(4) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(5) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(6) Balaban, A. T. *Theor. Chim. Acta* **1979**, *53*, 355.
(7) Razinger, M.; Chretien, J. R.; Dubois, J. E. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23.
(8) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(9) Program runs on Apple IIe home computer. Revised version involving prime number weights is available upon request to noncommercial users.
(10) Figueras, J., a preprint, private communication.
(11) Szymanski, K., Müller, W.R., Knop, J. V., Trinajstić, N., a preprint, private communication.
(12) Knop, J. V.; Müller, W. R. Jericević, Z.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1983**, *21*, 91.
(13) Triansjtić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vol. II, Table 4, 153.
(14) Randić, M., submitted for publication in *Croat. Chem. Acta.*

Compact Molecular Codes[†]

MILAN RANDIĆ*

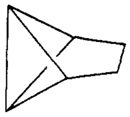
Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311, and Ames Laboratory, Iowa State University, Ames, Iowa 50011

Received May 22, 1985

In this paper we introduce structural codes that are easy to derive for most molecular skeletal forms of chemical interest yet satisfy numerous desirable properties, being linear, unique, reconstructable, derivable and decodable by hand, brief, based on familiar symbols, easily comprehensible, and efficient. Codes in general imply a resolution of the following problems: (1) canonical numbering of atoms; (2) graph isomorphism; (3) discernment of the symmetry of the structure (graph). Our approach resolves these problems in a remarkably simple way, at least for the examples selected. The approach is based on an extension of the *N*-tuple codes of Knop and co-workers, which apply only to trees (acyclic graphs). By excising selected vertices in a polycyclic graph, one arrives at subspanning trees for the polycyclic graph for which *N*-tuple codes of Knop et al. are adopted. Subsequently, such an incomplete code is augmented by the listing of adjacencies for the vertices, which represent ring closures. This paper presents numerous illustrations of the compact codes and discusses the rules that govern construction of the compact codes and the relative ease of the search for the codes. In order to more clearly show the relative simplicity of the new codes, we end with a comparison of the compact codes with a selection of alternative codes currently in use.

INTRODUCTION

The history of chemical nomenclature and the search for codes with desirable qualities is old and continuing. As early as 1881, Friedrich Konrad Beilstein[1] initiated a nomenclature system that is still of interest and serves as a basis for the naming of numerous structures. In 1900, Adolph von Bäyer[2] suggested the nomenclature for bridged bicyclic molecules, which is still the basis for the systematic naming of compounds like norbornane etc. Already, the extension of the nomenclature to tricyclic systems pointed to some difficulties. Besides *digits* used to indicate the number of carbon atoms in individual bridges, one needs *labels* to indicate the particular bridges in polycyclic structures. For example



is named (by IUPAC rules) tricyclo$[3.1.0.0^{2,6}]$hexane. Observe two *kinds* of uses of digits: 3.1.0.0 indicates *structural* data, the number of carbon atoms in the four branches of the structure, and 2,6 is a *label* referring to selected carbon atoms.

Table I. List of Requirements of Codes as Proposed by Read[4]

(1) codes should be a linear string of symbols
(2) coding algorithm should produce a unique code
(3) structure should be recoverable by a clearly defined process
(4) coding should be simple; preferably, it should be possible to code a compound by hand (without the use of a computer)
(5) decoding process should be simple, preferably one that can be carried out by hand
(6) coding process should not depend on chemical intuition or properties of chemicals
(7) coding should not depend on any list of names or other nonsystematic items
(8) codes should be brief
(9) codes should be pronounceable
(10) symbols used should be familiar (available on standard typewriter or computer keyboard)
(11) codes should be easily comprehensible
(12) coding and decoding algorithms should be efficient

Much progress followed the early interest in chemical nomenclature. Coding is important not only for chemical documentation but also for enumeration of isomers and the construction of graphs. Finally, structural codes are of interest in structure–property and structure–activity studies.[3] Recently, Read[4] reviewed desirable qualities for codes for chemical structures. These are listed in Table I. Various codes proposed in the past satisfy to some degree several of the suggested desirable features, but no code has been found that would satisfy all the requirements satisfactorily. Not all the requirements are, however, equally important, nor can they be resolved with similar efforts. Of the attributes required of codes, according to Goodson,[5] the ones most difficult to comply with are that names be based on linear character strings to permit lexicographic ordering and that names be brief. This means that codes should be short and that standard symbols (e.g., digits, letters, and other common mathematical or typographical symbols, such as brackets, slashes, asterisks, etc.)

[†] Dedicated to Professor Vladimir Prelog on the occasion of his 80th birthday. Ames Laboratory is operated by the Iowa State University for the U.S. Department of Energy, under Contract W-7405-Eng-82, Division of Basic Sciences. This work is supported in part by the Office of the Director.

[*] Address correspondence to the author at Drake University.

0095-2338/86/1626-0136$01.50/0

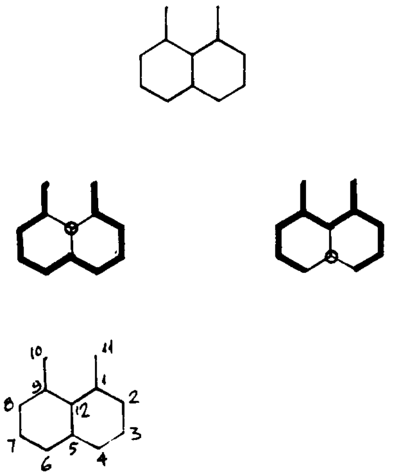


Figure 1. Bicyclic graph and two alternative ways of obtaining an acyclic subgraph by erasure of a *single* vertex.

should be used so that ordering of "words" is possible. Let us discuss these two attributes, which we can condense into a single requirement: codes ought to be simple. The requirement that one use as few symbols as possible (in order to make the codes short) and the requirement that the codes be unique oppose one another. It is not difficult to construct structural codes that will contain all the information on the connectivity of a molecule. What is apparently difficult is to design structural codes that will have all the structural information and yet be *brief.* To paraphrase Albert Einstein,[6] "The codes should be as simple as possible but not simpler."

OUTLINE OF THE PAPER

We will first consider the *N*-tuple codes of Knop, Müller, Jeričević, and Trinajstić[7] designed for the unique characterization of *trees*. Next, we will extend their codes to cyclic structures by selecting an "acyclic" fragment to which the *N*-tuple code applies. For the extension of the code to cyclic (polycyclic) structures, it is imperative that the "acyclic backbone" fragment can be found easily, as one of a *few* alternatives rather than one of many. This requirement immediately eliminates *spanning trees* as a practical basis for extending *N*-tuple codes, because there are too many spanning trees even in the case of simple graphs. For example, the simple bicyclic skeleton shown in Figure 1 already has 20 symmetry unrelated spanning trees. We will show that subgraphs obtained by erasing a vertex (or several vertices) offer an excellent starting point for extending the acyclic codes of Knop et al. Rules for selecting the vertex (or vertices) that when excised will produce an acyclic fragment have to be developed. In the case of the graph of Figure 1, following our rules (vide infra) one derives only two acyclic subgraphs to be further examined. We will discuss in some details how to select a unique subspanning tree among the few alternatives. The rest of this paper gives illustrations on a selected class of compounds. Examples have a twofold purpose. Besides illustrating similarities and differences among the codes of similar compounds, they also provide the reader with a sufficient experience that allows one to derive the codes for the compounds of his/her interest. Finally, the examples ought to convince skeptics that in most cases of chemical interest the code can be obtained after a *few* trials. We end with a comparison of our compact codes with several available alternatives. In addition, briefly and somewhat superficially, we outline how one can extend the codes to record heteroatoms and even molecular stereochemistry. These extensions, of prime importance in use for practical information systems, are outside the scope of this paper. Equally, the examples discussed and the illustrations shown in this paper do not exhaust all existent or conceivable structural variations of chemical interest. These important problems will be addressed in the subsequent publications of this series.[8,9] Hence, the title of the paper "Compact Molecular Codes" is not a misnomer, as it applies to the series of papers to follow, where besides molecular graphs also molecular structures will be considered. For the same reason, one may object that discussions of conciseness are somewhat beside the point, since the code is incomplete in its present form. But here we restricted attention to molecular graphs, and already for them the present codes are compact when compared to *alternatives*. After the present codes are augmented to take into account bond types and atom types, and even stereochemistry, they are likely to remain still relatively brief, because similar generalizations will augment other codes with which the comparison is made.

It should be recognized that deriving a code implies a *definite* numbering for the vertices. A task that is inherently *n*! in character is here dramatically reduced for most molecular graphs to less than *n* attempts! Finding the compact code thus became a problem that can be solved "on the back of an envelope", a task for which one does not necessarily need a computer, at least for a typical chemical molecule having several rings and a few dozen atoms.

TREES

A graph is fully determined by its adjacency matrix, but the adjacency matrix does not give a compact representation of a graph. For most molecular graphs and in particular for trees (acyclic structures), the matrix is sparse (i.e., has many zero entries). A tree can be fully characterized with fewer data. A desirable feature for codes is that their length is the same for graphs of the same size and similar complexity. This condition has not been explicitly stated by Read,[5] but we feel that it ought to be included. The length of the code should be a function of the number of vertices and the number of edges and should not depend strongly on a mode of their interconnections. Under such provisions, the length of a code also becomes a qualitative measure of the complexity of the compound. An example of "uniform" (i.e., of equal length for isomers) codes are those in which each vertex is represented by a single symbol in the code. Alternatively, one may identify each bond (edge) in a graph by a single symbol. Then, the lower bound for the length of a code is *n*, the number of vertices (or $n - 1$, the number of edges). We reviewed a number of codes for trees and found that the codes of Knop, Müller, Jeričević, and Trinajstić[7] have the desired feature: a full characterization of a tree having *n* vertices is accomplished by the use of *n* (and no more) digits. The code, referred to as *N*-tuple, consists of a string of numbers, each representing the valency of each vertex decreased by 1 ($d - 1$) except for the starting vertex (of the highest valency), which is represented by *d* (rather than $d - 1$). To obtain the *N*-tuple code, one first has to identify the vertices of the highest valency and select among them one that will result in a code that, lexically speaking, produces the largest number. After the starting vertex is located, that vertex is erased, and the disjoint fragments produced are examined. Each of the new fragments is a smaller tree. Hence, their codes are derived and combined in such a way that the result corresponds to the *maximal* number. For the tree shown in Figure 2, we thus obtain

3 ((1 1 1 2 1 1 0) (1 0)) (0) (0)

Here we inserted the parentheses only to make more visible *groupings* of vertices in a same branch. The parentheses are unnecessary (and have not been even suggested in the original work of Knop et al.). After they are omitted the code becomes

3 1 1 1 2 1 1 0 1 0 0 0

If we were to use the other trivalent vertex of the tree in Figure

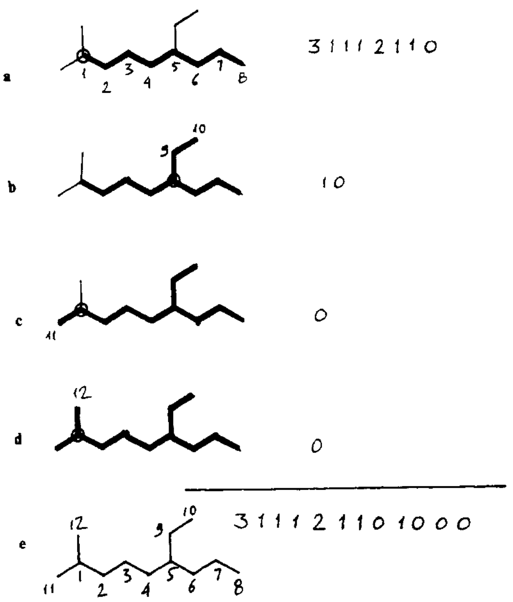


Figure 2. Illustration of the search for the N-tuple code of a tree: (a) locate vertices of the highest valency; (b) locate the longest path; (c) backtrack to the last past branching point to visit all vertices in that branch; (d) continue the process till all vertices branching from the longest path have been accounted for; (e) locate the next longest path and continue the process until all vertices in all paths have been recorded.

2 to start the code construction, we would obtain a code with a *smaller* lexical value:

3 1 1 1 2 0 0 1 1 0 1 0

It is discarded as noncanonical. The N-tuple code thus consists of n digits, each representing a single vertex. The *order* in which vertices appear in the code induces labels from 1 to n individual vertices. The code necessarily ends with a zero. Hence, we have $n - 1$ essential symbols.

The question is, can an N-tuple code for a tree be generalized, or augmented, so that it would characterize uniquely cyclic structures? Many graph theoretical algorithms cannot be extended from trees to cyclic structures. In general, extending an acyclic code to polycyclic structures is a novel task. Without a conceptual "break-through", one cannot expect significant progress, unless the extension is trivial.

SEARCH FOR COMPACT CODES FOR CYCLIC STRUCTURES

Before we outline a generalization of the N-tuple codes for trees to polycyclic structures based on a novel concept of subspanning trees, we will briefly review the coding recently developed by Herndon[10] because of some similarity with our codes. Herndon's scheme produces codes that are relatively brief. Herndon suggested molecular codes based on a canonical labeling. The essential ingredients of Herndon's approach are (a) introduction of canonical lables for the vertices (atoms) and (b) listing of atoms and their neighbors. Thus, Herndon's codes represent a canonization of vertex neighbor lists. A graph with N vertices and E edges has therefore $N + E$ entries, because each atom is listed by its label, followed by a list of its edges (each edge being listed only the first time). In the case of the graph of Figure 1, Herndon's code has 25 entries, 12 for labeling vertices and 13 for indicating edges. For canonical numbering, Herndon adopts a modification of the so-called extended connectivities[11] initially introduced by Morgan.[12] The coding requires some preliminary work, which in most cases can be performed without resorting to a computer. After deriving the modified connectivity values, one establishes ranks for individual vertices and assigns to them labels $1 - n$. Construction of the code is straightforward once labels are known. For the graph of Figure 1, one obtains

(01)020304(02)0506(03)0711(04)0812(05)09(06)10(07)0
9(08)10(09)(10)(11)(12)

The above can be contracted by leaving out parentheses and interlacing zeros to

1234256371144812596107981091011112

In all, 25 symbols (which Herndon further reduced to 24 by eliminating the leading "1" as redundant). Can we have a shorter code?

In Herndon's codes, every vertex is represented by its label, which is followed by a list of neighbors. In the N-tuple code, *no labels* for vertices are explicitly used, but the code holds only for *acyclic* structures. The contrast suggests shorter (i.e., simpler) codes by combining the features of the two schemes: use the N-tuple code for the acyclic part of the structure, to be considered as the "backbone", and complete the code with a list of neighbors for ring-closure bonds. The open problems are (1) how to select an acyclic fragment for the "backbone" of the code and (2) how to incorporate the information on the portion of the structure not included in the N-tuple part of the code.

SPANNING SUBTREES OR SUBSPANNING TREES

The task is to find a way to associate with a polycyclic structure an underlying tree. Spanning trees represent the case of *erasure* of qualified *bonds* of a polycyclic structure. Such an approach breeds difficulties, a polycyclic graph has as a rule numerous spanning trees. Instead of using spanning trees, we will consider *acyclic* subgraphs obtained by erasing *vertices*. When a vertex is "excised", one also eliminates all edges incident to that vertex. In this way, one also "opens" rings, as is illustrated in Figure 1. Clearly, there will be several choices for "excising" a vertex. The possibilities shown in Figure 1 involve erasure of a single vertex, additional possibilities would arise if for constructing subtrees one considers erasure of two vertices. This immediately calls for some rules that will reduce the number of viable alternatives to a single possibility. If we require excision of as few vertices as possible in order to "open" all rings, for the bicyclic graph of Figure 1 instead of 20 different spanning trees (the actual number is even larger if symmetry equivalent subgraphs are considered separately) there are only *two* subtrees to consider. For the N-tuple part of the codes we obtain

2 1 1 1 1 1 1 1 1 0 0 and 3 1 3 1 1 0 0 1 1 0 0

At this step, one selects the *simpler* N-tuple for further consideration. This we interpret as the preference of *unbranched* spanning subtree; hence, we eliminate the other subspanning tree from further considerations. The linear chain (path) has the smaller N-tuplet value because its leading (highest) valency is 2, while the branching tree has a higher maximal valency of 3. The N-tuple codes for a tree are based on the *maximal* binary value possible, but among *different* candidate trees, we select one with the *minimal* code. Informally, the selection rule can be referred to as *mini-max* rule. The code for the bicyclic graph of Figure 1 will be of the form

2 1 1 1 1 1 1 1 1 0 0 + information on ring closures

We will refer to the underlying tree as *spanning subtree* or *subspanning tree*, because it is a spanning tree of a subgraph (i.e., a graph in which a single vertex, or several vertices in more complex cases, has been removed).

We now have to consider how to augment the code to show the erased vertex and the missing edges. A simple way is to continue the code by indicating the *valency* of the missing vertex (which is 3 for the graph of Figure 1), followed by a list of *neighbors* of the added vertex. Remember that the N-tuple code *induces* the labeling of vertices by assigning k to the kth position in the code. Hence, each symbol in the N-tuple code belongs to a single vertex. Induced numbering of vertices of the subspanning tree represented by the N-tuple code is shown in Figure 1. The last vertex, *12*, initially "erased" has neighbors 1, 5, and 9. We now construct the code by *adding* on to the end of the N-tuple part of the code the valency of the "missing" vertex, 3, followed by numbers 1, 5, and 9, the neighbors of the "excised" vertex, obtaining

2 1 1 1 1 1 1 1 1 0 0 3 1 5 9

In all, 15 digits! This is a very compact code for the bicyclic graph considered. It is difficult to imagine a more compact code for a general structure. For *special* classes of compounds, such as benzenoid systems or polyadamantanes, one can derive a very brief notation. Smith,[13] in a study of physical properties of benzenoid systems, has already represented molecules by condensing each benzene ring to a single vertex, thus reducing the notation considerably. This idea has been rediscovered and developed considerably.[14] Graphs built from similar blocks allow a number of simplifications and can produce relatively short codes.[15] But these are *special*, highly regular graphs. These special cases should not be confused with the problem of a *general* structural code. For a general structure, the augmented N-tuple code may well be the most optimal. The spanning acyclic part is already in a *compact* form based on adopting the N-tuple notation of Knop et al. The N-tuple part of the code immediately induces labels for all "backbone" atoms. Then, we have to "close" all rings, and each time we "add" an excised vertex and its neighbors all, the accompanying rings are recorded.

No information has been lost in the coding process. This one can see by trying to *reconstruct* a graph from the code. Let us illustrate this on the code 2 1 1 1 1 1 1 1 1 0 0 3 1 5 9. The code begins with the N-tuple part, so 2 is the highest valency of the graph. N-tuple codes always end with zero, so it is not difficult to partition the code into the acyclic subspanning tree part and the ring closure part (i.e., the list of valencies and neighbors of excised vertices). The entry following zero is immediately recognized as not to be a part of the N-tuple, because the digit is larger than (or at best equal to) the leading digit in the code, which represents the maximal valency in the spanning subtree. The entries that follow the valency of the "missing vertex" are the *labels* for atoms. Entries beyond those identified as labels of vertices belong to the next erased vertex and its neighbors. Reconstruction is straightforward. One first derives the subspanning tree from the N-tuple portion of the code. Next, one assigns labels to vertices of the subtree, from $1 - m$ (m is the number of vertices in the subtree, $m < n$). One by one, each of the erased vertices is added, assigned its label, and connected to the neighbors as the list of neighbors in the code dictates.

THE HIERARCHICAL RULES

An unambiguous reconstruction of a graph from its code is not the evidence that the code is unique but establishes only that no information is lost in the coding process. For example, if we were to select the alternative subspanning tree for the code of the bicyclic graph of Figure 1 we would have

3 1 3 1 1 0 0 1 1 0 0 3 2 6 1 0

It is not difficult to convince oneself that this code also allows reconstruction. A code is unique only if it is not possible to find more than one code that represent a same structure, or leads to a given reconstruction. Can two different combinations of excised vertices result in the same code, barring symmetry equivalent cases? If there are such alternative choices of vertices, they will have to produce the same subspanning tree, because each code starts with the unique list of valencies representing the N-tuple of the subspanning tree. We have seen that induced labeling of vertices of the subspanning tree is unique. "Missing" vertices to be attached to the subspanning tree thus already have labeled sites to which they are linked in a ring-closure step. If rules are available as to how to order and link the information on the missing vertices to the N-tuple part of the code, then *different* pairs of excised vertices will necessarily produce different codes. The only way that the same code can result is when excised pairs of vertices are equivalent because then the label of vertices is not unique. This indicates that, in principle, one can detect the symmetry of a molecular graph from a close look at the codes and their construction.

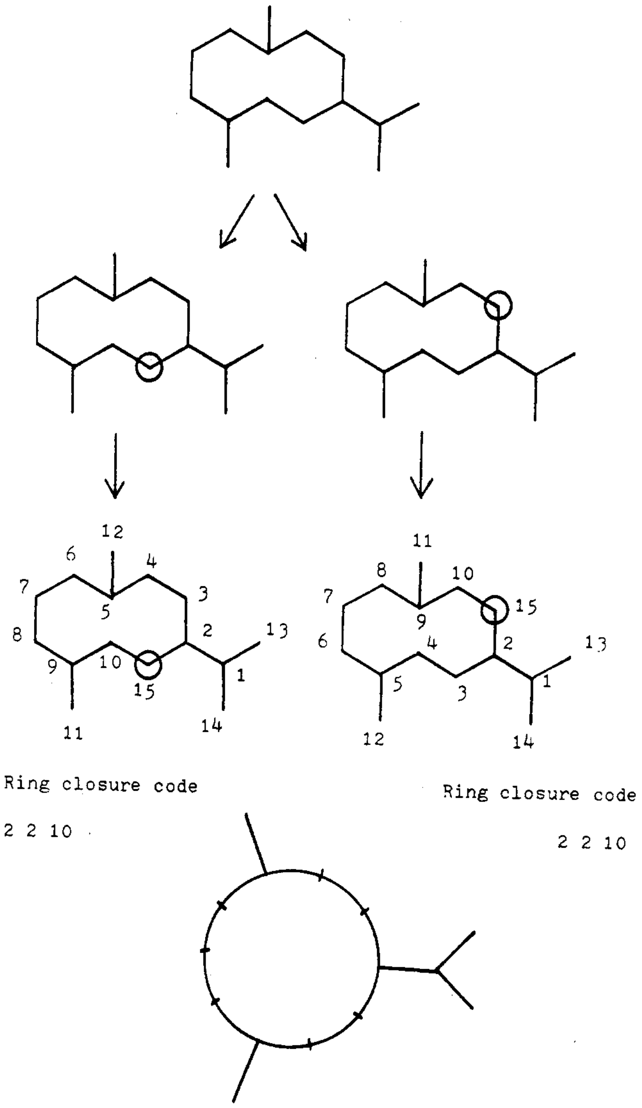


Figure 3. Identical code for seemingly different choices of excised vertex reveals symmetry of the molecular graph.

In summary, alternative routes in construction of a code will result in *different* codes, except if symmetry is present. Hierarchical rules have to ensure that only one of possible alternatives qualifies as the canonical code. Hence, nonunique codes, if found, would indicate a deficiency of the hierarchical rules, and possibly hint at a missing rule that ought to have been proposed and that would eliminate the deficiency. For example, consider the skeleton of a monocyclic terpene shown in Figure 3, which can be "opened" by excision of a single vertex. Consider the two alternatives shown in Figure 3, both of which lead to the N-tuple code

3 1 1 1 2 1 1 1 2 0 0 0 0 0

We will see that these conform to our hierachical rules, discussed later. The code induces in each structure a labeling of the vertices. If we adopt the labels shown in Figure 3, we can immediately construct the ring-closure portion of the code for each alternative. As we see in *both* cases, the excised vertex is of valency 2 and is adjacent to vertices labeled as 2 and 10; hence, the overall code for both alternatives of the structure of Figure 3 is

3 1 1 1 2 1 1 1 2 0 0 0 0 0 2 2 1 0

It appears that choice of a different vertex for ring opening produced the same code, but a closer look at the structure shows that the two excised vertices are *equivalent*. This is readily seen from the bottom diagram in Figure 3, where the graph has been redrawn to show its symmetry. *Chemically*, however, the two vertices will not be equivalent, but such nonequivalence emerges only when one goes *beyond* the atom connectivity. When (in the future) spatial nonequivalence is considered, it will be reflected in the codes, and the two alternatives will no longer be equivalent. The codes will differ then in some descriptors that are now ignored (i.e., not yet introduced). The nonuniqueness of such codes could *then* be resolved by considering *additional* rules that apply to such more general situations.

Here we will confine attention to atom connectivities and will propose rules that govern construction of the compact codes for molecular graphs. The rules, to be presented shortly, have a two-fold purpose: (1) to reduce the number of *alternatives* to be considered at each step in the development of the code in order to speed the search for the code and (2) to ensure that the derived code is *unique*. With this in mind and the guiding principle that *the codes* and *the search* for the codes be as simple as possible, after analyzing hundreds of structures we arrived at the following *hierachical* rules:

(A) Choose for erasing the vertices that produce as few as possible disconnected components (subtrees).

(B) Use as few "excised" vertices as possible.

(C) Initiate opening of ring with excision of vertices that "open" the largest number of rings (this is a typical "greedy algorithm" approach).

(D) Excise those vertices that produce trees having the smallest N-tuple code.

(E) Select the labels for the "backbone" tree that give the smallest labels for the vertex erased *last*.

(F) Order excised vertices by *inverse order* of excision (i.e., the last excised vertex comes first in the list of ring-closure part of the code).

(G) Order excised vertices by increasing valencies and when valencies are equal by increasing labels of neighbors, when ring "openings" are *independent* from one another.

The rules are ordered in a hierarchical manner, and when more than one applies the preference is given to those that precede. Rationale for the rules can be seen from specific applications. Consider the monocyclic graph of Figure 4 and the three alternative ways of "opening" the ring. The excision of the branching vertex leads to the code 2 1 0 0 0 3 3 4 5, having nine digits; the alternative codes have only eight digits. The increase in the length of the code is due to production of *disjoint* fragments, hence rule A, which informally suggests avoiding producing disconnected fragments. The two alternatives having eight-digit codes lead to *different* subspanning trees, hence rule D, which gives preference to the code that *appears* "simpler" (i.e., corresponds to a smaller N-tuple). The N-tuple parts of the two codes are 2 1 1 0 0 and 3 1 0 0 0, the former belonging to the linear chain and the latter to the branched tree. Clearly, the linear chain has a smaller numerical value for the N-tuple portion of the code, which is taken as the basis for the rule D. Now, we have selected a single subspanning tree, but numbering of vertices can run in two opposite directions. The two possibilities produce a different ring-closure parts of the code: 2 1 4 and 2 3 5. Here, 2 indicates the valency of the "missing" vertex, while 1 4 and 3 5 are alternative labels for the remaining vertices of the structure. Rule E resolves the above ambiguity by proclaiming smaller labels as desired.

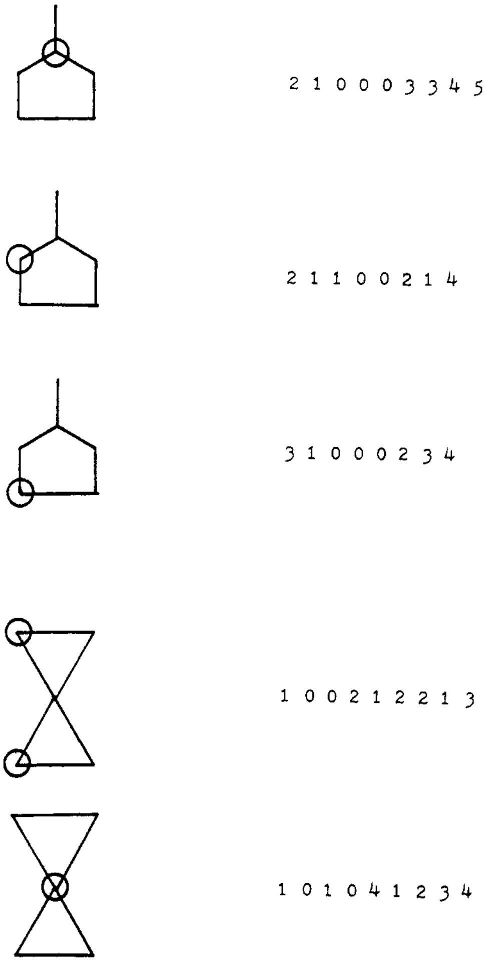


Figure 4. Alternative choices for excised vertices and the corresponding codes.

The bicyclic graph of spiropentane (Figure 4) may be "opened" by excising the central spiro vertex, which results in two disconnected components, or by excising two vertices resulting in a single component. Which is preferable? The corresponding codes are 1 0 1 0 4 1 2 3 4 and 1 0 0 2 1 2 2 1 3. The codes are of the same length and even have the same number of ring-closure labels. The latter has a smaller value and therefore we placed rule A above rule B. Clearly, other choices are possible, but at least in some simple situations, our selection produces what appears a simpler alternative. Rule B is natural one, as one can see from the bicyclic graph of Figure 1, which can also be "opened" by deleting two vertices instead of a single one. In that case, one will necessarily produce a code that is longer, and that is the reason for formulating rule B.

The cases discussed are simple, hardly requiring any deliberation. In Figure 5 we illustrate a relatively complex polycyclic graph for which it is not obvious how to select vertices to be erased. Hence, we have to approach the search for the compact code in a systematic way. The graph has three asymmetry-nonequivalent vertices, so in the first step we erased each of them separately. This results in three subgraphs shown in Figure 5 as *alternatives*. The first of these subgraphs has six distinctive sites for erasure of the *second* vertex, the second has three such sites, and the third has only one. Hence, after erasure of two vertices, we have 10 subgraphs to examine, of

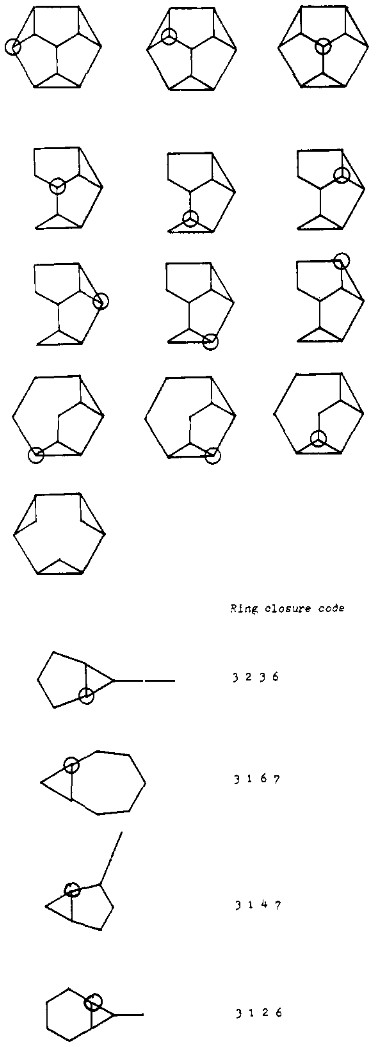


Figure 5. A highly cyclic graph that is gradually reduced to an acyclic subgraph.

which only six are nonisomorphic. Two of the six bicyclic subgraphs have disconnected rings, while four have fused (adjacent) rings. Only the latter can be "opened" in the next step by erasure of a *single* vertex; hence, only four need further examination. The four are shown at the bottom of Figure 5, and each of them results in subspanning linear chain having eight vertices. Hence, all alternative codes will have the same initial part (*N*-tuple): 2 1 1 1 1 1 1 0. In order to complete the codes, one has to indicate the ring-closure parts of the codes, which are 3 2 3 6, 3 1 6 7, 3 1 4 7, and 3 1 2 6 (when the correct numbering for vertices is assumed). From the four possibilities, one can immediately select the last one, 3 1 2 6, as being lexically smallest. The choice eliminates a need to complete the codes for discarded alternatives, which increases the efficiency of the construction. By selecting the last alternative as viable, we have also assigned labels 1–8 to the bicyclic spanning subgraph, the last label belonging to the ring-closure vertex. The remaining "missing" vertices can be assigned labels 9 and 10 in the reverse order of excision. This is the essence of rule F, which eliminates need for further analysis. Finally, when erased vertices are independent of one another (i.e., the same subspanning graph results regardless of the order in which the vertices were erased), rule G prescribes ordering of such vertices in the code.

ILLUSTRATIONS

We will illustrate the search for the compact codes on a number of terpenes (Figure 6). They appear to be a fairly representative class of compounds, having diverse structural

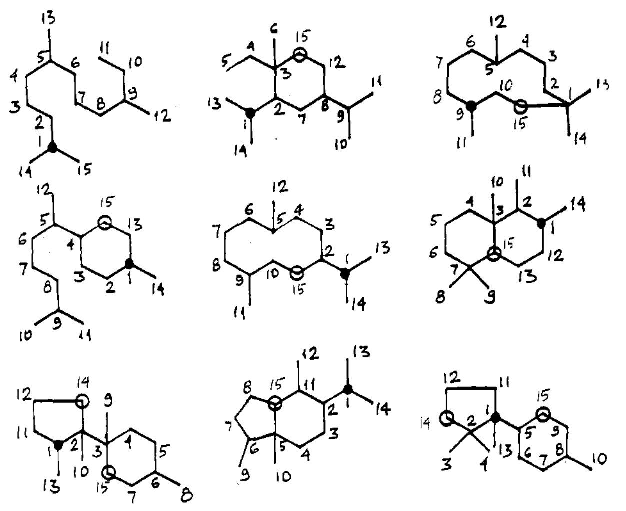


Figure 6. Search for compact codes for selected terpenes: If cyclic, locate vertex (or vertices) that needs to be erased to produce an acyclic subgraph. Such vertices are indicated by an "open" circle. Locate the vertex of highest valency (after removal of ring closure atoms) and start labeling atoms consistent with the *N*-tuple code rules. The initial atom is indicated by a "full" circle.

forms: acyclic, monocyclic, bicyclic, and tricyclic structures with pending fragments, bridges, and spiro carbon atoms. Yet all the skeletons have the same number of carbon atoms, which will make all the codes of comparable size. In the following discussions, we will use the labels shown in Figure 6.

Linear. Being acyclic, its code is given as *N*-tuple:

3 1 1 1 2 1 1 1 2 1 0 0 0 0 0

If we had selected the root at vertex *5* or vertex *9*, the codes of a lesser numerical value, i.e., noncanonical, would result:

3 1 1 1 2 1 0 0 1 1 1 2 0 0 0

3 1 1 1 2 1 1 1 2 0 0 0 1 0 0

Elemane. To "open" the ring, vertex *7*, *12*, or *15* has to be used. Use of vertex 2, 3, or 8 would introduce disjoint fragments, in violation of rule A. Among alternative subspanning trees, rule D dictates *15* as the choice, as this will reduce the highest valency and give the smallest code among alternatives (each of which, however, is represented by the maximal *N*-tuple code):

3 2 2 1 0 0 1 2 2 0 0 0 0 0 2 3 12

Humulane. To open the ring and at the same time reduce the maximal valency (rule D), erase *8* or *15*. For the *N*-tuplet part of the code, *15* gives 3 1 1 1 2 1 1 1 2 0 0 0 0 0, and *8* gives 3 1 1 2 1 1 1 2 1 0 0 0 0 0. Both codes have to be augmented with the information on ring closure, but *15* is smaller numerically, hence

3 1 1 1 2 1 1 1 2 0 0 0 0 0 2 1 10

If we selected numbering starting with 1 at site 9, the code would be

3 1 1 1 2 1 1 1 2 0 0 0 0 0 2 9 13

which would violate rule E.

Bisabolane. Removal of either *13* or *15* (we do not consider symmetrically equivalent sites *2* and *3*) will produce an acyclic case: *15* gives 3 1 1 1 2 1 1 1 2 0 0 0 0 0 0; *13* leaves as adjacent *two* vertices of degree 3, which consequently gives for the initial entries of the code 3 2 ... ruled out by rule D. Again, we have to select the site for labeling of vertices, the optimal choice leading to the code

3 1 1 1 2 1 1 1 2 0 0 0 0 0 2 4 13

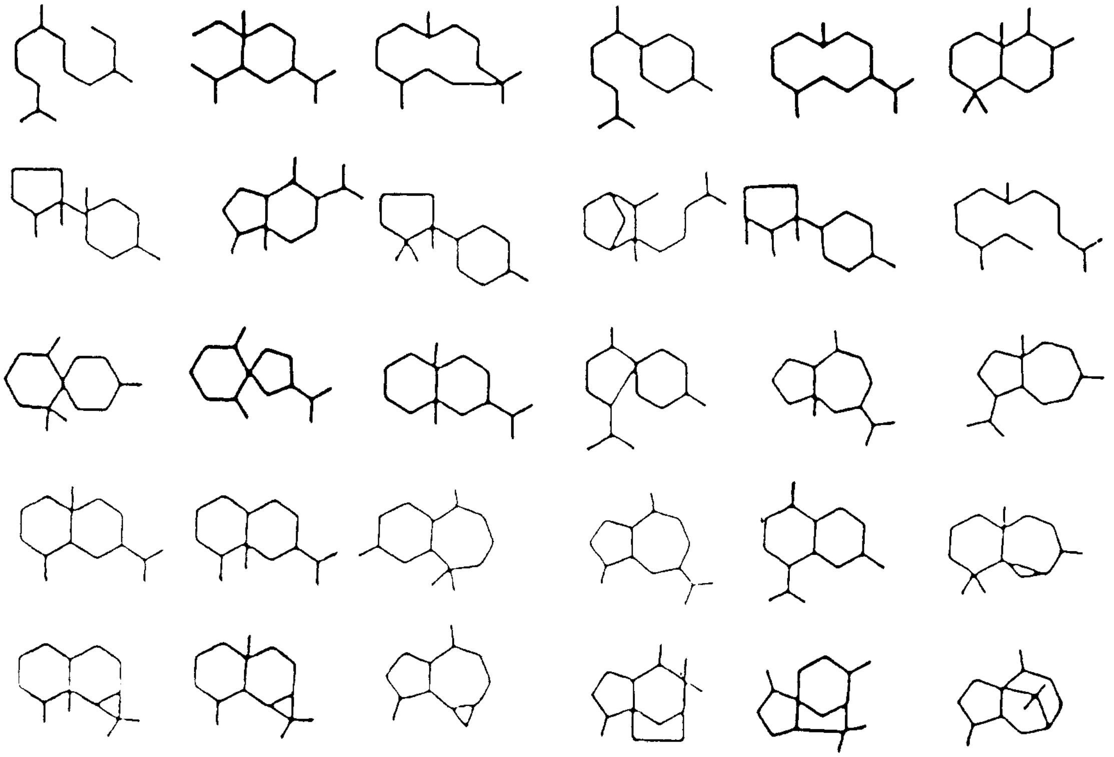


Figure 7. Molecular diagrams for the collection of terpenes. Observe structural variations (linear, monocyclic, bicyclic, tricyclic, spiro compounds, etc.).

Germacrane. Remove either *3* or *15* in order to reduce the valency of one of the two adjacent vertices of high valency. In either case the code is the same:

3 1 1 1 2 1 1 1 2 0 0 0 0 0

augmented by 2 2 10, the valency and list of neighbors of the missing vertex. Existence of two identical alternatives signifies automorphism (i.e., equivalence of *3* and *15*). This is the case already discussed. The equivalence of *3* and *15* will be lifted when the spatial environments of the two sites are considered.

Drimane. There is a unique vertex, *15*, that opens both rings giving

3 2 2 1 1 1 2 0 0 0 0 1 0 0 3 3 7 13

Trichothecane. Both rings are "opened" by excising vertices *14* and *15*, which also reduces the valencies of the adjacent vertices of highest valency. The code becomes

3 2 2 1 1 2 0 0 0 0 1 0 0 2 2 12 2 3 7

Rule G is used to order the "excised" vertices.

Tutin Group. The removal of a unique single vertex opens both rings (the other vertex common to both rings would produce a disjoint component, violating rule A). The numbering induced starting at *1* rather than 6 gives smaller labels for the excised vertex (rule E). The code is

3 2 1 1 2 2 1 0 0 0 1 0 0 0 3 5 8 11

Cuparane. The five-membered ring can be opened by either *11* or *14*, while the six-membered ring can be opened by erasing *9* or *15*. Because the *N*-tuple code will start in a five-membered ring (with a vertex of valency 4), *15* is preferred. This gives for the initial part of the code the following: *14*, 4 2 0 0 1 1 1 2 0 0 1 0 0; *11*, 4 2 1 1 1 2 0 0 0 1 0 0 0. Hence, *14* is to be used, and the code is

4 2 0 0 1 1 1 2 0 0 1 0 0 2 2 12 2 5 9

Table II. Compact Codes for Numerous Terpenes (Depicted in Figure 7)

name	compact code
linear	3 1 1 1 2 1 1 1 2 1 0 0 0 0 0
elemane	3 2 2 1 0 0 1 2 2 0 0 0 0 0 2 3 12
humulane	3 1 1 1 2 1 1 1 2 0 0 0 0 0 2 1 10
bisabolane	3 1 1 1 2 1 1 1 2 0 0 0 0 0 2 4 13
germacrane	3 1 1 1 2 1 1 1 2 0 0 0 0 0 2 2 10
drimane	3 2 2 1 1 1 2 0 0 0 0 0 1 0 3 3 7 14
trichothecane	3 2 2 1 1 2 0 0 0 0 0 1 0 1 1 13 2 3 7
tutin group	3 2 1 1 2 2 1 0 0 0 1 0 0 0 3 5 8 11
cuparane	4 2 0 0 1 1 1 2 0 0 1 0 0 2 2 12 2 5 9
santalane	3 2 2 1 1 1 2 0 0 0 0 1 0 0 3 3 13 14
laurane	3 2 2 1 1 1 2 0 0 0 0 0 0 2 3 12 2 4 8
caryophyllane	3 1 1 1 2 1 1 1 2 1 0 0 0 0 0
chamigrane	3 2 2 1 0 0 1 1 2 0 0 0 0 2 1 5 2 2 10
vetispirane	3 2 1 1 1 2 0 0 1 0 1 0 0 2 2 5 2 9 12
valerane	3 2 1 1 1 2 0 0 0 1 1 0 0 2 1 5 2 2 12
acorane	3 2 1 2 0 0 1 1 2 0 0 0 0 2 2 10 2 3 12
pseudoguaiane	3 1 1 2 1 1 1 2 0 0 0 1 0 0 3 4 7 13
carotane	3 1 1 2 1 1 1 2 0 0 0 1 0 0 3 4 7 11
eudesmane	3 2 1 1 2 1 1 1 1 0 0 0 0 0 3 5 9 12
eremophilane	3 2 1 2 2 1 1 0 0 0 1 0 0 0 3 4 8 12
himachalane	3 1 1 1 2 1 1 1 2 0 0 0 0 0 3 1 6 10
guaiane	3 2 1 1 2 1 0 0 1 1 1 0 0 0 3 4 7 11
cadinane	3 1 1 1 2 1 1 1 2 0 0 0 0 0 3 2 6 10
widdrane	3 1 1 1 2 1 1 2 1 0 0 0 0 0 4 1 5 9 10
aristolane	3 2 1 1 1 1 2 0 0 0 1 0 0 2 3 12 3 2 6 7
maaliane	3 1 1 1 2 1 1 1 1 0 0 0 0 3 1 2 15 3 5 9 14
aromadendrane	3 1 1 1 1 1 1 0 1 0 0 2 3 4 3 2 7 10
khusane	4 2 1 1 1 1 0 0 2 1 0 0 0 0 4 3 6 11 12
cedrane	4 2 2 1 0 0 0 1 1 1 1 0 0 0 4 5 7 8 11
patchoulane	3 2 1 1 2 1 0 0 1 1 1 0 0 0 4 1 7 11

The compact codes for some 30 terpenes (shown in Figure 7) are given in Table II. The above outline of the derivation of the codes answers many of the questions that occur in the construction of the compact codes. Observe that in most cases we arrived at the code after making *very few* tests; in some cases, a single attempt produced the code! Finding the code is tantamount to finding numerical labels for the carbon atoms of the molecular skeletons. The simplicity of the coding can be appreciated when one remembers that the total number of possible labelings for a graph with 15 vertices is 15! = 1 307 674 368 000. The cases considered are clearly not "the worst possible" cases, nor have they been selected because they would lead to a simple construction. We know of no other

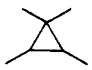
Table III. Compact Codes for All Monocyclic Graphs Having $n = 7$ Vertices and All Bicyclic Graphs having $n = 6$ Vertices[a]

graph	compact code
	Monocyclic ($n = 7$)
1	2 1 0 0 1 0 3 1 2 5
2	2 1 1 1 0 0 2 1 2
3	2 1 1 1 0 0 2 1 3
4	2 1 1 1 0 0 2 1 4
5	2 1 1 1 0 0 2 1 5
6	2 1 1 1 0 0 2 1 6
7	2 1 1 1 0 0 2 2 3
8	2 1 1 1 0 0 2 2 5
9	2 1 1 1 0 0 2 2 6
10	2 1 1 1 0 0 2 5 6
11	3 1 0 0 0 0 3 1 2 6
12	3 1 0 1 0 0 2 1 2
13	3 1 0 1 0 0 2 1 3
14	3 1 0 1 0 0 2 1 6
15	3 1 0 1 0 0 2 2 3
16	3 1 0 1 0 0 2 2 4
17	3 1 0 1 0 0 2 2 5
18	3 1 0 1 0 0 2 2 6
19	3 1 1 0 0 0 2 1 2
20	3 1 1 0 0 0 2 1 3
21	3 1 1 0 0 0 2 1 4
22	3 1 1 0 0 0 2 1 5
23	3 1 1 0 0 0 2 2 3
24	3 1 1 0 0 0 2 2 4
25	3 1 1 0 0 0 2 3 4
26	3 2 0 0 0 0 2 1 2
27	3 2 0 0 0 0 2 1 3
28	3 2 0 0 0 0 2 1 5
29	4 1 0 0 0 0 2 2 3
	Bicyclic ($n = 6$)
30	2 1 0 0 2 1 2 2 1 2
31	2 1 0 0 2 1 2 2 2 3
32	2 1 0 0 2 1 3 2 1 4
33	2 1 0 0 2 1 4 2 2 3
34	2 1 1 0 0 3 1 2 3
35	2 1 1 0 0 3 1 2 4
36	2 1 1 0 0 3 1 2 5
37	2 1 1 0 0 3 1 3 4
38	2 1 1 0 0 3 1 4 5
39	2 1 1 0 0 3 2 4 5
40	3 1 0 0 0 3 1 2 3
41	3 1 0 0 0 3 1 2 4
42	3 1 0 0 0 3 1 3 4
43	3 1 0 0 0 3 1 4 5
44	3 1 0 0 0 3 2 4 5
45	3 1 0 0 0 3 3 4 5
46	4 0 0 0 0 3 1 2 3

scheme that allows one to derive canonical labels and codes on comparable structures with such minimal effort!

MORE EXAMPLES

In order to further illustrate the construction of the compact codes, we consider first monocyclic graphs having $n = 7$ vertices and bicyclic graphs having $n = 6$ vertices (both shown in Figure 8). The codes are collected in Table III. The codes for both groups of graphs can be obtained merely by inspection! Choosing the vertex (or vertices) to be removed leads immediately to the compact code. The few rules that we stated suffice to resolve potentially ambiguous situations. For example, in the case of



one may wonder if a rule is required to prescribe which of the two nonequivalent ring vertices should be removed. One choice leaves *two* isolated fragments; the other introduces a single disjoint fragment. But the dilemma is resolved by the criterion that the code ought to be as short as possible. That rule alone eliminates the former alternative.

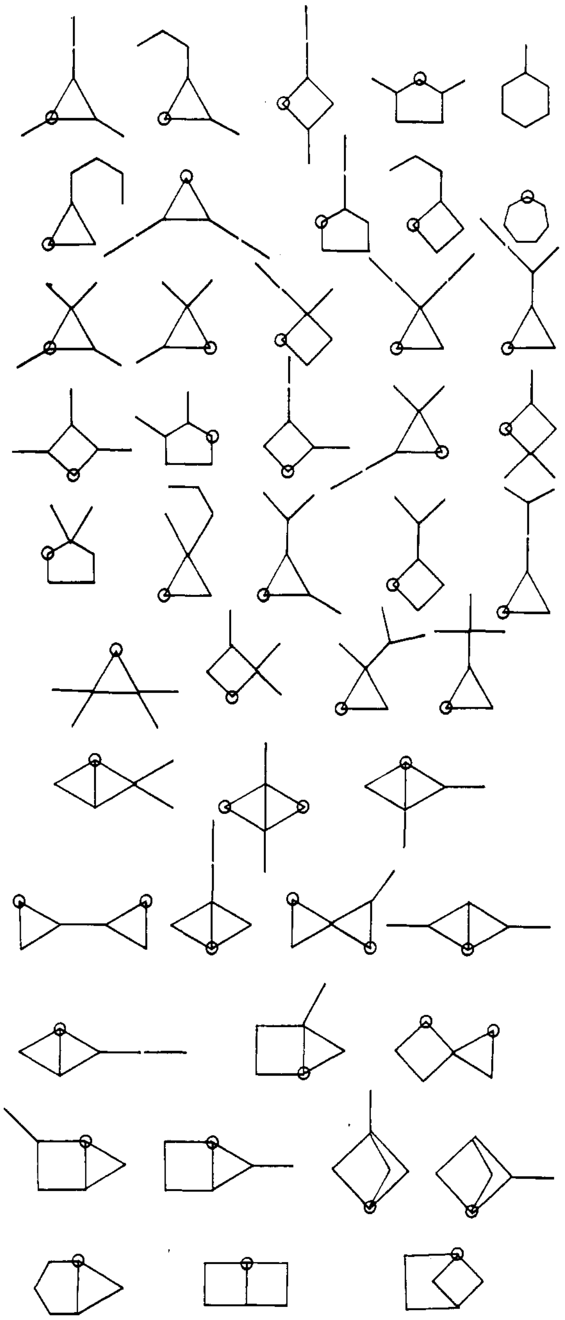
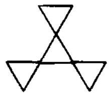


Figure 8. All monocyclic graphs having $n = 7$ vertices and bicyclic graphs having $n = 6$ vertices (the compact codes are shown in Table III).

Additional illustrations of compact codes (Table IV) include skeletons of organic compounds showing different ring structures, including unusual ring systems. The molecules, shown in Figure 9, have been selected by Hanack, Subramanian, and Eymann[16] for their study of monocyclic and polycyclic compounds and unconventional aromatic systems and offer wide structural variations. Hence, they provide a valid test for code constructions. Many compounds in this group allow one to select the excised vertex (or vertices) by inspection. A few, however, require a more careful examination. In the case of



one again cannot avoid disjoint fragments. By excising the three vertices, we obtain two linear fragments with the N-tuple codes 2 1 0 0 and 10, respectively. One then combines these

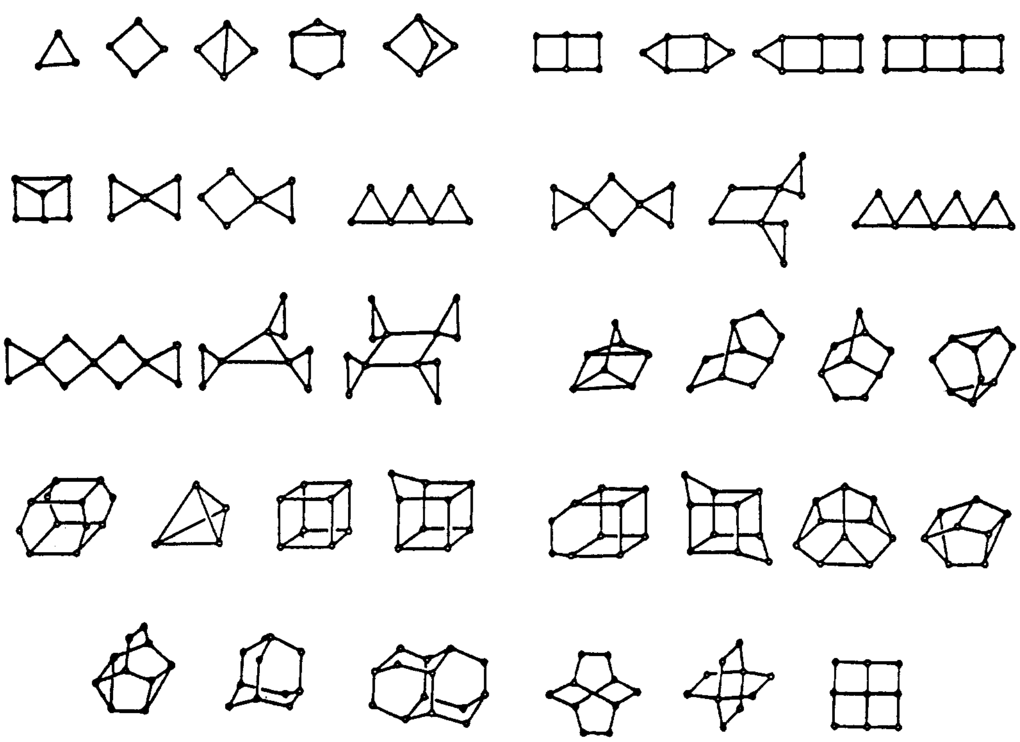


Figure 9. Skeletons of organic compounds showing different ring structure including unusual ring systems as selected in reference 16. The corresponding compact codes are listed in Table IV.

Table IV. Compact Codes for Cyclic Compounds Showing Different Ring Structure Including Unusual Ring Systems[a]

graph	compact code
1	1 0 2 1 2
2	2 0 0 2 2 3
3	2 0 0 3 1 2 3
4	2 1 1 0 0 3 1 4 5
5	3 0 0 3 2 3 4
6	2 1 1 0 0 3 2 4 5
7	2 1 0 0 2 1 4 3 2 3 4
8	2 1 1 0 0 2 1 5 3 1 4 5
9	2 1 1 1 0 0 2 2 6 3 1 3 5
10	2 1 0 0 2 1 2 3 3 4 5
11	1 0 0 2 1 2 2 1 3
12	2 1 0 0 2 1 3 2 1 4
13	2 1 0 0 2 1 2 2 1 4 2 2 3
14	2 1 1 0 0 2 1 3 2 1 5 2 3 4
15	3 1 0 0 0 2 1 4 2 2 3 2 2 5
16	2 1 1 0 0 2 1 2 2 1 5 2 2 3 2 3 4
17	2 1 1 1 1 0 0 2 1 3 2 1 7 2 3 5 2 5 6
18	2 1 0 0 1 0 2 1 4 2 2 3 4 1 2 5 6
19	3 1 0 1 0 0 1 0 2 1 5 2 2 3 2 4 5 4 2 4 6 7
20	2 1 1 0 0 3 1 2 4 3 3 4 5
21	2 1 1 1 1 0 0 2 2 6 3 3 6 7
22	2 1 1 1 1 0 0 2 1 6 3 3 6 7
23	2 1 1 1 1 0 0 3 1 2 6 3 4 5 7

[a] Molecular skeletons are depicted in Figure 9.

two parts into a single code: 2 1 0 0 1 0. The rules of Knop et al.[7] suffice here to decide how to order the fragments. The resulting code ought to be maximal. The final code can be immediately written by adding the list of erased connections, giving 2 1 0 0 1 0 2 1 4 2 2 3 4 1 2 5 6. Several of the polycyclic graphs may require half a dozen tests before one determines which vertices are to be excised.

SPARSE GRAPHS

"Sparse graphs" are graphs having a *sparse* adjacency matrix. Sparse matrices are those that have relatively few nonzero entries. Most chemical compounds are represented by sparse matrices (i.e., adjacency matrices that have very few nonzero entries). In such instances, after removing a few vertices we may obtain a long-chain structure or structures with long branches. It is possible in such cases to abbreviate the codes, make them even more condensed, and possibly make them easier to work with. For example, consider carbon skeletons of chrysene and that of its isomer benzophenanthrene (Figure 10). In both cases it is not difficult to locate the two vertices that, when excised, produce the subspanning tree on which the compact code is based. We obtain the following codes: for chrysene

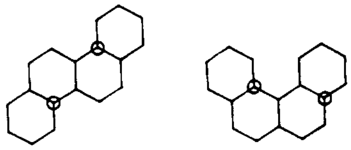


Figure 10. Molecular graphs of chrysene and its isomer benzophenanthrene.

2 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 3 4 8 16 3 7 11 15

for benzophenanthrene

3 1 1 1 1 1 1 0 1 1 1 1 1 1 0 0 3 4 8 11 3 11 15 16

The long strings of 1's can be contracted with a superscript, subscript, or other typographical sign. For example, we can write for the two codes

2 1**13 0 0 3 4 8 16 3 7 11 15

3 1**6 0 1**6 0 0 3 4 8 11 3 11 15 16

Here 1**13 and 1**6 indicate repetitive occurrences of the digit 1 13 and 6 times. The portion 1**6 0 in the case of benzophenanthrene appears twice. Hence, one can consider further contractions of the repetitive groups to shorten the code. However, contractions are inconsistent with the desideratum that the length of the code should be a function of the number of vertices and edges. Moreover, one may question that "simplicity" of a coding scheme that accumulates contractions, particularly contractions of contractions. Hence, we will avoid the use of contracted forms, but for special applications they may be of interest. The above illustrates yet another important

Table V. Compact Codes for All Catacondensed Polycyclic Conjugated Hydrocarbons Having the Empirical Formula $C_{14}H_{10}$ Shown in Figure 11[a]

structure	*N*-tuple part	label part
	2 1 1 1 1 1 1 1 1 1 0 0	
1		2 1 2 3 4 11 12
2		2 1 2 3 5 11 12
3		2 1 2 3 6 11 12
4		2 1 2 3 7 11 12
5		2 1 12 3 4 11 12
6		2 1 12 3 5 11 12
7		2 1 12 3 6 11 12
8		2 1 13 3 7 11 12
9		2 2 3 3 6 11 12
10		2 2 3 3 7 11 12
11		2 2 12 3 1 4 11
12		2 2 12 3 1 5 11
13		2 2 12 3 1 6 11
14		2 2 12 3 1 7 11
15		2 2 12 3 1 8 11
16		2 3 12 3 2 4 11
17		2 3 12 3 2 5 11
18		2 3 12 3 2 6 11
19		2 3 12 3 2 7 11
20		2 3 12 3 2 8 11
21		2 4 12 3 3 5 11
22		2 4 12 3 3 6 11
23		2 4 12 3 3 7 11
24		2 5 12 3 4 6 11
	3 1 1 1 0 1 1 1 0 1 1 0	
25		2 2 12 3 7 9
26		2 5 10 3 7 9 12
	3 1 1 1 1 0 1 1 0 1 1 0	
27		2 7 12 3 3 6 9
28		2 7 12 3 4 6 9
	3 1 1 1 1 0 1 1 1 0 1 0	
29		2 2 12 3 4 6 10
30		2 6 11 3 8 10 12
31		2 7 12 3 3 6 10
32		2 7 12 3 4 6 10
33		2 10 11 3 3 6 12
	3 1 1 1 1 1 0 1 1 0 1 0	
34		2 8 12 3 4 7 10
35		2 10 11 3 3 7 12

[a] The codes are split into two parts. The first part is the *N*-tuple portion uniquely characterizing the subspanning tree; the second part represents bonds (edges) that make ring closure, named label part.

property of *N*-tuple codes and *N*-tuple portions of compact codes: They can be subject to symbolic algebraic manipulations. Of course, labels (the end part of the code) have no such properties. In Table V we illustrate the compact codes for all the polycyclic catacondensed conjugated isomers of the formula $C_{14}H_{10}$.[17] Molecular graphs are shown in Figure 11. The codes have been shown in two parts, the first describing the *N*-tuple and the second describing the connectivities of the erased vertices. One sees that many compounds of Table V have identical *N*-tuple parts, which permits further simplification if one introduces an abbreviation for the repeating portion of the codes.

COMPARISONS

Before judging various codes, one should bear in mind different uses of codes. One should also remember that molecular notation (i.e., the codes) and molecular names serve different purposes. Nomenclature is defined as a set of names indicating the *composition*, as well as other relevant structural elements (e.g., configuration, substitution sites), while notation represents a list of qualified symbols (such as digits).[18] In comparing codes one should consider the work needed to derive the codes (i.e., how much or how little preliminary work has to be done in order to derive the code), the work needed to derive the numbering of vertices accompanying the code, the ease with which one can reconstruct the structure from the code, the conceptual base of the code as well as the ease of perceiving some important structural features from the code, etc. Our criterion for the selection of codes for comparison was the availability (in print) of codes for examples that illustrate diverse skeletons features.

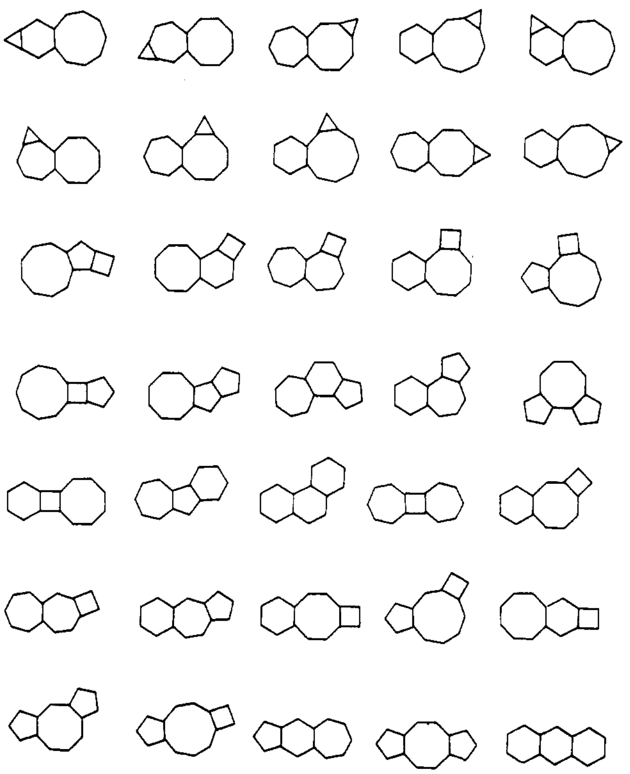
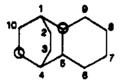
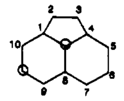
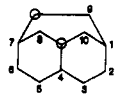


Figure 11. All catacondensed polycyclic conjugated systems having 14 carbon atoms and 10 hydrogen atoms. The compact codes are shown in Table V.

Table VI. Illustration of Compact Codes and WLN Codes for Selected Structures

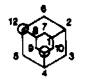
structure	WLN; compact code
	L 666/GL 2 AF LTJ; 2 1 1 1 1 1 1 0 0 2 4 10 3 1 5 9
	L 566 1A LTJ; 2 1 1 1 1 1 1 1 0 0 2 9 10 3 1 4 8
	L 76 B6 AC 1B LJ; 3 1 1 1 1 1 1 0 0 0 2 7 9 3 4 8 10

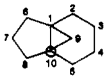
WLN

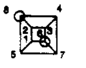
Wiswesser chemical line notation[19] represents a coding approach using alphanumeric symbols extensively. The WLN notation represents a hybrid of structural and conventional (i.e., by agreement) approaches. For various structural groups, monocyclic components, etc., one introduces special symbols. Encoding branching, perifused rings, bridged systems, etc. leads to numerous rules. In Table VI we illustrate the WLN codes as the corresponding compact codes for a few carbocyclic compounds. Clearly, the WLN codes are short, but at the *cost* of having numerous rules and special symbols. It is not an easy exercise for most chemists either to derive WLN codes or to interpret given codes for molecules of modest complexity. Few chemists have an idea what L 76 B6 AC 1B LJ represents, unless specially trained. In contrast, everyone can easily reconstruct 3 1 1 1 1 1 1 0 0 0 2 7 9 3 4 8 10 once a *single* such code is explained to him/her.

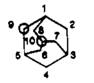
Table VII. Comparison between Compact Codes and Nodal Nomenclature of Lozac'h and Co-workers

structure	common name; nodal name; compact code

	iceane;
	tetracyclo[010/$1^{1,7}$/^{2,6}/$0^{4,9}$]dodecanodane;
	3 2 1 1 0 0 1 1 0 0 3 4 9 10 3 5 6 8

	[4.3.1]propellane;
	tricyclo[09/$1^{1,5}$/$0^{1,5}$]decanodane;
	3 1 1 1 0 1 1 0 0 4 1 5 8 9

	cubane;
	pentacyclo[08/$0^{1,4}$/^{2,7}/^{3,6}/^{5,8}]octanodane;
	2 1 1 0 0 2 1 3 3 2 4 5 3 4 5 6

	adamantane;
	tricyclo[08/$1^{1,5}$/^{3,7}]decanodane;
	3 1 1 0 1 1 0 0 2 3 6 3 4 7 8

NODAL NOMENCLATURE

In Table VII we illustrate a number of codes based on a proposal on the nomenclature of organic chemistry initiated by IUPAC and directed by Lozac'h and others.[20] The idea is to use a graph-theoretical approach yet retain as much as possible (in appearance) of traditional systematic notations of organic chemistry. For example, bicyclo[2.2.2]octane becomes bicyclo[06/$2^{1,4}$]octanodane. "Nodane" is a generic name. Hence, the difference between the two nomenclatures is in the coded section [2.2.2] vs. [06/$2^{1,4}$]. The former looks simpler *but* cannot be extended to tricyclic and other structures without further elaborations, while the nodal nomenclature *can* be so extended. If we compare the nodal codes with our compact code, a fair comparison would be 3 1 0 1 0 1 0 3 3 5 7 against bicyclo[06/$2^{1,4}$]octa. "Bicyclo" and "octa" are useful *descriptions* that are immediately visible. However, they can be easily *deduced* from our codes by counting the "valency" entries in the code 3 1 0 1 0 1 0 3—in all eight, i.e., "octa". Similarly, one can count the numbeɛ of "closure" bonds that are represented by atom "labels" (labels 3 5 7 in this case). In this case, we have three labels, so there are *three* regions of the molecular graph. But the "outside" region does not count as a cycle or ring, so the code represents a "bicyclo" structure. To appreciate the simplicity of our codes, one should be reminded of *lengthy* rules that govern construction of nodal (and many other) codes. It suffices to look at the *rules* for numbering atoms in various schemes to see how elaborate these can be. In fairness to other schemes, while our rules are few and simple, occasionally there may be some work involved in finding the correct subspanning tree and its labeling. But, it appears that the work involved in such constructions is not excessive.

STRUCTURE NOMENCLATURE NOTATION (SNN)

Not long ago Walentkowski[21] proposed a unique, unambiguous representation of chemical structures by computerization of a simple notation. A molecule is split into fragments by structure-determining vertices. This approach represents an advancement of the historical Beilstein system.[1] In selecting examples for comparison, we again restrict our attention to carbocyclic compounds. The codes are compared in Table VIII for a selection of compounds. Observe the use of a number of special symbols (and parentheses) in the case of SNN codes. Clearly, the symbol 6 6 & 6 (1R) is considerably shorter than our 2 1 1 1 1 1 1 1 1 1 1 0 0 4 4 8 12 13, but one has to know how to interpret & and (1R), two symbols out of a dozen or so similar symbols, like /, +, –, :, ′, &&, [], (), etc., that appear in the codes. The simplicity of the compact codes is already evident if one compares the SNN codes for an acyclic graph with the *N*-tuplet code of Knop et al.[7] In the case of the 16-carbon atoms acyclic molecular skeleton, we have – 9 3 (4) 1 (2′) 2 (5) 1 (2) vs. our 3 2 1 1 1 0 1 0 1 2 0 0 1 2 0 0.

Table VIII: Illustration of a Few SNN Codes and Corresponding Compact Codes (Observe the Use of Special Typographical Symbols for the Former)

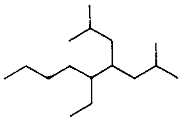
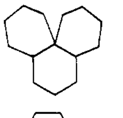
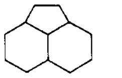
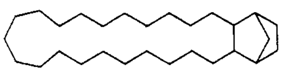
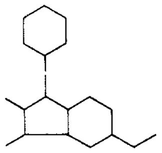
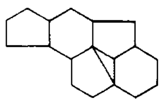
structure	SNN code; compact code
	–9 3 (4) 1 (2′) 2 (5) 1 (2); 3 2 1 2 0 0 1 2 0 0 1 1 1 0 1 0
	6 6 & 6 (1R); 2 1 1 1 1 1 1 1 1 1 1 0 0 4 4 8 12 13
	6 6 & 5 (1.2); 2 1 1 1 1 1 1 1 0 0 2 2 6 3 9 10 11

Table IX. Illustration of Graph-Based Nomenclature of Goodson and Corresponding Compact Codes

structure	graph based; compact code
	22,24(26)-binipenta-1-dicontatrigon; 2 1^{23} 0 0 2 3 26 3 1 4 25
	tricyclo(09.$0^{1,5}$)2:10(1)10:11(06)7:1 7(2)3:19(1)4:20(1)icosanodane; 3 2 1 1 2 1 0 1 0 0 1 1 1 1 1 1 0 0 2 13 17 3 3 7 11
	7(12)-tria-1,6,8(13)-ternipenta-4,12-(15)-binihexalane; 2 1^{12} 0 0 2 3 7 3 11 14 15 4 1 7 10 16

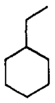
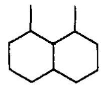
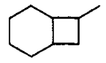
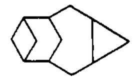
GRAPH-BASED NOMENCLATURE

The previously mentioned nodal nomenclature of Lozac'h and collaborators has lead to an alternative scheme by Goodson.[22] The main departure consists of abandoning the approach inherent in the von Baeyer nomenclature[2] in which a ring system is numbered by selecting a path through it. Instead, Goodwin follows the proposal of Taylor,[23] which is based on the smallest set of smallest rings.[24] In Table IX we list selected codes for the structures shown there. The last polycyclic structures having a half a dozen rings may appear as presenting a substantial increase in the amount of work needed to find the compact code. This, however, is not so. It takes less than a dozen alternative diagrams to find the optimal case.

UNIQUE LINEAR NOTATION

Herndon and Leonard,[25] extending the initial work of Herndon,[10] recently developed a scheme for a unique linear notation based on canonical numbering. Here the code consists of a canonical list of neighbors for all atoms. When reduced to essentials, the codes appear quite brief. However, in order to derive the codes, one first has to find the associated canonical numbering. In Table X, we show the comparison between the compact codes and those of Herndon for a selection of structures. In the case of sparse graphs, our codes are visibly shorter because we do not list atomic labels for atoms of the underlying acyclic subgraph for which the *N*-tuplet code is used. However, if a graph has many cycles, the difference in length between our compact code and Herndon's code decreases. In the case of the tetracyclic graph of Table X, Herndon's code has 26 entries (34 digits because there are labels using two digits), and our compact code has only 18 (19

Table X. Comparison of Herndon's Codes and Compact Codes for a Few Cyclic Structures

structure	Herndon's code; compact code
	2 3 6 2 4 3 5 4 7 5 7 6 8 7 8; 2 1 1 1 1 0 0 2 2 6
	2 3 4 2 5 6 3 7 11 4 8 12 5 9 6 10 7 9 8 10 9 10 11 12; 2 1 1 1 1 1 1 1 1 0 0 3 1 5 9
	2 3 5 2 4 6 3 4 9 4 5 7 6 8 7 8 8 9; 2 1 1 1 1 1 0 0 3 3 7 8
	2 7 8 2 7 9 3 5 8 10 4 6 9 10 5 11 12 6 11 12 7 8 9 10 11 12; 3 2 1 1 0 1 1 0 0 0 3 4 5 8 3 7 9 10

digits, because of a single two-digit entry). In the case of a cube, Herndon's code has 19 digits as compared to our 16. For a complete graph the two approaches produce codes of the same length. Herndon does not show the first digit of the coding scheme, because it is always 1 so his code appears shorter. However, we need not show the last *N*-tuple digit, which is always 0. It is when one considers sparse graphs, which represent the *vast majority* of chemical structures, that our compact codes are visibly simpler.

CONCLUDING REMARKS

There are numerous other proposals for chemical nomenclature and codes.[26] Many have been developed with the potential for expanding to incorporate information on heteroatoms, multiple bonds, and even stereochemistry. Some such extensions can be added to our compact codes without difficulty. We may mention in particular the canonical connection table representation of molecular structure of Bersohn and Esack[27] and related work of Beierbeck,[28] who follow the compact connection table by a number that tells how many pieces of stereochemical information are associated with an atomic site. If atom *i* is related to atom *j* by stereorelation number *K*, then this is coded as a strong *k i j*. At this stage of the development of the compact code, we wish only to indicate the *possibilities* for extending the code to any chemical compound. Before proposing final rules, one would like to write down very many codes for existing compounds, starting first with hydrocarbons, of course. The reason for such a conservative attitude is a pragmatic one: if the proposed rules and codes need any improvement, their application to carbocyclic compounds will tell, since carbocyclic compounds offer almost unlimited ring and structural variations. In this way, an unanticipated structural feature may show up, and one will have an opportunity to refine the approach before it becomes implemented on a large data bank. Once a scheme is implemented on a data base having some 10000–100000 or more structures, even minor changes, may become impractical, unless codes themselves can be automatically altered. It seems prudent therefore to gradually *scale up* the proposed scheme from the current illustrations on some 100–200 representative compounds to some 1000–2000 structures. At this stage, there should be no important nontrivial structural features omitted. In the next step, the approach should be extended to include compounds with unsaturation (multiple bonding) as well as compounds having heteroatoms. Some preliminary advance in this direction has been made.[8,9] At the same time, one should consider developing computer programs that will generate the compact code from input connectivity information. It should be stated that apparent simplicity of the compact codes for humans, the possibility, for instance, to derive the codes in many instances by simply inspecting the molecular graph, does not necessarily imply that computer approach will be equally simple. The computer algorithm faces the basic difficulty of finding the *right* vertices to excise to give optimal code. Finding such vertices by computer may be less tractable than it appears.[29] In this paper we outlined construction of the compact codes for chemists use, for use *without* computer. Even here we do not claim more than we *demonstrated* (by examples) that the approach is *practical* for *typical* chemical structure. The questions of computational complexity, of the behavior of "the worst" possible cases, are open, but they ought not to detract a user, because they are of lesser importance for practical applications (and are equally unresolved in other nomenclature schemes). The question of computer implementation, of course, is very important and will be considered in the near future (possibly by several groups) but should be preceded with some experience on a larger sample of structures in order to modify or supplement the present rules if necessary.

In summary, codes should be (1) linear, (2) unique, (3) reconstructable, (4) derivable by hand, (5) decodable by hand, (6) independent of properties, (7) exempt from conventional items, (8) brief, (9) pronouncable, (10) based on familiar symbols, (11) easily comprehensible, (12) efficient, and, finally, (13) uniform in length (size). This represents a tall order. No known codes satisfy an appreciable fraction of the listed properties. It is then a remarkable accomplishment that the proposed compact codes satisfy *all* relevant indicated requirements. In particular, they are brief and simple. By simple we mean that anyone can use them *without* any prior training and that anyone who tries will without difficulties derive codes for compounds of his/her interest. This latter quality should tempt everyone to try to use them. However, one should start with bicyclic and tricyclic compounds in order to gain some experience. The statement "the codes are simple" should be taken as relative to efforts involved in other schemes. Any coding system will become increasingly difficult as the number of rings increases. Hence, the need for a computer is anticipated in the case of large structures having numerous rings, even though numerous "...ring-bridges, spirolinked, polyfused, 'chickenwire' monsters lurking between the covers of *The Ring Index*"[30] can be tackled with moderate patience, possibly on the back of a (large-size) envelope!, as one will surprisingly find.

REFERENCES AND NOTES

(1) For a few historical remarks concerning the Beilstein system (F. K., Beilstein, 1881, see the introductory part of reference 21 (by R. Walentkowski).

(2) Baeyer, A. *Ber. Dtsch. Chem. Ges.* **1900**, *33*, 3771.

(3) For a review of computerized chemical structure handling techniques in structure–activity studies and molecular property predictions, see Bawden, D. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 14.

(4) Read, R. C. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 135.

(5) Goodson, A. L. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 167.

(6) The Einstein quote is "The theories should be as simple as possible but not simpler."

(7) Knop, J. V.; Müller, W. R.; Jeričević, Z.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 91.

(8) Randić, M. *Croat. Chem. Acta*, in press. Compact Codes, Part II of this series.

(9) Randić, M. submitted for publication in *J. Chem. Inf. Comput. Sci.* "Compact Codes", Part III of this series.

(10) Herndon, W. C. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983. Herndon, W. C.; Davis, M. L.; Ellzey, M. L. Jr., unpublished results.

(11) Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107. Razinger, M. *Theor. Chim. Acta* **1982**, *61*, 581.

(12) Wipke, W. T.; Dyott, T. M. *J. Am. Chem. Soc.* **1974**, *96*, 4834.

(13) Smith, F. T. *J. Chem. Phys.* **1961**, *34*, 793.

(14) Balaban, A. T.; Harary, F. *Tetrahedron* **1968**, *24*, 2505.

(15) Balaban, A. T.; Schleyer, P. v. R. *Tetrahedron* **1978**, *34*, 3599. Balaban, A. T. *Tetrahedron* **1969**, *25*, 2949. Lederberg, J. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *53*, 134. Silk, J. A. *J. Chem. Doc.* **1961**, *1*, 58. Tanaka, N.; Iizuka, T.; Kan, T. *Chem. Lett.*, **1974**, 539. Wenchen, H.; Wenjie, H. *Theor. Chim. Acta*, in press.

(16) Hanack, M.; Subramanian, L. R.; Eymann, W. *Naturwissenschaften* **1977**, *64*, 397.

(17) Dias, J. R. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 124.
(18) For further clarification consult Goodson, A. L. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 142.
(19) Smith, E. G. *The Wiswesser Line Formula Chemical Notation*; McGraw-Hill: New York, 1968.
(20) Lozac'h, N.; Goodson, A. L.; Powell, W. H. *Angew. Chem., Int. Ed. Engl.* **1979**, *18*, 887.
(21) Walentowski, R. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 181.
(22) Goodson, A. L. *Croat. Chem. Acta* **1983**, *56*, 315. Also see reference 16.
(23) Taylor, F. L. *Ind. Eng. Chem.* **1948**, *40*, 734.
(24) Zamora, A. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40. Schmidt, B.; Fleischhauer, J. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 204.
(25) Herndon, W. C.; Leonard, J. E. *Inorg. Chem.* **1983**, *22*, 554.
(26) Bonchev, D. *Pure Appl. Chem.* **1983**, *55*, 221. Nakayama, T.; Fujiwara, Y. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 23. Lin, C.-H. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 47. Dubois, J. E.; Chretien, J. *J. Chormatogr. Sci.* **1974**, *12*, 811. Kudo, Y.; Sasaki, S.-I. *J. Chem. Doc.* **1974**, *14*, 200. Quadrelli, L.; Bareggi, V.; Spiga, S. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 37. Jochum, C.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113. Bonchev, D.; Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 223. Hendrickson, J. B.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171.
(27) Bersohn, M.; Esack, A. *Chem. Scr.* **1974**, *6*, 122.
(28) Beierbeck, H. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 215.
(29) A referee expressed opinion that finding the "right" vertices to excise to obtain the optimal code with a computer will turn out to be an NP complete problem. The problem of graph isomorphism is known to belong to NP problems, *but* it is not yet known if it is NP complete. At the same time, the problem of subgraph isomorphism is known to be NP complete. Our approach requires (rule C) one to examine numerous *rings*, not perception of all cycles, which would lead to a NP complete problem as this would include also a search for Hamiltonian cycles, which is a known NP complete problem. Hence, the complexity of our approach is not worse than that of a graph isomorphism problem, but as demonstrated on chemical examples, it is practical for structures of intermediate size and complexity (e.g., having half a dozen rings).
(30) Gibson, G. W.; Granito, C. E. *Am. Lab. (Fairfield, Conn.)* **1972**, *4*, 27.

LETTERS TO THE EDITOR

Dear Sir:

Drs. D. F. Zaye and W. V. Metanomski, in their splendid JCICS article [**1986**, *26*(2), 43–44], enumerated well the channels by which science is conveyed. However, in technology not only are the channels enumerated important, but there is another called "reverse engineering". That is a fancy way of describing the things a child learns the first time he takes apart a clock. (When clocks had gears instead of transistors.) There is much to be learned by looking at a competitor's factory or formulation or gadget. As a matter of fact, one of the most expensive chemicals sold today...about $800 a pound...is used for "potting" electronic devices to frustrate anyone who wants to know what is "inside" them. (Potting is embedding in a thermosetting resin.)

Although I cannot cite specific examples, I am sure that there is a lot of science that can be learned by reverse engineering, and so I suggest that that method be added to our list of means of transferring the culture. One of the things one learns early on in technology is, first, that you do not necessarily need to understand all the scientific principles to make something useful. Second, there is a tremendous advantage when one is attacking a problem in knowing that a solution exists...as, for example, in another nation.

B. J. Luberoff
CHEMTECH
Summit, New Jersey 07901