Under such conditions, students may be tempted to search for elements that are likely to change oxidation states between compounds before establishing what the oxidation numbers are within each compound. This frequently causes them to miss changes that they do not expect or to assign values without adequate thought. The discipline of following the rules developed for the oxidation expert should reduce these occurrences.

FUTURE WORK

One question that remains in step 3 and step 8 is how to accept or assign an oxidation number for a species which can only occur under specific conditions. For example, in a compound such as Mn_3C, Mn has an oxidation state not seen in other of its compounds. Similarly, although metals may have an oxidation state of zero in metal carbonyls, it may not be useful to include zero in the expert as a legal oxidation state for metals in compounds. A feasible approach might be to build each possible oxidation state for each element as a program object containing named slots which restrict its applicability. This is being explored as part of a general structure for chemical information which will be built to support CHEMPROF experts.

Since names are frequently dependent on oxidation states, the general framework for constructing the oxidation expert also appears to be applicable to building naming experts. Given a chemical input, these experts would either convert inorganic names to formulas or formulas to names. Such experts, which would need to allow for a certain amount of spelling abberration by the students, would also be effective tools for student communication with CHEMPROF. It is also our intent to build an expert to solve redox equations.

ACKNOWLEDGMENT

We acknowledge an equipment grant from IBM under Project Trochos which has given us tools to develop both this expert and CHEMPROF as a whole.

REFERENCES

(1) Eggert, A. A.; Kean, E.; Middlecamp, C. CHEMPROF—An Intelligent Tutoring System. *Proceedings of the Tenth Biennial Chemical Education Conference*; ACS Division of Chemical Education: West Lafayette, IN, 1988.

(2) Eggert, A. A.; Middlecamp, C.; Kean, E. CHEMPROF—An Intelligent Tutor for General Chemistry. Submitted to *J. Chem. Educ.*

(3) Koln, D. The Chemical Equation Part II: Oxidation-Reduction Reactions. *J. Chem. Educ.* **1978**, *55*, 326-333.

(4) Woof, A. A. Oxidation Numbers and Their Limitations. *J. Chem. Educ.* **1988**, *65*, 45-46.

(5) Holleran, E. M.; Jespersen, N. D. Elementary Oxidation-Number Rules. *J. Chem. Educ.* **1980**, *57*, 670.

(6) Lederberg, J. DENDRAL-64: A system for computer construction, enumeration and notation of organic molecules as tree structures and cyclic graphs. Report No. CR-57029, 1964; NASA, Washington, DC.

(7) Shortliffe, E. H. *Computer-based Medical Consultations: MYCIN*; American Elsevier: New York, 1976.

(8) Doyle, J. A Model for Deliberation, Action, and Introspection. Technical Report AI-TR-581, 1980; AI Lab, Massachusetts Institute of Technology, Cambridge, MA.

(9) Burton, R. R.; Brown, J. S. An Investigation of Computer Coaching for Informal Learning Activities. *Int. J. Man-Machine Stud.* **1979**, *11*, 5-24.

(10) Kauffman, J. M. Simple Method for Determination of Oxidation Numbers of Atoms in Compounds. *J. Chem. Educ.* **1986**, *63*, 474-475.

(11) Dickerson, R. E.; Gray, H. B.; Darenbourg, M. Y.; Darenbourg, D. J. *Chemical Principles*, 4th ed.; Benjamin/Cummings: Reading, MA, 1984; p 227.

(12) Kroschwitz, J. I.; Winokur, M. *Chemistry, A First Course*; McGraw-Hill: New York, 1980; p 430.

Computer Perception of Constitutional (Topological) Symmetry: TOPSYM, a Fast Algorithm for Partitioning Atoms and Pairwise Relations among Atoms into Equivalence Classes

GERTA RÜCKER and CHRISTOPH RÜCKER*

Institut für Organische Chemie und Biochemie, Universität Freiburg, Albertstrasse 21, D-7800 Freiburg, FRG

Received February 12, 1990

An algorithm for the perception of constitutional symmetry in molecules (graphs) is presented, which partitions not only atoms (vertices) but also all pairwise relations among skeleton atoms into equivalence classes. The method works without canonical numbering, essentially by raising the connectivity matrix of the arbitrarily numbered molecule (graph) to its second, third, etc. power and evaluating the entries in these higher order matrices.

When developing a computer program for the machine generation of systematic (IUPAC) names for polycyclic organic compounds,[1] we encountered the more fundamental problem of computer perception of symmetry. For example, how many different pairs of potential bridgeheads are present in the bis- and trissecododecahedranes **1-4** (Figure 1): in particular, which pair is equivalent to which other pair? This question is obviously much harder to answer than the question of how many different kinds of atoms are present in these same compounds. Whereas several computer methods exist for the purpose of partitioning atoms in a molecule into equivalence classes[2-8] and some work has also been done on the detection of the identity of bonds,[4,9] no general method appears to be available to treat pairwise (or even higher) relations among the atoms.[10]

The symmetry properties of a molecule (as well as all other properties) are obviously encoded in its structure; the difficulty lies in the decoding process. Since the structure (more precisely, the constitution) can be represented by a connectivity (adjacency) matrix[11] and since the constitutional symmetry is a very fundamental and simple property, we expected that it could be derived by some simple mathematical manipulation of the connectivity matrix. This turned out to be the case, and we report herein on an algorithm (and the computer program TOPSYM based on it) that finds equivalence classes of atoms and pairs of atoms by a purely mathematical approach. That is, we do not need a canonical numbering,[3,5,6] nor do we need to assign numbers to atomic properties (as was done in a somewhat arbitrary manner in some recent solutions of the atom equivalence problem[8]).

0095-2338/90/1630-0187$02.50/0

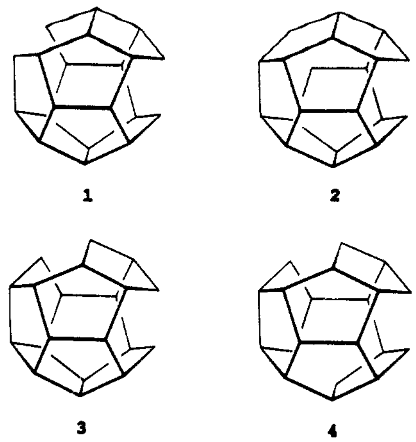
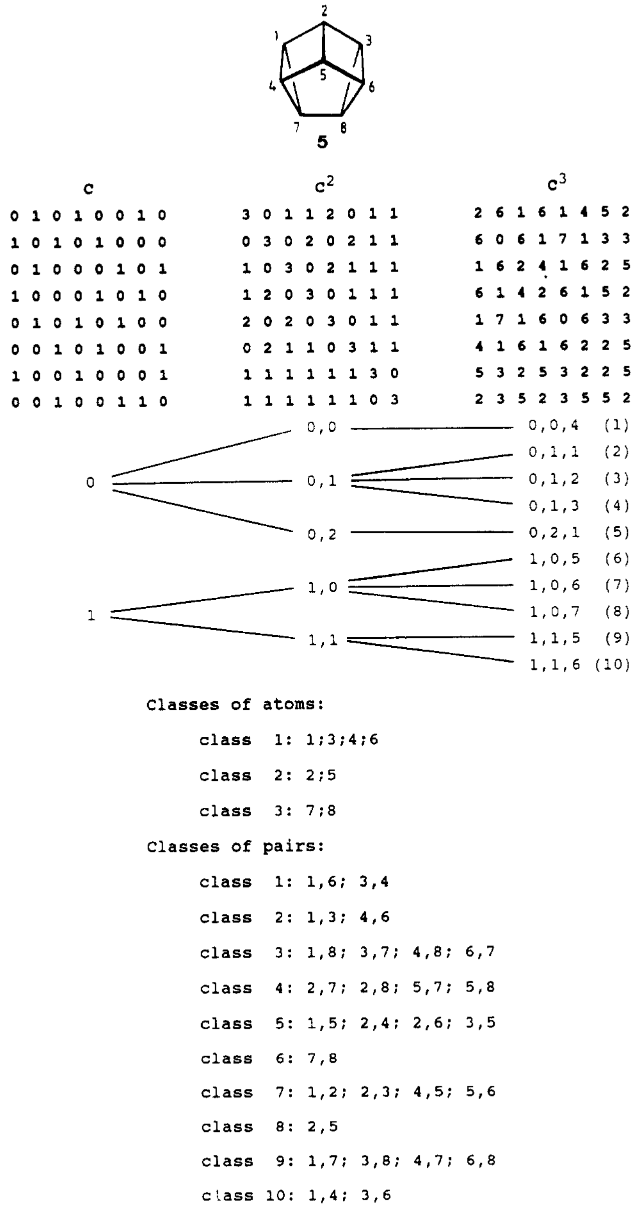


Figure 1. Bis- and trissecododecahedranes.



```
       C                          C²                         C³
0 1 0 1 0 0 1 0          3 0 1 1 2 0 1 1          2 6 1 6 1 4 5 2
1 0 1 0 1 0 0 0          0 3 0 2 0 2 1 1          6 0 6 1 7 1 3 3
0 1 0 0 0 1 0 1          1 0 3 0 2 1 1 1          1 6 2 4 1 6 2 5
1 0 0 0 1 0 1 0          1 2 0 3 0 1 1 1          6 1 4 2 6 1 5 2
0 1 0 1 0 1 0 0          2 0 2 0 3 0 1 1          1 7 1 6 0 6 3 3
0 0 1 0 1 0 0 1          0 2 1 1 0 3 1 1          4 1 6 1 6 2 2 5
1 0 0 1 0 0 0 1          1 1 1 1 1 1 3 0          5 3 2 5 3 2 2 5
0 0 1 0 0 1 1 0          1 1 1 1 1 1 0 3          2 3 5 2 3 5 5 2
```

```
Classes of atoms:

     class  1: 1;3;4;6

     class  2: 2;5

     class  3: 7;8

Classes of pairs:

     class  1: 1,6; 3,4

     class  2: 1,3; 4,6

     class  3: 1,8; 3,7; 4,8; 6,7

     class  4: 2,7; 2,8; 5,7; 5,8

     class  5: 1,5; 2,4; 2,6; 3,5

     class  6: 7,8

     class  7: 1,2; 2,3; 4,5; 5,6

     class  8: 2,5

     class  9: 1,7; 3,8; 4,7; 6,8

     class 10: 1,4; 3,6
```

Figure 2. (Top) First three powers of the connectivity matrix for arbitrarily numbered cuneane (**5**). (Middle) Evolution of the different classes of pairs contained therein. (Bottom) List of the classes of atoms and pairs in **5** as they appear in the computer output.

Table I. Results for the Structures in Figures 1–3

structure (graph)	n	classes of atoms	classes of pairs	CPU time,[a] s	CPU time,[a] s, atoms only	CPU time,[b] s, (ref 5b)
1	20	10	100	0.124	0.035	
2	20	10	100	0.137	0.035	
3	20	12	102	0.128	0.035	
4	20	6	39	0.085	0.037	
5	8	3	10	0.023	0.010	0.93
6	15	9	57	0.055	0.020	
7	20	6	51	0.077	0.033	
8	22	8	76	0.137	0.049	
9	54	4	79	3.249	2.383	
10	60	4	96	4.223	3.091	
11	20	1	5	0.097	0.069	
12	60	6	165	5.047	2.868	
13	32	2	20	0.388	0.300	
14	35	3	40	0.485	0.277	
15	40	7	100	0.969	0.474	
16	60	1	23	7.212	6.456	
17	60	1	23	6.331	5.814	
18	8	1	3	0.017	0.011	
19	8	2	7	0.017	0.010	
20	10	3	10	0.024	0.012	5.25
21	11	3	12	0.024	0.012	0.67
22	12	2	10	0.026	0.015	2.57
23	12	12	66	0.042	0.013	0.92
24	12	7	32	0.033	0.015	1.18
25	12	3	16	0.033	0.017	1.66
26	16	3	23	0.061	0.029	1.38
27	16	16	120	0.087	0.018	0.56
28	18	8	52	0.079	0.030	
29	18	8	52	0.079	0.030	1.26
30	18	8	42	0.080	0.035	3.61
31	20	10	58	0.099	0.046	6.94
32	28	27	352	0.929	0.125	3.93
33	17	17	136	0.103	0.024	1.44
34	9	1	2	0.016	0.010	12.31
35	16	1	4	0.049	0.034	
36	15	1	3	0.037	0.024	
37	17	17	136	0.135	0.020	13.98
38	18	6	46	0.079	0.033	16.03
39	20	1	5	0.097	0.071	
40	20	3	18	0.089	0.049	36.44

[a] IBM 3090. [b] IBM 370/158.

The basic idea is very simple: The information contained but hidden in the connectivity matrix becomes evident in the higher powers of this matrix.[12] Consider Figure 2(top) which gives, as an example, the connectivity matrix **C** of arbitrarily numbered cuneane (**5**) as well as the result of multiplying the matrix once and twice by itself, i.e., the second and third powers of the matrix, $\mathbf{C}^2$ and $\mathbf{C}^3$. Information on properties of atom i is located in the ith element of the main diagonal of these matrices and in the pattern of entries in the ith row (or column) of these matrices. Correspondingly, information on the properties of the atom pair i,j is located in the nondiagonal elements i,j of these matrices.

Our method consists of five procedures (i–v), four of which (i–iv) are performed after each matrix multiplication (after each step), while the fifth (v) is used in later steps only to check whether the goal is reached and to remedy any remaining inconsistencies.

(i) The first thing to do is to count how many different entries are contained in the main diagonals of the matrices; at least as many equivalence classes of atoms will be found. Thus, the main diagonals of **C** and $\mathbf{C}^2$ each contain one entry, while in $\mathbf{C}^3$ there are two different entries in the main diagonal.

(ii) In **C** there is only one pattern of entries in a row (three 1's and five 0's), while in $\mathbf{C}^2$ there are three different patterns (a 3 and a 2 and four 1's in rows 1, 3, 4, 6; a 3, two 2's and two 1's in rows 2 and 5; a 3 and six 1's in rows 7 and 8). Thus, three equivalence classes of atoms are found in the second matrix, while $\mathbf{C}^3$ and all higher matrices do not give rise to any further discrimination.

(iii) To gain as much information as possible from each step, one asks after each step whether all the neighbors of two atoms

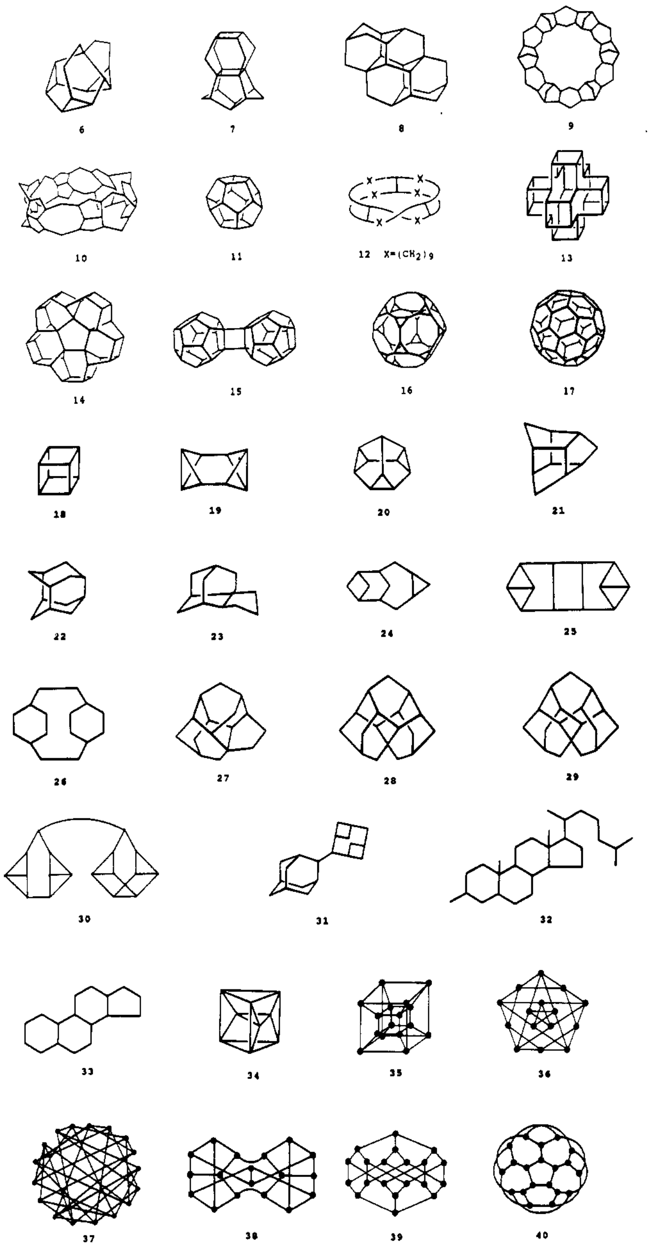


Figure 3. Some saturated polycyclic structures (monochromic graphs), the results of which are shown in Table I.

i and *j* found to be equivalent are themselves pairwise equivalent. If this is not the case, *i* and *j* are obviously not equivalent, and from now on they will be treated as nonequivalent.

(iv) With regard to the pairs, in **C** there are only two entries in the nondiagonal elements (0 and 1 for nonneighbors and neighbors, respectively), in $\mathbf{C}^2$ there are three (0, 1, 2), whereas in $\mathbf{C}^3$ there are already seven entries (1, 2, ... 7). Notice, however, that entries differentiated in one matrix may become alike in a higher matrix and that it is therefore necessary to trace the "history" of the elements through the matrices as shown in Figure 2(middle). The number of different classes of pairs found in $\mathbf{C}^3$ in our example is thus not 7 but 10. Again, higher matrices do not give rise to further discrimination: cuneane has 3 classes of atoms and 10 classes of pairs. The resulting partition of atoms and pairs into classes is explicitly shown in Figure 2(bottom).

Note that all the above is true independent of whether or not the matrix elements have any physical meaning. Actually, the elements have a meaning; the entry *i*,*j* in the *k*th power matrix gives the number of different walks of length *k* (bonds) from *i* to *j*.[12,13]

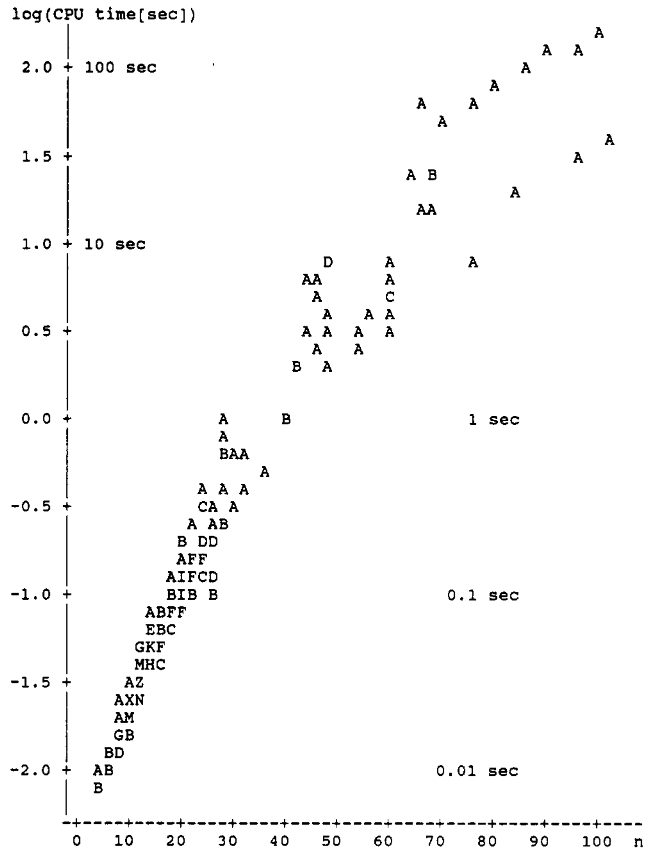


Figure 4. Plot of log (CPU time [s]) vs size *n* for saturated polycyclic structures (monochromic graphs). A = 1 observation, B = 2 observations, etc.

Generally, calculating and evaluating more and more matrices as described will partition both the atoms and pairs into more and more equivalence classes, until all such classes have been found. A problem with our approach is knowing after how many steps one can stop. We are not able to give a mathematically sound "stop" criterion, and we therefore tackled this problem in a pragmatic way (which has proved to be satisfactory in all cases treated to date). The interpretation of the matrix elements' meaning given above suggests a reasonable stop criterion. There should be at least as many steps performed so that the whole structure has been taken into account, i.e., so that every atom is connected to every other one by at least one walk. Therefore, as soon as for each element *i*,*j* an entry different from 0 has appeared at least once in the matrices $\mathbf{C}^1$–$\mathbf{C}^k$ (that is, when the power equals the diameter of the graph), one further step is made: this is the earliest time to stop.[14]

(v) Since the final number of pairs is often not yet found at this stage, it is now checked for every two pairs *i*,*j* and *k*,*l* found to be equivalent whether one of the atoms *k*,*l* is in fact equivalent to *i* and the other one equivalent to *j*. If this is not the case, the pairs *i*,*j* and *k*,*l* are separated into different classes, and further steps are performed until such a contradiction can no longer be detected.[15]

The program TOPSYM, written in FORTRAN77, performs the described procedure using the constitutional formula of the compound under investigation as the only input. Though the program was tailor-made to serve as an aid to the IUPAC nomenclature program POLCYC,[1] its scope is much broader. In fact, it is a program for symmetry perception (vertices and pairwise relations) in colored finite graphs in general; that is, in chemical terms, one can treat acyclic, monocyclic, and polycyclic compounds of any complexity which may or may not contain multiple bonds and/or heteroatoms. The logic of the program does not put any restrictions on the size or to-

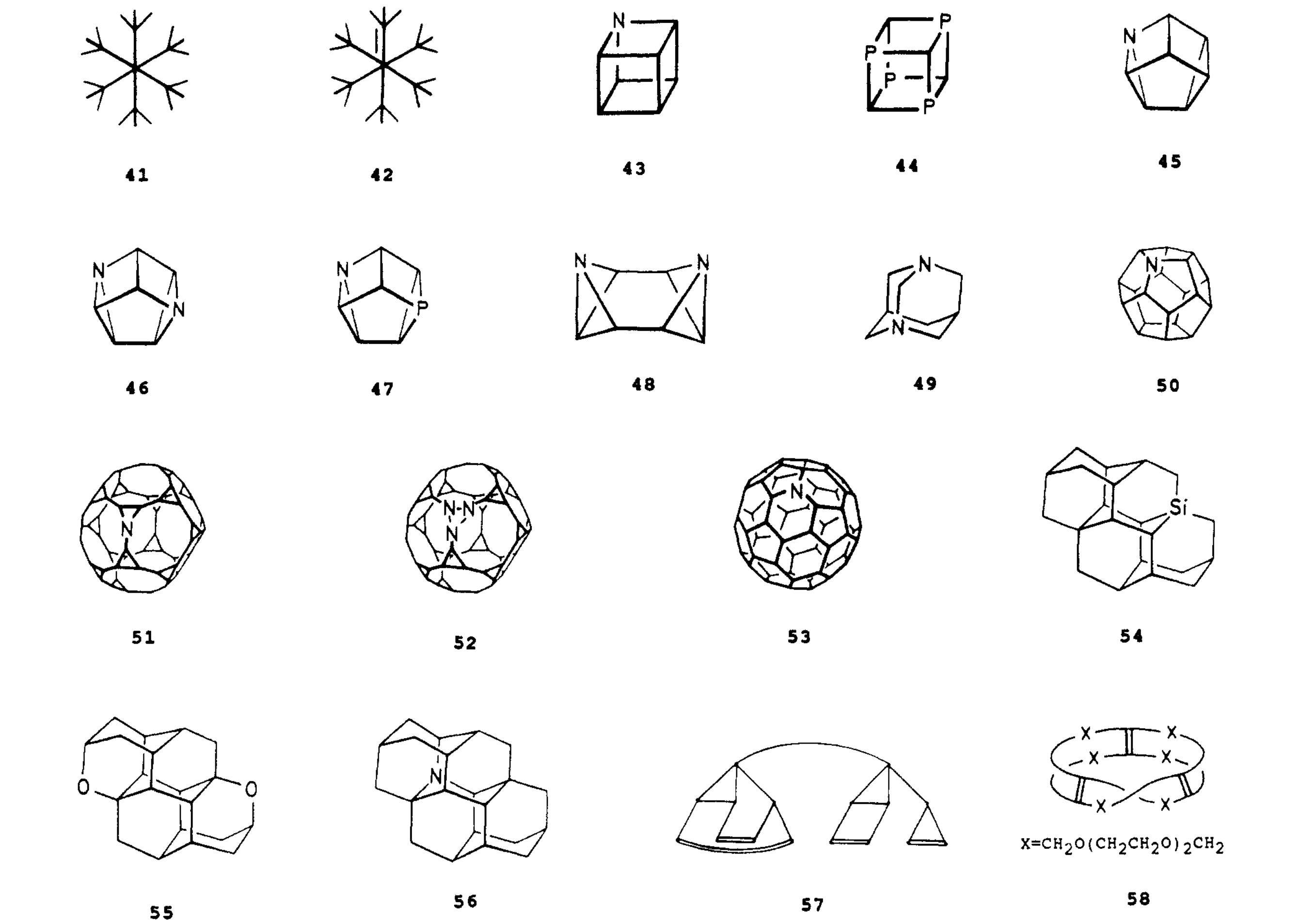


Figure 5. Some acyclic and cyclic structures containing multiple bonds and/or heteroatoms (colored graphs), the results of which are shown in Table II.

Table II. Results for the Structures in Figure 5

structure (graph)	n	classes of atoms	classes of pairs	CPU time, s	CPU time, s, atoms only
41	25	3	7	0.120	0.081
42	25	5	15	0.110	0.067
43	8	4	9	0.016	0.010
44	8	2	4	0.014	0.011
45	8	8	28	0.019	0.010
46	8	4	16	0.016	0.010
47	8	8	28	0.019	0.009
48	8	3	11	0.015	0.010
49	10	5	17	0.024	0.013
50	20	6	38	0.111	0.050
51	60	32	902	10.686	2.759
52	60	12	312	6.541	2.650
53	60	32	902	10.837	2.462
54	22	16	141	0.160	0.055
55	22	8	76	0.133	0.058
56	22	22	231	0.217	0.041
57	16	10	55	0.069	0.022
58	60	6	165	4.712	2.862

pology of the molecule (graph) or on the number of different kinds of heteroatoms or multiple bonds (colors of vertices or edges). A sample computer output [for cuneane (**5**)] is shown in Figure 2(bottom).

The program correctly processed all graphs fed to it hitherto, in particular all the recalcitrant graphs of refs 2–9. Even the so-called endospectral graphs[16] (graphs having two or more nonequivalent vertices that cannot be differentiated by their diagonal entries in all the powers of the adjacency matrix) do not pose any problems.[17,18]

Table I summarizes the results for several saturated carbocyclic compounds (monochromic graphs) treated by TOPSYM, the structures of which are depicted in Figures 1–3. Together with the numbers of equivalence classes of atoms and pairs we give the elapsed CPU time (seconds). For purposes requiring the partitioning of atoms only (and for comparison with an existing method[5b]) a shortened version of the program was prepared (procedures i–iii) which treats the atoms (vertices) only: the CPU times for this version are also given. Although part of the gain in performance relative to ref 5b is due to our use of a newer computer (IBM 3090 vs IBM 370/158[19]), much of the gain is obviously produced by the lower computational effort required by our method.

Figure 4 gives a plot of the decimal logarithms of the CPU time required for TOPSYM treatment (atoms and pairs) vs the size n of the more than 300 compounds (graphs) processed hitherto, including all those of Table I. The data can be described by a least-squares straight line with $r^2 = 0.94$.[20]

To demonstrate the capability of the program, the data for some acyclic compounds (graphs) and compounds containing multiple bonds and/or heteroatoms (colored graphs) are given in Table II. The corresponding structures appear in Figure 5. It is easily seen that the CPU time required for such compounds is in the same range as that for the saturated carbocycles of Table I.

Immediate chemical applications of the program (apart from its prime purpose[1]) may be the enumeration of mono- and disubstituted derivatives of a parent compound and the enumeration of long-range NMR couplings. It is expected that the program, when used as an aid to other structure-processing procedures (such as ring perception algorithms[22]), will result in considerable time savings. A copy of the program is available upon request from the authors.

Note Added In Proof. Another method for symmetry perception (atoms only) came to our knowledge recently: Davis, M. I.; Ellzey, M. L., Jr. *J. Comput. Chem.* **1983**, *4*, 267.

ACKNOWLEDGMENT

We thank Prof. M. Randić for information on endospectral graphs and on the Cayley–Hamilton theorem and Dr. A. Bömelburg (IBM Deutschland) for a useful programming hint.

REFERENCES AND NOTES

(1) Rücker, G.; Rücker, Ch. *Chimia*, in press.

(2) Shelley, C. A.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 110; **1979**, *19*, 247.

(3) Jochum, C.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113; **1979**, *19*, 49.

(4) Schubert, W.; Ugi, I. *J. Am. Chem. Soc.* **1978**, *100*, 37. Schubert, W. *MATCH* **1979**, *6*, 213.

(5) (a) Randić, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 36. (b) Randić, M.; Brissey, G. M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 52.

(6) Hendrickson, J. B.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171.

(7) Balaban, A. T.; Mekenyan, O.; Bonchev, D. *J. Comput. Chem.* **1985**, *6*, 538, and references cited therein.

(8) Bersohn, M. *Comput. Chem.* **1987**, *11*, 67.

(9) Ihlenfeldt, W. D.; Gasteiger, J. In *Software-Entwicklung in der Chemie 2*; Gasteiger, J.; Ed.; Springer-Verlag: Berlin, 1988; pp 13–33.

(10) Gray, N. A. B. *Computer-assisted structure elucidation*; Wiley: New York, 1986; Chapter 9.

(11) For the kind of symmetry discussed in this paper (which is clearly not the usual geometric symmetry in three-dimensional space) the terms "constitutional symmetry" and "topological symmetry" have both been used in the literature[2-5] since the information contained in the constitution (not configuration or conformation) is considered exclusively; i.e., geometric properties like bond lengths and angles or cis/trans relationships are disregarded. However, the term topological symmetry may be misunderstood since topological *isomers* [as defined earlier, e.g., a simple macrocycle and its knotted isomer (Frisch, H. L.; Wasserman, E. *J. Am. Chem. Soc.* **1961**, *83*, 3789) or the pair **28**/**29** in Figure 3] are *indistinguishable* in terms of the symmetry under discussion here (the two connectivity matrices of such a pair are identical).

(12) The higher powers of the adjacency (or a similar) matrix were used earlier, but their full potential for symmetry recognition was not exploited: Randić, M. *J. Comput. Chem.* **1980**, *1*, 386. Uchino, M. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 116. Golender, V. E.; Drboglav, V. V.; Rosenblit, A. B. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 196. Razinger, M. *Theor. Chim. Acta* **1982**, *61*, 581. Randić, M.; Woodworth, W. L.; Graovac, A. *Int. J. Quantum Chem.* **1983**, *24*, 435.

(13) A walk is a pathway with repetition: The entry 6 in element 1,2 in C^3 in our example cuneane means there are six different walks of length 3 bonds from atom 1 to atom 2, namely, 1-2-1-2, 1-4-1-2, 1-7-1-2, 1-2-3-2, 1-2-5-2, 1-4-5-2.

(14) As a rule of thumb, all classes of atoms are found at this stage. In fact, in most cases considerably fewer steps are required. We know of only one exception to this rule (Figure 5 in ref 7).

(15) The Cayley–Hamilton theorem in this connection states that if two entries are not differentiated in all matrices up to the *n*th power, then they will not be differentiated in higher matrices either. While this is theoretically pleasing, the stop criterion suggested thereby (stop after the *n*th power matrix has been evaluated) is of little practical value: In the majority of graphs the diameter is considerably less than *n*.

(16) Randić, M. *SIAM J. Alg. Disc. Meth.* **1985**, *6*, 145. Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N.; Kleiner, A. F.; Randić, M. *J. Math. Phys.* **1986**, *27*, 2601. Randić, M.; Kleiner, A. F. *Ann. N. Y. Acad. Sci.* **1989**, *555*, 320.

(17) Though the critical vertices in endospectral graphs obviously are not segregated by procedure i, they are differentiated either by procedure ii (the graphs in ref 16, which exhibit different patterns of entries in the rows corresponding to the critical vertices in low power matrices already) or by procedure iii, e.g., graphs **30** and **31** in Figure 3.

(18) The only known case where our algorithm stops prematurely is a graph of 18 equivalent vertices of degree 3 with diameter 4 (Figure 1.1 in: Coxeter, H. S. M.; Frucht, R.; Powers, D. L. *Zero-Symmetric Graphs*; Academic Press: New York, 1981) where in the fifth step only 9 different pairs are perceived (procedure v obviously cannot trigger further steps here). Forcing the program to do two further steps results in the correct perception of 13 different pairs.

(19) A factor of ca. 30 is to be expected (information given by IBM Deutschland).

(20) A plot of the tenth root of the CPU time vs *n*, however, looks almost the same and results in a least-squares straight line with $r^2 = 0.95$. Thus, it is by no means clear, either theoretically[5b,21] or experimentally, whether the symmetry perception effort increases exponentially or polynominally with increasing *n*.

(21) Wirth, K. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 242.

(22) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172.

Substructure Search Systems. 1. Performance Comparison of the MACCS, DARC, HTSS, CAS Registry MVSSS, and S4 Substructure Search Systems

MARTIN G. HICKS* and CLEMENS JOCHUM

Beilstein Institute, Varrentrappstrasse 40–42, 6000 Frankfurt 90, West Germany

Received December 1, 1989

A comparison of the performance of the substructure search systems MACCS, DARC, HTSS, and S4 has been carried out in-house at the Beilstein Institute, and that of the CAS Registry MVSSS system on STN International at FIZ Karlsruhe was carried out on-line. Included in the comparison were the hit sets, screening efficiency, task times, and elapse times. The results showed that all systems gave similar results in terms of retrieved hit sets, but S4 dramatically out-performed the other systems in terms of task and elapse times. A subsequent test of S4 with a very much larger file showed the search time/file size relationship to be very much less than linear.

Effective management of the information associated with the ca. 10 million chemical compounds known to date is of major importance to chemists in industry and at universities alike.[1-5] The ability of the computer to handle vast amounts of information has brought it to center stage in chemical information management. Recent advances in computer technology, fast mainframes with large storage capacities and cheap personal computers, have led to dramatic changes in chemical information handling. The user friendly interfaces, made possible by the graphical capabilities of the PC, have given easy access to this information to the lay chemist.

A chemical compound can be described and defined in various ways; irrespective of the method adopted, effective searching can only be achieved if a unique compound has a unique description. Moreover, effective storage of the information for a compound is only possible if the description of