WORKING NOTE AC46

Alan Creak
5 March 1986

COMPUTER-READABLE CHEMICAL FORMULAE

This note collects together for easy reference some of the things I have written about/done with chemical formulae. All the material is quite old, and I have only been able to estimate the dates.

(This version was reformatted in 2000 from the original working note, much of which was a barely legible photocopy of the ancient typed original copied by Gestetner duplicator. The essence of the original is preserved throughout, but I have made one or two minor changes to improve the readability.)

1. WORK DONE ON THE IBM1130.

This was my first attempt, carried out around 1970 to 1973. It was written mostly in Fortran, with bits of IBM1130 Assembly Language here and there. I wrote a large number of routines to manipulate the formulae : most of them worked, but there was always a problem with canonicalising cyclic molecules which I never managed to track down. I'll list the routines, with comments, in a later Working Note.

The text follows, headed "A LINE-FORMULA NOTATION".

(The original for this part was the photocopied Gestetner version mentioned. It was readable, but I was unable to get a sufficiently good scanned version for automatic conversion. Much of it is therefore retyped.)

2. WORK DONE ON THE BURROUGHS 6700.

This all happened from about 1975 to 1978; it had to stop when the B6700 went away. It was very dependent on B6700 Algol, which was perhaps a mistake. This version incorporated a number of improvements suggested by the earlier work, and made possible largely by the much larger word size of the B6700 (48 bits instead of 16).

Some notes follow, under the heading "B6700 WORK".

(Most of the text for this part was machine-readable, but the tabular material was copied from rather poor line-printer output and further photocopied. This has essentially all been retyped; in so doing I introduced gaps between the bit fields in the long words for easier reading, but the bits should be right.)

A LINE FORMULA NOTATION

In systems designed for the storage and retrieval of chemical information, it is usually necessary to establish conventions governing the representation of molecular formulae. While this requirement poses comparatively minor problems in a system intended for manual operation, the same problems become considerably more acute if the system is to be implemented on a computer. Two kinds of difficulty can be distinguished. They arise from the incompatibility of the linear storage organization of the computer with, on the one hand, the 3-dimensional nature of molecules, and, on the other, the absence of any inherent absolute ordering in the formulae. By providing suitable software, more complex forms of organization can be imposed on the store; but this advantage must be paid for in storage space for the software, and in extra execution time, and it is therefore worthwhile to seek an alternative solution.

The problem can be restated thus : we require a notation in which we can write the formula of a molecule in a linear form, while preserving the important features of the molecule's 3-dimensional structure, and rules which resolve possible ambiguities (such as that exemplified by the two equivalent formulae $CH_3.CO.CH_3$ and $CO.(CH_3)_2$).

These requirements can be met by a line-formula notation. Several such notations have been developed, mainly in connection with work on information retrieval and documentation; perhaps the best-known variant is that due to Wiswesser. These systems are not in general designed to facilitate extensive manipulation of the formulae by the computer, and it is usually assumed that the coding and decoding of the formulae will be performed by people; in consequence, rather rich "vocabularies" of symbols are provided, in which the emphasis lies on the compactness and comprehensibility of the coded formula.

The notation described below was developed in a different context. It is designed for use in a study of the application of machine intelligence to the field of gas kinetics, where the first priority is for easy mechanical manipulation of the formulae. The requirement that the formulae should be capable of being expressed in a standard form remains, as the system must be able to identify a compound in its files. In addition, it should be possible to represent unstable species (in particular, free radicals) in a natural way; this is not usually an urgent requirement in a conventional information storage and retrieval system.

DEFINITION OF THE LINE-FORMULAE

As well as the obvious requirement that the line-formulae should satisfactorily represent the structure of any molecule likely to be encountered, the context in which the formulae are to be used suggests other desirable features; these appear from considerations of the types of molecule likely to be important, and of the nature of their reactions. The size of computer used in the study (IBM1130, with 8K 16-bit words) provides an extra constraint on the final choice; internal storage space is at a premium, and a reasonably compact notation is essential.

DEFINITION OF THE GROUP-WORD

The requirement of compactness is to a certain extent alleviated by the fairly small molecules involved; it seems likely that an average-sized molecule will be about the size of hexane. Hexane contains 20 atoms, but can more usefully be represented as 6 groups of the form CH_n. The utility of a representation based on groups rather than on atoms is amply demonstrated by the functional-group concept of classical organic chemistry, and one computer word in the line-formula therefore represents a group of this type, specifying the type of central atom and its pendant atoms – with a salutary economy in storage requirements. Such a computer word will be called a *group-word*. This technique also has advantages in formula manipulation; thus, to break an atom-hydrogen bond, it is only necessary to remove one hydrogen atom from the appropriate group-word in the line-formula. The size of the functional group represented in one word is limited both by the word size and by the need to avoid special cases, which would require additional programming as far as possible. The convention adopted is to include in the group hydrogen atoms and oxygen atoms if doubly-bonded to the central atom of the group; methyl acetate is therefore represented as $CH_3.CO.O.CH_3$, with 4 groups. This formula also illustrates another special case which, somewhat unfortunately, arises out of the convention itself : two distinct types of oxygen atom are allowed, which may be distinguished as ketonic and etheric oxygen atoms. This drawback is accepted in view of the undoubted importance of the ketone functional group, and the necessary extra programming supplied as required.

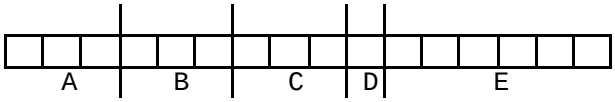
The group-word also includes an indication of the number of external bonds formed by the group. The two CH_2 groups in the n-propyl radical are therefore distinguishable in the formula, which could be written as CH3,l.CH2,2.CH2,1, where the bond counts are written after their groups.

This method again makes for straightforward manipulation; thus, to break a $CO\text{-}CH_2$ bond in acetylacetone, all that is required is to delete a bond and reduce the bond counts in the disconnected group-words by 1 :

$$CH3,l.CO,2.CH2,2.CO,2.CH3,l \rightarrow CH3,1.CO,2.CH2,1 + CO,l.CH3,1$$

The final factor affecting the structure of the group-word is isotopic substitution. Within the restricted field of application of this study, deuterium labelling has been particularly important, and it seemed desirable to make explicit provision for its inclusion in the formulae. At the same time, the important feature from the point of view of chemical identification is the total number of hydrogen atoms in the group, of whatever isotope, so the stratagem adopted is to retain in the group-word the total number of hydrogen atoms, and to qualify this by the number of deuterium atoms. The molecule CH_2D_2 is thus described essentially as "a central carbon atom with 4 hydrogen atoms, two of which are deuterium". Other forms of isotopic substitution are at present ignored.

These considerations lead to the pattern shown in Figure 1 for the group-word; examples are presented in Figure 2. The order of the fields in the group-word is not significant. Procedures, in the form of subroutines, are provided to extract any field from a group-word.



Field A : number of external bonds
B : total number of hydrogen atoms
C : number of deuterium atoms
D : 1 for a ketonic oxygen atom
E : central atom identification

Figure 1

CH2D2	...	1..	.1.	.	1
-CH3	..1	.11	...	.	1
-NO	..1	...	...	1	....1.
-CD=	.1.	..1	..1	.	1
| -CD-	.11	..1	..1	.	1
-OH	..1	..1	...	.	11
OH	...	..1	...	.	11

Figure 2

ARRANGEMENT OF GROUP-WORDS INTO LINE-FORMULAE

Having defined the group-word, we may now represent essentially linear molecules by arranging the appropriate group-words in the obvious sequence; but this method is clearly inadequate to represent molecules containing branched chains or rings. A straightforward extension of the linear method can be used to handle branches : we introduce the convention that every group after the first forms exactly one backward bond to the closest group with unsatisfied bonds. The simple formulae for the linear molecules conform to this convention, and it is adequate to handle a branched molecule of any complexity, provided that the molecule contains no rings. The convention has the advantage that many formulae can be written in rather familiar ways : thus, glycerol can be written as CH2.OH.CH.OH.CH2.OH. It is also now possible to rearrange formulae : acetone may be written either as CH3.CO.CH3 or as CO.CH3.CH3, a fact which will become important when seeking a standard form for the line-formulae.

Rings are slightly more difficult to accommodate in the scheme, as, having established the convention that each group forms exactly one backward bond to the closest unsatisfied bond, we now require that one atom in each ring shall form two backward bonds, one to an unsatisfied bond which may be anywhere in the preceding part of the formula. This can be accomplished without modifying the conventions so far laid down by defining a dummy group-word, the function of which is to take a bond from the closest preceding group and to direct it backwards to a specified earlier group. The convention adopted here is to represent the dummy group by the negative integer which, when added to thc ordinal number of the immediately preceding group-word, gives the ordinal number of the group-word corresponding to the other end of the bond; the ordinal numbers of the group-words are calculated omitting any dummy group-words. Thus, 1,2-dimethylcyclobutane could be coded as

CH.CH3.CH2.CH2.CH.-4.CH3,

where a dummy group-word –4 represents a branch from the CH group immediately before it (with ordinal number 5) to the group-word with ordinal number 1 the first CH group. The ordinal number of the final CH_3 group is 6. There is no possibility of confusing the dummy group-word with a normal group-word; as negative numbers are expressed in two's-complement form, a dummy group-word would be interpreted by the rules laid down above as a group XD_7O, forming 7 external bonds (where X is some central atom). The probability of any central atom in fact having a valency of 16 is considered small enough to be ignored.

This completes the definition of the line-formula. It will be noted that, while the formulae described are adequate to represent the topology of a molecule, there is no way of representing

stereochemical properties, or bond orders. The stereochemistry is omitted because it is not conspicuously important in the field of application of the formulae, and because its inclusion would have been somewhat expensive in terms of storage space. The bond orders are omitted for a more fundamental reason, which can best be explained by an example. Consider the radical $CH_3.CH.CH.CH_2$, by no means improbable in the context of gas-phase reactions of small organic molecules This radical is resonance-stabilised, and can exhibit free-radical character at either the second or the fourth group. any specification of a pattern of single and double bonds would clearly select one of the possible resonance structures, and would represent only one of the two possible groups as the free-radical centre, thus misrepresenting the actual structure of the radical.

THE EXTERNAL FORM OF THE LINE-FORMULAE

Although the formulae discussed so far have occasionally been written in a familiar chemical form, this has been understood as a convenient representation of the line-formulae coded into computer words as has been described. As the binary representation itself is less than immediately comprehensible, it is clearly desirable to define another line-formula notation in which the chemical formulae can be expressed for input and output. This notation is closely similar in structure to the binary notation, but the groups are written in terms of their constituent atoms in a readily comprehensible way. Subroutines are provided to translate from external to internal forms and vice versa.

Each group is written as the symbol for the central atom, followed by up to three extra terms : an H part, an O part, and a bonds part. Any or all of those extra terms may be omitted if appropriate, but, if present, must appear in the order specified. The H part defines the number and nature of the hydrogen atoms present; this is written flexibly as H and/or D, optionally followed by a count. For example, the molecule CH_2D_2 could be written as CH2D2, or CD2H2, or CHD2H, and so on. The O part consists of the letter O, and denotes a ketonic oxygen atom : the molecule CH_2O could be written CH2O (but not COH2, as this would contravene the ordering rule). The bonds part consists of a comma followed by an integer giving the number of external bonds formed by the group; this may be omitted on input (and will be omitted on output) if the total number of bonds formed by the central atom is equal to its conventional valency. With this convention, l-butene could be written as CH2,1.CH,2.CH2.CH3, and the acetonyl radical as CH3.CO.CH2,1.

Groups are separated from each other in a complete formula by full stops. Rings are indicated by appending labels to the groups joined by the extra required bond; a label is any character acceptable to the programme (including decimal digits and the element symbols) enclosed in brackets, and separated from adjacent groups by full stops. 1,2-dimethylcyclobutane could thus be written as CH.(3).CH3.CH2.CH2.CH.(3).CH3, which would be translated into the internal formula already given for this compound, and spiropentane as C.(1).(2).CH2.CH2.(1).CH2.CH2.(2)

No mechanism is provided for handling repeated groups, as in the formula $CO.(CH_3)_2$. While it would not be too difficult to incorporate such a mechanism, the advantage gained would be marginal with the small molecules involved, and the extra programme would occupy valuable storage space. Two other simplifications are introduced to conserve storage : the set of permissible elements (C, H, D, N, O, F, Cl, Br, I) is at present small enough to allow each element to be denoted by a single character, the letters G and J denoting chlorine and bromine; and a rule is established which requires that every group in a formula except the first must form a normal backward bond. The effect of the rule is to prohibit representations such as

CH2.(1).CH2.(2).CH2.(3).CH2.C.(1).(2).(3)

for spiropentane; in practice, the rule requires that every group written after the first must be attached to the closest bond still unsatisfied.

$CH_3COCH_2COCH_3$	..1	.11	...	.	1
	.1.	...	...	1	1
	.1.	.1.	...	.	1
	.1.	...	...	1	1
	..1	.11	...	.	1
$Cl.CH_2.CO.OH$	..1	...	...	.	...1.1
	.1.	.1.	...	.	1
	.1.	...	...	1	1
	..1	..1	...	.	11
$CH_2:CH.CDO$	..1	.1.	...	.	1
	.1.	..1	...	.	1
	..1	..1	..1	1	1
$CH_3.CH:N.OH$	..1	.11	...	.	1
	.1.	..1	...	.	1
	.1.	...	...	.	....1.
	..1	..1	...	.	11
$CH_3.CH_2.$	..1	.11	...	.	1
	..1	.1.	...	.	1

Figure 3

THE CANONICAL LINE-FORMULA

It has already been pointed out that a line-formula can usually be written in several distinct forms; even methanol can be written as CH3.OH or as OH.CH3. For purposes of identification, which includes filing, it is clearly useful to define a standard form into which any formula can be rearranged, and which will be unique for each formula. Such a standard representation is called a canonical line-formula.

As the line-formula contains a finite number of group-words, it can only be rearranged in a finite number of ways. By interpreting each computer word of the formula as a digit in the number system with base 2^{16}, each of these arrangements can be assigned a numerical value; and the canonical line-formula is defined as the arrangement with the highest associated numerical value. The rules governing the formation of the line-formulae ensure that every rearrangement is the same length; if a molecule is represented as n groups with m bonds (of any order) between groups, the requirement that each group-word after the first form one normal backward bond accounts for n-l bonds, and the remaining m-n+1 (if any) must be represented by exactly that number of dummy group-words.

A subroutine is provided for performing the rearrangement. It operates by first selecting the numerically largest group-word in the formula; if two or more group-words tie for the largest, all but one are saved for later consideration. A line-formula is then constructed by repeatedly selecting the largest group-word bonded to the most recent unsatisfied bond in the developing line formula, any ties being dealt with as before. Any necessary dummy group-words are inserted as required. After constructing the first line-formula in this way, the routine backtracks to the last saved alternative, if any, and constructs a new line-formula using this alternative. If the new line-formula is numerically greater than the previous one, it is accepted in place of the earlier version, and the process continues until all alternatives have been investigated. At the end of this process, the formula remaining is accepted as the canonical line-formula.

The efficiency of this proccss is enhanced in two ways. The first is obvious : a line-formula under development is abandoned as soon as it turns out to be numerically smaller than the currently accepted version. The second depends on a useful by-product of the subroutine, the symmetry vector. If the molecule contains elements of symmetry, then different arrangements of its groups may give identical line-formulae. By comparing the two formulae, pairs of groups which are symmetrically related to each other can be identified, and the subroutine maintains a record of such relationships. If now, immediately after backtracking, an attempt is made to replace a group with one of the same symmetry, the attempt can be abandoned on the grounds that it must of necessity lead to the same formula as that previously developed.

The symmetry vector is returned to the calling programme along with the canonical line-formula, and can be used in a variety of ways. For example, it shows that, when considering the reactions of acetone, it is only necessary to consider one methyl group, thus effecting economies in processing time.

B6700 WORK

My move from Derby to Auckland parted me from my trusty 1130, and I had to consider how best to continue on the B6700 now at my disposal. The choice was between carrying on with essentially the same system as I had been using at Derby, working in Fortran, or to start again from scratch, hoping to learn something from my earlier experience, and with a free choice of new facilities. I chose to start again, largely because the Fortran programmes which I had constructed had worked well enough when they did work, but had proved extraordinarily hard work to diagnose when anything went wrong, and I hoped that a more highly structured language would give me a better chance of making progress.

I think that was the right decision, but it gave me a lot more work to do – and it really never quite got finished. Certainly the Burroughs' Extended Algol procedures were far easier to understand than their Fortran equivalents, and recursion was sometimes very useful. The main improvements, though, were in the representation of the formula itself. These were of two sorts :

• Within the group words :

The structure of a group word is illustrated by these field definitions :

```
,  ATOMBITS    =  [ 6 : 7 ]#     % ATOMIC NUMBER OF CENTRAL ATOM.
,  HPENDBITS   =  [ HFLD : 3 ]# % PENDANT H ATOMS.
,  OPENDBITS   =  [ OFLD : 3 ]# % PENDANT O ATOMS.
,  NPENDBITS   =  [ NFLD : 3 ]# % PENDANT N ATOMS.
,  FPENDBITS   =  [ FFLD : 3 ]# % PENDANT F ATOMS.
,  FREEBITS    =  [ 21 : 3 ]#    % UNUSED BONDS.
,  SFREEBITS   =  [ 22 : 1 ]#    % SIGN OF FREEBITS ( 1 = -VE ).
,  COORDBITS   =  [ 25 : 3 ]#    % COORDINATION NUMBER OF THE GROUP.
,  ALLATOMBITS =  [ 25 : 26 ]#   % IDENTIFIES THE GROUP.
,  SYMBITS     =  [ 39 : 6 ]#    % SYMMETRY CLASS OF THIS GROUP.
,  FIRSTBITS   =  [ 40 : 1 ]#    % FIRST GROUP OF THIS SYMMETRY CLASS.
,  UNIQUEBITS  =  [ 41 : 1 ]#    % ONLY GROUP OF THIS SYMMETRY CLASS.
,  FLOOPBITS   =  [ 44 : 3 ]#    % NUMBER OF FORWARD LOOPS.
,  BLOOPBITS   =  [ 47 : 3 ]#    % NUMBER OF BACKWARD LOOPS.
,  ALLLOOPBITS =  [ 47 : 6 ]#    % BOTH LOOP COUNTS.
```

This new representation differs from the simpler IBM1130 version in several ways. The main such differences are :

- the central atom is identified by its atomic number, rather than by an arbitrary code;
- the group word is extended to include the numbers of all the common "first row" pendant atoms – nitrogen, oxygen, and fluorine – as well as hydrogen;
- the number of free bonds at a group (possibly negative if the central atom exceeds its normal maximum coordination number, as may be the case in activated complexes or hydrogen bonds) is explicitly specified, whereas it had previously been necessary to work it out from other information. This makes it much easier to identify potentially reactive centres, be they free radical centres or multiple bonds;
- the numbers of forward and backward loops which impinge on the group are explicitly included in the "LOOPBITS" fields;
- the "SYMBITS" and related fields provide information on the symmetry properties of the molecule. The information given corresponds to that in the old symmetry vector. The symmetry properties are byproducts of the canonicalisation operation; the canonical form of the formula is defined as that permutation thereof which comes last in "alphabetical" order, and (in effect, though not in fact) all permutations must be generated and compared to determine this form. A knowledge of the molecular symmetry is of great assistance in

operations such as "find all distinct modes of decomposition of the molecule into two fragments".

- the new representation does not include any means of distinguishing between different isotopes of the pendant atoms. The identification of deuterium atoms possible in the original scheme was very much a special case; no provision was made for identifying any other isotopic variation, and it was not even possible to represent tritium labelling. The whole idea was clumsy, and – as the great majority of gas kinetics research manages very well without bothering about isotopes – it seemed better to forget it unless I could think of a much cleaner way of doing the job.

• The formula as a whole :

Each formula was preceded by a header word, containing summary information. These are the field definitions for that word :

```
,  NATOMBITS   =  [ 23 : 6 ]#   % NUMBER OF ATOMS.
,  NGROUPBITS  =  [ 17 : 6 ]#   % NUMBER OF GROUPS.
,  NLOOPBITS   =  [ 11 : 6 ]#   % NUMBER OF LOOPS.
,  LENGTHBITS  =  [ 5 : 6 ]#    % NUMBER OF WORDS IN THE FORMULA.
```

The representation of loops was separated from the line formula proper, and appended as a suffix : within the line formula, each group was labelled with the number of loops in which it was involved, but the actual links were appended to the formula as a sequence of number pairs (eight bits per number, three pairs to a word), each pair identifying the two groups which were to be joined by a bond not represented within the formula. The representation of spiropentane in this convention would be something like this (where each line represents one formula word) :

5 words of line formula + 2 loop indicators
C with 2 forward loops
CH_2
CH_2 with 1 backward loop
CH_2
CH_2 with one backward loop
a loop joining 1 and 3
a loop joining 1 and 5

(Some bit-by-bit illustrations appear later.)

Some extensions were also made to the external form of the line formula.

- double bonds could be explicitly marked, denoted as ":" (not previously allowed : double bonds originally had to be implied by the coordination numbers);

- free bonds (as at free radical centres) could be denoted as "*" (also previously implied by the coordination numbers);

- loops were denoted by "@" followed by the loop identifier (rather than as a number in brackets).

Double bonds and radical centres were still not explicitly represented in the line formula, of course : these concessions are simply aids to presenting the external formulae in easy and familiar style.

SOME EXAMPLES :

There follow some examples of formulae as they were handled by the B6700 programmes. In each case the items given are :

The input formula;

The line beginning "FORMULA CONTAINS", which summarises the generated header word;

The formula (except for the header word) written out in binary;

The TOPOLOGICAL MATRIX;

The symmetry classification of the group words.

(Notice that not all details of the formulae are correct : the appearance of strings of question marks after some of the text formulae reflects a failure to clear a buffer to blanks before using it; and the symmetry numbers given for spiropentane are wrong.)

```
======================================================================

       CH2,1.CO.CH3

FORMULA CONTAINS 9 ATOMS, 3 GROUPS, 0 LOOPS, AND 3 WORDS.

000 000 0 0 000000 00000000 001 0 001 000 000 000 010 0000110
000 000 0 0 000000 00000000 010 0 000 000 000 001 000 0000110
000 000 0 0 000000 00000000 001 0 000 000 000 000 011 0000110

TOPOLOGICAL MATRIX:

       0      1      0
       1      0      1
       0      1      0

RECONSTRUCTED: CH2,1.CO.CH3|?????

CANONICAL: CO.CH2,1.CH3

FORMULA CONTAINS 9 ATOMS, 3 GROUPS, 0 LOOPS, AND 3 WORDS.

000 000 1 1 000001 00000000 010 0 000 000 000 001 000 0000110
000 000 1 1 000010 00000000 001 0 001 000 000 000 010 0000110
000 000 1 1 000011 00000000 001 0 000 000 000 000 011 0000110

TOPOLOGICAL MATRIX:

       0      1      1
       1      0      0
       1      0      0

                              01     SYMMETRY 01, FIRST, UNIQUE.
                              02     SYMMETRY 02, FIRST, UNIQUE.
                              03     SYMMETRY 03, FIRST, UNIQUE.
```

```
=====================================================================

       CH3.CO.CH2,1

FORMULA CONTAINS 9 ATOMS, 3 GROUPS, 0 LOOPS, AND 3 WORDS.

000 000 0 0 000000 00000000 001 0 000 000 000 000 011 0000110
000 000 0 0 000000 00000000 010 0 000 000 000 001 000 0000110
000 000 0 0 000000 00000000 001 0 001 000 000 000 010 0000110

TOPOLOGICAL MATRIX:

       0      1      0
       1      0      1
       0      1      0

RECONSTRUCTED: CH3.CO.CH2,1|?????

CANONICAL: CO.CH2,1.CH3

FORMULA CONTAINS 9 ATOMS, 3 GROUPS, 0 LOOPS, AND 3 WORDS.

000 000 1 1 000001 00000000 010 0 000 000 000 001 000 0000110
000 000 1 1 000010 00000000 001 0 001 000 000 000 010 0000110
000 000 1 1 000011 00000000 001 0 000 000 000 000 011 0000110

TOPOLOGICAL MATRIX:

       0      1      1
       1      0      0
       1      0      0

                            01     SYMMETRY 01, FIRST, UNIQUE.
                            02     SYMMETRY 02, FIRST, UNIQUE.
                            03     SYMMETRY 03, FIRST, UNIQUE.

=====================================================================

       CO.CH3.CH2,1

FORMULA CONTAINS 9 ATOMS, 3 GROUPS, 0 LOOPS, AND 3 WORDS.

000 000 0 0 000000 00000000 010 0 000 000 000 001 000 0000110
000 000 0 0 000000 00000000 001 0 000 000 000 000 011 0000110
000 000 0 0 000000 00000000 001 0 001 000 000 000 010 0000110

TOPOLOGICAL MATRIX:

       0      1      1
       1      0      0
       1      0      0

RECONSTRUCTED: CO.CH3.CH2,1|?????

CANONICAL: CO.CH2,1.CH3

FORMULA CONTAINS 9 ATOMS, 3 GROUPS, 0 LOOPS, AND 3 WORDS.

000 000 1 1 000001 00000000 010 0 000 000 000 001 000 0000110
000 000 1 1 000010 00000000 001 0 001 000 000 000 010 0000110
000 000 1 1 000011 00000000 001 0 000 000 000 000 011 0000110

TOPOLOGICAL MATRIX:

       0      1      1
       1      0      0
       1      0      0

                            01     SYMMETRY 01, FIRST, UNIQUE.
                            02     SYMMETRY 02, FIRST, UNIQUE.
                            03     SYMMETRY 03, FIRST, UNIQUE.
```

==========

CH2@1.CH2.C0@1

FORMULA CONTAINS 8 ATOMS, 3 GROUPS, 1 LOOPS, AND 4 WORDS.

```
000 001 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
000 000 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
001 000 0 0 000000 00000000 010 0 000 000 000 001 000 0000110
00000001 00000011    000000 000 0 000 000 000 000 000 0000000
```

TOPOLOGICAL MATRIX:

```
0      1      1
1      0      1
1      1      0
```

RECONSTRUCTED: CH2@1.CH2.C0@1|?????

CANONICAL: C0@1.CH2.CH2@1

FORMULA CONTAINS 8 ATOMS, 3 GROUPS, 1 LOOPS, AND 4 WORDS.

```
000 001 1 1 000001 00000000 010 0 000 000 000 001 000 0000110
000 000 0 1 000010 00000000 010 0 000 000 000 000 010 0000110
001 000 0 0 000010 00000000 010 0 000 000 000 000 010 0000110
00000001 00000011    000000 000 0 000 000 000 000 000 0000000
```

TOPOLOGICAL MATRIX:

```
0      1      1
1      0      1
1      1      0
```

```
                        01     SYMMETRY 01, FIRST, UNIQUE.
                        02     SYMMETRY 02, FIRST.
                        03     SYMMETRY 02.
```

==========

CH2@l.CO.CH2@1

FORMULA CONTAINS 8 ATOMS, 3 GROUPS, 1 LOOPS, AND 4 WORDS.

```
000 001 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
000 000 0 0 000000 00000000 010 0 000 000 000 001 000 0000110
001 000 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
00000001 00000011    000000 000 0 000 000 000 000 000 0000000
```

TOPOLOGICAL MATRIX:

```
0      1      1
1      0      1
1      1      0
```

RECONSTRUCTED: CH2@l.CO.CH2@1|?????

CANONICAL: C0@1.CH2.CH2@1

FORMULA CONTAINS 8 ATOMS, 3 GROUPS, 1 LOOPS, AND 4 WORDS.

```
000 001 1 1 000001 00000000 010 0 000 000 000 001 000 0000110
000 000 0 1 000010 00000000 010 0 000 000 000 000 010 0000110
001 000 0 0 000010 00000000 010 0 000 000 000 000 010 0000110
00000001 00000011    000000 000 0 000 000 000 000 000 0000000
```

TOPOLOGICAL MATRIX:

```
0      1      1
1      0      1
1      1      0
```

```
                        01     SYMMETRY 01, FIRST, UNIQUE.
                        02     SYMMETRY 02, FIRST.
                        03     SYMMETRY 02.
```

==========

C@l@2.CH2.CH2@2.CH2.CH2@l

FORMULA CONTAINS 13 ATOMS, 5 GROUPS, 2 LOOPS, AND 6 WORDS.

```
000 010 0 0 000000 00000000 100 0 000 000 000 000 000 0000110
000 000 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
001 000 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
000 000 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
001 000 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
00000001 00000011 00000001 00000101       000 000 000 0000000
```

TOPOLOGICAL MATRIX:

```
0      1      1      1      1
1      0      1      0      0
1      1      0      0      0
1      0      0      0      1
1      0      0      1      0
```

RECONSTRUCTED: C@l@2.CH2.CH2@1.CH2.CH2@2|?????

CANONICAL: C@l@2.CH2.CH2@1.CH2.CH2@2

FORMULA CONTAINS 13 ATOMS, 5 GROUPS, 2 LOOPS, AND 6 WORDS.

```
000 010 1 1 000001 00000000 100 0 000 000 000 000 000 0000110
000 000 1 1 000010 00000000 010 0 000 000 000 000 010 0000110
001 000 1 1 000011 00000000 010 0 000 000 000 000 010 0000110
000 000 0 1 000100 00000000 010 0 000 000 000 000 010 0000110
001 000 0 0 000100 00000000 010 0 000 000 000 000 010 0000110
00000001 00000011 00000001 00000101       000 000 000 0000000
```

TOPOLOGICAL MATRIX:

```
0      1      1      1      1
1      0      1      0      0
1      1      0      0      0
1      0      0      0      1
1      0      0      1      0
```

```
                        01     SYMMETRY 01, FIRST, UNIQUE.
                        02     SYMMETRY 02, FIRST, UNIQUE.
                        03     SYMMETRY 03, FIRST, UNIQUE.
                        04     SYMMETRY 04, FIRST.
                        05     SYMMETRY 04.
```

==========

CH2@1.C@2.CH2@1.CH2.CH2@2

FORMULA CONTAINS 13 ATOMS, 5 GROUPS, 2 LOOPS, AND 6 WORDS.

```
000 001 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
000 001 0 0 000000 00000000 100 0 000 000 000 000 000 0000110
001 000 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
000 000 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
001 000 0 0 000000 00000000 010 0 000 000 000 000 010 0000110
00000001 00000011 00000010 00000101       000 000 000 0000000
```

TOPOLOGICAL MATRIX:

```
0      1      1      0      0
1      0      1      1      1
1      1      0      0      0
0      1      0      0      1
0      1      0      1      0
```

RECONSTRUCTED: CH2@1.C@2.CH2@1.CH2.CH2@2|?????

CANONICAL: C@l@2.CH2.CH2@1.CH2.CH2@2

FORMULA CONTAINS 13 ATOMS, 5 GROUPS, 2 LOOPS, AND 6 WORDS.

```
000 010 1 1 000001 00000000 100 0 000 000 000 000 000 0000110
000 000 1 1 000010 00000000 010 0 000 000 000 000 010 0000110
001 000 1 1 000011 00000000 010 0 000 000 000 000 010 0000110
000 000 0 1 000100 00000000 010 0 000 000 000 000 010 0000110
001 000 0 0 000100 00000000 010 0 000 000 000 000 010 0000110
00000001 00000011 00000001 00000101       000 000 000 0000000
```

TOPOLOGICAL MATRIX:

```
0      1      1      1      1
1      0      1      0      0
1      1      0      0      0
1      0      0      0      1
1      0      0      1      0
```

```
                       01     SYMMETRY 01, FIRST, UNIQUE.
                       02     SYMMETRY 02, FIRST, UNIQUE.
                       03     SYMMETRY 03, FIRST, UNIQUE.
                       04     SYMMETRY 04, FIRST.
                       05     SYMMETRY 04.
```