On Finding Nonisomorphic Connected Subgraphs and Distinct Molecular Substructures

Gerta Rücker and Christoph Rücker*

Institut für Organische Chemie und Biochemie, Universität Freiburg, Albertstrasse 21, D-79104 Freiburg, Germany

Received July 11, 2000

The problem of finding all nonisomorphic subgraphs of a given graph (all distinct substructures of a given molecular structure) is discussed. A computer program is introduced that first generates all connected subgraphs and then uses a combination of well-discriminating graph invariants to eliminate duplicates. The program is broadly applicable, in particular for molecular graphs which may or may not contain unsaturation or heteroatoms. The number of distinct substructures (N_s), proposed earlier as a measure of a compound's complexity which takes into account its symmetry, is thus automatically obtained. As was to be expected, due to the nature of the problem the computational effort increases exponentially with problem size, whence in most cases complexity measures other than N_s are to be preferred.

INTRODUCTION

Every chemist is trained in perceiving, by inspection, molecular substructures thought to be crucial for a particular purpose. However, such a search is never exhaustive, and computationally the problem of exhaustively enumerating all connected substructures or all distinct connected substructures present in a molecular structure is nontrivial. Obviously this task corresponds to finding all connected subgraphs or all nonisomorphic connected subgraphs of a vertex- and edge-colored graph.

Why should one be interested in knowing molecular substructures? A few possible answers are that often a particular substructure is thought to be responsible for a desired or undesired property or activity,[1,2] that a substructure may correspond to a key building block for synthesis of a target structure,[3] that the number of distinct substructures contained in a substance library may serve as a measure of its diversity,[4] or that the total number of substructures or the number of distinct substructures may be used as measures of a compound's complexity, as recently proposed.[5−7]

An inverse problem is known as substructure search and has often been addressed. There for a given structure all superstructures present in a structure database are sought.[8]

On the contrary, few approaches to finding all substructures of a given structure are available.[1−3a,4]

In the pioneering work of Friedrich and Ugi subgraphs were generated from a hydrogen-containing graph by a breadth-first top-down procedure by successively removing edges. To keep the task manageable assumptions on the special characteristics of molecular graphs were made (e.g. planar graphs only, vertex degree ≤ 4), and some heuristics were used to exclude small fragments thought to be trivial from being generated, stored, and searched.[1,3a] To minimize the generation of duplicates a procedure was used which relies on the partition of graph vertices according to symmetry.

In Klopman's method linear substructures only of limited size (2 to 13 atoms) having at most one-atom branches are generated.[2]

The recent work of Bone and Villar in contrast to its ambitious title does not exhaustively enumerate the molecular substructures.[4] Rather it generates those substructures only which correspond to induced subgraphs,[9] that is to subsets of vertices connected by all edges as present between them in the original graph. As a consequence, for the structure given as a worked example in ref 4, 3-methyltetrahydrofuran, six distinct substructures out of 23 were lost (those of diethyl ether, ethyl propyl ether, butyl methyl ether, isobutyl methyl ether, 2-methylbutan-1-ol, and 3-methylbutan-1-ol). Further, for reasons unknown to us, even some of the induced subgraphs are missing in the other worked example in ref 4, methylcyclopentane, i.e., those coded as 1236, 1245, 12346, 23456 in the numbering used there. Similarly, for cubane Bone and Villar found a total of 167 substructures and but 13 distinct substructures, while in fact there are no less than 2441 subgraphs, 64 of which are distinct, as was found using the method reported here.

The present work was undertaken in the context of molecular complexity measuring. Several measures of a graph's or compound's complexity have recently been proposed, both such that take into consideration the graph's symmetry as a simplifying feature and such that do not.[5−7,10] Among those of the first category the number of distinct connected subgraphs, N_s, was introduced.[5b,c,6] N_s has special appeal in that it does not require explicit input of symmetry information and thus is independent of the (often misleading[10]) visual or algorithmic perception of graph symmetry.[7c] However, no computer program was available hitherto for N_s, so that numerical values had to be found by inspection.

RESULTS AND DISCUSSION

The problem with counting distinct connected subgraphs is 2-fold: It is first required to generate truly all connected subgraphs, both small and large, and second to eliminate exactly the duplicates. We recently reported how for graphs

* Corresponding author phone: +49-761-203-6034; fax: +49-761-203-5987; e-mail: cruecker@organik.chemie.uni-freiburg.de.

10.1021/ci000092b CCC: $20.00
Published on Web 02/06/2001

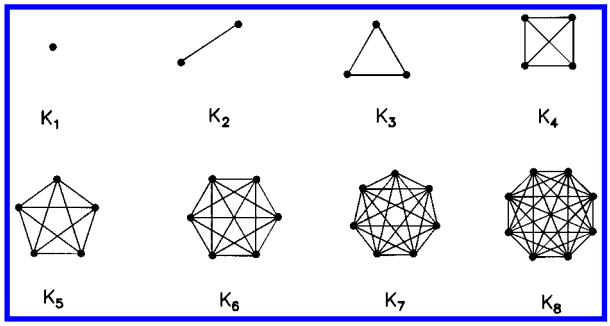

Figure 1. The complete graphs of up to eight vertices.

Table 1. Total Number of Subgraphs (N_t) and Number of Distinct Subgraphs (N_s) of the Complete Graphs up to Eight Vertices

	N_t	N_s
K_1	1	1
K_2	3	2
K_3	10	4
K_4	64	10
K_5	973	31
K_6	31743	143
K_7	2069970	996
K_8	267270040	12113

without multiple bonds or loops (simple undirected graphs) our program SUBGRAPH solved the first problem, which rendered the total number of connected subgraphs, N_t, a viable measure.[11] In short, our procedure is a depth-first bottom-up construction of subgraphs, working on the (hydrogen-suppressed) graph's edge adjacency matrix by successively adding an adjacent edge to a previously found connected subgraph, starting with an edge.[12] No restrictions apply as to the vertex degrees or any other characteristics of the graph, not even connectedness is required. So the program is completely general. As a path-tracing procedure, however, it is plagued by CPU time increasing notoriously exponentially with increasing problem size.

We now tested the program by running it on the complete graphs of n vertices, K_n, whose total numbers of connected subgraphs can mathematically be deduced. In Figure 1 we show the complete graphs up to K_8, in Table 1 the corresponding N_t values are given, as calculated using eq 1 in Appendix 1. Note the exponential increase in N_t for increasing n. Gratifyingly, the constructive program SUBGRAPH for the K_n graphs found exactly these numbers of connected subgraphs.

OUTLINE OF THE PROCEDURE TO FIND NONISOMORPHIC SUBGRAPHS

Here we report on the second step, the problems associated with distilling the set of distinct connected subgraphs out of all connected subgraphs free of redundancy. Strictly speaking, this task requires a general solution of the graph isomorphism problem. Since such is not available, it was clear at the outset that an approximate solution only would be achievable. In all but the smallest molecular graphs there is a total of thousands or millions of connected subgraphs, and to perform pairwise atom-by-atom and bond-by-bond isomorphism tests among these is obviously unfeasible, even if done within classes of constant numbers of atoms and bonds only. For the distinction between subgraphs we therefore had to rely on graph invariants. For the astronomic number of subgraphs to be processed, we could not afford to use any graph invariants but those extremely easily computable. We decided to concentrate on Balaban's J index[13] and on the eigenvalues of the graph's adjacency or distance matrix, $\lambda_1, \lambda_2, ..., \lambda_n$ or $\delta_1, \delta_2, ..., \delta_n$, respectively. J is the most discriminating among the simple topological indices[14] and at the same time is easily calculated, requiring nothing but the graph distance matrix as input, which in turn is easily obtained from the adjacency matrix.[15] Eigenvalues are easily calculated from the adjacency or distance matrix and also are rather well-discriminating invariants, though many pairs of isospectral graphs (graphs having identical adjacency matrix eigenvalues) are known.[16] Distance-isospectral graphs (graphs having identical distance matrix eigenvalues), on the other hand, are scarce. While among the 35 nonanes (4-trees of $n = 9$) there are four pairs of isospectral structures, the first pair of distance-isospectral trees is found in the $n = 17$ family.[17]

Preliminary tests showed that the pairs of graphs known to be J-equivalent[13,14] are, as a rule, resolved by adjacency or distance eigenvalues and that conversely pairs of isospectral and even of distance-isospectral graphs, as a rule, are resolved by J.[16]

For real number graph invariants such as J and matrix eigenvalues, the number of decimal places used for comparisons is critical. We calculated J and the eigenvalues as double precision numbers. To get an idea of a reasonable number of decimals to be used for J comparisons we determined for the subgraphs of K_6 ($N_t = 31743$) the number of distinct J values as a function of the number of decimal places, so ignoring for a moment the eigenvalues. The result was constant (138) in the broad range from 5 to 13 decimal places, while four decimals gave low, 14 decimals gave spuriously high numbers of distinct J values, wherefrom it was decided to generally use eight decimal places for J comparisons.

Similar results were obtained for the number of decimal places of the eigenvalues, e.g. for λ_2 and the subgraphs of K_6 a constant number of distinct values (124) was found in the range of 4 to 11 decimal places. It was decided to generally use seven decimal places for eigenvalue comparisons.

Since sorting by several variables is time-consuming, we do not use all eigenvalues. We first experimented with λ_1 and λ_2 but found that the combination λ_2 and λ_3 is more discriminating. (Since λ_1 is an average degree of a graph's vertices, λ_1 values of similar graphs tend to be similar, in particular λ_1 of all regular graphs of degree d equals d.) For the higher discriminating power of the distance spectrum as compared to the adjacency spectrum we now routinely use the first and last distance matrix eigenvalues δ_1 and δ_n. Few applications in chemistry of eigenvalues other than the leading eigenvalue of any matrix were reported previously.[16f]

The procedure performed by the new program (NIMSG, NonIsoMorphic SubGraphs) thus is the following. For each connected subgraph found by SUBGRAPH,[11] J and the distance eigenvalues are calculated, and the number of vertices (n), edges (m), J, δ_1, and $-\delta_n$ are written in a file, together with the subgraph's structure. After all subgraphs are processed, this file is sorted by n, m, J, δ_1, and $-\delta_n$, finally from the sorted file every entry having all these five variables identical with the previous one is discarded.

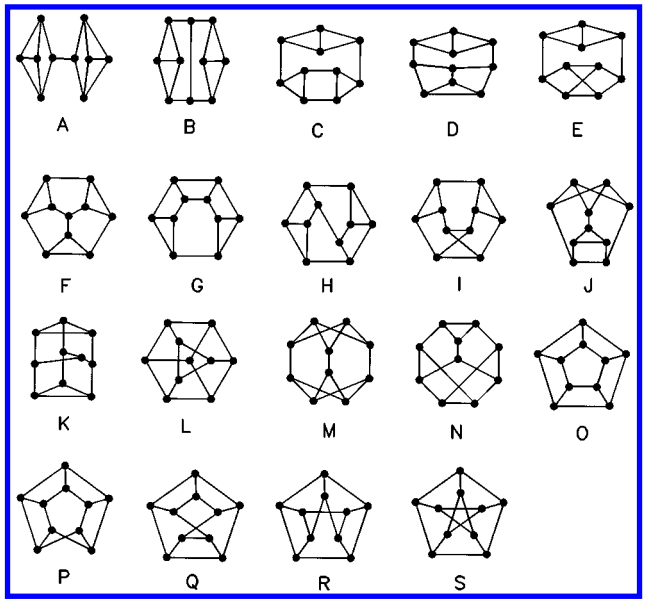

Figure 2. The 19 saturated $(CH)_{10}$ structures.

The program was thoroughly tested on the complete graphs K_n, which are ideal test cases since every simple connected graph of up to n vertices is a subgraph of K_n. The numbers of distinct connected simple graphs of n vertices are known and tabulated, ordered by n and m,[18a] see Appendix 1 also. When the program was run on the complete graphs up to K_7, it exactly reproduced the tabulated numbers (see Table 1, last column, where for each K_n the numbers of distinct connected subgraphs are summed over all n and m).

APPLICATIONS

Complexities of Graphs and Molecules. Numerical values of complexity indices N_t and N_s of some *n*-alkanes, branched hexanes, cycloalkanes, acyclic through polycyclic butanes, and pentacyclic octanes as obtained by the program are given in ref 7c. In Appendix 2 of the present paper all 64 distinct substructures of cubane are listed. Other complexity indices such as the total walk count[7a,c] (twc) are more easily obtained than N_t and N_s, so that the latter are useful mostly for regular graphs, for which twc is constant within a family of constant size and degree. Therefore we illustrate here the usefulness of the program in comparing the complexities of $(CH)_{10}$ molecular graphs. In Figure 2 the structures of all 19 saturated $(CH)_{10}$ isomers[18b] (including the very hypothetical ones) are arranged in increasing order of N_t. Since size, branching, and cyclicity are constant throughout this graph sample, the considerable differences in N_t reflect some more subtle structural differences such as the number of small-size rings and the line-connectivity. The line-connectivity is the minimal number of edges which when cleaved cause the graph to fall apart. Table 2 gives N_t, line-connectivity, numbers of three- and four-membered rings, N_s values, and the number of graph-theoretically different classes of vertex pairs (a measure of symmetry) for these structures. It is easily seen that N_t increases with increasing line-connectivity and with decreasing numbers of small rings. Index N_s can be understood as derived from N_t by the influence of symmetry, in that it increases with decreasing symmetry (increasing number of pair classes). The Peterson graph (**S**) for example is 3-connected and does not contain small rings; its N_t is therefore highest, but due to its extremely high symmetry its N_s is lowest among these isomers. These trends are the same as seen earlier for the five saturated $(CH)_8$ isomers.[7c]

Table 2. Some Invariants of the 19 Saturated $(CH)_{10}$ Structures.

	N_t	line-connectivity	3-rings	4-rings	N_s	classes of pairs
A	6306	1	4	6	319	11
B	9542	2	4	2	488	11
C	10318	2	4	2	679	15
D	10885	2	3	3	987	22
E	11220	2	2	6	521	14
F	12796	3	3	0	635	10
G	12847	3	2	3	906	15
H	13021	3	2	2	1155	25
I	13518	3	2	1	745	12
J	13613	3	1	4	990	22
K	13697	3	2	0	540	8
L	14124	3	1	3	775	12
M	14296	3	0	6	275	7
N	14310	3	1	2	1162	25
O	14770	3	0	5	390	5
P	14819	3	0	5	305	5
Q	15167	3	0	3	731	15
R	15367	3	0	2	592	12
S	15770	3	0	0	165	2

Systematic Search for *J*-Equivalent Graphs. The program can be used for systematic searches for *J*-equivalent or for (distance-)isospectral graphs. The difference between the set of distinct connected subgraphs of K_n found by the complete procedure and the set resulting from switching off one or the other criterion is obviously the set of graphs distinguished from others only by the switched-off criterion.

Thus it is known that there exist 143 connected simple graphs of up to six vertices,[18a] and the program found exactly 143 distinct connected subgraphs of K_6. When the eigenvalues were switched off, only 138 distinct subgraphs remained, so five graphs were lost for being *J*-equivalent to other graphs of the same number of edges and vertices. Four pairs of *J*-equivalent simple graphs of $n = 6$ were published,[13,14] so it was obvious that another such pair must exist. This pair of graphs was then easily identified by inspection of the two sets, it is the pair of pentacyclic hexanes shown in Figure 3.

Figure 3. Two *J*-equivalent graphs of six vertices, $J = 2.53509699$.

Similarly among the distinct connected subgraphs of K_7 ($N_s = 996$) there are 900 distinct *J* values only, so (in addition to the above five pairs of $n = 6$) 46 pairs, 10 triplets, 7 quadruplets, and even a quintuplet of *J*-equivalent simple connected tetracyclic through decacyclic graphs of $n = 7$ were identified, all not known previously to the best of our knowledge. Figure 4 shows the five simple graphs of $n = 7$ and $m = 12$ that are *J*-equivalent.

GENERALIZATION FOR EDGE- AND VERTEX-COLORED GRAPHS

Multigraphs, Unsaturation. A double, triple, or aromatic bond connecting atoms i and j can be represented in the molecule's adjacency matrix as $a_{ij} = a_{ji} = 2$, 3, or 1.5, respectively. The corresponding distance matrix element then

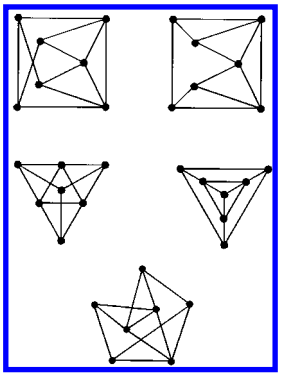

Figure 4. Five *J*-equivalent graphs of seven vertices, $J = 2.42646877$.

is $d_{ij} = d_{ji} = 0.5$, 0.333, or 0.667, respectively.[19] Thus both the eigenvalues and index *J* contain information on multiple bonds (unsaturation). Again it was checked that the known pairs of isospectral multigraphs[20] are resolved by their *J* values.

Graphs with Loops, Heteroatoms. The presence of a heteroatom in position *i* is represented in the adjacency matrix as a_{ii} = color(atom *i*) ≠ 0. The adjacency matrix eigenvalues thus will respond to the presence of heteroatoms. To introduce the information on heteroatoms into *J* we preferred not to touch the distance matrix (after all the graph-theoretical distances in a molecule are not changed on substitution of a heteroatom for a carbon atom), rather following Balaban we multiply the distance sum of a heteroatom by an (arbitrary) factor which is a function of its nature.[21] For our purpose, the discrimination of nonisomorphic graphs, the particular values used for the colors of atoms and for the factors are uncritical, we obtained good results using values such as color(carbon) = 0, color(oxygen) = 1, color(nitrogen) = 2, and factors $2^{0.1 \cdot \text{color(atom)}}$, though it is perfectly possible to use noninteger values for both the colors and the factors, as proposed by Balaban.[21] The particular numbers used here result in *J* values for heteroatom-containing substructures similar to those of the all-C analogues, so that in the sorted file substructures differing only in their heteroatom composition appear in close neighborhood, e.g. *J* values of three-atom substructures are as follows:

C−C−C 1.63299316, C−C−O 1.60518029, C−O−C 1.57736742, C−C−N 1.57831483, C−N−C 1.52363649, C−C═C 2.18749610, C−C═O 2.14816276, C−C═N 2.11016926, C−N═C 2.04100603.

The distance eigenvalues are in a similar manner enriched with information on heteroatoms: Each row and column of the distance matrix is multiplied by the factor for the nature of the respective atom, as given above, to result in the "heteroenriched distance matrix", eigenvalues of which are then calculated und used as before.

A test case for treatment of heteroatom information are heteroatom-containing molecular graphs derived from endospectral graphs. A graph is endospectral if identical perturbations at one or the other of two special vertices (endospectral vertices) result in two nonisomorphic but isospectral graphs.[22] A classical case of an endospectral graph is the ethylcyclohexane graph, with ring position 4 and the side chain α position being endospectral positions. Its two oxygen analogues bearing O in either of these positions (two isospectral graphs shown in Figure 5) do not pose any problem to our procedure since they are neither distance-isospectral nor *J*-equivalent.

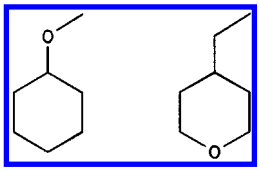

Figure 5. Two colored graphs that are isospectral but not distance-isospectral nor *J*-equivalent.

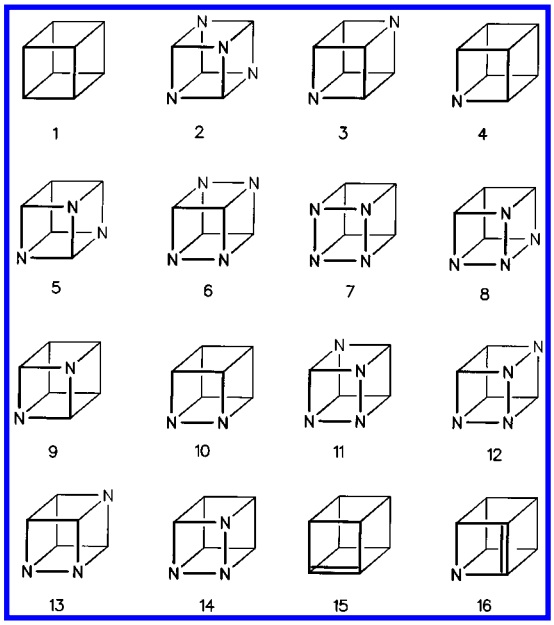

Figure 6. Sixteen colored cubane derivatives.

Table 3. Some Invariants of the Cubane Derivatives in Figure 6

	N_s	classes of pairs	NN bonds
1	64	3	0
2	101	4	0
3	214	6	0
4	302	9	0
5	309	9	0
6	328	8	2
7	331	7	4
8	392	9	3
9	433	12	0
10	469	11	1
11	856	18	2
12	895	16	3
13	904	18	1
14	918	18	2
15	318	11	0
16	1244	28	0

APPLICATIONS

In Figure 6 all nitrogen derivatives of cubane (azacubanes) containing zero to four nitrogen atoms are shown in the order of increasing N_s (N_t throughout is 2441). For comparison cubene and an azacubene are also included. In Table 3 the N_s values, the number of graph-theoretically different classes of vertex pairs (a measure of symmetry), and the number of nitrogen−nitrogen bonds (a rough measure of closeness of the heteroatoms) are given. The same N_s values were found independent of the particular number chosen to represent the color of a nitrogen atom. Note that N_s increases with

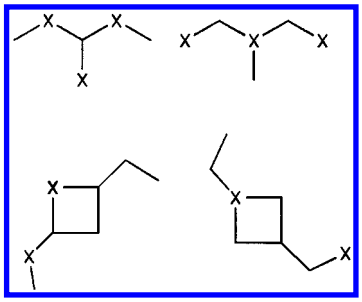

Figure 7. Two pairs of colored isospectral and *J*-equivalent but not distance-isospectral graphs.

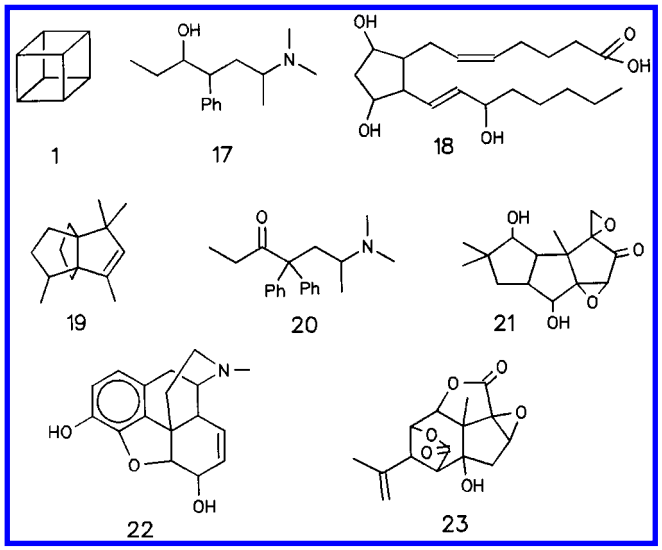

Figure 8. Some molecular graphs whose N_t and N_s appear in Table 4.

Table 4. N_t and N_s Values for the Structures in Figure 8 and CPU Times [s] Used To Obtain This Information

	N_t	N_s	n	m	μ	CPU time [s]
1	2441	64	8	12	5	3
17	2007	897	17	17	1	5
18	6460	5620	25	25	1	30
19	17232	4152	15	17	3	42
20	44548	6552	23	24	2	204
21	887784	439267	20	24	5	4281
22	826764	599560	21	25	5	4772
23	1408347	1087080	21	25	5	7704

decreasing symmetry (increasing number of pair classes) and with increasing closeness of the heteroatoms.

As a spin-off this study generated several pairs of small isospectral colored graphs. Thus as subgraphs of triazacubane **5** two acyclic isospectral colored graphs of $n = 6$ were found (Figure 7, top) by comparison of the sets of subgraphs resulting from the procedure as described and the analogous procedure using adjacency eigenvalues. Similarly, two monocyclic isospectral colored graphs of $n = 8$ were found as subgraphs of diazacubane **9** (Figure 7, bottom). These pairs of graphs are isospectral independent of the nature of the heteroatoms, i.e., independent of the particular number used as heteroatom color. We do not know of any previous results on isospectral colored graphs.

To illustrate the scope of the program in Figure 8 a few typical molecular structures encountered in Organic Chemistry are given, while Table 4 lists their N_t and N_s values, the numbers of non-hydrogen atoms (n), of bonds (m), and of cycles (μ) as indicators of problem size, and the CPU times used (on a SG Indigo workstation, 150 MHz, R5000 coprocessor). The examples, given in the order of increasing CPU times, include cubane (**1**), prostaglandin $F_{2\alpha}$ (**18**), modhephene (**19**), methadone (**20**), coriolin (**21**), morphine (**22**), and picrotoxinin (**23**). Note that the number of cycles ($\mu = m - n + 1$) along with m is the decisive descriptor of problem size.

LIMITATIONS

The principal limitation is the run time limitation. As said above and as seen in the last column of Table 4 finding nonisomorphic subgraphs is a task of computer time exponentially increasing with the size of the problem.

The program in the form described obviously cannot discriminate graphs which are *J*-equivalent *and* have identical δ_1 and δ_n values. Therefore if a graph contains two such graphs as subgraphs, one of them will erroneously be discarded. In such cases use of *all* distance eigenvalues may be helpful. Graphs which at the same time are *J*-equivalent and distance-isospectral will still not be discriminated. Such graphs do exist, e.g. the highly regular graphs which cause problems to most graph isomorphism or graph automorphism programs, as compiled by Weisfeiler[23] and Mathon.[24,25] Such graphs either exhibit high vertex degrees[23,24] or are rather large (a pair of regular graphs of degree 3 and $n = 40$[26]) and so are highly improbable to be ever encountered in chemistry.

Of course there are single-number graph invariants more discriminant than *J*; however, they typically require knowledge of all paths in a (sub)graph.[27] To avoid an exponentially increasing task (path-tracing) within an exponentially increasing task (finding subgraphs), we decided not to use such invariants.

APPENDIX 1

The number of all connected subgraphs of K_n is derived as follows: This number is the sum of the counts of all connected graphs on k vertices for $k = 1, ..., n$. For each k there are $\binom{n}{k}$ possibilities to select k out of n vertices. Assume that there are C_k connected graphs on k labeled vertices, then

$$N_t(K_n) = \sum_{k=1}^{n} \binom{n}{k} C_k \qquad (1)$$

The numbers C_k are obtained using the following recursion formula which is a transcription of an equation given by Gilbert.[28]

$$C_{k+1} = 2^{\binom{k+1}{2}} - \sum_{i=0}^{k-1} \binom{k}{i} C_{i+1} \cdot 2^{\binom{k-i}{2}} \qquad (2)$$

C_1 obviously is 1, so eq 2 yields the sequence $C_2 = 1$, $C_3 = 4$, $C_4 = 38$, $C_5 = 728$, $C_6 = 26704$, $C_7 = 1866256$, $C_8 = 251548592$, ... This sequence and the reference to Gilbert can be found in Sloane's Online Encyclopedia of Integer Sequences.[29]

Insertion of these numbers into (1) yields the sequence of N_t given in Table 1.

The number of distinct connected subgraphs of K_n is the sum of the counts of all (connected simple undirected) distinct graphs on k vertices for $k = 1, ..., n$. Thus,

$$N_s(K_n) = \sum_{k=1}^{n} G_k \quad (3)$$

where G_k is the number of connected graphs on k unlabeled vertices.

The sequence of G_k is found in the above Online Encyclopedia as the sequence 1, 1, 2, 6, 21, 112, 853, 11117, ...

Insertion of these numbers into (3) yields the sequence of N_s given in the last column of Table 1.

APPENDIX 2

Here the 64 distinct substructures of cubane are listed in the order of increasing n, m, and J: methane, ethane, propane, *n*-butane, 2-methylpropane, cyclobutane, *n*-pentane, 2-methylbutane, methylcyclobutane, *n*-hexane, 2-methylpentane, 3-methylpentane, 2,3-dimethylbutane, cyclohexane, ethylcyclobutane, 1,3-dimethylcyclobutane, 1,2-dimethylcyclobutane, bicyclo[2.2.0]hexane, *n*-heptane, 2-methylhexane, 3-methylhexane, 3-ethylpentane, 2,3-dimethylpentane, *n*-propylcyclobutane, 1-ethyl-3-methylcyclobutane, methylcyclohexane, isopropylcyclobutane, 1-ethyl-2-methylcyclobutane, 1,2,3-trimethylcyclobutane, bicyclo[3.1.1]heptane, 2-methylbicyclo[2.2.0]hexane, tricyclo[3.1.1.$0^{3,6}$]heptane, *n*-octane, 3-methylheptane, 2,5-dimethylhexane, 3-ethylhexane, 2,3-dimethylhexane, 3,4-dimethylhexane, *n*-butylcyclobutane, cyclooctane, 1,3-diethylcyclobutane, 1-methyl-2-propylcyclobutane, ethylcyclohexane, *sec*-butylcyclobutane, 1-isopropyl-3-methylcyclobutane, 1,2-diethylcyclobutane, 1,4-dimethylcyclohexane, 1-ethyl-2,3-dimethylcyclobutane, 1,2-dimethylcyclohexane, 1,2,3,4-tetramethylcyclobutane, bicyclobutyl, 2-ethylbicyclo[2.2.0]hexane, bicyclo[4.2.0]octane, 2,5-dimethylbicyclo[2.2.0]hexane, 2-methylbicyclo[3.1.1]heptane, 2,3-dimethylbicyclo[2.2.0]hexane, 6-methylbicyclo[3.1.1]-heptane, bicyclo[2.2.2]octane, tricyclo[4.2.0.$0^{2,5}$]octane, 2-methyltricyclo[3.1.1.$0^{3,6}$]heptane, tricyclo[3.1.1.$1^{2,4}$]octane, tricyclo[4.2.0.$0^{3,8}$]octane, secocubane, and cubane.

REFERENCES AND NOTES

(1) (a) Friedrich, J.; Ugi, I. Substructure Searching and Structure Property Locating by Means of Subgraph Generation. *MATCH − Commun. Math. Comput. Chem.* **1979**, *6*, 201−211. (b) Friedrich, J.; Ugi, I. Substructure Retrieval and the Analysis of Structure−Activity Relations on the Basis of a Complete and Ordered Set of Fragments. *J. Chem. Res. (S)* **1980**, 70.

(2) (a) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321. (b) Klopman, G.; Balthasar, D. M.; Rosenkranz, H. S. Application of the Computer-Automated Structure Evaluation (CASE) Program to the Study of Structure-Biodegradation Relationships of Miscellaneous Chemicals. *Environ. Toxicol. Chem.* **1993**, *12*, 231−240.

(3) (a) Ugi; I.; Bauer, J.; Bley, K.; Dengler, A.; Fontain, E.; Knauer, M., Lohberger, S. Computer-Assisted Bilateral Synthesis Design: A Status Report. *J. Mol. Struct. (THEOCHEM)* **1991**, *230*, 73−82. (b) Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2613−2633.

(4) Bone, R. G. A.; Villar, H. O. Exhaustive Enumeration of Molecular Substructures. *J. Comput. Chem.* **1997**, *18*, 86−107.

(5) (a) Bertz, S. H.; Herndon, W. C. Similarity of Graphs and Molecules. In *Artificial Intelligence Applications in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; American Chemical Society: Washington, DC, 1986; pp 169−175. (b) Bertz, S. H.; Sommer, T. J. Rigorous Mathematical Approaches to Strategic Bonds and Synthetic Analysis Based on Conceptually Simple New Complexity Indices. *Chem. Commun.* **1997**, 2409−2410. (c) Bertz, S. H.; Wright, W. F. The Graph Theory Approach to Synthetic Analysis: Definition and Application of Molecular Complexity and Synthetic Complexity. *Graph Theory Notes of New York* **1998**, *35*, 32−48. (d) Bertz, S. H.; Zamfirescu, C. M. New Complexity Indices Based on Edge Covers. *MATCH − Commun. Math. Comput. Chem.* **2000**, *42*, 39−70. (e) Bertz, S. H. Complexity of Molecules and their Synthesis. In *Complexity in Chemistry*; Bonchev, D., Rouvray, D. H., Eds.; Gordon & Breach: Reading, U.K. In press.

(6) (a) Bonchev, D. Novel Indices for the Topological Complexity of Molecules. *SAR QSAR Environ. Res.* **1997**, *7*, 23−43. (b) Bonchev, D. Overall Connectivity and Topological Complexity: A New Tool for QSPR/QSAR. In *Topological Indices and Related Descriptors in QSAR and QSPR*, Devillers, J., Balaban, A. T., Eds.; Gordon & Breach, 1999; pp 361−401. (c) Bonchev, D.; Gordeeva, E. Hierarchical Partially Ordered Sets Based on Topological Complexity. *MATCH − Commun. Math. Comput. Chem.* **2000**, *42*, 85−117.

(7) (a) Rücker, G.; Rücker, C. Walk Counts, Labyrinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules. *J. Chem. Inform. Comput. Sci.* **2000**, *40*, 99−106. (b) Nikolić, S.; Tolić, I. M.; Trinajstić, N. On the Complexity of Molecular Graphs. *MATCH − Commun. Math. Comput. Chem.* **1999**, *40*, 187−201. (c) Nikolić, S.; Trinajstić, N.; Tolić, I. M.; Rücker, G.; Rücker, C. On Molecular Complexity Indices. In *Complexity in Chemistry*; Bonchev, D., Rouvray, D. H., Eds.; Gordon & Breach: Reading, U.K. In press.

(8) (a) Barnard, J. M. Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532−538. (b) Willett, P. Matching of Chemical and Biological Structures Using Subgraph and Maximal Common Subgraph Isomorphism Algorithms. In *IMA Vol. Math. Appl.*; Springer: New York, 1999; Vol. 108, pp 11−38.

(9) (a) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969. (b) Gould, R. *Graph Theory*; Benjamin, Menlo Park, CA, 1988.

(10) Randić, M. On the Concept of Molecular Complexity. *Croat. Chem. Acta* In press.

(11) Rücker, G.; Rücker, C. Automatic Enumeration of All Connected Subgraphs. *MATCH − Commun. Math. Comput. Chem.* **2000**, *41*, 145−149.

(12) To obtain numbers identical to those given in refs 5 and 6 we now have the single-vertex subgraphs counted, too.

(13) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(14) (a) Balaban, A. T.; Quintas, L. V. The Smallest Graphs, Trees, and 4-Trees with Degenerate Topological Index J. *MATCH − Commun. Math. Comput. Chem.* **1983**, *14*, 213−233. (b) Razinger, M.; Chrétien, J. R.; Dubois, J. E. Structural Selectivity of Topological Indexes in Alkane Series. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23−27.

(15) Müller, W. R.; Szymanski, K.; Knop, J. V., Trinajstić, N. An Algorithm for Construction of the Molecular Distance Matrix. *J. Comput. Chem.* **1987**, *8*, 170−173.

(16) For leading references and examples of pairs of isospectral graphs that are resolved by J see the following: (a) Balaban, A. T.; Harary, F. The Characteristic Polynomial Does Not Uniquely Determine the Topology of a Molecule. *J. Chem. Doc.* **1971**, *11*, 258−259. (b) Mihalić, Z.; Veljan, D.; Amić, D.; Nikolić, S.; Plavsić, D.; Trinajstić, N. The Distance Matrix in Chemistry. *J. Math. Chem.* **1992**, *11*, 223−258. (c) Hosoya, H.; Nagashima, U.; Hyugaji, S. Topological Twin Graphs. Smallest Pair of Isospectral Polyhedral Graphs with Eight Vertices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 428−431. (d) Cvetković, D.; Rowlinson, P.; Simić, S. *Eigenspaces of Graphs*; Cambridge University Press: Cambridge, 1997. (e) Balasubramanian, K.; Basak, S. C. Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 367−373. (f) Randić, M.; Vracko, M.; Nović, M. Eigenvalues as Molecular Descriptors. In *QSAR/QSPR Studies by Molecular Descriptors*; Diudea, M. V., Ed.; Nova Science Publ.: Commack, NY, in press.

(17) Cvetković, D. M.; Doob, M.; Gutman, I.; Torgasev, I. *Recent Results in the Theory of Graph Spectra*; Elsevier: Amsterdam 1988; p 128.

(18) (a) Read, R. C.; Wilson, R. J. *An Atlas of Graphs*; Clarendon Press: Oxford, 1998; Table on p 7. (b) Read, R. C.; Wilson, R. J. *An Atlas of Graphs*; Clarendon Press: Oxford, 1998; p 127.

(19) Balaban, A. T.; Filip, P. Computer Program for Topological Index J (Average Distance Sum Connectivity). *MATCH − Commun. Math. Comput. Chem.* **1984**, *16*, 163−190.

(20) Randić, M.; Baker, B. Isospectral Multitrees. *J. Math. Chem.* **1988**, *2*, 249−265.

(21) Balaban, A. T., Topological Index J for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties. *MATCH − Commun. Math. Comput. Chem.* **1986**, *21*, 115−122.
(22) (a) Randić, M.; Kleiner, A. F. On the Construction of Endospectral Graphs. *Ann. New York Acad. Sci*. **1989**, *555*, 320−331. (b) Randić, M.; Barysz, M.; Nowakowski, J.; Nikolić, S.; Trinajstić, N. Isospectral Graphs Revisited. *J. Mol. Struct. (THEOCHEM)* **1989**, *185*, 95−121. (c) Rücker, C.; Rücker, G. Understanding the Properties of Isospectral Points and Pairs in Graphs: The Concept of Orthogonal Relation. *J. Math. Chem.* **1992**, *9*, 207−238.
(23) Weisfeiler, B. *On the Construction and Identification of Graphs*; Lecture Notes in Mathematics No. 558; Springer: Berlin 1976.
(24) Mathon, R. Sample Graphs for Isomorphism Testing. *Proc. 9th S.-E. Conf. Combinatorics, Graph Theory and Computing* **1978**, 499.
(25) Rücker, G.; Rücker, C. On Using the Adjacency Matrix Power Method for Perception of Symmetry and for Isomorphism Testing of Highly Intricate Graphs. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 123−126.
(26) Cai, J.; Fürer, M.; Immerman, N. An Optimal Lower Bound on the Number of Variables for Graph Identification. *Combinatorica* **1992**, *12*, 389−410.
(27) Hu, C.-Y.; Xu, L. Developing Molecular Identification Numbers by an All-Path Method. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 311−315. See also references therein.
(28) Gilbert, E. N. Enumeration of Labelled Graphs. *Can. J. Math*. **1956**, *8*, 405−411.
(29) http://www.research.att.com/%7Enjas/sequences.

CI000092B