

reveal that overall α -alkylation of a ketone can be retrieved by means of the above substructure as a clue.

CONCLUSION

Two reaction strings combine with each other in three ways, i.e., degeneration, accumulation, and compensation, which are useful concepts to describe synthetic pathways. Multiplication of complex bond numbers, synthesis spaces, and synthesis classes are now introduced in order to apply the ITS concept to the description of synthetic pathways.

REFERENCES AND NOTES

- (1) Fujita, S. *J. Chem. Inf. Comput. Sci.* preceding papers of this series in this issue.

- (2) Only Ugi's system solves this problem in terms of "isomeric ensemble of molecules". See: Jochum, C.; Gasteiger, J.; Ugi, I. *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 495 and references cited therein.
- (3) Drake, N. L.; Cooke, C. B. *Org. Synth.* **1943**, *Coll. Vol. 2*, 406.
- (4) Diels, O.; Alder, K. *Justus Liebigs Ann. Chem.* **1931**, *490*, 243.
- (5) Gall, M.; House, H. O. *Org. Synth.* **1972**, *52*, 39.
- (6) An ITS bond can be denoted by a set of three integers (o, i, p), wherein integer o represents a number of out-bonds, i is that of in-bonds, and p represents that of par-bonds. The condition implied by this expression is $oi = 0$. The following two equations are obtained easily: $a = p + o$ and $b = i - o$.
- (7) The ITS (R_2R_1) denotes the multiplication of the first ITS (R_1) by the second ITS (R_2). In general, $R = R_nR_{n-1}\dots R_1$ means the multiplication of successive ITS's, R_1, \dots , and R_n .
- (8) It is to be noted that all nodes are numbered commonly throughout the reaction pathway.
- (9) Conia, J. M.; Girard, G. *Tetrahedron Lett.* **1973**, 2767.

Coding of Relational Descriptions of Molecular Structures

KARL WIRTH

Fachbereich Mathematik, KME, CH-8001 Zürich, Switzerland

Received November 27, 1985

This is an approach to a naming procedure for molecular structures, which are understood to be stereochemical models of molecules. The procedure is both general and uniform; i.e., it provides any aspect of any molecular structure and codes by using a fully unified criterion. Structural elements of such aspects as, for instance, atoms, bonds, connection angles, dihedral angles, and orientations, are represented by tuples to arrive at relational descriptions of molecular structures. The coding of those relational descriptions is expounded, including the presentation of a canonization algorithm in particular, which is based on minimalization. The result of the coding, which is a systematic name for a described molecular structure, comprises the absolute configuration and enables the symmetry group to be determined.

INTRODUCTION

Why develop a new naming procedure for chemical entities if several already exist?¹ I agree that many of these procedures work very well, which, however, cannot conceal the fact that they are, in a way, patchwork. This is quite understandable as they gradually developed from daily requirements of applied chemistry. As a result existing procedures have been affected in two ways. They are bound to special aspects and classes of molecules and are therefore not general enough. On the other hand, new requirements led to modifications that made these procedures less and less uniform.

Owing to its deductive approach, this paper offers a solution to these two disadvantages. To achieve this, stereochemical models are resorted to. In these models molecules are understood to be (mobile or rigid) arrangements of atoms in space: The atoms are represented by points, and the relationships between them by internal coordinates, e.g., bond lengths, bond angles, and dihedral angles. To avoid any possible misunderstanding it must be pointed out that in this article the term *molecular structure* refers to these limited idealized models of molecules.

We consider a naming procedure of molecular structures to consist of two steps: a first one, which we call *describing*, followed by a second one, which we call *coding*. Describing means representing a molecular structure by a linear sequence of symbols; coding then brings the resulting description into canonical form. The advance of the present procedure over others can now be specified as follows: It is *general*, which means that, in principle, the suggested describing aims to take into account any conceivable aspect of any conceivable molecular structure; and it is *uniform*, which means that the expounded coding functions for any resulting description regardless of whatever structural aspects it represents.²






In order to achieve generality and uniformity describing and coding are completely separated, which the existing naming procedures hardly ever try for. In those procedures describing and coding are mixed, coding, moreover, frequently being based on a different criterion for each structural aspect (e.g., IUPAC Nomenclature³). Such mixing may be very efficient with specific classes of fairly simple molecular structures. However, not only does it prevent generality and uniformity, with more complex molecular structures it may also be a source of errors with regard to uniqueness.

Generality is achieved in the present naming procedure by using relational descriptions as the result of describing. The term relational description corresponds to that of a finite relational system in mathematics.⁴ Such a system is based on tuples, which we use to describe relationships between atoms⁵ of molecular structures. A tuple can consist of any number of atoms in any order, and even repeated atoms are admitted. This enables any element of structural aspects to be expressed: constitutional elements such as atom types, bonds, chains, rings, etc. and stereochemical ones such as bond angles, dihedral angles, orientations, topicities,⁶ etc.

Uniformity is achieved by consistent use of minimalization for all structural aspects as the basis of coding. This criterion was applied by other naming procedures that, however, usually involve only the constitution of molecular structures and operate with adjacency matrices.⁷ A mathematical examination shows that, besides minimalization, any other criterion may be possible.⁴ It ought to be an interesting question whether certain criteria have chemical relevance according to the represented molecular structures.

This article is primarily about coding and not describing, but presupposes relational descriptions as the result of describing. It is, of course, impossible to solve all the problems

Table I. Manner of Describing Usual Structural Elements by Tuples

structural element			describing tuple		
name	drawing	dimension	name	symbol	length
atom		0	single	a	1
bond		1	pair	either ab or ba	2
connection angle		2	triple	either abc or cba	3
dihedral angle		3	quadruple	either abcd or dcba	4
orientation		3	quadruple	one of the following: abcd, acdb, adbc, badc, bcad, bdca, cabd, cbda, cdab, dacb, dbac, dcba.	4

involved in a fully systematic naming procedure at once. As far as describing is concerned, this paper merely makes suggestions and provides ideas as to how molecular structures might be represented by relational descriptions. It is up to chemists to deal with the largely unsolved problems of describing by relational descriptions and to decide which relational descriptions of molecular structures may be useful.

Understandably, it may appear rather strange not to elaborate on describing, the first step of a naming procedure, and, indeed, to focus largely on coding, the second. There are two reasons for this. First, the purely mathematical problem of coding is easier than the largely unsolved one of describing. Second, and quite promisingly, coding can have a rewarding effect on describing, inasmuch as a coding computer program could be used for experimenting: Various relational descriptions of a molecular structure are coded, and the results may allow conclusions to be drawn as to their suitability. Concerning such conclusions, the symmetry group and the absolute configuration obtained in addition to the name provide a relevant test.

An important question in connection with coding is that of the complexity of its algorithm. It may be seen in the light of the famous unsolved graph isomorphism problem:⁸ Is there an algorithm capable of detecting isomorphism between two graphs for which the worst-case time required does not depend exponentially but polynomially on the length of the input? The existence of a polynomial coding algorithm would be an affirmative answer to this question: All one has to do is to check whether the codes of two graphs are identical or not. Since graphs are represented by specific relational descriptions, it is in no way to be expected that the canonization algorithm, which realizes the essence of the coding in this paper, will be polynomial. Nevertheless, owing to the small order of the symmetry groups of molecular structures, worst-case times are not likely to appear when chemically relevant relational descriptions are handled. In fact, for those relational descriptions which represent conceivable molecular structures, existing programs operate within very practical time limits.

The main goal of a naming procedure is to get a structurally informative name; i.e., the properties of the molecular structure should be contained in the name, which includes the symmetry group. Several approaches for symmetry perception have already been developed.⁹ In the coding presented here the proper symmetry group of a molecular structure, or better, the automorphism group of its relational description, is produced in quite a natural way. If configurational aspects are considered, the absolute configuration can be defined when the code of the original molecular structure is compared with that of its reflection image. Additionally, in the case of achiral

molecular structures, the improper elements of the symmetry group, which are the antimorphism of the description, are obtained. Hence, as an overall result, the coding brings forth the whole symmetry group.

The idea of this paper is to provide examples followed by comments, without strictly mathematical terms being applied. To represent molecular structures by relational descriptions, structural elements have to be described in a specific way, as is suggested in section 1. The coding now consists of three steps, which are dealt with in section 2. Of these three the central step is the canonization, an algorithm of which will be expounded in section 3. The structural features in connection with symmetry are determined in the final section, section 4.

1. ASPECTS OF DESCRIBING

This section explains how the elements of the usual structural aspects can be described by tuples to get relational descriptions of molecular structures. Questions of redundancy and arbitrariness are raised but not dealt with. Finally, the kinds of variability that have to be eliminated by coding once arbitrariness has been avoided are pointed out.

It may seem reasonable to describe the elements of the usual structural aspects with tuples as will follow. Nevertheless, it must be emphasized that the proposed describing and the additional remarks as well as the questions raised only serve to whet the appetite for further investigations. I by no means wish the suggestions to be understood as definite.

As has already been said, in principle, any aspect of molecular structures can be represented by relational descriptions. Thus the following limitation to the usual aspects is not in any way meant to restrict what has been emphasized as the generality of the naming procedure.¹⁰ Not mentioning topicities, for instance, does not imply that they cannot be described.

Structural Elements and Describing Tuples. The concepts behind the chemical view of molecular structures are composition, constitution, conformation, and configuration.¹¹ These are based on what we call *structural elements*: atoms, bonds, connection and dihedral angles,¹² and, in addition, orientations, which serve particularly to determine chirality. These structural elements can be described by *tuples* of atom labels. A tuple, more precisely a tuple of length n or in short an n -tuple, is a sequence of n labels where the order is essential. In the special cases $n = 1, 2, 3$, and 4 , which are of interest here, the following expressions are used: single = 1-tuple, pair = 2-tuple, triple = 3-tuple, and quadruple = 4-tuple.

As depicted in the upper part of Table I, an atom is described by a single, a bond by a pair, a connection angle by

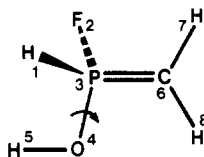


Figure 1. Numbered molecular structure M1.

a triple, and a dihedral angle by a quadruple. The two possible pairs, triples, or quadruples describing such a structural element are related by reflection. In an obvious way, each of these structural elements can be assigned a dimension that is smaller by 1 than the length of the describing tuple.

The lower part of Table I shows how an orientation of dimension three is described by a quadruple. As we can see in the drawing, it results by following the line of a right-handed spiral through four atoms not situated in a plane. It can be shown that whenever another right-handed spiral through the four atoms is considered, one of the twelve quadruples related by even permutations results.

Although three-dimensional orientations are dealt with here, lower dimensional ones could also be considered. An orientation of dimension one can be described by a pair, one of dimension two by one of three triples related by even permutations. With such orientations linear or planar substructures can be characterized.

Apart from the usual structural elements just mentioned, others can be described by tuples. Tuples of any length may serve to describe chains, rings, topologies, or whatever else there may be.¹⁰ Create, for instance, a ring list headed by R_5 containing tuples of length 5, each describing a 5-ring.

Relational Descriptions. A relational description, in the following called a *description*, now consists of *lists* in a given order, each representing a structural aspect. Such a list contains tuples describing structural elements. A list heading specifies the represented structural aspect: A stands for atoms, B for bond, C for connection angle, D for dihedral angle, and O for orientation.

The description starts with A-lists, each list representing atoms of the same kind. The atom kind symbol is noted as an index of the heading in the appropriate list. These indices allow the A-lists to be ordered (e.g., decreasing atomic numbers). Then the description continues with B-, C-, and D-lists, where each list represents structural elements of the same probable range of metric values. That means that a B-list contains pairs describing bonds of the same type such as, for instance, single, double, or triple bonds and so on between atoms of the same kind (CC or CH etc.). A C-list includes triples describing connection angles of the same size such as, for instance, those around 120° for trigonal planar triligant (tp) and 109° for tetrahedral tetraligant (th) or nonlinear biligant (nl) arrangements. A D-list, finally, comprises quadruples describing dihedral angles of the same size such as, for instance, those around 180° for antiperiplanar (ap), 120° for anticlinal (ac), 60° for synclinal (sc), or 0° for synperiplanar (sp) arrangements. The abbreviations in brackets are noted as indices of the headings in the appropriate lists. The ranges of metric values may be used to define an order inside each of the three types of lists B, C, and D (e.g., decreasing values). The description is accomplished by the O-list, which represents orientations of dimension three. With the O-list the index is omitted because right-handedness of the spiral is chosen as standard of orientations.

Example M1. To exemplify what we have said, we use the molecular structure M1, modeled in Figure 1, where each atom is arbitrarily labeled by a natural number. The description of M1 is now self-explanatory. The lists have been ordered as mentioned above:

A_p:3; A_F:2; A_O:4; A_C:6; A_H:1,5,7,8;
B_{PC}:36; B_{PO}:34; B_{PF}:32; B_{PH}:31; B_{CH}:67,68; B_{OH}:45;
C_{tp}:367,368,768; C_{th}:431,432,436,132,136,236; C_{nl}:543;
D_{ap}:4367; D_{ac}:1368,2368; D_{sc}:1367,2367; D_{sp}:4368;
O:3124,6124,7124,8124,6123,7123,8123,7126,8216,8217.

Note that some dihedral angles have not been described because free rotation around the bond 34 (curved arrow in Figure 1) can be expected under usual conditions of observation. Moreover, the only orientations considered are those involving both atoms 1 and 2 in combination with all others (lying in the paper plane) except movable atom 5.

Redundancy and Arbitrariness. Of course, one condition that must be fulfilled by a description is that it should be brief. It should not contain more information than chemists need in order to regenerate the given labeled structure. Indeed, the above description of M1 is highly redundant. The presence of a number of tuples can be derived from the presence of other tuples.

Redundancy in the above description of M1 is particularly found in the B-, C-, and D-lists. Separate B-lists for different bonds are not necessary because the constitution of M1 is already defined by a neighborhood of bonded atoms. In other words, one B-list suffices since numerical values of bond length can be introduced by using the information contained in the A-lists. Furthermore, the kind and number of bonded atoms around atoms 3, 4, and 6 define the size of the connection angles. Thus the C-lists are superfluous since their information is already given in the A- and B-lists. To arrive at the information of the D-lists, additional information, which can be found in the O-list, is needed. If we use only the information of the A- and B-lists, atoms 1 and 2 as well as 7 and 8 are still interchangeable, but they are fixed if any two orientations described in the O-list not including atom 4 are used. In this case the D-lists are superfluous. Moreover, these two orientations are quite sufficient because the remaining eight in the O-list are redundant.

In view of what has been said, the above description of M1 is reduced to a description that still unambiguously characterizes the molecular structure:

A_p:3; A_F:2; A_O:4; A_C:6; A_H:1,5,7,8;
B:36,34,32,31,67,68,45;
O:6123,8217.

A general treatment of questions in connection with redundancy leads to delightful and largely unsolved problems. Moreover, the problem of avoiding redundancy becomes even more complicated when it is considered together with the problem of avoiding arbitrariness. To avoid arbitrariness, we have to consider two questions: What structural aspects are to be described in a list, and in what order are the lists to be arranged?

Variability. Even if these last two questions can be answered, three kinds of variability in describing still remain. They depend on (a) the choice of atom labels; (b) the choice of tuples, when describing structural elements (there are two possibilities for bonds and connection and dihedral angles and twelve for orientations, as can be seen in Table I); and (c) the choice of order of tuples within the lists.

How to cancel this variability is the central task of this paper. Among a class of all descriptions that differ only in variabilities a, b, and c our coding puts out a uniquely determined representative, which we call a *name* of the molecular structure described. It is found from any given representative of the class. For our example of the nonredundant description of M1, let us anticipate the resulting name:

A_p:1; A_F:2; A_O:3; A_C:4; A_H:5,6,7,8;
B:12,13,14,15,36,47,48;
O:1254,2578.

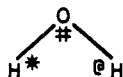


Figure 2. Labeled molecular structure M2.

2. CODING STEPS

Coding lends itself to computer implementation, where the description of the molecular structure would be the input, its name the output. First, two examples are presented which, owing to the fact that they can easily be handled manually, serve to elucidate the three steps involved in coding. Once these examples have been given, the ensuing general presentation should be evident.

The first and third of the three coding steps refer to describing, as suggested in the preceding section, so they are limited to a special class of descriptions, for which they are actually trivial but necessary. The second and central step, however, is purely mathematical and not restricted to any special class of descriptions.

Additionally, it must be mentioned that the simplicity of the given examples must not mislead the reader to mix describing and coding and thereby to neglect the required uniformity of the naming procedure. This happens, for instance, when, in describing a bond, preference is given to one of the two possible pairs by using the order of the A-lists and, in consequence, the first coding step is considered to be superfluous. Procedures of this kind cannot be generalized: Coding would be limited to a specific class of descriptions and thus, to stress the point again, not be uniform.

Example M2. Let us consider the molecular structure M2 of water with the arbitrarily assigned labels *, @, and #, as shown in Figure 2. We use labels that do not suggest an order as would be the case if numbers or letters were used. A nonredundant description is given as follows:

DESC[M2] $A_O: \#; A_H: *, @; B: \# *, \# @.$

In order to derive a name for M2, we eliminate, in a first step, variability b, which results from the choice of the pair when describing a bond. This is achieved by expansion, i.e., by adding the pairs related by reflection in the B-list. Thus we get an expanded description, in short, an EXP-description:

EXP[M2] $A_O: \#; A_H: *, @; B: \# *, * \#, \# @, @ \#.$

In a second step the core of coding, which will be called the canonization, is applied. This cancels variabilities a and c, which arise from the choice of labels and the choice of the order of tuples within the lists. To illustrate this, we first consider all $3! = 6$ numberings of the labels:

	*	@	#
ν_1	1	2	3
ν_2	1	3	2
ν_3	2	1	3
ν_4	2	3	1
ν_5	3	1	2
ν_6	3	2	1

For each of these numberings the labels in EXP[M2] are replaced by the corresponding natural numbers. In this way EXP[M2] is transformed to the numeral EXP-descriptions, in short, NUM-descriptions:

NUM1	$A_O: 3; A_H: 1, 2; B: 31, 13, 32, 23.$
NUM2	$A_O: 2; A_H: 1, 3; B: 21, 12, 23, 32.$
NUM3	$A_O: 3; A_H: 2, 1; B: 32, 23, 31, 13.$
NUM4	$A_O: 1; A_H: 2, 3; B: 12, 21, 13, 31.$
NUM5	$A_O: 2; A_H: 3, 1; B: 23, 32, 21, 12.$
NUM6	$A_O: 1; A_H: 3, 2; B: 13, 31, 12, 21.$

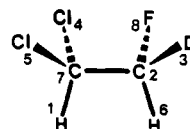


Figure 3. Numbered molecular structure M3.

In each of these NUM-descriptions the tuples within each list are now ordered lexicographically. Ordering tuples of natural numbers lexicographically means arranging them in their numerical order. This leads to the ordered NUM-descriptions, in short, ORD-descriptions:

ORD1	$A_O: 3; A_H: 1, 2; B: 13, 23, 31, 32.$
ORD2	$A_O: 2; A_H: 1, 3; B: 12, 21, 23, 32.$
ORD3	$A_O: 3; A_H: 1, 2; B: 13, 23, 31, 32.$
ORD4	$A_O: 1; A_H: 2, 3; B: 12, 13, 21, 31.$
ORD5	$A_O: 2; A_H: 1, 3; B: 12, 21, 23, 32.$
ORD6	$A_O: 1; A_H: 2, 3; B: 12, 13, 21, 31.$

Two of these ORD-descriptions, namely ORD4 and ORD6, which are identical, are the smallest in terms of lexicographical order. They are generated by two numberings, so-called *minimum numberings*, which we put together in the *minimum class*:

MIN[M2]	*	@	#
ν_4	2	3	1
ν_6	3	2	1

We call the ORD-description that results from any minimum numbering the canonized description, in short, CAN-description:

CAN[M2] $A_O: 1; A_H: 2, 3; B: 12, 13, 21, 31.$

In a third and last step CAN[M2] is reduced by compression back to the size of DESC[M2] as follows: Of the two pairs related by reflection, the lexicographically smaller is always chosen. In this way we finally get a uniquely determined description, which is taken as the name of M2:

NAME[M2] $A_O: 1; A_H: 2, 3; B: 12, 13.$

Example M3. The molecular structure M3, which is depicted in Figure 3, is considered. Here the eight atoms are labeled arbitrarily by natural numbers. We take the following description, which is one of the nonredundant ones:

DESC[M3] $A_{Cl}: 4, 5; A_F: 8; A_C: 7, 2; A_D: 3; A_H: 1, 6;$
 $B: 75, 74, 71, 72, 26, 28, 23;$
 $O: 7415, 2836.$

The name of M3 is derived in the same way as in the preceding example of M2, namely by an expansion-, canonization-, and compression-step. Here the presence of the O-list is of special interest.

In the expansion-step the B-list is treated as in M2. However, variability b in the O-list is removed by adding all quadruples related by even permutations. Thus the EXP-description is obtained:

EXP[M3] $A_{Cl}: 4, 5; A_F: 8; A_C: 7, 2; A_D: 3; A_H: 1, 6;$
 $B: 75, 57, 74, 47, 71, 17, 72, 27, 26, 62, 28, 82, 23, 32;$
 $O: 7415, 7154, 7541, 4751, 4517, 4175, 1745,$
 $1457, 1574, 5714, 5147, 5471, 2836, 2368,$
 $2683, 8263, 8632, 8326, 3286, 3862, 3628,$
 $6238, 6382, 6823.$

To get CAN[M3] by canceling variabilities a and c in the canonization-step, one does not have to generate all $8! = 40320$ numberings and the assigned ORD-descriptions. A considerable reduction of the numberings results from the fact that

the beginning of CAN[M3], which is determined by the A-lists, looks apparently as follows:

A_{Cl}:1,2; A_F:3; A_C:4,5; A_D:6; A_H:7,8;

There remain $2^3 = 8$ numberings:

	1	2	3	4	5	6	7	8
ν_1	7	5	6	1	2	8	4	3
ν_2	8	5	6	1	2	7	4	3
ν_3	7	4	6	1	2	8	5	3
ν_4	8	4	6	1	2	7	5	3
ν_5	7	5	6	2	1	8	4	3
ν_6	8	5	6	2	1	7	4	3
ν_7	7	4	6	2	1	8	5	3
ν_8	8	4	6	2	1	7	5	3

From a study of the B-list it follows that some of these numberings may be eliminated. To show this, the B-lists of the assigned NUM-descriptions are written down:

NUM1 B:42,24,41,14,47,74,45,54,58,85,53,35,56,65;
 NUM2 B:42,24,41,14,48,84,45,54,57,75,53,35,56,65;
 NUM3 B:52,25,51,15,57,75,54,45,48,84,43,34,46,64;
 NUM4 B:52,25,51,15,58,85,54,45,47,74,43,34,46,64;
 NUM5 B:41,14,42,24,47,74,45,54,58,85,53,35,56,65;
 NUM6 B:41,14,42,24,48,84,45,54,57,75,53,35,56,65;
 NUM7 B:51,15,52,25,57,75,54,45,48,84,43,34,46,64;
 NUM8 B:51,15,52,25,58,85,54,45,47,74,43,34,46,64;

Lexicographical ordering within each B-list produces the B-lists of the ORD-descriptions:

ORD1 B:14,24,35,41,42,45,47,53,54,56,58,65,74,85;
 ORD2 B:14,24,35,41,42,45,48,53,54,56,57,65,75,84;
 ORD3 B:15,25,34,43,45,46,48,51,52,54,57,64,75,84;
 ORD4 B:15,25,34,43,45,46,47,51,52,54,58,64,74,85;
 ORD5 B:14,24,35,41,42,45,47,53,54,56,58,65,74,85;
 ORD6 B:14,24,35,41,42,45,48,53,54,56,57,65,75,84;
 ORD7 B:15,25,34,43,45,46,48,51,52,54,57,64,75,84;
 ORD8 B:15,25,34,43,45,46,47,51,52,54,58,64,74,85;

The B-lists of ORD1 and ORD5 are the lexicographically smallest and thus two numberings remain:

	1	2	3	4	5	6	7	8
ν_1	7	5	6	1	2	8	4	3
ν_5	7	5	6	2	1	8	4	3

We now turn to the O-list, which leads to the last numbering elimination. With respect to the two remaining numberings, the O-lists of the assigned NUM-descriptions are as follows:

NUM1 O:4172,4721,4217,1427,1274,1742,7412,7124,
 7241,2471,2714,2147,5368,5683,5836,3586,
 3865,3658,6538,6385,6853,8563,8635,8356.
 NUM5 O:4271,4712,4127,2417,2174,2741,7421,7214,
 7142,1472,1724,1247,5368,5683,5836,3586,
 3865,3658,6538,6385,6853,8563,8635,8356.

The resulting O-lists of the ORD-descriptions are as follows:

ORD1 O:1274,1427,1742,2147,2471,2714,3586,3658,
 3865,4172,4217,4721,5368,5683,5836,6385,
 6538,6853,7124,7241,7412,8356,8563,8635.
 ORD5 O:1247,1472,1724,2174,2417,2741,3586,3658,
 3865,4127,4271,4712,5368,5683,5836,6385,
 6538,6853,7142,7214,7421,8356,8563,8635.

The O-list of ORD5 is lexicographically smaller, and thus the minimum class contains only one surviving numbering:

MIN[M3]	1	2	3	4	5	6	7	8
ν_5	7	5	6	2	1	8	4	3

The complete ORD-description ORD5 is now CAN[M3]. It has been constructed in stages and consists of the obvious A-lists and the B- and O-lists of ORD5:

CAN[M3] A_{Cl}:1,2; A_F:3; A_C:4,5; A_D:6; A_H:7,8;
 B:14,24,35,41,42,45,47,53,54,56,58,65,74,85;
 O:1247,1472,1724,2174,2417,2741,3586,
 3658,3865,4127,4271,4712,5368,5683,
 5836,6385,6538,6853,7142,7214,7421,
 8356,8563,8635.

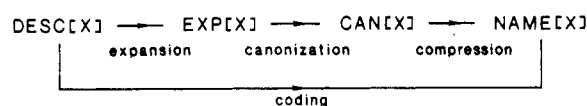
What remains is the compression-step, which always chooses the lexicographically smallest of two pairs related by reflection and of twelve quadruples related by even permutations. The result is the name of M3, which is as long as DESC[M3], in particular, having two quadruples in the O-list:

NAME[M3] A_{Cl}:1,2; A_F:3; A_C:4,5; A_D:6; A_H:7,8;
 B:14,24,35,45,47,56,58;
 O:1247,3586.

Steps of Coding in General. As illustrated with molecular structures M2 and M3, NAME[X] of any molecular structure X can be found from DESC[X] in the following three steps:

1. *Expansion.* The B- and, if they appear in DESC[X], the C- and D-lists are extended by adding all tuples related by reflection and the O-list is extended by adding to each quadruple the eleven related by even permutations. Thus we get the *EXP-description* EXP[X], which expresses the symmetry of each structural element.
2. *Canonization.* The labels of EXP[X] are repeatedly numbered, and for each numbering a *NUM-description* is obtained. The tuples of the latter are lexicographically ordered within each list, and thus we get an *ORD-description*. All ORD-descriptions are compared by again using lexicographic ordering. The smallest one, which may appear more than once, is taken as the *CAN-description* CAN[X].
3. *Compression.* CAN[X] is reduced by deleting all tuples except the lexicographically smallest of those that are related by reflection or even permutations. The result is NAME[X], which is of equal size as the original description DESC[X].

The sequence of these three steps, which are schematically represented below, forms what we call coding. As has been declared, coding picks out exactly one of all descriptions that differ only in variabilities a, b, and c, no matter which description one starts from.



3. CANONIZATION ALGORITHM

A procedure that realizes the canonization in an efficient way is called a *canonization algorithm*. The one presented here is not restricted to EXP-descriptions, but can be used for what we term general descriptions, in short, *GEN-descriptions*, which correspond to that of finite relational systems in mathematics.⁴ These include any number of lists and, within a list, tuples of any fixed length where the labels need not differ. Those who are mainly interested in the chemical side of the problem may wish to skip this purely theoretical section.

First we define the canonization algorithm generally, restricting ourselves, for reasons of simplicity, to GEN-descriptions with just one list. Then two examples follow from which the algorithm becomes evident; the second of these illustrates its application to GEN-descriptions with several lists.

Steps of the Canonization Algorithm. Given a GEN-description GEN[X] with one list L, we define a *q-string* as a

Table II. Canonization Algorithm Applied to GEN[S1]^a

number <i>q</i> of the step	<i>q</i> -string	assigned canonized <i>q</i> -string
1	ab, cb, dc, de,	12, 12, 12, 12,
2	ab, cb, ab, dc, ab, de, cb, ab, cb, dc, cb, de, dc, ab, dc, cb, dc, de, de, ab, de, cb, de, dc,	12,32, 12,34, 12,34, 12,32, 12,31, 12,34, 12,34, 12,23, 12,13, 12,34, 12,34, 12,13,
3	dc, de, ab, dc, de, cb, de, dc, ab, de, dc, cb,	12,13,45, 12,13,24, 12,13,45, 12,13,34,
4	dc, de, cb, ab.	12,13,24,54.

^a Remaining *q*-strings are boldfaced.

sequence of *q* tuples of L. By replacing the labels of a *q*-string with natural numbers in such a way that the smallest in terms of lexicographical order is achieved, we arrive at what is called a *canonized q-string*. With these terms the canonization algorithm is described recursively. It includes *r* steps, *r* being the number of tuples of the list L:

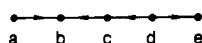
In the first step select all 1-strings of GEN[X] (they are the tuples of L) which lead to the lexicographically smallest canonized 1-string.

After *q* - 1 steps there is a set of (*q* - 1)-strings, all of which lead to the same canonized (*q* - 1)-string. At the *q*th step extend these given (*q* - 1)-strings to all still possible *q*-strings and select for survival those which lead to the lexicographically smallest canonized *q*-string.

After the *r*th step a set of *r*-strings remains. They all lead to the same canonized *r*-string, which is CAN[X].

Example S1. We consider a GEN-description with one list representing the structure S1, visualized by the pictured graph:

GEN[S1] L:ab,cb,dc,de.



The canonization algorithm, including *r* = 4 steps, works according to Table II. After the last step there is only one surviving 4-string. With the assigned canonized 4-string the CAN-description is achieved, and the replacing of labels by natural numbers corresponds to the minimum numbering:

CAN[S1] L:12,13,24,54.
MIN[S1]

a	b	c	d	e
5	4	2	1	3

Example S2. We shall now exemplify the canonization algorithm with a GEN-description containing several lists. The lists will be handled one at a time; after exhausting one list, the remaining *q*-strings will be continued in the following list. We consider an "abstract" structure S2 with three lists, L₁, L₂, and L₃:

GEN[S2] L₁:ae,ea,dd;
L₂:ae,de;
L₃:bec,cef,feb.

Table III. Canonization Algorithm Applied to GEN[S2]^a

<i>q</i>	<i>q</i> -string	canonized <i>q</i> -string
L ₁ 1	ae, ea, dd,	12, 12, 11,
2	dd, ae, dd, ea,	11,23, 11,23,
3	dd, ae, ea; dd, ea, ae;	11,23,32; 11,23,32;
L ₂ 4	dd, ae, ea; ae, dd, ae, ea; de, dd, ea, ae; ae, dd, ea, ae; de,	11,23,32,23, 11,23,32,13, 11,23,32,32, 11,23,32,12,
5	dd, ea, ae; de, ae;	11,23,32,12,32;
L ₃ 6	dd, ea, ae; de, ae; bec, dd, ea, ae; de, ae; cef, dd, ea, ae; de, ae; feb,	11,23,32,12,32,425, 11,23,32,12,32,425, 11,23,32,12,32,425,
7	dd, ea, ae; de, ae; bec, cef, dd, ea, ae; de, ae; bec, feb, dd, ea, ae; de, ae; cef, bec, dd, ea, ae; de, ae; cef, feb, dd, ea, ae; de, ae; feb, bec, dd, ea, ae; de, ae; feb, cef,	11,23,32,12,32,425,526, 11,23,32,12,32,425,624, 11,23,32,12,32,425,624, 11,23,32,12,32,425,526, 11,23,32,12,32,425,526, 11,23,32,12,32,425,624,
8	dd, ea, ae; de, ae; bec, cef, feb. dd, ea, ae; de, ae; cef, feb, bec. dd, ea, ae; de, ae; feb, bec, cef.	11,23,32,12,32,425,526,624. 11,23,32,12,32,425,526,624. 11,23,32,12,32,425,526,624.

^a Remaining *q*-strings are boldfaced.

As can be seen from Table III, the algorithm requires eight steps, three for L₁, two for L₂, and three for L₃. Note that it fluctuates; i.e., the number of remaining *q*-strings increases or decreases with each step. At the end of the algorithm three 8-strings remain. They all lead the same assigned canonized 8-string, which is the CAN-description. In each of the three cases the replacing of the labels by natural numbers corresponds to a minimum numbering:

CAN[S2] L₁:11,23,32;
L₂:12,32;
L₃:425,526,624.

MIN[S2]

a	b	c	d	e	f
3	4	5	1	2	6
3	6	4	1	2	5
3	5	6	1	2	4

Remarks. Note that if, in dealing with descriptions of molecular structures, the A-lists are treated first, the given canonization algorithm must be slightly modified to avoid factorial effects in the case where several atoms of the same kind appear. An additional remark to be made is that the canonization algorithm produces the CAN-description by adding a tuple at each step. Algorithms that work differently are quite conceivable. Their discussion, however, would lead too far in this context. An important criterion in choosing an algorithm is certainly its efficiency.

4. CHIRALITY AND SYMMETRY

As opposed to the previous section, where we have considered general descriptions, the rest of this paper is again restricted to descriptions of molecular structures. Furthermore, we assume that the descriptions are symmetry consistent, i.e., that they correctly represent the symmetry of the described molecular structures. Some details in connection with symmetry consistency have already been studied.¹³

To begin with, two more examples are given to illustrate what will be considered. It is shown how the absolute configuration of a molecular structure may be defined. Then we

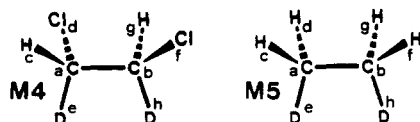
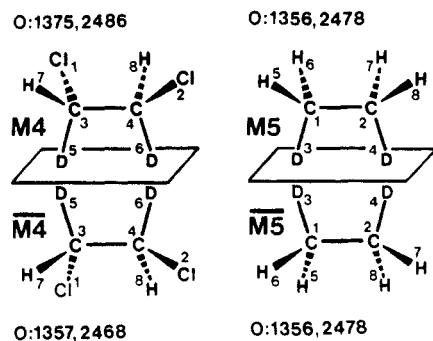


Figure 4. Labeled molecular structures M4 and M5.

Figure 5. Minimum numbered molecular structures M4, $\overline{M4}$, M5, and $\overline{M5}$ and the resulting O-lists of their names.

demonstrate how the proper and improper elements of the symmetry group are to be determined.

Examples M4 and M5. The two molecular structures M4 and M5, shown in Figure 4, differ in comparison with M3 only in their ligands of the C atoms. To avoid ambiguity when these examples are applied, letters should be chosen as labels instead of numbers. It is left to the reader to verify the following names of M4 and M5, which can be done manually, either analogously to M3 in section 2 or by consistent use of the canonization algorithm of section 3:

NAME[M4] $A_C:1,2; A_C:3,4; A_D:5,6; A_H:7,8;$
 $B:13,24,34,35,37,46,48; O:1375,2486.$
 NAME[M5] $A_C:1,2; A_D:3,4; A_H:5,6,7,8;$
 $B:12,13,15,16,24,27,28; O:1356,2478.$

Absolute Configuration. By reflection of a molecular structure X we get a molecular structure \bar{X} . To each description DESC[X] there is a related description DESC[\bar{X}], which is obtained by replacing every quadruple of the O-list of DESC[X] with a quadruple related by an odd permutation. Thus NAME[X] and NAME[\bar{X}] differ only in the O-list, if at all.

When NAME[X] and NAME[\bar{X}] are compared lexicographically, it is possible to define the *signum* of X, in other words, to define its absolute configuration: We say that X is of signum +1 resp. 0 resp. -1 if NAME[X] is lexicographically larger resp. equal resp. smaller than NAME[\bar{X}]. The signum leads to three classes of molecular structures, an achiral one and two chiral ones. A molecular structure is taken as *chiral* when its signum is +1 or -1 and as *achiral* when its signum is 0.

Figure 5 illustrates the situation in our examples of the molecular structures M4 and M5. Only the O-lists are mentioned since M4 and $\overline{M4}$ as well as M5 and $\overline{M5}$ differ at most in these. M4 is of signum +1 and therefore chiral, its enantiomer $\overline{M4}$ is of signum -1. M5 and $\overline{M5}$ are both of signum 0 and thus achiral.

A general recipe for a classification of molecular structures was first given by Cahn, Ingold, and Prelog.¹⁴ This rule is handy and effective because it mixes describing and coding in a most skillful way, but for this very reason it is also limited. Other approaches that are quite similar in nature to the present one lead to the absolute configuration, too.¹⁵

Proper Elements of the Symmetry Group. In dealing with the examples in the preceding sections, we have seen that coding leads to the minimum class. The fact that there are several minimum numberings has to do with symmetry of the

considered molecular structure X, or more mathematically, with its automorphisms.

A permutation of the labels which transforms an element of MIN[X] into an element of MIN[X] is called an *automorphism* of the molecular structure X. More precisely, if ν and ν' are minimum numberings of X, then $\nu' = \nu\alpha$ and thus $\alpha = \nu^{-1}\nu'$ is an automorphism of X.¹⁶ It follows that the number of automorphisms is the same as the number of minimum numberings. An automorphism is taken now as a proper element of the symmetry group. Thus the class AUT[X] of all automorphisms forms the *proper symmetry group* of X.

In our examples of molecular structures M4 and M5 we get the minimum numberings:

MIN[M4]	a	b	c	d	e	f	g	h
ν_1	3	4	7	1	5	2	8	6
ν_2	4	3	8	2	6	1	7	5

MIN[M5]	a	b	c	d	e	f	g	h
μ_1	1	2	5	6	3	8	7	4
μ_2	2	1	7	8	4	6	5	3

From these the proper elements of the symmetry group result, which are the same for M4 and M5. According to the drawing in Figure 4, they correspond to the identity and to a 180° rotation around a vertical C_2 axis in the paper plane bisecting the bond *ab*:

AUT[M4]	a	b	c	d	e	f	g	h
$\nu_1^{-1}\nu_1$	a	b	c	d	e	f	g	h
$\nu_1^{-1}\nu_2$	b	a	g	f	h	d	c	e

AUT[M5]	a	b	c	d	e	f	g	h
$\mu_1^{-1}\mu_1$	a	b	c	d	e	f	g	h
$\mu_1^{-1}\mu_2$	b	a	g	f	h	d	c	e

Improper Elements of the Symmetry Group. Let X be an achiral molecular structure. A permutation of the labels which transforms a numbering of MIN[X] into a numbering of MIN[\bar{X}] is called an *antimorphism* of X. More precisely, if ν and $\bar{\nu}$ are minimum numberings of X and \bar{X} , then $\bar{\nu} = \nu\beta$ and thus $\beta = \nu^{-1}\bar{\nu}$ is an antimorphism of X. The number of antimorphisms is the same as the number of minimum numberings and thus of automorphisms. An antimorphism is taken now as an improper element of the symmetry group. While AUT[X] forms a group, this is not the case for the class ANT[X] of antimorphisms.

In our example of the achiral molecular structure M5 we get the minimum numberings of $\overline{M5}$:

MIN[$\overline{M5}$]	a	b	c	d	e	f	g	h
$\bar{\mu}_1$	1	2	6	5	3	7	8	4
$\bar{\mu}_2$	2	1	8	7	4	5	6	3

From this the improper elements of the symmetry group result. According to the conformation drawn in Figure 4, they correspond to plane reflections at the paper plane and a plane orthogonal to it bisecting bond *ab*:

ANT[M5]	a	b	c	d	e	f	g	h
$\mu_1^{-1}\bar{\mu}_1$	a	b	d	c	e	g	f	h
$\mu_1^{-1}\bar{\mu}_2$	b	a	f	g	h	c	d	e

The *symmetry group* of a molecular structure now consists only of proper elements if it is chiral and of proper and improper elements if it is achiral.

ACKNOWLEDGMENT

I would like to thank A. S. Dreiding, M. K. Huber, and D. Pazis for helpful discussions. Special thanks are due to E. Karafiat and F. Siegerist, without whose invaluable help this article would hardly have been possible.

REFERENCES AND NOTES

- (1) A select bibliography up to 1980 is found in ref 4. For new studies see: Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. "Condensed Structure Identification and Ring Perception". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 195-203. Fella, A. L.; Nourse, J. G.; Smith, D. H. "Conformation Specification of Chemical Structures in Computer Programs". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 43-47.
- (2) The two attributes "general" and "uniform" can be added to those given in the list of desirable qualities chemical coding systems ought to have: Read, R. C. "A New System for the Designation of Chemical Compounds. 1. Theoretical Preliminaries and the Coding of Acyclic Compounds". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 135-149.
- (3) *IUPAC, Nomenclature of Organic Chemistry*; Pergamon: Oxford, 1979; Sections A-F, H.
- (4) Wirth, K.; Huber, M. K. "Numbering of Finite Relational Systems". *MATCH* **1981**, *12*, 3-14.
- (5) For certain purposes electrons, kernels, entire subunits, etc. may be chosen instead of atoms.
- (6) Topicity is used as a general term including homotopic, enantiotopic, and diastereotopic cases (see ref 11).
- (7) The idea of minimalization or the quite similar one of maximalization already arises in the IUPAC numbering practice. The same idea is implicitly contained in: Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, *5*, 107-113. Furthermore see: Hendrickson, J. B.; Toczko, A. G. "Unique Numbering and Cataloguing of Molecular Structures". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171-177.
- (8) There are comprehensive annotated bibliographies up to 1978 in: Read, R. C.; Corneil, D. G. "The Graph Isomorphism Disease". *J. Graph Theory* **1977**, *1*, 339-363. Gati, G. "Further Annotated Bibliography on the Isomorphism Disease". *J. Graph Theory* **1979**, *3*, 95-109. For later developments see: Corneil, D.; Goldberg, M. "A Non-Factorial Algorithm for Canonical Numbering of a Graph". *J. Algorithms* **1984**, *5*, 345-362.
- (9) For a bibliography see: Randic, M.; Brissey, G. M.; Wilkins, C. L. "Computer Perception of Topological Symmetry via Canonical Numbering of Atoms". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 52-59.
- (10) All the same the best thing might still be to limit describing to the usual aspects presented in Table I, i.e., to try to represent the whole molecular structure by these. In this case further aspects, e.g., chains, rings, topicities, etc., would be derivable properties of the structurally informative name as symmetries are, too.
- (11) Mislow, K. In *Introduction to Stereochemistry*; Benjamin: New York, 1966; pp 24, 50, 82.
- (12) We use "connection angle" instead of the usual "bond angle" to get the first four letters of the alphabet (bold face), which is purely accidental. The term "dihedral angle" is used according to ref 11, p 4.
- (13) Floersheim, P.; Wirth, K.; Huber, M. K.; Pazis, D.; Siegerist, F.; Haegi, H. R.; Dreiding, A. S. "From Mobile Molecules to Their Symmetry Groups: A Computer-Implemented Method". *Stud. Phys. Theor. Chem.* **1983**, *23*, 59-80.
- (14) Cahn, R. S.; Ingold, C. K.; Prelog, V. "Specification of Molecular Chirality". *Angew. Chem., Int. Ed. Engl.* **1966**, *5*, 385-415.
- (15) For a selection of publications see: Beierbeck, H. "Simple Stereochemical Structure Code for Organic Chemistry". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 215-222. For quite a different approach see: Ruch, E. "Homochiralität als Klassifizierungsprinzip von Molekülen spezieller Molekülklassen". *Theor. Chim. Acta* **1968**, *11*, 183-192. In this study a different way of considering molecular structures leads to the concept of homochirality, a principle which he considers to be suitable for a physically relevant classification. He shows that the latter cannot exist for certain classes of molecular structures.
- (16) If ξ maps a into b and η maps b into c, the composition $\eta\xi$ will map a into c. If ξ maps a into b, the inverse ξ^{-1} will map b into a.

BOOK REVIEWS

Chemometrics. By Muhammad A. Sharaf, Deborah L. Illman, and Bruce R. Kowalski (University of Washington). Wiley, New York, 1986. xi + 332 pp. \$50.00.

The authors of *Chemometrics* are currently, or previously in the case of Sharaf, associated with the chemometrics group at the University of Washington. The purpose of this book is to present "a logical and systematic introduction to chemometrics as incorporated into analytical chemistry as well as other areas of experimental chemistry". The general organization of the book involves seven chapters. Some of the chapters contain applications of the techniques to data from the literature. All chapters have appropriate references through 1985 and supplemental readings which are annotated, at least through Chapter 5. After Chapter 5 the notes on the contents of the references disappear. A page count analysis shows that 38% of the text is devoted to exploratory data analysis, including pattern recognition (Chapter 6); 17% to signal detection and manipulation (Chapter 3); approximately 10% each to design and optimization (Chapter 2), calibration and analysis (Chapter 4), and signal resolution (Chapter 5); and approximately 5% each to sampling (Chapter 1) and control of systems (Chapter 7). There are some 113 figures and 48 tables effectively spread throughout the book.

The first chapter provides a brief introduction to univariate statistics and sampling errors. Chapter 2 briefly describes experimental design, analysis of variance, and response surfaces. The next chapter has a very clear and informative discussion of the often confusing detection limit problem as well as shorter discussions of signal/noise enhancement, filtering, transforms, and data smoothing. Chapter 4 contains a very good

discussion of the calibration and analysis process, calibration curve fitting, and associated errors. It also discusses the generalized standard addition method. The following chapter deals with general response curve fitting, factor analysis use for number of component determinations, and regression and other techniques for curve resolution. Chapter 6, the largest and the best in the book, provides an excellent overview of chemical pattern recognition, display methods, and applications for these techniques. This chapter also contains a very good description of SIMCA pattern recognition, which is difficult to find elsewhere. Partial least squares, a type of regression analysis involving blocks of data that is currently being used in multivariate calibration, is also described. The last chapter provides a brief introduction to system control and optimization, including simplex optimization.

This book provides a very good introduction to the chemometrics topics described. It is nicely produced and edited with only a few typographical errors. However, it is relatively short and could not possibly cover all the topics that are included, sometimes by default, in the chemometrics category. For example, there is no coverage of information theory, which could have been easily introduced in discussions of the chemical analysis system. I also would have preferred some discussion of multivariate statistics including the correlation and covariance matrices and matrix algebra in the first chapter. For the chemist who is interested in an overview of this rapidly expanding field or who wishes to read the more detailed chemometrics literature, this book should definitely be read. The chapter on pattern recognition and data analysis alone is very useful. At present this is the only introductory book in this field.

Donald R. Scott, U.S. Environmental Protection Agency