Simple Stereochemical Structure Code for Organic Chemistry

HELMUT BEIERBECK
Département de Chimie, Université de Montréal, Montréal, Québec, Canada H3C 3V1

Received April 9, 1982

A simple stereocode for the computer representation of molecular structure in organic chemistry is described. For each nonhydrogen atom a line code is generated, which consists of arbitrary numeric indices for that atom and its α substituents. Stereochemistry is defined by the sequence of substituent indices. The matrix of atom codes is converted into a canonical stereochemical description of structure by rearranging and interchanging atom codes according to atomic numbers and connectivities, replacing the chemical indices by ranks, and incorporating the atomic numbers into the code matrix. This matrix is premultiplied by its transpose, and the determinant of the resulting square matrix is an abbreviated stereocode.

An ever increasing number of papers are dealing with the computer storage and retrieval of chemical information and the machine representation of molecular structure, the principal component of any chemical information system. The first structure codes were two-dimensional,[1-4] and topological representations continue to be used and developed.[5,6] There can be no doubt, however, that only a three-dimensional code can fully describe chemical structure and molecular properties, and various stereochemical codes have already been developed.[7-9]

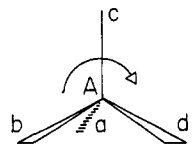
I propose here a three-dimensional structure representation for organic chemistry, which is simple to use and produces a very compact stereocode. For every nonhydrogen atom in the molecule an atom code is generated, which consists of arbitrary chemical indices for that atom and its α substituents. The stereochemical information is contained in the sequence of the substituent indices, encoded by the chemist according to a few simple rules. This starting matrix is converted by machine to a canonical description of molecular structure by rearranging and interchanging the atom codes according to priority criteria based on atomic numbers and connectivities. The chemical indices are then replaced by ranks, and the atomic numbers are added to the matrix to give the full stereocode. This matrix is premultiplied by its transpose, and the determinant of the resulting square matrix is the abbreviated stereocode.

First the coding rules are explained, and then the conversion rules and procedure are given and illustrated with examples.

CODING RULES

The coding of molecular structures requires the identification of each nonhydrogen atom and the definition of its structural surroundings. Atomic numbers are used for identification, since they will be used for ranking the atoms in the conversion routine. The atomic numbers could of course be generated from the elemental symbols. The structural environment of an atom is defined by a line code, which consists of arbitrary numeric indices for that atom and its α substituents. Hydrogens are represented by zeroes. The stereochemistry is defined by the sequence of the substituent indices in the atom code.

The configuration at sp^3 centers is defined by the convention that substituent c in A abc(d) is clockwise from b, looking from



A to a, and substituent d, where it exists, is counterclockwise. Any choice of a and b which conforms to this rule is acceptable. For example, any of these codes might represent C_{13} in morphine enantiomer I (Chart I)

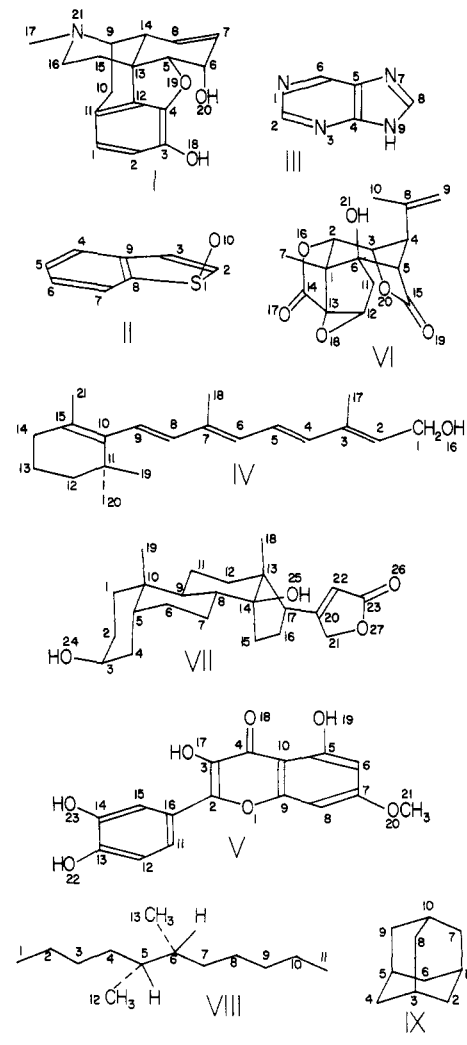
13 15 12 5 14 13 15 5 14 12 13 15 14 12 5
13 12 15 14 5 13 12 14 5 15 13 12 5 15 14
13 5 14 15 12 13 5 15 12 14 13 5 12 14 15
13 14 5 12 15 13 14 12 15 5 13 14 15 5 12

or S_1 in benzothiophene *S*-oxide II

1 2 8 10 1 8 10 2 1 10 2 8

A choice will be made in the conversion routine, after examination of all possible combinations. Where only the relative stereochemistry is required, the choice of enantiomer is immaterial. The second enantiomer code will be generated in the conversion routine, and a decision will be made there. The

Chart I



0095-2338/82/1622-0215$01.25/0

Table I. Input Code for Morphine I

AN	misc	AC
6	sp^2-1	1 2 11 0
6	sp^2-1	2 3 1 0
6	sp^2-1	3 4 2 18
6	sp^2-1	4 19 12 3
6		5 6 19 13 0
6		6 7 20 5 0
6	sp^2-2	7 8 6 0
6	sp^2-2	8 14 7 0
6		9 21 10 14 0
6		10 11 9 0 0
6	sp^2-1	11 1 12 10
6	sp^2-1	12 11 4 13
6		13 15 12 5 14
6		14 0 13 8 9
6		15 13 16 0 0
6		16 15 21 0 0
6		17 21 0 0 0
8		18 3 0
8		19 4 5
8		20 6 0
7	inv	21 9 16 17

choice of configuration is arbitrary for sp^3 centers with configuration inversion. In this case a representative code will be chosen from the complete set of codes for both configurations. The recognition of centers with uncertain configuration could in principle be incorporated into the conversion algorithm, since they only occur for certain combinations of atomic numbers and code lengths. For the moment, however, these cases have to be identified in the conversion routine (e.g., N_{21} of morphine in Table I).

The stereochemistry of trigonal systems is encoded by the convention that the substituents of all conjugated sp^2 centers are read in the same sense, clockwise or counterclockwise, from a given topological map of the molecule. The ambiguity arises from the fact that the molecule may be viewed from either of two sides. The problem is resolved in the conversion routine, where the other series of codes is generated and an indepenent decision is made for each conjugated system. Therefore the simpler rule, that the substituents at all sp^2 centers are read clockwise from any topological map, serves the same purpose. For example, N_1 and C_2 in purine representation III would be coded as

N_1	1 6 2
C_2	2 1 3 0 or 2 3 0 1 or 2 0 1 3

The choice of topological representation is not completely arbitrary, however, since cis and trans 1,3-dienes give different codes. An additional rule is therefore introduced which stipulates that the coding of sp^2 stereochemistry must be based on a topological structure in which all conjugated double bonds are trans, except endocyclic ones. Where this rule makes no sense, conjugation is considered cut. For example, all five double bonds in vitamin A (IV) are trans for coding purposes, and all ten sp^2 carbons belong to one conjugated system. The orientation of the phenyl group in rhamnetin (V), on the other hand, cannot be established by the trans rule, and conjugation is considered cut between C_2 and C_{16}. All sp^2 centers must presently be assigned to their respective systems of unsaturation (e.g., Table I). Eventually this attribution and the implementation of the conformational trans rule could be incorporated into the conversion algorithm.

The coding rules may be summarized as follows.

Every nonhydrogen atom is given an arbitrary numeric index. Hydrogens are represented by zeroes.

The choice of enantiomer for the representation of relative stereochemistry is arbitrary. The coding of sp^2 stereochemistry is based on a topological structure representation, in which all conjugated double bonds are trans, except endocyclic ones.

Each nonhydrogen atom is characterized by its atomic number and atom code, which consists of the indices for that atom and its α substituents. Centers with configuration inversion are identified, and sp^2 centers are assigned to their respective conjugated systems.

The sp^3 stereochemistry is encoded by the convention that substituent c in Aabc(d) is clockwise from b, and d, where it exists, is counterclockwise. sp^2 stereochemistry is encoded by the convention that the substituents at all trigonal centers are read clockwise from the topological representation of the molecule.

A sample input code is shown in Table I.

CONVERSION RULES

The morphine code in Table I is only one of many possibilities. In order to arrive at a canonical description of molecular structure it is necessary to bring atom codes into standard form, rank them according to some measure of priority, choose one enantiomer to represent relative stereochemistry, and standardize atomic numbering. In addition, unique representations for sp^3 centers with configuration inversion and for sp^2 centers have to be found. This conversion is carried out by machine, according to the following rules.

Atom codes are rearranged in such a way that the substitution index (v.i.) is a maximum. If different atom codes have identical substitution indices, the code with the lowest rank index (v.i.) is retained. Atom codes for sp^3 centers with configuration inversion are chosen from the complete set of codes for both configurations.

Atoms are ranked in the order of increasing atomic numbers. Atoms with identical atomic numbers are ranked in the order of decreasing substitution indices. If the substitution indices are identical, atoms are ranked in the order of increasing rank indices.

That series of sp^2 codes is retained which has the higher of the first unequal pair of substitution indices. If the substitution indices for the two series are identical, the series with the lower of the first unequal pair of rank indices is chosen.

That enantiomer represents relative stereochemistry, which has the higher of the first unequal pair of substitution indices. If the two matrices of substitution indices are identical, the enantiomer with the lower of the first unequal pair of rank indices is taken.

The atomic indices are replaced by the atomic ranks.

The substitution index (SI) of an atom code is composed of the connectivities of the atoms in that code. For example, atom code 13 14 5 12 15 for C_{13} of morphine has substitution index 43332, since C_{13} is quaternary, C_{14}, C_5, and C_{12} are tertiary, and C_{15} is secondary.

The rank index (RI) of an atom code consists of the ranks of the atoms in that code, two digits per rank (it is assumed that the number of nonhydrogen atoms does not exceed 99). For example, if the ranks of carbons 13, 14, 5, 12, and 15 of morphine are 1, 2, 3, 10, and 6, respectively, then the rank index for the same C_{13} code will be 0102031006. The ranks, and consequently the rank indices, change in the course of the conversion process. The rank indices are used to minimize the value of the final rank matrix, read row by row, within the limits imposed by the atomic numbers and substitution indices.

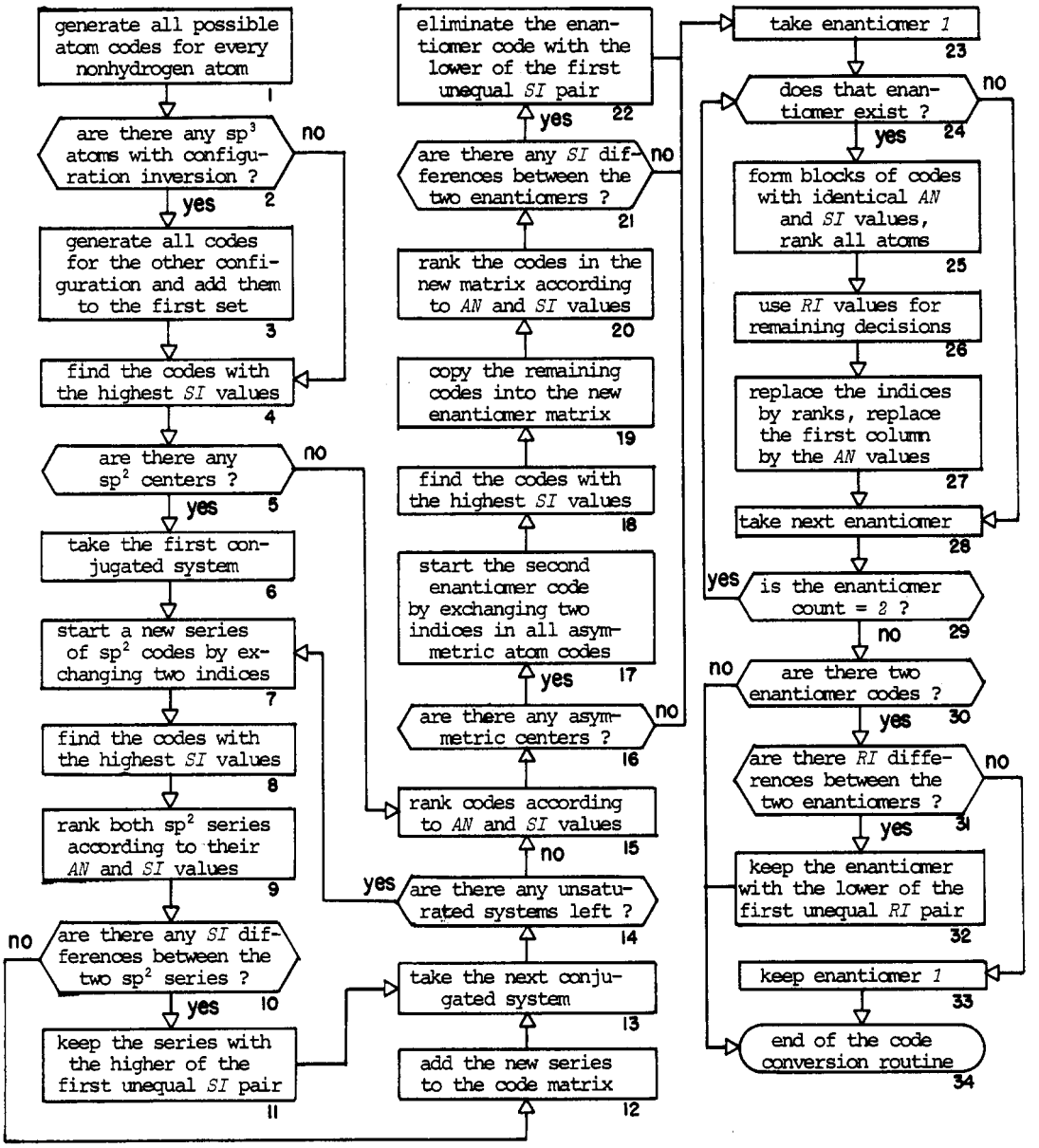
An algorithm for the implementation of these conversion rules, programmed and tested in Fortran, is described and illustrated in the next section.

CONVERSION ALGORITHM

The transformation of the input code to a canonical description of molecular structure is carried out in the sequence of steps shown in Chart II.

First the starting codes are brought into standard form by

Chart II. Code Conversion Algorithm



maximizing the substitution indices (steps 1–4 in Chart II). For example, the highest substitution index for C_{13} of morphine is 43332, shared by three different codes, and all three are retained at this stage.

SI	AC
42333	13 15 12 5 14
42333	13 15 5 14 12
42333	13 15 14 12 5
43233	13 12 15 14 5
43332*	13 12 14 5 15
43323	13 12 5 15 14
43323	13 5 14 15 12
43233	13 5 15 12 14
43332*	13 5 12 14 15
43332*	13 14 5 12 15
43323	13 14 12 15 5
43233	13 14 15 5 12

For N_{21} of morphine configuration inversion is indicated, and the codes for both configurations are generated.

configuration a		configuration b	
SI	AC	SI	AC
3321*	21 9 16 17	3312	21 9 17 16
3213	21 16 17 9	3123	21 17 16 9
3132	21 17 9 16	3231	21 16 9 17

Here the input code is found to have the highest SI value, 3321.

After the standardization of the atom codes a second series of sp² codes is generated for each unsaturated system, both series are ranked and examined for SI differences (steps 5–14). If a difference is found, as for the benzene ring in morphine

series a		series b	
SI	AC	SI	AC
3433	12 13 11 4	3433	12 13 4 11
3332	4 12 3 19	3332	4 3 12 19
3322	11 12 10 1	3322	11 12 1 10
3321*	3 4 2 18	3312	3 4 18 2
2320	2 3 1 0	2320	1 11 2 0
2302	1 11 0 2	2302	2 3 0 1

the series with the lower of the first unequal SI pair, series b in this example, is eliminated. If there are no SI differences, as in the case of C_7 and C_8 of morphine

series a		series b	
SI	AC	SI	AC
2320	8 14 7 0	2320	7 6 8 0
2302	7 6 0 8	2302	8 14 0 7

the new series of codes is added to the matrix.

Next all codes are ranked according to atomic numbers and substitution indices (step 15). At this stage the morphine code

Table II. Morphine Code Prior to Minimization of the Rank Matrix

AN	sp²	SI	AC
6		43332	13 12 14 5 15 13 5 12 14 15 13 14 5 12 15
6		34320	5 13 6 19 0
6		34320	14 13 9 8 0
6		33320	9 14 21 10 0
6		33210	6 5 7 20 0
6		24200	15 13 16 0 0
6		23300	10 11 9 0 0 10 9 11 0 0
6		23200	16 21 15 0 0
6		13000	17 21 0 0 0
6	1-1	3433	12 13 11 4
6	1-1	3332	4 12 3 19
6	1-1	3322	11 12 10 1
6	1-1	3321	3 4 2 18
6	1-1	2320	2 3 1 0
6	2-2	2320	7 6 8 0
6	2-1	2320	8 14 7 0
6	1-1	2302	1 11 0 2
6	2-1	2302	7 6 0 8
6	2-2	2302	8 14 0 7
7		3321	21 9 16 17
8		233	19 4 5 19 4 5
8		130	18 3 0
8		130	20 6 0

Table III. Structure Code for Picrotoxinin VI

AN	sp²	SI	AC
6		44431	1 6 13 2 7
6		44332	13 1 12 14 18
6		44321	6 1 5 11 21
6		34330	5 6 4 15 0
6		34302	2 1 3 0 16
6		34220	12 13 18 11 0
6		33330	4 3 5 8 0 4 5 8 3 0 4 8 3 5 0
6		33320	3 4 2 20 0
6		24300	11 6 12 0 0
6		14000	7 1 0 0 0
6		13000	10 8 0 0 0
6	2 2	3421	14 13 16 17
6	3-2	3321	15 5 20 19
6	1-1	3311	8 4 10 9
6	1 2	3311	8 4 9 10
6	1-1	1300	9 8 0 0
6	1-2	1300	9 8 0 0
8		243	18 13 12
8		233	16 2 14 16 14 2
8		233	20 3 15 20 15 3
8		140	21 6 0
8		13	17 14
8		13	19 15

of Table I has become the matrix in Table II.

Finally, a second enantiomer code is generated and ranked, and the two series are compared for SI differences (steps 16–22). If a difference is found, as in the case of morphine

enantiomer a		enantiomer b	
SI	AC	SI	AC
43332	13 12 14 5 15	43332	13 12 5 14 15
	13 5 12 14 15		13 5 14 12 15
	13 14 5 12 15		13 14 12 5 15
34320*	14 13 9 8 0	34302	14 13 9 0 8
...	...	...	...

the enantiomer code with the lower of the first unequal SI pair is eliminated, enantiomer b in this case. Where no decision is possible at this time, both enantiomer codes are retained for the moment.

No further differentiation on the basis of atomic numbers and substitution indices is possible, and the minimization of the rank matrix begins (steps 23–33). The codes are divided into blocks with identical atomic numbers and substitution indices, and the atoms are ranked at the bottom of their respective blocks, or first of two blocks. For instance, C_{13} of morphine has rank 1, C_5 and C_{14} both have rank 3, and C_2, C_7, and C_8 all have rank 16. Furthermore, the number of atoms with known code and rank is determined. In the case of morphine there are ten such atoms: 9, 6, 15, 16, 17, 12, 4, 11, 3, and 21. This parameter serves as the counter in the minimization routine.

The minimization of the final rank matrix, subject to the constraints imposed by the atomic numbers and substituent indices, is carried out in the following way. The codes are examined row by row. If the correct entry for row i is known, the substituents of that code are brought to the top or next available row in their respective blocks. If the correct code for row i is unknown, the rank indices are derived for all unranked codes in that block, and the code with the lowest rank index, i.e., highest-ranked substituents, is placed in row i. If there is no minimum, the rank indices for the remaining codes are derived, and the ranks are updated. This cycle is repeated until a code for row i is found. If the entry in row i is one of two codes for a trigonal center, the codes for that system can at that point be chosen as well.

The details of the minimization routine are shown in Chart III, where the parameters have the following meaning:

n_a	number of nonhydrogen atoms in the molecule
n_c	number of atom codes; $n_c = n_a$ when all sp² codes are assigned
n_r	number of atoms for which code and rank are established
n_x	number of rank updates in one passage through the code selection routine
AC_{ikl}	index l of atom code k in row i
B_{1j}, B_{2j}	blocks containing codes for atom j; $B_{2j} = 0$ if j only has one code
U_j	unsaturated system containing atom j
M_i	atom code in row i, for which the rank index is a minimum; initially $M_i = 1$ or 0, depending on whether or not the substitution index uniquely defines a code for atom j
R_j	rank of atom j; initially set to the end of the first block containing a code for atom j
A_j	2 when rank and code for atom j are known, 1 when only the rank is known, 0 when atom j is unranked
RI_i	rank index for the code in row i
N_m	next available rank in block m; initially set to the block start
S_n	indicates the series, 1 or 2, chosen to represent unsaturated system n; $S_n = 0$ until a decision is reached

The parameters M_I, AC_{IKL}, and RI_I have row subscripts I rather than atom subscripts J, since their values differ for the two possible codes of sp² centers. They may have to be reassigned to other rows, as rank changes occur within blocks. Consider, for example, the structure code for picrotoxinin (VI) in Table III. The codes and ranks for 14 atoms are known, and double bond C_8–C_9 is represented by two series of codes. The counters thus have the following values

$$n_r = 14$$

$$n_c = n_a + 2 = 23$$

The C_1 code in row 1 is unique

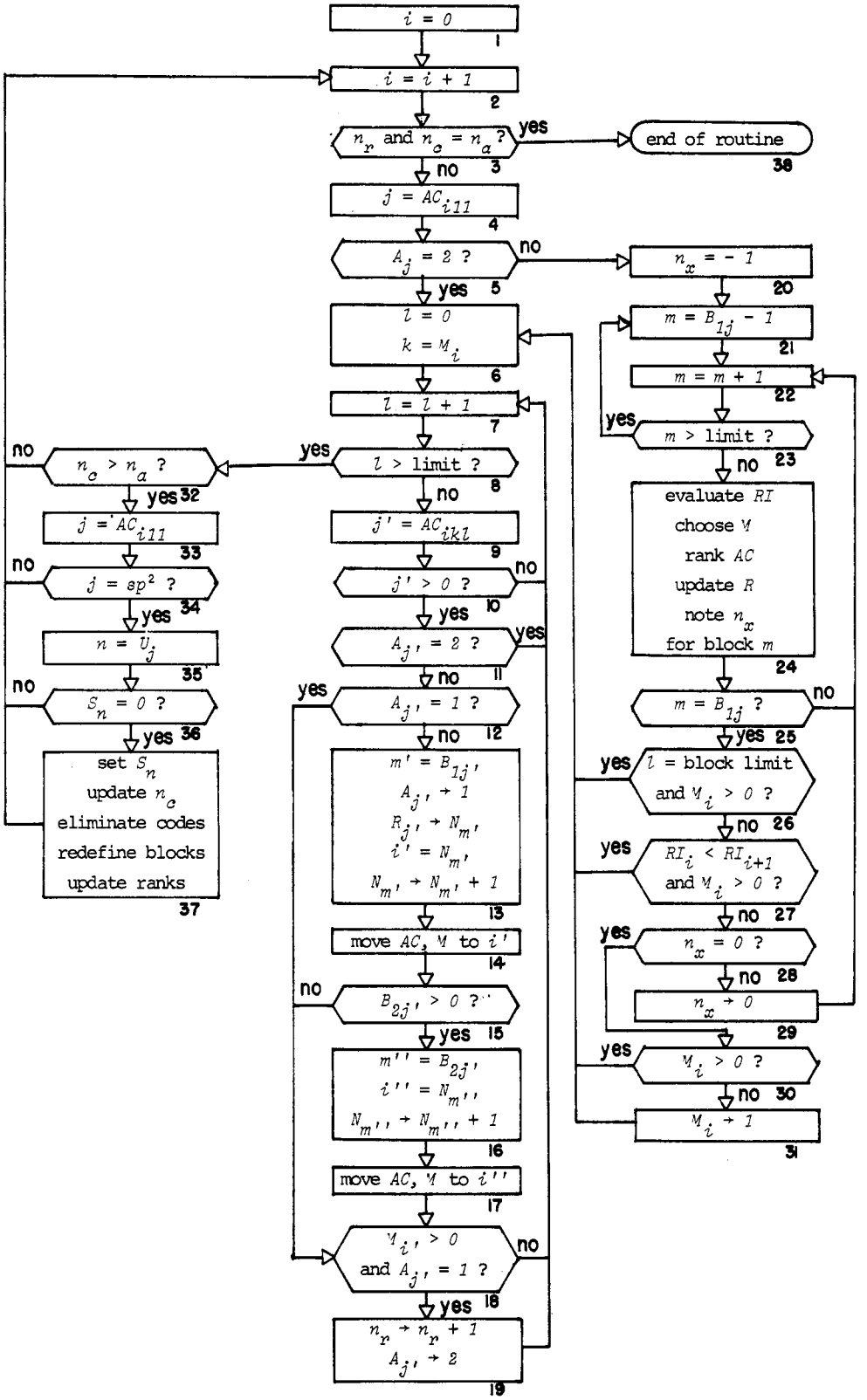
$$i = 1$$

$$j' = AC_{111} = 1$$

$$A_1 = 2$$

and the substituent ranking routine is entered (step 5 → 6). The reason for starting with $I = 1$, rather than the first sub-

Chart III. Step 26 of the Code Conversion Algorithm



stituent $I = 2$, will become apparent shortly. All C_1 substituents are already ranked, and it is only in row 5 that the first unranked substituent is encountered

$$i = 5$$
$$k = M_5 = 1$$
$$l = 5$$
$$j' = AC_{515} = 16$$
$$A_{16} = 0$$

The assign flag is set, C_{16} is given the first available rank in its block, and that parameter is incremented

$$A_{16} \rightarrow 1$$
$$m' = B_{116} = 17$$
$$R_{16} \rightarrow N_{17} = 19$$
$$N_{17} \rightarrow N_{17} + 1 = 20$$
$$B_{216} = 0$$

Table IV. Structure Code for Digitoxigenin VII

AN	sp²	SI	AC
6		44321	13 14 17 12 18
6		44321	14 13 8 15 25
6		43321	10 5 9 1 19
6		34320	8 14 9 7 0
6		34320	17 13 20 16 0
6		34302	9 10 8 0 11
6		34220	5 10 4 6 0
6		32210	3 2 4 24 0
6		24200	1 10 2 0 0
6		24200	11 10 12 0 0
6		24200	12 13 11 0 0
6		24200	15 14 16 0 0
6		23300	4 5 3 0 0 4 3 5 0 0
6		23200	2 3 1 0 0
6		23200	6 5 7 0 0
6		23200	7 8 6 0 0
6		23200	16 17 15 0 0
6		23200	21 20 27 0 0
6		14000	18 13 0 0 0
6		14000	19 10 0 0 0
6	1-1	3322	20 17 22 21
6	1-2	3322	20 17 21 22
6	1-1	3221	23 27 22 26
6	1-2	3221	23 22 27 26
6	1-1	2330	22 23 20 0
6	1-2	2330	22 20 23 0
8		232	27 23 21
8		140	25 14 0
8		130	24 3 0
8		13	26 23

Since the atom code for C_{16} in row 19 is known, the rank counter is incremented

$$M_{19} = 1$$

$$n_r \rightarrow n_r + 1 = 15$$

$$A_{16} \rightarrow 2$$

The code in row 1 of the morphine matrix in Table II is ranked but is degenerate ($M_1 = 0$), and the code selection routine is entered (step 5 → 20). The rank indices are evaluated, and code $k = 3$ is found to minimize the RI value

k	AC	RI
1	13 12 14 5 15	110030306
2	13 5 12 14 15	103100306
3	13 14 5 12 15	103031006*

Since

$$i = 1 = \text{block limit}$$

$$M_1 = 3$$

control returns to the substituent ranking routine (step 26 → 6).

The code in row 1 of the matrix for digitoxigenin (VII) in Table IV is unranked, since block 1 contains two rows, and again the code selection routine is entered (step 5 → 20). C_{13} is found to have the lower rank index

SI	AC	RI
44321	13 14 17 12 18	202051220*
44321	14 13 8 15 25	202051228

and since

$$RI_1 < RI_2$$

$$M_1 = 1$$

control returns to the substituent ranking routine (step 27 → 6). This example also shows why the first atom in every code has to be reexamined

$$i = 1$$

$$l = 1$$

$$j' = AC_{111} = 13$$

$$A_{13} \rightarrow 1$$

$$R_{13} \rightarrow N_1 = 1$$

$$N_1 \rightarrow N_1 + 1 = 2$$

$$l = 2$$

$$j' = AC_{112} = 14$$

$$A_{14} \rightarrow 1$$

$$R_{14} \rightarrow N_1 = 2, \text{ etc}$$

Had the first substituent, C_{14}, been taken right away, step $N_1 \rightarrow N_1 + 1 = 2$ would have been missed, with the result that $R_{13} = R_{14} = 1$.

Table V. Structure Code for 5(*S*),6(*S*)-Dimethylundecane VIII

AN	SI	AC
6	33210	5 6 4 12 0
6	33210	6 5 7 13 0
6	23200	4 5 3 0 0
6	23200	7 6 8 0 0
6	22200	3 2 4 0 0 3 4 2 0 0
6	22200	8 7 9 0 0 8 9 7 0 0
6	22200	9 8 10 0 0 9 10 8 0 0
6	22100	2 3 1 0 0
6	22100	10 9 11 0 0
6	13000	12 5 0 0 0
6	13000	13 6 0 0 0
6	12000	1 2 0 0 0
6	12000	11 10 0 0 0

The atom code in row 1 of the matrix for 5(*S*),6(*S*)-dimethylundecane (VIII) in Table V is also unranked, but this time an evaluation of the rank indices for block 1 alone does not lead to a differentiation

SI	AC	RI	j	R_j
33210	5 6 4 12 0	202041100	5	2
33210	6 5 7 13 0	202041100	6	2
...		...		

Since this is the first passage through the code selection routine ($n_x = -1$), rank indices are derived for the remaining blocks and then again for block 1, and ranks are updated where possible. There is still no difference between RI_1 and RI_2, but there are RI differences, and therefore rank changes, in block 3 ($n_x > 0$)

SI	AC	RI	j	R_j
33210	5 6 4 12 0	202041100	5	2
33210	6 5 7 13 0	202041100	6	2
23200	4 5 3 0 0	402070000	4	4
23200	7 6 8 0 0	402070000	7	4
22200	8 7 9 0 0	704070000	8	5
22200	3 4 2 0 0	704090000	3	6
22200	9 8 10 0 0	707090000	9	7
...		...		

The cycle is repeated, and this time there is an RI difference in block 1

SI	AC	RI	j	R_j
33210	6 5 7 13 0	202031100	6	1
33210	5 6 4 12 0	202041100	5	2
23200	7 6 8 0 0	402050000	7	3
23200	4 5 3 0 0	402060000	4	4
22200	8 7 9 0 0	504070000	8	5
22200	3 4 2 0 0	604080000	3	6
22200	9 8 10 0 0	705090000	9	7
...		...		

Table VI. Structure Code for Adamantane IX

AN	SI	AC
6	32220	1 2 7 6 0 1 7 6 2 0 1 6 2 7 0
6	32220	3 2 4 8 0 3 4 8 2 0 3 8 2 4 0
6	32220	5 6 9 4 0 5 9 4 6 0 5 4 6 9 0
6	32220	10 7 8 9 0 10 8 9 7 0 10 9 7 8 0
6	23300	2 3 1 0 0 2 1 3 0 0
6	23300	4 5 3 0 0 4 3 5 0 0
6	23300	6 1 5 0 0 6 5 1 0 0
6	23300	7 10 1 0 0 7 1 10 0 0
6	23300	8 10 3 0 0 8 3 10 0 0
6	23300	9 10 5 0 0 9 5 10 0 0

Table VII. Complete Stereocode for Morphine

AN	ranked code matrix	rank matrix	stereocode
6	13 14 5 12 15	1 2 3 10 6	6 2 3 10 6
6	14 13 9 8 0	2 1 4 14 0	6 1 4 14 0
6	5 13 6 19 0	3 1 5 19 0	6 1 5 19 0
6	9 14 21 10 0	4 2 18 7 0	6 2 18 7 0
6	6 5 7 20 0	5 3 16 20 0	6 3 16 20 0
6	15 13 16 0 0	6 1 8 0 0	6 1 8 0 0
6	10 9 11 0 0	7 4 12 0 0	6 4 12 0 0
6	16 21 15 0 0	8 18 6 0 0	6 18 6 0 0
6	17 21 0 0 0	9 18 0 0 0	6 18 0 0 0
6	12 13 11 4	10 1 12 11	6 1 12 11
6	4 12 3 19	11 10 13 19	6 10 13 19
6	11 12 10 1	12 10 7 17	6 10 7 17
6	3 4 2 18	13 11 15 21	6 11 15 21
6	8 14 7 0	14 2 16 0	6 2 16 0
6	2 3 1 0	15 13 17 0	6 13 17 0
6	7 6 0 8	16 5 0 14	6 5 0 14
6	1 11 0 2	17 12 0 15	6 12 0 15
7	21 9 16 17	18 4 8 9	7 4 8 9
8	19 5 4	19 3 11	8 3 11
8	20 6 0	20 5 0	8 5 0
8	18 3 0	21 13 0	8 13 0

(the rank indices are composed of the ranks of the previous cycle). Now

$$RI_1 < RI_2$$

$$M_1 = 1$$

and control returns to the substituent ranking routine (step 27 → 6).

The entry in row 1 of the code for adamantane (IX) in Table VI is also unranked, but because of the symmetry properties of the molecule, cycling through the code matrix cannot produce any changes in rank. All rank indices are either 410101000 (tertiary carbons) or 1004040000 (secondary carbons) and

$$n_x = 0$$

$$M_1 = 0$$

Now M_1 is arbitrarily set to 1, and control returns to the main routine (step 31 → 6). The code which happens to be in row 1 is automatically ranked first:

$$i = 1$$

$$k = M_1 = 1$$

$$l = 1$$

$$j' = AC_{111} = 1$$

$$A_1 \rightarrow 1$$

$$R_1 \rightarrow N_1 = 1$$

Any other code in block 1 could have been ranked first, without changing the final result.

The first sp^2 code in the morphine matrix is found in row 10, and the sp^2 code selection routine is entered (step 8 → 32).

$$i = 10$$

$$n_c = n_a + 2$$

$$j = AC_{1011} = 12$$

$$C_{12} = sp^2$$

$$n = U_{12} = 1$$

$$S_1 = 1$$

C_{12} is part of the benzene ring, for which codes have already been chosen, and no further action is taken (step 36 → 2). The first code of double bond C_7–C_8 is encountered in row 14. At that point the matrix has the following appearance:

AN	sp^2	SI	AC	j	R_j
	...		...		
6	1-1	3321	3 4 2 18	3	13
6	2-1	2320	8 14 7 0	8	14
6	2-2	2320	7 6 8 0	7	15
6	1-1	2320	2 3 1 0	2	16
6	2-2	2302	8 14 0 7		
6	2-1	2302	7 6 0 8		
6	1-1	2302	1 11 0 2	1	19
7		3321	21 9 16 17	21	20
8		233	19 5 4	19	21
8		130	20 6 0	20	22
8		130	18 3 0	18	23

The representation for C_7–C_8 is still undecided at this point:

$$i = 14$$

$$n_r = 21$$

$$n_c = n_a + 2 = 23$$

$$j = AC_{1411} = 8$$

$$C_8 = sp^2$$

$$n = U_8 = 2$$

$$S_2 = 0$$

Since the code in row 14 belongs to series 2-1, series 2-2 is eliminated

$$S_2 \rightarrow 1$$

$$n_c \rightarrow n_c - 2 = n_a$$

the gaps are closed, and the block limits and ranks are redefined:

AN	sp^2	SI	AC	j	R_j
	...		...		
6	1-1	3321	3 4 2 18	3	13
6	2-1	2320	8 14 7 0	8	14
6	1-1	2320	2 3 1 0	2	15
6	2-1	2302	7 6 0 8	7	16
6	1-1	2302	1 11 0 2	1	17
7		3321	21 9 16 17	21	18
8		233	19 5 4	19	19
8		130	20 6 0	20	20
8		130	18 3 0	18	21

Since $n_r = n_c = n_a$, the transformation of the morphine matrix is complete (Table VII).

If there are still two enantiomer codes at this stage, the one with the higher of the first unequal RI pair is eliminated. For example 5(*S*),6(*R*)- and 5(*R*),6(*S*)-dimethylundecane have identical substitution indices (Table VIII). The rank indices of the first four rows are identical as well. It is the ranks of

Table VIII. Code Matrices for 5(*S*),6(*R*)- and 5(*R*),6(*S*)-Dimethylundecane

SI	5(*S*),6(*R*)	5(*R*),6(*S*)
33210	6 5 7 13 0	5 6 4 12 0
33201	5 6 4 0 12	6 5 7 0 13
23200	7 6 8 0 0	4 5 3 0 0
23200	4 5 3 0 0	7 6 8 0 0
22200	8 7 9 0 0	3 4 2 0 0
22200	3 4 2 0 0	8 7 9 0 0
22200	9 8 10 0 0	9 8 10 0 0
22100	2 3 1 0 0	2 3 1 0 0
22100	10 9 11 0 0	10 9 11 0 0
13000	13 6 0 0 0	12 5 0 0 0
13000	12 5 0 0 0	13 6 0 0 0
12000	1 2 0 0 0	1 2 0 0 0
12000	11 10 0 0 0	11 10 0 0 0

the second substituents of the fifth-ranked carbons which decide in favor of the 5(*S*),6(*R*) enantiomer:

	5(*S*),6(*R*)			5(*R*),6(*S*)	
j	R_j	RI	*j*	R_j	RI
6	1	102031000	5	1	102031000
5	2	201040011	6	2	201040011
7	3	301050000	4	3	301050000
4	4	402060000	7	4	402060000
8	5	503070000*	3	5	503080000
3	6	604080000	8	6	604070000
9	7	705090000	9	7	706090000
2	8	806120000	2	8	805120000
		...			...

With the choice of enantiomer the conversion process is complete.

FULL AND ABBREVIATED STEREOCODES

The converted matrix is a unique stereochemical record of molecular structure, but it does not contain the elemental composition. The complete stereocode is derived from this matrix by replacing the first column, which only consists of the ordered sequence 1, 2, ..., *n*, by the atomic numbers. The full stereocode for morphine is shown in Table VII.

An abbreviated stereocode is generated in the following way. The full stereocode is zero filled to a rectangular matrix. Since zeroes represent hydrogens, and this information should be preserved, the indices are first augmented, i.e., hydrogens take label 1, atom 1 becomes atom 2, etc. The atomic numbers remain unchanged, e.g.:

```
...                ...
6 18  6  0 0  →  6 19  7  1 1
6 18  0  0 0  →  6 19  1  1 1
6  1 12 11    →  6  2 13 12 0
6 10 13 19    →  6 11 14 20 0
...                ...
```

The zero-filled matrix is then premultiplied by its transpose. The determinant of the nonsingular matrix of highest rank contained in the resulting square matrix is the abbreviated stereocode. The absolute configuration is added to the abbreviated stereocode in the form of the subscripts or fractions 0.1 and 0.2, where 0.1 refers to the enantiomer representing relative stereochemistry. The abbreviated stereocodes for substrates I and VI–IX are given in Table IX.

Table IX. Abbreviated Stereocodes for Structures I and VI to IX

no.	compd	stereocode
I	morphine	33 644 803 832 407.1
VI	picrotoxinin	154 162 990 804 468.1
VII	digitoxigenin	5 423 562 705 300 408.2
VIII	5(*S*),6(*S*) dimethylundecane	1 686 388 248.2
IX	adamantane	2 366 208.2

Coding these molecular structures required only a few simple rules, and even some of these rules can eventually be eliminated by making the appropriate modifications to the conversion algorithm. Thus sp^3 centers with configuration inversion may be recognized by their atomic numbers and code lengths, and the assignment of sp^2 codes to conjugated systems and the generation of topological maps with all-trans conjugated chains could be incorporated into the conversion routine as well. The remaining rules can be summarized in a few words. For each nonhydrogen atom the chemist must supply the atomic number (or elemental symbol) and atom code. The code consists of arbitrary numeric indices for that atom and its α substituents, where hydrogens are represented by zeroes. The configuration at sp^3 centers is encoded by the convention that substituent c in Aabc(d) is clockwise from b, looking from A to a. The stereochemistry of sp^2 systems is defined by coding the substituents of sp^2 centers in a clockwise manner.

The coding of inorganic chemical structures may be approached in the same way. It is merely necessary to define appropriate coding rules and make the necessary modifications in the conversion algorithm.

ACKNOWLEDGMENT

I thank the Université de Montréal for the use of their computing facilities.

REFERENCES AND NOTES

(1) Ray, L. C.; Kirsch, R. A. "Finding Chemical Records by Digital Computers". *Science* **1957**, *126*, 814.
(2) Gluck, D. J. "A Chemical Structure Storage and Search System Developed at DuPont". *J. Chem. Doc.* **1965**, *5*, 43.
(3) Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structure—A Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, *5*, 107.
(4) Smith, E. G.; Baker, P. A. "The Wiswesser Line-Formula Chemical Notation", 3rd ed.; Chemical Information Management Inc.: Cherry Hill, NJ, 1975.
(5) Dubois, J. E. "Darc System in Chemistry" In: Wipke, W. T.; Heller, S.; Feldmann, R.; Hyde, E., Eds. "Computer Representation and Manipulation of Chemical Information". Wiley: New York, 1974.
(6) Bremser, W. "Hose—a Novel Substructure Code". *Anal. Chim. Acta* **1978**, *103*, 355.
(7) Petrarca, A. E.; Lynch, M. F.; Rush, J. E. "A Method for Generating Unique Computer Structural Representations of Stereoisomers". *J. Chem. Doc.* **1967**, *7*, 154.
(8) Wipke, W. T.; Dyott, T. M. "Simulation and Evaluation of Chemical Synthesis. Computer Representation and Manipulation of Stereochemistry". *J. Am. Chem. Soc.* **1974**, *96*, 4825.
(9) Gray, N. A. B.; Nourse, J. G.; Crandell, C. W.; Smith, D. H.; Djerassi, C. "Stereochemical Substructure Codes for ^{13}C Spectral Analysis". *Org. Magn. Reson.* **1981**, *15*, 375.