high for a one-quarter course. Several important areas do not receive the in-depth treatment that they really need. Such areas include patents, government documents, *Chemical Abstracts* supplemental sources, strategies for searching, and the use of *Beilstein*. It would be possible to change the order in which the material is presented, discussing some of the more general sources of information first, which would be somewhat more logical. However, this would limit the use of a retrospective search as an assignment in the course. The overall response of the students is that while they found the course to be very time-consuming, they, indeed, consider it to have been very beneficial to their training, and despite the restricted coverages, it served the purpose of providing a base from which they can operate in the future.

REFERENCES AND NOTES

(1) "Chemical Abstracts Service Source Index", CASSI; Chemical Abstracts Service: Columbus, OH, 1979.
(2) "Chemical Abstracts Index Guide"; Chemical Abstracts Service: Columbus, OH, 1982.
(3) Maizell, R. E. "How to Find Chemical Information"; Wiley: New York, 1979.
(4) Bottle, R. T. "Use of Chemical Literature"; Butterworths: London, 1979
(5) Herner, S. "A Brief Guide to Sources of Scientific and Technical Information", 2nd ed.; Information Resources Press: Arlington, VA, 1980.
(6) Woodburn, H. M. "Using the Chemical Literature"; Marcel Dekker: New York, 1974.
(7) Mellon, M. G. "Chemical Publications", 5th ed.; McGraw-Hill: New York, 1982.
(8) "Information Tools"; Chemical Abstracts Service: Columbus, OH, 1978.
(9) "Searching CA"; Chemical Abstracts Service: Columbus, OH, 1981.
(10) "How to Use Beilstein"; Springer-Verlag: New York, 1981.
(11) "This is Gmelin"; Springer-Verlag: New York, 1981.

Condensed Structure Identification and Ring Perception[1]

JAMES B. HENDRICKSON,* DAVID L. GRIER, and A. GLENN TOCZKO
Edison Chemical Laboratory, Brandeis University, Waltham, Massachusetts 02254

Received February 10, 1984

For acyclic graphs, the full connectivity is uniquely represented by a T-list of the row sums from the maximal adjacency matrix, requiring only $2n$ bits in length. Polycyclic graphs are represented by the T-list of the maximal spanning tree as well as an R-list of the row sums of ring-closure bonds on that tree, also requiring only $2n$ bits for storage. The combined T/R-list ($4n$ bits) provides for unique reconstruction of the graph in most cases, and in all cases when an offset number (ON) is appended to identify the rank order in the few instances for which duplication occurs. A procedure for reconstructing the matrix, and hence the graph itself, from the T/R-list (and ON) is presented. A rapid protocol for perception of the smallest set of smallest rings (SSSR) in the graph also derives from the maximal matrix. All the procedures are contained in a program (TRGEN) written for minicomputer.

We recently showed that a chemical structure, or any graph, can be uniquely numbered by creating its maximal adjacency matrix.[2] This maximal matrix is created by assigning numbers to the atoms, or graph points, in such a way that each row in the matrix, considered as a binary number, must be the maximum possible number. The full binary number representing the molecular skeleton is obtained by stringing out each successive matrix row of 1/0 entries to the right of the diagonal into a single list, to be seen as a binary identification number for that skeleton. This number is unique because it is the maximum possible binary number among all the possible $n!$ matrices representing different numberings of the skeleton. When the skeleton, or graph, has isomorphic atoms or points, the same maximal matrix represents each equivalent numbering, and it was shown that such isomorphic points can be easily identified via equivalency classes.[2]

The nature of the maximal adjacency matrix has some further properties, explored here, which allow an unusually condensed form of representation of the matrix for compact storage and also a very simple basis for determining the smallest set of smallest rings (SSSR)[3] in the molecule. In general, the adjacency matrix is a symmetrical $n \times n$ matrix, of 1 entries indicating bonds or connections between atoms and of 0 entries for no connection. Since the maximal matrix is created row by row to make each row a maximum binary number, the effect is to press the 1 entries as far up and to the left as possible (above the diagonal) and so to leave an extensive region of 0 entries to the upper right in the matrix. It is this effect that allows a much shorter representation of the matrix to be written and also allows the smallest rings to be identified readily and rapidly.

Acyclic Molecules: The T-List. Let us take first the case of the acyclic molecules, or tree graphs. The procedure for creating the maximal matrix is to assign the numbered rows in the (empty) matrix successively to atoms, in such a way as to make each row a maximum. This requires that each row

0095-2338/84/1624-0195$01.50/0

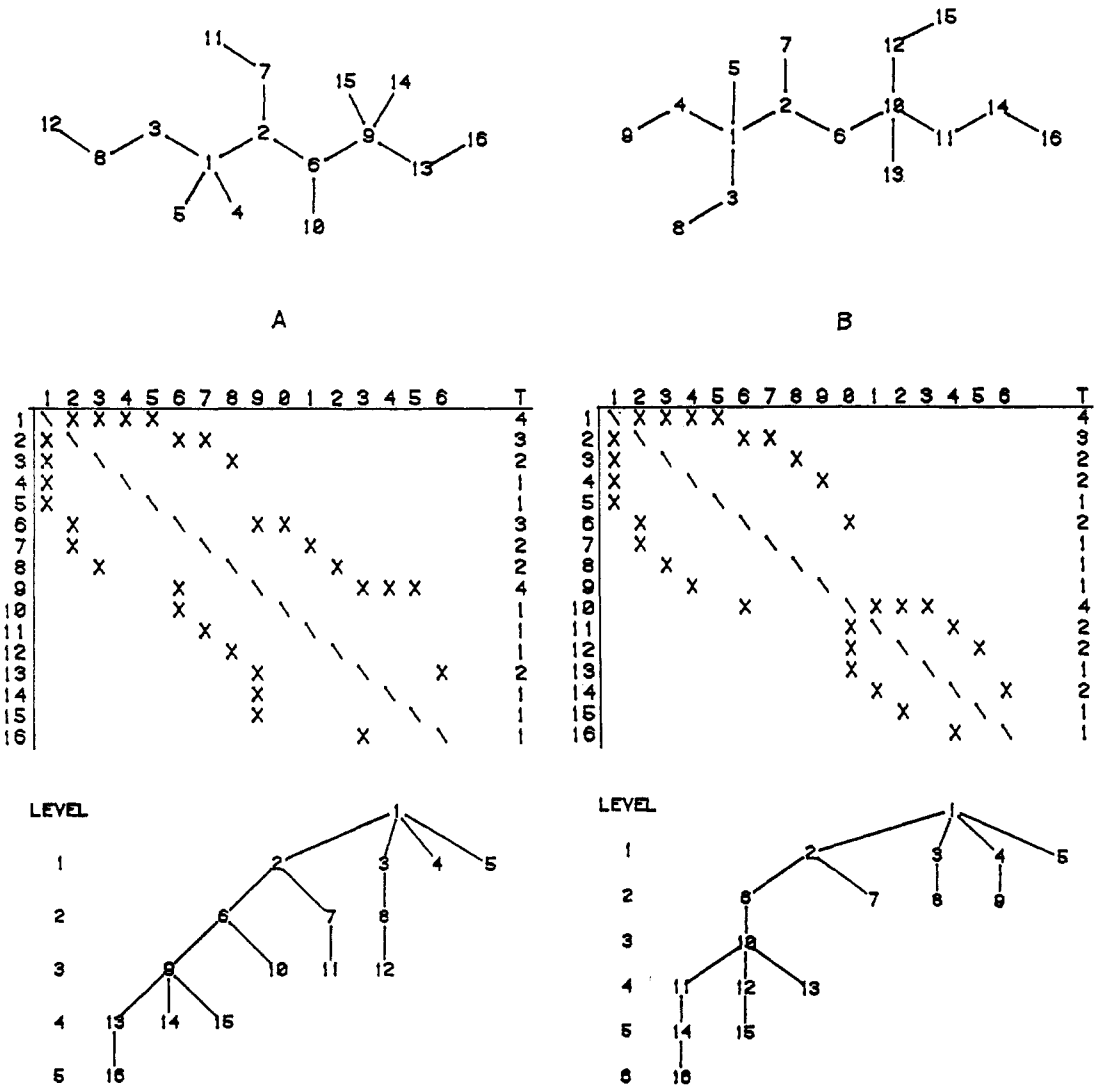


Figure 1. Maximal matrix and growing tree for acyclic molecules.

will have 1 entries as far to the left as possible and be assigned to the highest valent atom available. Hence, atom 1 will have the highest valence (v) and be attached to atoms 2, 3, 4, and 5 (for $v_1 = 4$) so that row 1 will read 1111000... (to the right of the diagonal). Hence, the atoms numbered from 2 to 5 are all attached to atom 1. This process is described in detail in reference 2 but in general, the maximization requires that each newly assigned atom be attached to the lowest numbered previously assigned atom, regarded as its parent atom. This places 1 entries as far left in the parent row as possible. In simple linear skeletons, this will place atom 1 in the center; hence, the numbering for n-heptane is 7531246.

This process of successively assigning atoms in the matrix is paralleled structurally by a stepwise generation of a growing tree, isomorphic with the given skeleton, illustrated with the examples in Figure 1. In the first example (A), atom 1 is tetravalent, hence attached to atoms 2, 3, 4, and 5 in valence order. These represent the first level of the growing tree below. Atoms 6 and 7 are then assigned to atom 2, 8 to 3, and none to atoms 4 and 5. This completes the second level of the tree, i.e., all atoms grown from the first-level parents. Assignment continues in this fashion until the full matrix is created; the growing tree is complete to five levels of generation and isomorphic with the given skeleton of 16 atoms. Initially, a second matrix generation would have been begun, in competition, numbering the other tetravalent atom as 1 but eliminated at row 7 (with no new entry) as less than maximal.[2] Another, isomeric, skeleton is maximally numbered in example B, with its matrix and isomorphic tree grown row by row in the same way.[4]

The form of the maximal matrix for an acyclic skeleton has important consequences for more condensed representation. In the growing tree each point (atom) has one and only one connection back to its (lower-numbered) parent in the prior level. Hence, each column has one and only one entry above the diagonal, and conversely, each row has one and only one entry below the diagonal. The atom numbers on each level are successive to and higher than all those on the previous level. The valence (v_i) of each atom (i) is shown by the row sum, and an ordered list of these valences (v_i) for a tree graph is then called a T-list, indicated vertically to the right of each matrix in Figure 1. The numbers in the T-list have values of 1–4, and no number in the list may exceed the first. The sum of the T-list is twice the number of bonds or $\sum_i v_i = 2(n-1)$, where n = number of atoms. *This T-list is fully characteristic of any acyclic skeleton*, and the skeleton may be reconstructed from the T-list. Since each matrix row has only one 1 entry below the diagonal, the number above the diagonal is $(v-1)$ for each row except for row 1, which has all entries above.

To recreate the matrix from the T-list, then, we place v_1 1 entries in the first columns of row 1 and then successively add $(v_i - 1)$ 1 entries to each row i, filling successive columns with one entry each from left to right. The corresponding entries (one per row) are then filled in below the diagonal to create a symmetrical $n \times n$ matrix. In the example of Figure 1A, the T-list 4321132241112111 fully characterizes the structure;

the matrix is created from it by entering four 1 entries in the leftmost four places after the diagonal in row 1 (ending in column 5), then two entries for row 2 (in columns 6 and 7), one for row 3 (column 8), none for rows 4 and 5, then two for row 6 in columns 9 and 10, etc., until all columns have received one entry each, placed as high and to the left as possible in accord with maximization. The skeleton or graph itself may then of course be regenerated from the matrix, most easily in the isomorphic form of the growing tree, illustrated in Figure 1.

This T-list therefore contains all the information of the acyclic skeleton itself and so represents an unusually compact identification number to characterize it uniquely. Each row sum has values of 1–4, hence requires only two binary digits (bits) for expression. The whole T-list is then only $2n$ bits in length compared to $(1/2)n^2$ bits necessary to list the whole matrix; i.e., it is a linear function rather than a square. It is not only linear in n but also much more compact than an ordinary connectivity table, which requires $20n$ bits to represent a skeleton of under 32 atoms. It may further be noted that the string of 1 values at the end of any T-list represents the last numbered terminal atoms and may be left off for an even shorter list, but one for which the list length does not tell the number of atoms. However, in an acyclic skeleton, the total number of terminal atoms ($v = 1$) is related to the number of tri- and tetravalent atoms by $N_1 = 2N_4 + N_3 + 2$, where N_v is the number of atoms of valence v. Hence, in a given truncated T-list the number of missing univalent atoms at the end of the list can be calculated. Alternatively, the number of atoms in the skeleton is given by $n = 3N_4 + 2N_3 + N_2 + 2$ and so the whole skeleton is still characterized uniquely by a shorter, truncated T-list in which the terminal 1 digits are omitted.

Polycyclic Molecules: The R-List. In polycyclic graphs or molecular skeletons the ring closures affect the maximal numbering assignments as well as the valency and connectivity seen in the acyclic ones. The matrix is generated by the same rowwise maximization procedure for assigning maximal numbering,[2] but the ring-closure entries now affect the ordering. However, the skeletal graph may be divided into a maximal spanning tree and a set of ring-closure links on it so that the number of these ring-closure links equals the number of rings in the skeleton. This division is reflected in the overall matrix (**M**), in which the spanning tree entries and the ring-closure entries each occupy a defined space in the matrix and can be easily separated into a **T** matrix and an **R** matrix such that $\mathbf{M} = \mathbf{T} + \mathbf{R}$. The **T** matrix is that of a unique spanning tree (linking all atoms) and like the matrix of any acyclic graph is represented above the diagonal by the top entry in each column and symmetrically below it by the first entry in each row. These **T** matrix entries may be separated from the rest by two demarcation lines dividing the space in matrix **M** into the **T** matrix outside the demarcation lines and the **R** matrix between them. The numbers of rings is then the number of entries in the **R** matrix space above the diagonal.

The row sums across the **T** matrix are now the same as before, constituting a T-list of the valences (t_i) of the maximal spanning tree above, without the ring-closure bonds. The remaining set of ring-closure bonds, i.e., the **R** matrix, is then represented by a separate R-list of row sums (r_i) such that the sum of the two digits for any row represents the full valence (v_i) of atom i and constitutes a V-list, the list of valences of the ordered atoms ($v_i = t_i + r_i$). Thus, the sums of list digits are $\sum_i t_i = 2(n-1)$, $\sum_i r_i = 2\rho$ (where ρ = number of rings), and $\sum v_i = 2b$ (where b = number of bonds), i.e., $b = \rho + n - 1$.

This division of entries is shown in Figure 2 for a polycyclic skeleton: the **T** matrix is shown as X entries, the **R** matrix is shown as 0 entries, and the two heavy demarcation lines are drawn just within each **T** matrix entry. The maximization procedure[2] assigns the numbering to the atoms rowwise and concurrently grows the maximal spanning tree shown to the right (solid lines). Each atom in the spanning tree has only one link back to a parent atom in the previous level, as before. Any link on the same level or a new link down to the next level (to a higher numbered atom than the parent) is a ring-closure bond and is shown dotted on the initial given skeleton and dotted and overlaid on the generated spanning tree. These ring-closure bonds always occur between atoms in the same level or in adjacent levels.

In the stepwise growing of the spanning tree from a given matrix, the tree bonds are established before the ring-closure bonds, but the actual atom identities (numbers) on each level will be affected by the ring-closure positions. Thus, a set of four bonds from atom 1 is generated to atoms 2–5, but the assignment of numbers among the four now depends on the 2–3 ring closure, giving the maximum number to row 2. A three-membered ring containing a maximum-valence atom in the skeleton will always appear as a ring closure in row 2 and conversely force the numbering of atoms 1, 2, and 3.[6] In the example (Figure 2), it is this smaller ring which gives precedence to the tetravalent atom labeled 1 over the other one labeled 7; this precedence is established by the second row maximization.[2] Similarly, the presence of any ring will at some stage in the maximization affect the canonical atom numbering and so affect the final choice of the maximal spanning tree.

Once the maximization of the matrix is complete, the maximal spanning tree of any cyclic skeleton is uniquely characterized by the T-list (as with any acyclic skeleton) and may be directly generated from it, as described above. Examination of Figure 2 shows, however, that the spanning tree is not necessarily numbered maximally for that acyclic molecule standing alone. The spanning tree in Figure 2 has a T-list of 4223123112111, but maximal numbering for that skeleton without its ring closures would have resulted in a T-list of 4322111321121, the same digits but ordered differently (atom 4 becomes 2, etc.). Thus, the nature and placement of ring closures affect the choice of spanning tree.

The stepwise maximization is readily done by hand with the recognition that atom 1 must always have the highest valence and be contained in the largest number of smallest rings. In most cases, the remainder of the numbering is then rapid and straightforward.[2]

Reconstruction of Skeletons from T/R-List. The R-list in most cases is enough to regenerate the ring-closure positions across the spanning tree unambiguously. In some cases, however, multiple possibilities exist. With a simple added rank order number developed below, however, the two lists create a very compact identification number uniquely characterizing any polycyclic skeleton. Reconstruction of a cyclic skeleton or graph from its T/R-list requires a manipulation procedure, the price paid for the compactness of this skeletal identification. In most common molecular cases, this manipulation is trivial, easily carried out by hand, and unambiguous.

Given the T/R-list, the empty $n \times n$ matrix is first set up with the T- and R-lists vertically at the right as in Figure 2, and the spanning tree entries are placed in the matrix from the T-list as for any acyclic skeleton (above). The demarcation lines are then added just inside the tree entries above and below the diagonal to block off the **R** matrix space between them. The levels of the spanning tree are then drawn in across the matrix and across the T- and R-lists as well. Level 0 is row 1; level 1 is rows $2 - (v_i + 1)$, i.e., the rows of atoms attached to atom 1; level 2 is the next group of rows of atoms attached to level 1 atoms; level 3 rows are those attached to atoms of level 2; etc.

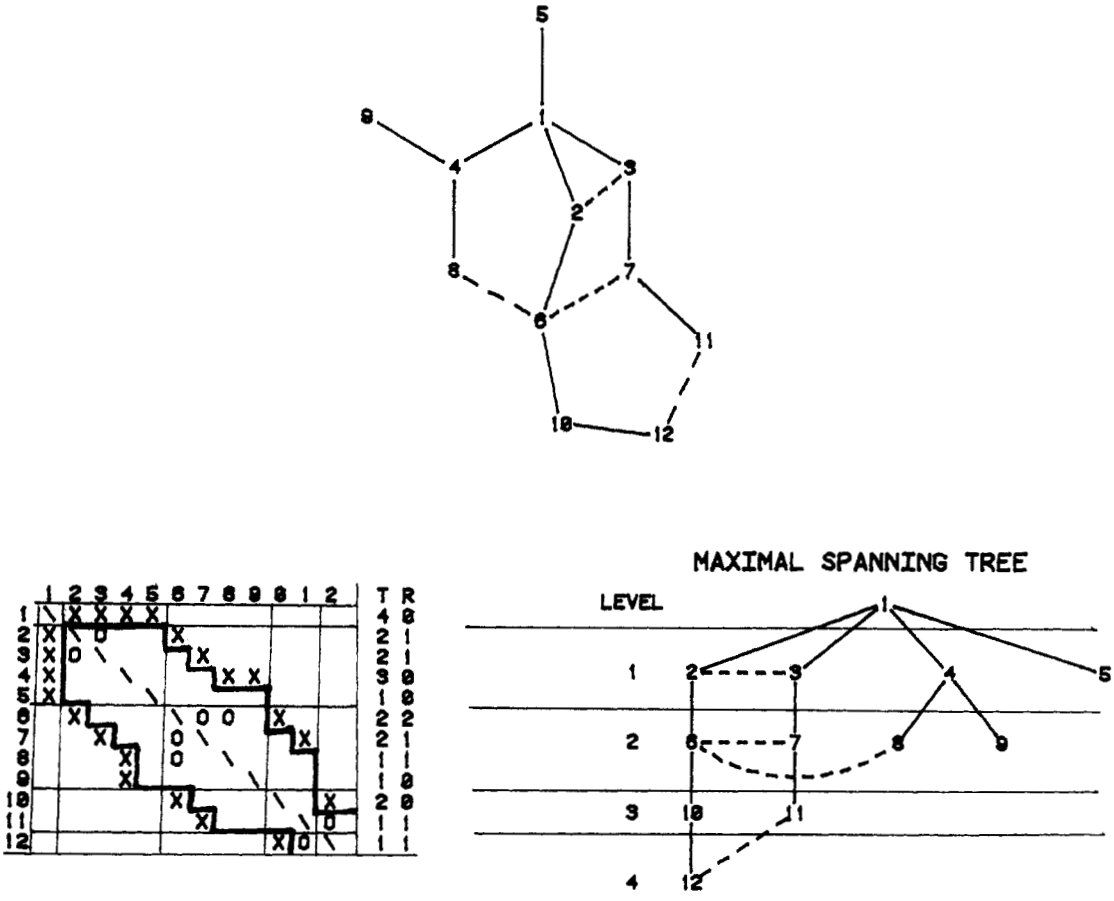


Figure 2. Maximal matrix and spanning tree for a polycyclic molecule.

To fill in the **R** matrix, we proceed stepwise down the R-list from the second row ($r_1 = 0$ always), i.e., r_2, r_3, r_4, ..., examining non-zero r_i numbers. Each such value of r_i is the number of **R** matrix entries a_{ij} in that row i and above the diagonal $j > i$. The available places for new entries in any row are those between the diagonal and the upper demarcation line. Each **R** matrix entry a_{ij} removes one unit from r_i and one from r_j in the R-list, and no entries in row i may be placed in any column j with $r_j = 0$. This pairing must occur only between two r numbers in the same or adjacent levels, the former preferred for maximal matrix numbering. If $r_2 = 0$ all r_i numbers in level 1 must be paired with r_i numbers in level 2, however.[6] Entries a_{ij} are placed in the matrix in the leftmost available positions above the diagonal in each row i, the symmetrical a_{ji} is also entered below the diagonal, and a unit is removed from each of the paired r_i and r_j numbers in the R-list. In this way, the R-list numbers are successively paired and assigned to the matrix as entries a_{ij} (and a_{ji}) until all have been successfully paired, and the **R** matrix entries are placed as high and to the left as possible in accord with matrix maximization.

Thus, in reconstructing Figure 2 from its T/R-list, after the T-list entries and demarcation lines and levels have been added to the matrix, the first R-list numbers are $r_2 = r_3 = 1$, which are paired as matrix entry 23 (and 32), upper left in level 1. The next number ($r_6 = 2$) requires pairing with $r_7 = 1$ and $r_8 = 1$ for level 2, leaving only two remaining R-list numbers, $r_{11} = r_{12} = 1$ to pair across levels 3 and 4 for the final ring closure. In Figure 3A is shown a stepwise reconstruction from the T/R-list of a pentacyclic six-atom skeleton, starting with T-list entries (X), levels, and demarcation lines drawn in. Since $r_2 = 2$, two matrix entries (A) are added at 23 and 24, cancelling one each from r_3 and r_4 to give a remaining R-list R_A = 001131. Going to row 3, we can pair $r_3 = r_4 = 1$ to assign the leftmost matrix entry (34), shown as B on the matrix. This leaves an R-list R_B = 000031, which cannot be further paired. Hence, the leftmost entry (B) in row 3 cannot be used, and the next (35) is selected, shown as entry C on the matrix. This in turn leaves an R-list R_C = 000121, which can now be fully paired only as 45 and 56, entries D on the matrix, and C and D are also entered below the diagonal. The final matrix (X, A, C, D) results in the skeleton shown at the right, first by growing the spanning tree and then identifying the ring-closure bonds on it (dotted). This is then redrawn as a planar isomorphic graph. Other examples are illustrated in Figure 3 with the T/R-list and the uniquely generated corresponding skeleton, or graph.

Offset Numbers. The procedure so far is simple, and correct for most skeletons, placing each new **R** matrix entry as far left on its row as is consistent with pairing of all R-list numbers and so achieving a maximized row. This implies that the first row entry (a_{ij}) for $r_i > 0$ will always be next to the diagonal, i.e., $j = i + 1$, as long as $r_j > 0$ and all other R-list numbers can be paired. However, in some cases a leftmost entry in some row will result in a matrix that on remaximizing will reorder its atom numbering to afford a different matrix with a higher binary list and a different T/R-list altogether, i.e., a different maximal spanning tree. Hence, the procedure for reconstruction must include a check that remaximizing yields the same matrix. Procedurally, it is enough to remaximize only down through each new row when its **R** matrix entry is assigned.

Furthermore, there are occasionally instances in which more than one maximal graph can be generated from one T/R-list. These will have the same maximal spanning tree and will exhibit pairwise swapping of ring closures low enough on the matrix so as not to alter the row sums of the R-list either. Such cases appear to be uncommon. Among the 78 connected graphs that can be drawn on six points[7] (with $V_{max} = 4$), only one case emerges, shown at the top of Figure 4. Row 2 has two entries maximally placed next to the diagonal. Row 3 has

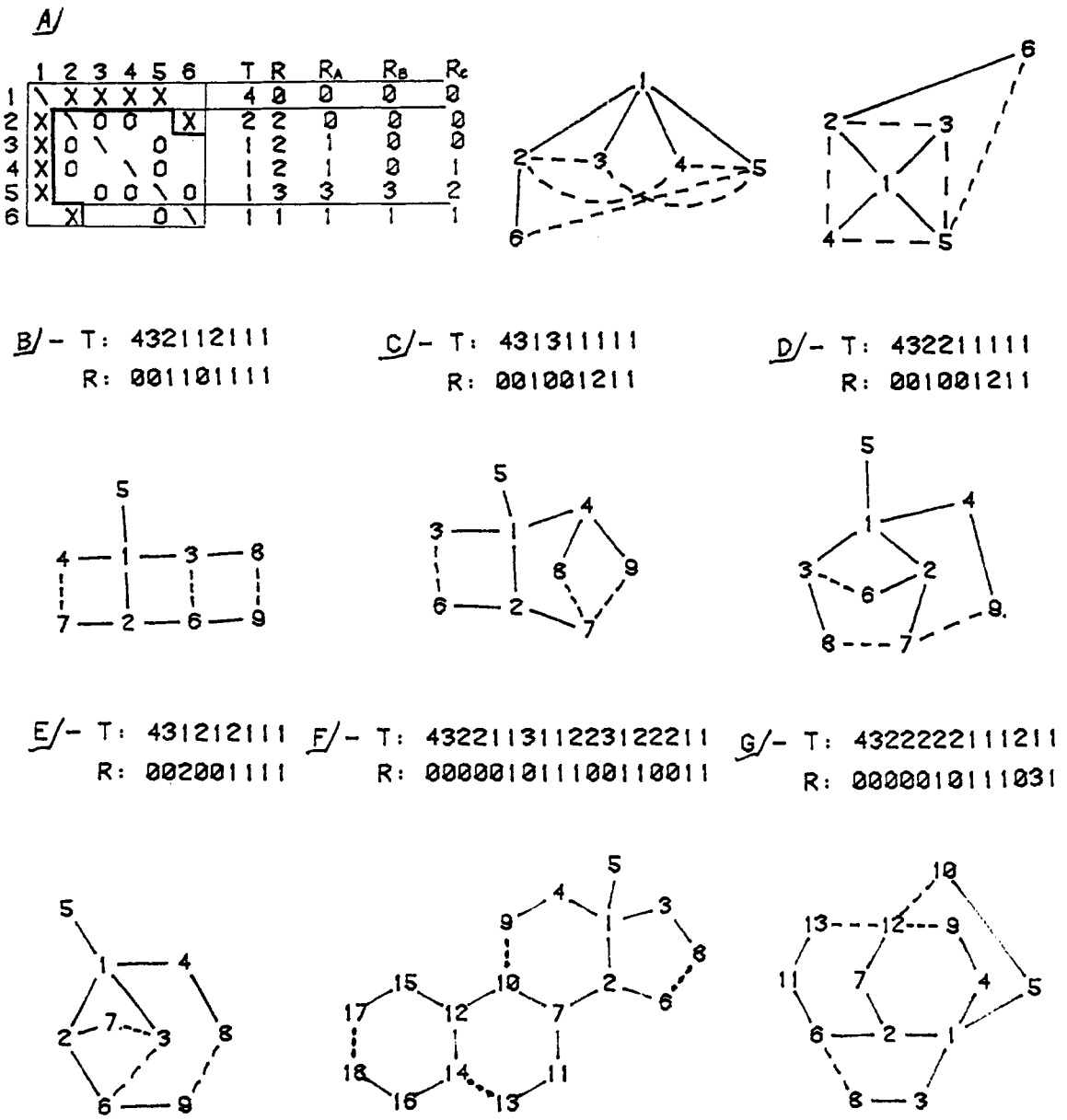
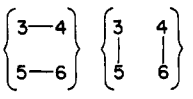


Figure 3. Creation of unique skeletons from T/R-lists.

three viable choices, 34, 35, and 36, leaving the remaining ring at 56, 46, and 45, respectively. The first two show pairwise swapping of



and this alters neither the T- nor R-list. Both are viable structures (Figure 4) from the same T/R-list. The first, however, is distinguished by a matrix with a larger binary list number than the second; i.e., row 3 is 100 in the first and 010 in the second. The third choice, with **R** matrix entries at 36 and 45 has row 3 equal to 001 and represents another, less than maximal, numbering of the second one, hence is not a viable (maximal) matrix.

These cases of multiple skeletons for one T/R-list are generated successively from their T/R-list in the procedure described above and so turn up in descending numerical order of their binary list numbers. Hence their specific identification can be attained by assigning them rank orders, or *offset numbers* (ON), indicating the order in which they appear in a full reconstruction procedure (all viable a_{ij} permutations in the **R** matrix). As in the example above from Figure 4, some that appear are not maximally numbered and are rejected when the remaximizing does not recreate the same matrix. Usually, there is only one skeleton corresponding to a T/R-list, as in the Figure 3 samples, and so ON = 1. The T/R(ON) identifications for the two examples at the top of Figure 4 are 421111/022211(1) and 421111/022211(2), respectively. When a given graph has once been identified by a T/R(ON), however, it is sufficient to reconstruct it uniquely from that T/R(ON) by stopping when the procedure has arrived at the ONth viable matrix. Most common molecules (with five rings or less) have ON = 1, and offset numbers above 5 are exceedingly rare since the number of possible viable ring-closure permutations is very limited. The stepwise procedure is summarized in Table I.

In the second part of Figure 4 is shown a T/R-list for a hexacyclic 10-atom skeleton. On reconstruction of the matrix from the given T/R-list, an initial, leftmost assignment of the row 5 entry to 56 and remaximizing yield a different matrix with a T/R-list of 3222231111/011102222. The next assignment of row 5 to 57 yields the same matrix on remaximizing, as does further assignment of row 6 to either 67 or 68. The former has only one possible further pairing, 67, 89, and 810, shown as entries A in the matrix (Figure 4). The latter (68) has only two further pairings, shown as entries B and C, respectively. A further assignment of row 5 to entry 58 yields

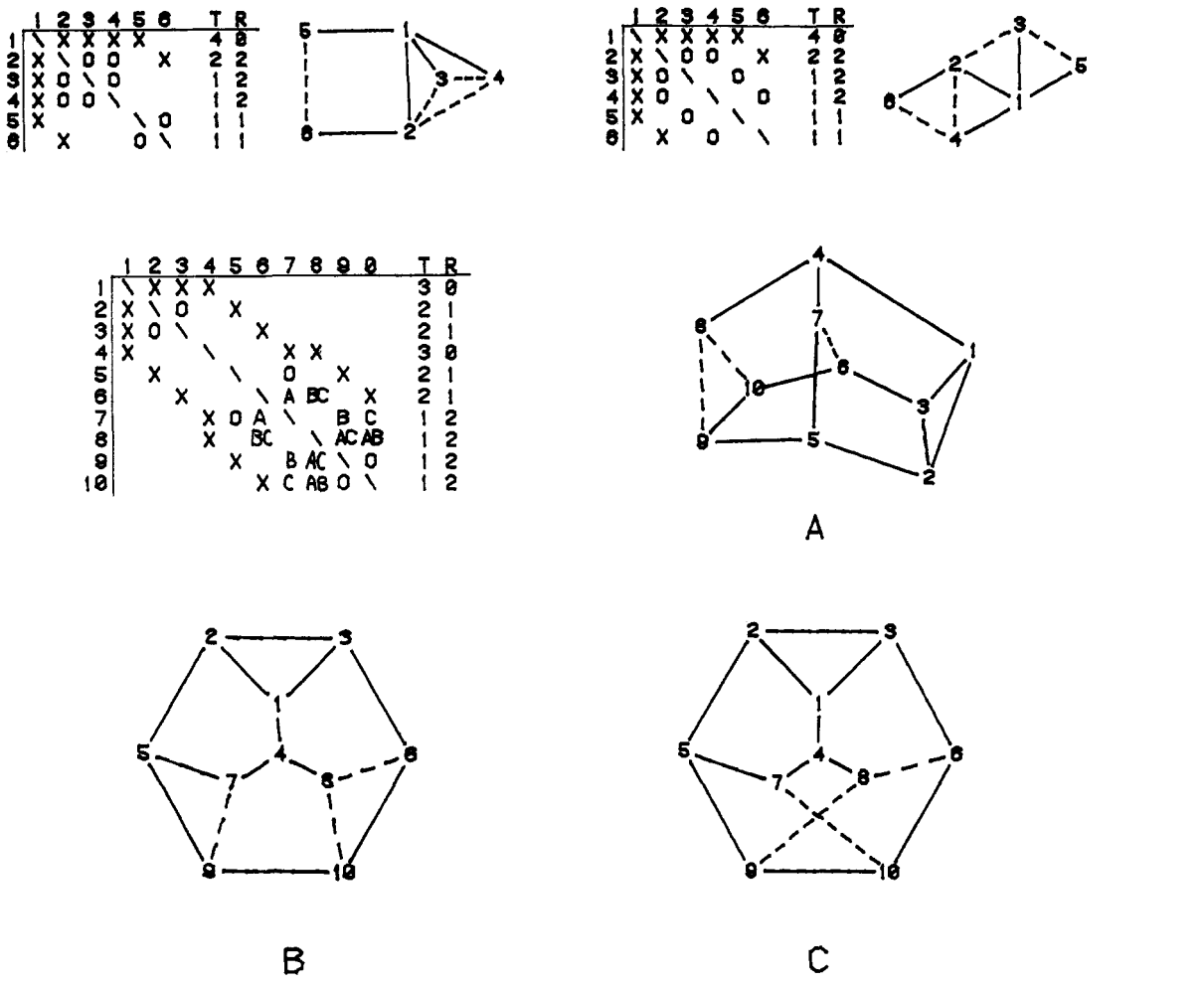


Figure 4. Multiple isomers with single T/R-lists.

only the same skeletons again, less than maximally numbered and so not viable. The order in which matrices **A**, **B**, and **C** appear in the procedure corresponds to the numerical order of their binary list numbers and so to their offset numbers: ON = 1(A); ON = 2(B); ON = 3(C). The three corresponding skeletons are drawn out with only the variable ring-closures dotted.

In a catalog of structures by T/R(ON), the offset number (ON) serves to indicate which of the ordered isomers with that T/R-list is identified. The T-list requires only $2(n-1)$ bits for computer storage, and the R-list is the same; so that with an appended offset number of 1–16 (4 bits) to cover multiple isomers, the entire storage of any polycyclic skeleton can be accommodated in only $4n$ binary bits. Even shorter storage is required if terminal units in the T-list are truncated, as described above, although slightly more computation is then required.

A simple and fast program (TRGEN) has been written in FORTRAN to maximize a given molecular skeleton or graph (of $v_{max} = 4$), to create its T/R-list, then to rebuild its matrix, remaximizing row by row to secure only maximal matrices and check their identity, then to append the ID-number if multiple isomers appear, and finally to draw both the matrix/matrices and generated graphs as ring-closed maximal spanning trees. The illustrations shown were drawn by this program. The computation times on our relatively slow DEC 11/23 minicomputer are generally less than 0.5 s for common structures, either to establish the T/R(ON) from the structure or to reconstruct the matrix and structure from the list. On the faster DEC PDP-10 computer, these times are on the order of 30 times faster. When several isomeric structures are possible for one T/R-list, the search is longer, as much as 10 s on the minicomputer. Specific times for the examples shown are as follows: 0.20 s, Figure 2; 0.11–0.31 s, Figure 3; 0.62–1.88 s for the hexacyclodecanes, Figure 4.

Table I. Summary of Procedure: Structure T/R(ON)

(A) Determine T/R(ON) (from Structure)

1 given structure, develop maximal matrix, maximizing row by row as atoms are assigned[2]
2 draw two demarcation lines under top element/each column; to right of first element/each row
3 tree matrix = outside demarcation lines: above upper line, below lower line. T-list = list of row sums/T matrix; sum of list entries = $2(n-1)$.
4 ring closure matrix = between demarcation lines; R-list = list of row sums/**R** matrix; sum of list entries = 2ρ
5 offset number (ON): reconstruct full matrices from T/R-list as in (B), retaining all with same matrix on maximization
6 assign ON to candidates with same T/R-list, in numerical order of their appearance in the reconstruction, i.e., their binary list number order

(B) Reconstruct Matrix from T/R(ON)

1 in first row of matrix fill in successive elements equal to first T-list entry (t_1)
2 in successive rows fill in $t_i - 1$ elements successively, starting with first empty column
3 fill in (one per row) corresponding elements below diagonal
4 group rows by levels: first level = all rows of atoms attached to 1; second level = all rows of atoms attached to first-level atoms; etc.
5 pair R-list entries (down from top) to enter leftmost elements/each row with all R-list entries fully paired
6 remaximize down through each new row as it is assigned; retain only if original matrix is produced
7 the first such viable matrix is ON = 1; successive ones have successive ON values in order

Ring Perception: Smallest Set of Smallest Rings (SSSR). There are many rings, or cyclic paths, to be found in any polycyclic molecule, but many are simply envelopes of smaller rings. A fundamental set of rings is one that incorporates all ring bonds and atoms and consists of a minimum number of rings easily defined by $\rho = b - n + 1$. When these are also the smallest rings, which are not envelopes of any others, then this is the smallest set of smallest rings (SSSR). This problem has been addressed by a number of authors, and many systems have been developed for definition of these rings.[8-17] In general, these systems begin from an arbitrary ring or set of rings and come to a final set by examining all possibilities, either via path tracing or by *exclusive or* comparisons.

A common start is to define an arbitrary spanning tree and then the fundamental set of rings from the ring-closure bonds across that tree. However, the maximization procedure used here orders the atom numbering and the matrix and so defines a single, unique maximal spanning tree and its attendant ring-closure bonds. This creates an ordering of rings as well, so that an arbitrary starting place for search is unnecessary.

The maximization procedure obliges all bonds to appear as matrix elements as far up and to the left as possible above the matrix diagonal. The maximization proceeds by rows and in any row (above the diagonal) defines the ring-closures first, i.e., before, or to the left of, the spanning tree bonds. The smallest rings are those that link previously defined atoms soonest and, hence, are favored by the maximization procedure, which puts their ring-closure entries as far up and to the left as possible. Thus, it is a property of the maximal matrix that with each row an encountered ring-closure entry usually defines one of the smallest rings in terms of entries above and to the left in the matrix.

A basis set of rings is one in which no ring may be created by linear combinations of any other rings in the set. Such a basis set is first generated from the ring closure bonds on the maximal spanning tree. For each ring-closure bond, taken in order, the smallest ring enclosing that bond is generated from tree bonds and any prior ring-closure bonds, i.e., moving back up the spanning tree from the ring-closure bond until a (smallest) ring is defined. Although this is a basis set, the rings defined are not all necessarily the smallest rings, since at the time of definition of any ring the later ring-closure bonds in the order are not yet available to it.

A matrix ($\rho \times b$) is thus created of the numbered rings (ρ) as rows against the list of all ring bonds (b) as columns, each ring row of the matrix annotated (1/0) with the bonds which that ring contains. The unique bonds in each ring are those that do not appear in any other rings, i.e., columns of this matrix with sums of 1. Now, passing through the rings in order, we define the *smallest* ring containing all the unique bonds of a given matrix row. If this is the same as the ring previously defined by the row (as it generally is), that ring is defined. If a smaller ring can be so defined, it replaces the original, and the unique bonds of subsequent rings in the order are then redetermined. When this iterative examination is complete, the smallest set of smallest rings has been found, as long as every ring has at least one unique bond. This procedure serves in nearly all molecular graphs. The three 10-atom systems of Figure 4 (A–C) thus produced the final rings sets (SSSR) displayed in Figure 5. For each ring, the first two atom numbers identify its corresponding ring-closure bond, numbered in order from the maximal matrix.

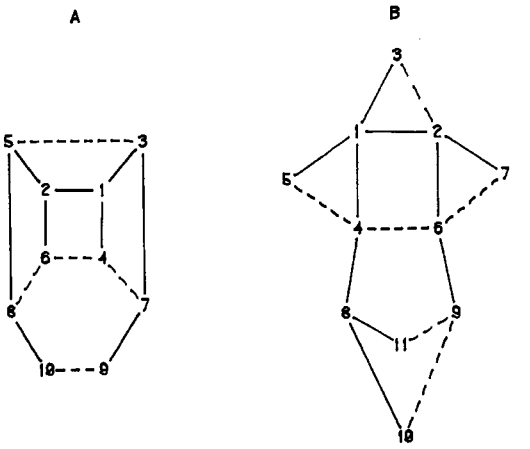
In this procedure, each fundamental ring is redefined by its unique bonds and is in fact the smallest circuit containing those unique bonds. It is possible that this circuit may no longer contain the original ring-closure bond that first defined it. However, the number of ring-closure bonds is the same as the number of rings and so it is easy to reshuffle their numbering

```
STRUCTURE - A
RING NO:  1  ATOMS:  2  3  1
RING NO:  2  ATOMS:  5  7  4  1  2
RING NO:  3  ATOMS:  6  7  4  1  3
RING NO:  4  ATOMS:  8  9  5  7  4
RING NO:  5  ATOMS:  8 10  6  7  4
RING NO:  6  ATOMS:  9 10  8

STRUCTURE - B
RING NO:  1  ATOMS:  2  3  1
RING NO:  2  ATOMS:  5  7  4  1  2
RING NO:  3  ATOMS:  6  8  4  1  3
RING NO:  4  ATOMS:  7  9  5
RING NO:  5  ATOMS:  8 10  6
RING NO:  6  ATOMS:  9 10  8  4  7

STRUCTURE - C
RING NO:  1  ATOMS:  2  3  1
RING NO:  2  ATOMS:  5  7  4  1  2
RING NO:  3  ATOMS:  6  8  4  1  3
RING NO:  4  ATOMS:  7 10  9  5
RING NO:  5  ATOMS:  8  9  5  7  4
RING NO:  6  ATOMS:  9 10  6  8
```

Figure 5. Smallest set of smallest rings. Three examples from Figure 4.




```
FINAL RING SET
RING NO:  1  ATOMS:  3  5  2  1
RING NO:  2  ATOMS:  4  6  2  1
RING NO:  3  ATOMS:  4  7  3  1
RING NO:  4  ATOMS:  6  8  5  2
RING NO:  5  ATOMS:  9 10  8  5  3  7
```

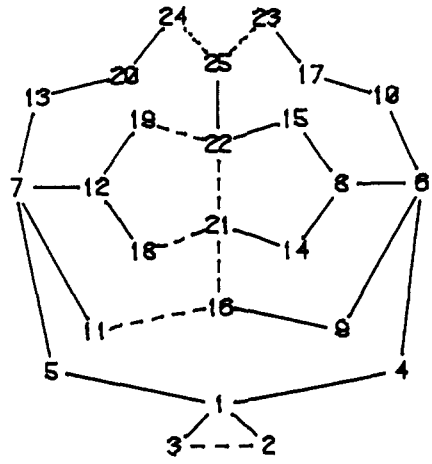
```
FINAL RING SET
RING NO:  1  ATOMS:  2  3  1
RING NO:  2  ATOMS:  4  5  1
RING NO:  3  ATOMS:  4  6  2  1
RING NO:  4  ATOMS:  6  7  2
RING NO:  5  ATOMS:  9 10  8  4  6
RING NO:  6  ATOMS:  9 11  8 10
```

Figure 6. Ring solutions (SSSR) for embedded ring.

now to recreate a one-to-one identifying correspondence of each ring with one ring-closure bond that it contains.

In those cases for which at least one ring has no unique bonds, the procedure may not create the correct answer, but such cases are of course recognized by the program and may be solved by the traditional method of carrying out $2^{\rho} - 1$ linear combinations (*exclusive or* operations), on the basis set, but with many rings this is time consuming. We did in fact use this to check the results of our faster procedure (below) in all cases. In Figure 5, example C is such a case in which the sixth ring has no unique bonds and is only solved by linear combinations. The other two solved directly. Example C is not a planar graph. In planar graphs, a ring with no unique bonds is one embedded fully in other rings (cf. Figures 6 and 7); i.e., the unique bonds are the outer periphery.

In most cases of rings with no unique bonds, however, the problem can still be rapidly reduced by a further procedure that provides for the removal of some rings already fixed. Usually, this will fully solve the problem without resorting to




```
INITIAL BASIS SET
RING NO:  1  ATOMS:  2  3  1
RING NO:  2  ATOMS: 11 16  9  6  4  1  5  7
RING NO:  3  ATOMS: 16 21 14  8  6  9
RING NO:  4  ATOMS: 18 21 16 11  7 12
RING NO:  5  ATOMS: 19 22 15  8 14 21 18 12
RING NO:  6  ATOMS: 21 22 15  8 14
RING NO:  7  ATOMS: 23 25 22 15  8  6 10 17
RING NO:  8  ATOMS: 24 25 22 19 12  7 13 20

*****PROBLEM IS REDUCIBLE*** IGNORE RING      1
*****PROBLEM IS REDUCIBLE*** IGNORE RING      7
*****PROBLEM IS REDUCIBLE*** IGNORE RING      8
RINGSET PRIOR TO REDUCTION
RING NO:  1  ATOMS:  2  3  1
RING NO:  2  ATOMS: 11 16  9  6  4  1  5  7
RING NO:  3  ATOMS: 16 21 14  8  6  9
RING NO:  4  ATOMS: 18 21 16 11  7 12
RING NO:  5  ATOMS: 19 22 15  8 14 21 18 12
RING NO:  6  ATOMS: 21 22 15  8 14
RING NO:  7  ATOMS: 23 25 22 15  8  6 10 17
RING NO:  8  ATOMS: 24 25 22 19 12  7 13 20

FINAL RING SET
RING NO:  1  ATOMS:  2  3  1
RING NO:  2  ATOMS: 11 16  9  6  4  1  5  7
RING NO:  3  ATOMS: 16 21 14  8  6  9
RING NO:  4  ATOMS: 18 21 16 11  7 12
RING NO:  5  ATOMS: 19 22 21 18 12
RING NO:  6  ATOMS: 21 22 15  8 14
RING NO:  7  ATOMS: 23 25 22 15  8  6 10 17
RING NO:  8  ATOMS: 24 25 22 19 12  7 13 20
```

Figure 7. Full ring perception in an embedded-ring example.

linear combinations. Even when the removal of some rings does not allow a solution, it at least will reduce the number of remaining rings for the linear combinations operation.

Thus, if any path of unique bonds from point A to point B is longer than, or the same as, any other path from A to B, it can be removed from the graph, since the smallest rings remaining can always be defined by using the other, shorter path. In this way, the $\rho \times b$ matrix is reduced by the removal of rings already fixed as defined by those unique paths. In the smaller polycycle that remains, there may now be found new bonds that are unique (i.e., had previously been shared only with rings now removed). These in turn may be excised if they also fit the criterion of being longer than (or the same as) any other path across their termini. This procedure can be seen as stripping the outer rings off a polycycle to bare, and so define, any inner, embedded rings. In those few, generally highly symmetric, cases in which this reduction cannot be done, i.e., no unique paths are longer than alternate ones, then one is obliged to go to the $2^{\rho} - 1$ linear combinations for the final solution(s). Cubane and dodecahedrane are molecular examples of such ultimately intractable systems for which no unique bond paths can be stripped off.

The higher homologue of cubane shown in Figure 6A allows the reduction procedure to simplify and solve the ring perception. The unique bond path 7–9–10–8 is now as long as the alternative 7–4–6–8. Its removal now allows the further removal of unique paths 6–8–5 (=6–2–5) and 4–7–3 (=4–1–3), and the remaining bicyclic is simply solved. The program output is thus shown below the ring in Figure 6A. Here, the initial and final sets are the same, but the reduction of the system is necessary to prove it since one ring (1) has no unique bonds. Another example of an embedded ring is shown in Figure 6B. Here, deleting the outer (two) unique bonds of each three-membered ring, as longer than the inner path (length = 1), bares a tricyclic residue that now has unique bonds in all rings and so is soluble.

One further example with embedded rings is detailed in Figure 7 and shows a case in which the initial ring set definition was found to be incorrect. The initial basis set is shown first and contains four eight-membered rings; reexamination for smallest rings from the unique bond paths next yields the same set. However, not all rings have unique bonds (some are embedded), so that the program now undertakes reduction of the rings by removing rings 1, 7, and 8 and so bares the inner, embedded five-membered rings and solves the correct ring set (SSSR) as shown in the third set of Figure 7. This set has replaced one eight-membered ring (5) with a five-membered ring and is now the smallest set of smallest rings.

Utility of the Procedure. The essence of the system here is to allow storage of molecular skeletons as T/R-lists in an unusually small space and in a linear binary form allowing for very rapid numerical search techniques. As such, it appears to offer a powerful challenge to the ubiquitous Wiswesser Line Notation[18] (WLN) for storage and retrieval of chemical systems. The memory requirements for a molecule of n atoms is essentially $4n$ bits for the skeleton, followed by a list of na bits to designate the atom types and other atom attributes, where a is the number of bits required per atom entry, i.e., 2^a possible atom-type distinctions. Allowing for 128 such atomic distinctions ($a = 7$), the total bit requirements are $4n + 7n = 11n$. Hence, a 20-atom molecule (without hydrogens) requires only 28 bytes of storage and a 30-atom one only 42 bytes, irrespective of complexity. It is perhaps surprising that this system requires fewer bits for designating the molecular skeleton than it does to specify the atoms in it.[19]

Here, the storage requirement is a function only of size (n), whereas in the Wiswesser notation it is a free function of both skeletal and functional complexity and so quite variable in size. However, it is common to find substances of reasonable complexity requiring 50 alphanumeric characters in WLN, i.e., around 50 bytes of storage. Hence, the system here compares very favorably. Furthermore, the system is also infallible and easily manipulated by hand or very rapidly by computer from graphics input and so constitutes a strong competitor for the traditional Wiswesser Line Notation, which is hard to learn for noncomputer use and is sometimes incorrect in complex cases.

Finally, of course, the procedure is also very rapid in perceiving all rings in a structure and establishing the smallest fundamental set of rings (SSSR). In our program, the rings are identified by size and by atom-number sequence. As previously reported,[2] the system can also easily identify skeletally equivalent atoms (isomorphic points on the graph). It seems likely that the T/R-lists can further be manipulated to generate and enumerate all possible isomers of any skeleton of n atoms and ρ rings;[7] this procedure is currently under study.

ACKNOWLEDGMENT

We are grateful to the National Science Foundation for generous support of this work through a research grant (CHE-8102972).

REFERENCES AND NOTES

(1) This is paper 11 in the series "Systematic Synthesis Design". For paper 10, see footnote 2.
(2) Hendrickson, J. B.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171.
(3) Zamora, A. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40.
(4) The Morgan algorithm[5] affords similar but not idential numbering since it also numbers out from atom 1, assigning its adjacent atoms as 2, 3, 4, ..., but the algorithm does not necessarily assign atom 1 to a tetravalent one but rather to that which on iteration develops the most extended connectivity. The Morgan numbering for Figure 1A is compared as an example: (maximal) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16; (Morgan) 3 1 7 8 9 2 4 14 5 6 10 16 11 12 13 15.
(5) Morgan, H. L. *J. Chem. Soc.* **1965**, *6*, 107.
(6) An entry in row 2 requires a three-membered ring on atom 1. The earliest row in which a ring of ρ atoms can appear is row $\rho - 1$.
(7) All graphs on six points are drawn out in Appendix 1 of Harary, F. "Graph Theory"; Addison-Wesley: Reading, MA, 1969.
(8) Fugmann, R.; Dolling, U.; Nickelsen, H. *Angew. Chem., Intl. Ed. Engl.* **1967**, *6*, 723.
(9) Plotkin, M. *J. Chem. Doc.* **1971**, *11*, 60.
(10) Corey, E. J.; Petersson, G. A. *J. Am. Chem. Soc.* **1972**, *94*, 460.
(11) Bersohn, M. *J. Chem. Soc., Perkin Trans. 1* **1973**, 1239.
(12) Esack, A. *J. Chem. Soc., Perkin Trans. 1* **1975**, 1120.
(13) Wipke, W. T.; Dyott, T. M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140.
(14) Schmidt, B.; Fleischhaver, J. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 204.
(15) Randic, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 36.
(16) Gasteiger, J.; Jochum, C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43.
(17) Roos-Kozel, B. L.; Jorgensen, W. L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 101.
(18) Smith, E. G.; Baker, P. A. "The Wiswesser Line-Formula Chemical Notation (WLN)", 3rd ed.; Chemical Information Management: Cherry Hill, NJ, 1975.
(19) For skeletons containing atoms of connectivity greater than 4, more bits are required since the maximal row sum in either T-list or R-list is presently 4 and so requires only 2 bits. For atom connectivities with $v_i \leq 8$, 3 bits per list entry would be required, or $6n$ bits per skeleton.

Derivation of the Principle of Smallest Set of Smallest Rings from Euler's Polyhedron Equation and a Simplified Technique for Finding This Set

SEYMOUR B. ELK

Department of Computer Science, The William Paterson College of New Jersey, Wayne, New Jersey 07470

Received November 16, 1983

Because the arrangement of atoms in a chemical compound often involves many overlapping rings, for purposes of taxonomy and nomenclature, the description that has traditionally been chosen as the geometrical model of this compound is the smallest set of smallest rings (SSSR) that includes all of the chemical bonds present. This paper first shows that SSSR is merely a reformulation of Euler's Polyhedron Equation in topology. Next, the "art" of finding this SSSR is reexamined and a theorem proposed in order to expedite finding the SSSR for a given compound. Finally, limitations, inherent in the topology of Euler's equation, are shown to have their corresponding limitations in the description of three-dimensional chemical molecules.

Chemical compounds have traditionally been classified according to the number and type of closed connected pathways that exist in a given molecule. However, in compounds that contain more than one single continuous closed path between atoms, there often exists many such circuits—which are frequently overlapping.[1] In order to circumvent the descriptive problems associated with overlapping rings, various "quick-fix" methods have been devised that convert such ring systems into topologically simpler systems. The one traditionally chosen,[2] the number of rings in a system is equal to the minimum number of scissions required to convert the cyclic system into an open-chain compound, forms the cornerstone of the method in present usage in most systems of nomenclature, referred to as the smallest set of smallest rings (SSSR).

Such methods give excellent results for compounds that are essentially planar such as polybenzenes,[3] e.g., coronene (Figure 1). In fact, the use of this method removes one element of ambiguity (Zamora's[4] type II and type III ring systems[5]) by mandating that the inner ring of coronene be considered a part of the figure rather than a part of the boundary. This is a highly desirable property for coronene inasmuch as the coplanar "inner" ring contributes to the aromatic stability of the compound. On the other hand, it is of dubious chemical merit for all of the circulenes[6] except coronene.

At this point, however, it should be noted that there exists a much simpler technique for determining the number of rings in a chemical compound, a technique that does not use the idea of bond scission. Instead, only the incidence relationships of the individual vertexes are important: (theorem 1) The number of rings in a chemical compound, R, equals 1 plus half the incidence excess;[9] i.e., $R = 1 + (1/2)n$, when n is the summation over all of the vertexes of their incidence minus 2 [$n = \sum_v (i_v - 2)$].

The mathematical basis for this principle lies in the metonymic[10] usage of "ring" as (1) a closed electronic pathway in a molecule and as (2) the boundary of a face of the geometrical model that represents this molecule.[12] Because we are interested in the number of faces of a simply connected[13] geometrical model in three dimensions, Euler's Polyhedron Formula[14] ($F + V = E + 2$) is applicable. Also, because the projection of a three-dimensional object onto a planar surface, the so-called Schlegel projection,[15] causes exactly one of the planar faces of the polyhedron to become the outer perimeter of the entire figure, $R = F - 1$. Next, we note that each edge of a polyhedron contributes 1 to the sum of the incidence at each edge; i.e., $2E = \sum_v i_v$. Subtracting $2V$ from each side of this equation yields $2E - 2V = \sum_v i_v - 2V = \sum_v (i_v - 2)$, which may be given the symbol n and the name "incidence excess". Finally, by substitution into Euler's equation, this yields $R = 1 + (1/2)n$.

From application of this theorem to coronene, there exists 12 vertexes with incidence = 3 (darkened) and 12 vertexes with incidence = 2 (Figure 2). Each incidence = 3 vertex contributes 1 to the value of n and each incidence = 2 vertex contributes 0. Thus

$$R = 1/2 \times [12 \times (3 - 2) + 12 \times (2 - 2)] + 1 = 7$$

Similarly, for the acyclic compound 2,2,3,3-tetramethylbutane (Figure 3), two carbon atoms have incidence = 4 and six have incidence = 1. Thus

$$R = 1/2 \times [2 \times (4 - 2) + 6 \times (1 - 2)] + 1 = 0$$

Because the traditional technique for finding SSSR is based on finding an appropriate topological simplification, Dyson[16] advises: "Some skill must be exercised in detecting the smallest

0095-2338/84/1624-0203$01.50/0