Interim Report to the
National Aeronautics and Space Administration
Grant NsG 81-60

DENDRAL-64

A SYSTEM FOR COMPUTER CONSTRUCTION, ENUMERATION AND NOTATION OF ORGANIC MOLECULES AS TREE STRUCTURES AND CYCLIC GRAPHS

Part I.	Notational Algorithm for Tree Structures
II.	Topology of Cyclic Graphs
III.	Notational Algorithm for Chemical Graphs
IV.	Generator Algorithms
V.	Directions for Further Analysis

submitted by

Joshua Lederberg
Professor of Genetics
School of Medicine
Stanford University
Palo Alto, California

Studies related to this report have been supported by research grants from the National Aeronautics and Space Administration (NsG 81-60), National Science Foundation (NSF G-6411), and National Institutes of Health (NB-04270, AI-5160 and FR-00151).

PART I. December 15, 1964

Tables Referred to in Part I

1. DENDRAL Primer.
2. Canons of DENDRAL Valuation.
3. Character Codes for Electronic Computation.
4. Notational Abbreviations.
5. Some Examples of DENDRAL Codes: Responses to NAS Test List.[1]

References Cited in Part I

1. Survey of Chemical Notation Systems, National Academy of Sciences, National Research Council Publication 1150 (1964), 467 pp.
2. H. R. Henze and C. M. Blair, The Number of Isomeric Hydrocarbons of the Methane Series, J. Am. Chem. Soc. 53: 3077-3085 (1931).
3. Polish notation is a device to avoid the use of nested parentheses in algebraic expressions and other tree structures. It depends on an inference (a) of the valence of any operator and (b) the syntax of a complete operand. It is the basis of most algebraic compilers for the translation of algebraic expressions into a series of computer instructions. The advantages of Polish notation for chemical structures (along somewhat different lines than here), has also been illustrated by S. H. Eisman, A Polish-Type Notation for Chemical Structures, J. Chem. Doc. 4:186-190 (1964), and H. Hiz, A Linearization of Chemical Graphs, J. Chem. Doc. 4:173-180 (1964).
4. The discussion of stereoisomerism very closely follows the remarkably clear exposition by E. L. Eliel, Stereochemistry of Carbon Compounds McGraw-Hill Book Company, Inc., New York (1962).

DENDRAL-64

A SYSTEM FOR COMPUTER CONSTRUCTION, ENUMERATION AND NOTATION OF ORGANIC MOLECULES AS TREE STRUCTURES AND CYCLIC GRAPHS

FOREWORD

DENDRAL-64 is a preliminary version of a proposed system of topological ordering of organic molecules as tree structures, hence dendritic algorithm. In computer applications to analytical work in biochemistry, a system was needed for scanning hypothetical structures to be matched against experimental data and prior constraints. DENDRAL also proved to be an unusually simple basis for computable notations, and equally for human-oriented indexing of molecular structures.

Proper DENDRAL includes a certain detail by way of precise rules to maintain the uniqueness, as well as the non-ambiguity, of its representations. However, to read DENDRAL, or to write vernacular (non-unique though unambiguous) forms in the same notation requires very little indoctrination. A primer of basic DENDRAL is therefore included as Table 1. Computer programs are being completed to conventionalize vernacular forms. Thus it should be possible for relatively unskilled workers to produce computable input or to interpret dictionaries. Programs are also being tested to generate graphic displays from DENDRAL codes.

As DENDRAL-64 implies, this proposal is regarded as a provisional version, subject to substantial improvement on the basis of wider experience. Therefore,

this report will be circulated in its present tentative form before a definitive version is prepared for more extensive publication.

It might be expected that the general treatment of complex rings poses many problems. I believe most of these have been met; however, to program the generating algorithms is a formidable task, probably much more costly than the interpretative ones. It would be prudent to postpone this commitment until the general utility of DENDRAL has been evaluated, and some assessment made of the depth to which the general topological treatment should be carried. Furthermore, the mathematical approach presented here is quite crude. This may help to provoke deeper interest in or application of this branch of topology, the isomorphism of graphs, a theory which might supersede all previous efforts in the taxonomy of molecules.

Fortunately, chemical notational systems have been extensively reviewed to date by a committee of the National Academy of Sciences,[1] which saves the need to classify them here. In that report unique, unambiguous notations are represented only by the efforts of Wiswesser and of Dyson (the IUPAC-61 report) on whose pioneering work any further discussions must lean heavily.

The principle distinction of DENDRAL is its algorithmic character. Each structure has an ordered place, regardless of its notation. DENDRAL was intended primarily for the systematic <u>generation</u> of unique structures on the computer. Only incidentally, but by no means accidentally, does this prove advantageous for notation and classification.

Operationally, DENDRAL avoids the use of locant numerals as far as possible. Instead the emphasis is on topological uniqueness. Functional groups are, in general, analyzed rather than named. The exceptions, -COOH and $-COCH_2-$, are optional and could be discarded. Alternatively, a user

could introduce others at his own convenience without seriously inconveniencing his correspondents and with no embarrassment at all to the computer. In principle, it should be possible to program a translator from any unambiguous notation to any other unique unambiguous notation. The tree-structural representation of DENDRAL may furnish a particular facility for this purpose. Codification in the various systems should, therefore, be interconvertible with rather little effort as soon as the basic programming for the interpretation of each of the other systems has been accomplished.

DENDRAL-64

Many constructive applications of computer programs to organic and biological chemistry await systems for efficient representation of molecular structures. The DENDRAL-64 system outlined here stems from an effort to program the analysis of mass spectra, but may also be applicable to other areas of structure analysis, and to general problems of classification and retrieval. It is presented as a first effort and, if it survives at all, will surely benefit from future revisions.

The present objective is simply a computer program to make an exhaustive, nonredundant list of all the structural isomers of a given formula. Existing notations[1] proved, at least in our hands, to be poorly oriented for this purpose. The work of Henze and Blair (1931)[2] on the enumeration of hydrocarbons suggested a more satisfactory system. DENDRAL aims (1) to establish a unique (i.e., canonical) description of a given structure; (2) to arrive at the canonical form through mechanistic rules, minimizing repetitive searches and geometrical intuition; (3) to facilitate the ordering of the isomers at any point in the scan, and thus also the enumeration of all of them.

The treatment of ring structures is deferred to a later section. Up to that point, the following account applies only to unringed molecules, with no important restrictions on composition or branching. A ring is defined as a set of atoms which can be separated only by cutting at least two links. Hence an unringed structure is defined as one that can be separated by cutting any link. A ringed structure will have one or more rings, and perhaps some additional links and atoms.

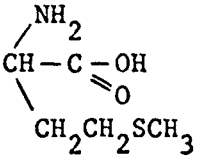
Notation.

A linear notation can be conveniently superimposed on the model-building as a means of communicating results in a form mutually convenient to the chemist and the computer.

A parenthesis-free system, analogous to reverse "Polish" notation for algebra, is suggested.[3] The "operators" are valence bonds, represented by dots issuing from an atom. Each dot looks for a single complete operand. Operand is (recursively) defined as an undotted atom, or an atom whose following dots are each satisfied in turn by an operand. The form .H is generally omitted, and may be freely inserted up to the valence limits of each atom. As a rule, H atoms are not specified. Instead the "unsaturation" of the molecule is examined or calculated and we write one "U" for each pair of hydrogen atoms by which the molecule falls short of saturation (2 + 2C + N). In the final formula, " : " stands for " .U. ", and " ! " for " .U.U. " ; the punch card representations are = and $ respectively.

The explanation is almost more complex than its use. Thus methionine, $C_5H_{11}NO_2S$ becomes an isomer of C_5NO_2SU , and its structure, if written as $CH_2(SCH_3)CH_2CHNH_2COOH$, becomes " C..S.CC.C..NC.:OO ". Note that the whole molecule is a valid operand, the initial C.. being satisfied by the operands S.C and C.C..NC.:OO , the second C.. by N and C.:OO , and the third C.. by O and U.O (represented as C.:OO).

As we shall see, this is not the canonical form, which is instead:



or C...NC.:OOC.C.S.C

See Tables 3 and 4 for a summary of character codes and further details on an operational form of the notation which includes some abbreviations.

Canonical Forms.

To write all possible representations of structures built from a given set of atoms is an elementary but tedious exercise in permutations. However, the more demanding problem is to ensure a unique representation of each isomer. The key to this approach is the recognition of a unique center of any tree structure.[2] Once this is established, the ordering of successive branches is relatively straightforward.

The program has two aspects: (1) A notational algorithm, the transposition of a stated structure to its canonical form, or (2) a generative algorithm, the successive building of each of the hypothetical structural isomers for a given composition. That is, (1) standardize the representation of a given structure to confer its unique location in the dictionary, or (2) generate a dictionary of all possible structures. While (2) is the motif of the study, its principles are best illustrated by application to (1). In the development of the system, notational exercises have also furnished an indispensable discipline to test each facet. This exposition will therefore demonstrate DENDRAL notation in detail, followed by an outline of the generator.

Notational Algorithm: Linear DENDRAL.

A tree structure is analyzed in two stages: (1) the unique centroid is located, and used to root the tree; (2) where two or more branches or radicals stem from a node, they are listed in ascending DENDRAL order. At any point in the analysis the remaining graph can be regarded as a choice among the possible partitions of the atoms not yet accounted for (see Table 2).

1. Locate the Centroid: Primary Partition of the Molecule.

This is the link or node that most evenly divides the tree. The molecule

must fall into just one of the following categories, tested in sequence. Let V be the count of skeletal atoms (CNOS).

A. Two central radicals of equal count are either (1) united by a leading link (V is even), or (2) sister branches from an apical node (V is odd).

B. Three or more central radicals, each counting $< V/2$, stem from a unique apical node.

2. The Radicals are then Arranged in DENDRAL Order.

If two radicals have the same composition but different structure, the structures must be analyzed. To implement the canons of Table 2, each radical is dissected into its apex (i.e., the next node of the tree) plus 1, 2 or 3 radicals. The system order of a radical is determined by the rules of DENDRAL order (Table 2). The radicals are arranged canonically at each step. When every atom has been scanned, the analysis is complete.

"DENDRAL order", synonymously "vector value" or simply "weight", is an evaluation procedure used incessantly in this exposition.

An expression may be treated as a compound number (that is to say vector) with cells x_{ij} in a designated hierarchical sequence in j. Thus we may have $V_1 = (....v_{14}, v_{13}, v_{12}, v_{11})$. The most significant cell is written first, like the digits of an integer. Similarly, to compare two vectors, corresponding cells are scanned from left to right. The first inequality determines which of the two vectors is senior (synonymously heavier, later, larger, greater). Note that any cell may be itself a vector. When terms are missing, for example when vectors of different dimensions are compared, the expression is right-justified, i.e., empty cells are freely supplied according to the context.

This procedure corresponds precisely to numerical order of integers. It also corresponds to common dictionary order if each letter is regarded as a cell, the words are left-justified, and blanks are taken as null-valued cells.

When a cell designates a vector, the procedure is recursive, but can lead to a valuation either if the value of a vector can be obtained by any other rule, or if a vector is ultimately resolved into a set of numbers.

Table 2 is the gist of DENDRAL order. At this stage, the references to rings may be deferred. The weight of a radical is evaluated by the criteria (descending significance is understood): Count, Composition, Unsaturation, Next Node, Its Attached Substructures. In general, each of these may be a vector. For example, in the complete system, Count will be separated into (Rings, Other Atoms). For the moment, Rings = 0, and "Other Atoms" = Count. (H is omitted.) In general node may be "ring" or "atom"; here we will discuss only atoms.

Composition is a list of species and their frequency. Their significance is proportional to atomic number. Hence (S,P,O,N,C) which by coincidence is also alphabetic; implied zeroes are overlooked unless a species is present in only one of the radicals. The priority deviates from most chemical dictionaries, usually (C,H,N,O,...) disregarding count. The rule here favors a greater weight to the more complex structures, so that complexity will run with count not against it.

Unsaturation (degree of) has already been explained.

When the items so far will not decide between two radicals, their own structure must be examined point by point, and if need be substructure by substructure, until an inequality is found or the tree has terminated and the radicals inferred to be equivalent.

Apical Node. This is the node linked to the preceding node (or central apex or central link) of the analysis. The valuation is a vector with components:

Ring Value -- zero, now.

Degree -- number of efferent radicals. This will be uniformly one within a straight chain; hence these will always be junior to branched structures. For the same number of branches, those nearer the central root will be seen earlier, hence add more weight.

Composition -- as above. Hence terminal heteroatoms add less weight than central ones, being seen later.

Afferent Link -- . < : < ⁝

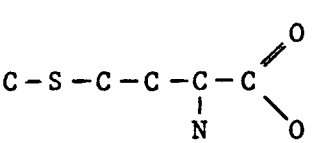
Again, terminal unsaturations add less weight than central ones.

When the next nodes have the same degree and are the same atomic species (e.g., both tertiary carbons), it may be necessary to evaluate the vector of attached radicals, e.g., the set of three joined to a tertiary carbon. For this purpose, each of the radicals must itself be evaluated, and the set placed in ascending order, before the vectors can be compared. The dissymmetry of the apex is a value added only at this point, junior to all previous considerations even though the codes + , - are to replace dots written before the radicals.

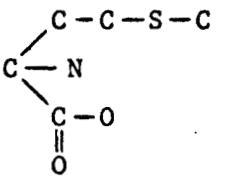
The process is quicker done than explained. Thus for methionine,

$$C_5H_{11}NO_2S \longrightarrow C_5NO_2SU \qquad V = 9 \text{ (i.e., odd).}$$

Try Centroid Rule A. From any terminal, count down to a prospective centroid, atom #5, to try for

X.. (4)(4)  This fails.

Try B. The center of count is quickly found:



C...(4)(1)(3)

and the canonical ordering is already given by the criterion of count

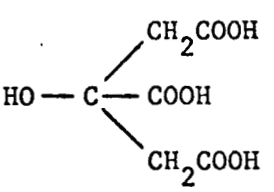
C...(1)(3)(4) namely C...(N)(CO_2U)(C_3S)

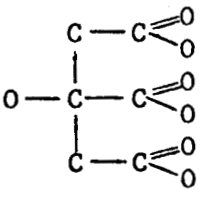
which immediately expands into C...(N) (CO_2U) (C.C.S.C.)

The subdivision of CO_2U becomes C..(O)(OU) , since O < OU , and is abbreviated C.:OO . Thus, the canonical form, after dropping parentheses:

C...N C.:OO C.C.S.C

Another example, citric acid:

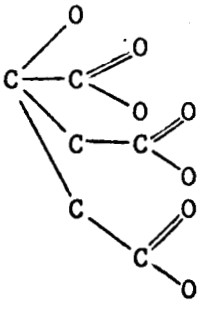




$C_6O_7U_3$

V = 13

The symmetry here is obvious, and to obtain the canonical form we need only order the radicals C....(1)(3)(4)(4)



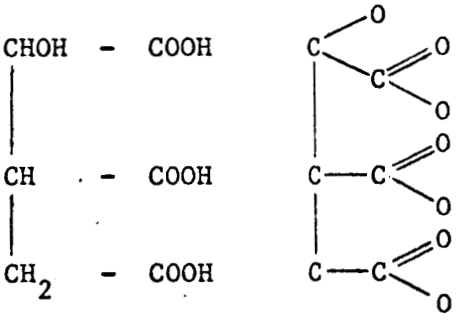
and we immediately write C....(O) (C.:OO) (C.C.:OO) (C.C.:OO)

or C....OC.:OOC.C.:OOC.C.:OO

or, not unhappily, the abbreviated C..O Y C.Y

(Y standing for -COOH, see Table 4).

By comparison, the isomer isocitric acid:



gives the partition C...(3)(4)(5)

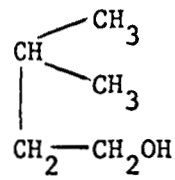
which already places isocitric before citric in system order. This is then

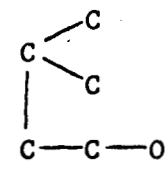
$$C...(CO_2U)\ (C.(CO_2U))\ (C..(O)(CO_2U))$$

and the canonical form is C...Y C.Y C..OY

To turn to an even example:

isopentanol C_5O





is divisible . (3)(3)

the radicals being ordered by composition

. $(C_3)(C_2O)$

quickly reducing to .C..CC C.C.O

which can be abbreviated .C./C 2.0

Additional examples of coded structures, printed from punch cards, are shown in Tables 1 and 5, the latter illustrating the generator algorithm for isomers of alanine.

Significance and Extension of Symbol Codes.

The basic character set given in Table 3 provides a reference to the most familiar atoms and their chemical behavior. How far basic DENDRAL should go into special connotations is a subject of further discussion. Each user inevitably will add his own definitions and elaborations. This need not disturb communication within the system if some care is taken to facilitate algorithmic translation on the computer.

The treatment of rings (v.i.) shows how nodes can be taken as variables having been defined initially. This device could be extended to a variety of special situations. In addition, the characters " Q " and " R " have been reserved for special bonds and nodes, respectively. By rule, the two characters following Q or R are read as part of the same symbolic code. in DENDRAL-64 the combinations R02 - R99 are reserved for otherwise unspecified elements by atomic number. Other letter combinations are available for other conventions, e.g., isotopic substitutions, but have not been rigidified at this stage. The

combinations R.. , R.+ , and R.- are recommended to mark terminations as free-radicals, cations, and anions, respectively. Thus ethyl radical could be coded .C C.R.. . Ethyl^{+} would be .C C.R.+ . Butyrate would be .3 Y.R.- . For Ammonium, see below.

Q codes allow for possible specifications of such non-covalent bonds as coordination complexes, hydrogen bonding, ring-interlocking, as may be needed.

The elements may occur in other than their canonical valence states (e.g. 2: S,O ; 3: N,P ; 4: C ;). The generator algorithm must take account of these variations, but can multiply connotations without enlarging the character set. In notational DENDRAL, ambiguities arise in filling in implicit H's. Bivalent carbon can be treated as a biradical. N is assumed to be trivalent unless more than three links are shown, in which case it is read as quadrivalent ammonium without further notation. If the ammonium ion has less than quaternary substitution, however, requisite .H should be supplied. Thus Tetramethyl ammonium is simply N.///C . Trimethylamine is N.//C . Trimethylammonium is N..//H C .

Salts might best be treated by prefacing a special declaration that two or more species following are to be correlated. For example

QQ2 N..HC Y.R.-

would signify binary salt, methylammonium, formate.

O and S are taken to be bivalent, and H will be filled in accordingly. However, if a higher valence is already encoded, no additional H will be provided unless explicitly noted. The same for B (R07), N, P, and As (R33) at valence 3. No presumptions are made for the valence of other elements.

Asymmetric Carbons.[4]

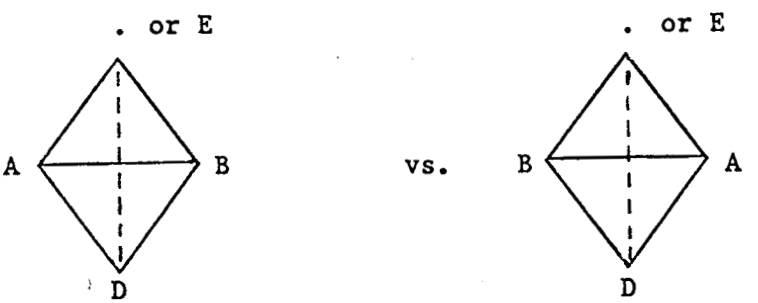
Since DENDRAL assigns a unique path through the structural tree, the specification of stereoisomerism is simplified, although the hierarchy may

differ in detail from other conventions. Once a C atom or ammonium ion is recognized as having four distinct substituents, it should be marked as unspecified, D- , L- , or explicitly a racemic pair, DL- . This is done by inserting nothing, replacing one dot with a " + " or " - ", or two dots with " +- ".

1. We will <u>define</u> as dextro and levo, respectively:

.C+.-A B D vs. .C-..A B D or, when A = " H ",

.C+.B D vs. .C-.B D . according to



dextro C+ vs. levo C-

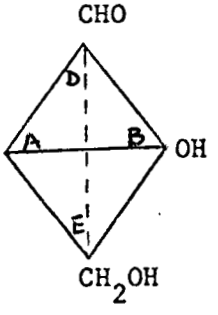
2. Similarly, we distinguish

C+...A B D E vs. C-...A B D E

or,

C+..B D E C-..B D E

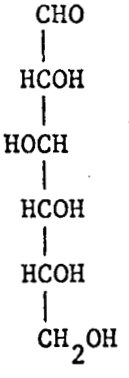
D-glyceraldehyde



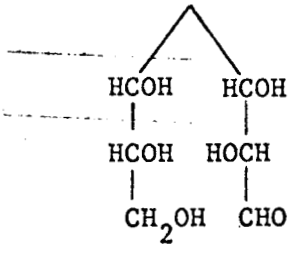
is C+.. O C.O C:O

However, since the DENDRAL path or hierarchy is not always the same as in other conventions, there will be no general correspondence with D.L nomenclature. Thus

D-Glucose (aldose form)



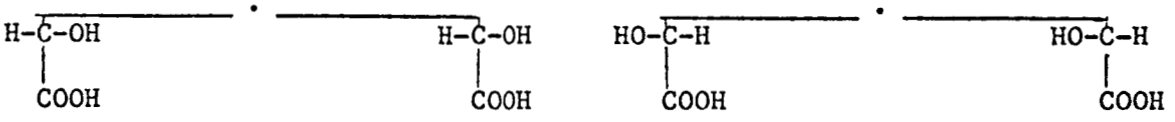
becomes



.C+.OC+.OC.O C+.OC-.OC.C:O

Meso Forms.

The divisibility canons make meso forms easy to recognize. Thus the tartaric acids are dissected

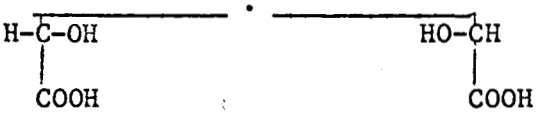


D-tartaric

/C+.O Y or[1] .C+.OY C+.OY

L-tartaric

/C-.O Y or[1] .C-.OY C-.OY



meso-tartaric

.C-.OY C+.OY

Racemic Forms.

The notation allows explicit denotation of racemic pairs as C+- on the indicated carbon. In this context, C.. would imply indifference to (or generalization or ignorance of) the stereoisomerisms. For example,

DL-tartaric acid /C+-O Y interpreted[1] as /C+.O Y plus /C-.O Y ;

mixed tartaric acids /C..O Y .

System Order.

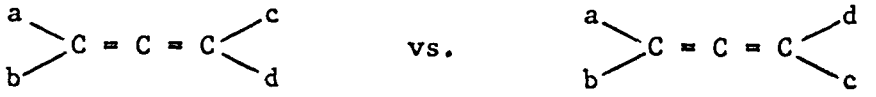
Dissymmetry modifies the DENDRAL value of the apical node

C+- > C+ > C- > C .

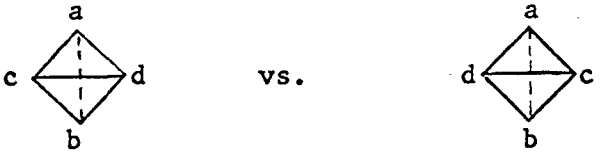
Allenes can be treated in similar fashion. In

---

[1] See Table 4 for significance of " / " .

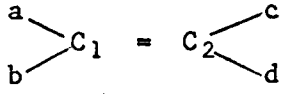


Enantiomers may occur if a ≠ b or c ≠ d. We can visualize



We orient the figure so that a is the senior radical (or afferent link). Then if d > c, we have a "dextral" isomer. Notationally, the enantiomers can be distinguished by writing VCV (dextro) or WCW (levo) in place of the indiscriminate =C= .

Cis-Trans Isomerism must be considered for every double bond unless two identical substituents (or 2H) appear at each of the bonded atoms. The symbol : may be replaced by V for cis or W for trans where indicated. The following rule conforms to conventional practice.[4] If we have



a condition for cis-trans isomerisms is a ≠ b and c ≠ d. The bond is cis if a = c or b = d; and trans if a = d or b = c. I.e., if two equal radicals are on the same side, the bond is cis, where all the radicals are different. If, in DENDRAL order, a > b and c > d, i.e., the senior radical of each pair is on the same side, the bond is " cis ", otherwise " trans ".

The afferent radical linked to a non-central double bonded pair is, of course, senior. Thus, in the construction [a = rest of molecule] a.C.:bC..xy , we already know a > b and the double bond is cis if x (which is by definition junior to y), is on the same side as b.

In DENDRAL order, trans > cis > indifferent, evaluated not with the bond but as a junior element of the set of appended radicals at an apex. This is to retain cis-trans isomers as adjacent values in the scan.

Some of these examples, together with vernacular forms, are repeated in Table 1. The vernacular forms are quite unambiguous. They are, however, highly redundant (non-unique), since the root node of the tree is arbitrarily chosen as is the order of radicals. However, there is no reason why the chemist who has access to a computer need bother himself about the details of DENDRAL order since a computer program can reformat any unambiguous input into a unique form. A program to accomplish this has been written by Mr. Larry Tesler (an undergraduate student in Computer Science at Stanford University), and operated in ALGOL on a Burroughs B-5000 computer. The recursive-procedure facilities of ALGOL are especially useful in the evaluation of a radical by its substructure.

The chemist will, of course, need to know DENDRAL order for any off-line applications such as manual search through a dictionary.

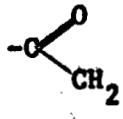
Part I. Table 1.

A DENDRAL PRIMER

1. Acyclic (tree) structures.

Rules for reading DENDRAL are very simple and can be applied directly to writing formulas. To be sure these are in canonical form, however, additional rules must be observed which can be implemented either by a trained analyst or the computer. Vernacular forms are unambiguous, but generally not unique.

To read DENDRAL , it is sufficient to know the principal symbolism for unringed structures:

A..BC		Link(s)	
.A B		Leading Link	
/A		Repeat	/A stands for .AA A./B stands for A..BB
Y		Carboxyl	replaces .C.:OO
G		Aceto	replaces .C.:CO
Integers, e.g., 3	C.C.C	Alkyl	
* e.g., 3*	C:C.C:	Conjugated	

A SUMMARY OF FORMAL DENDRAL

1. Skeletonize formula: Strip H's; shrink rings to nodes.
2. Define rings, if any: see parts 2 and 3.
3. Count skeletal atoms.
4. Identify central apex (unique point of division or central branching).
5. List attached radicals in dictionary order.
6. Dissect each radical, node by node, according to the same rules. Note repeats and use "/" notation.
7. Replace strings representing alkyl, carboxy, and aceto radicals by integers, "Y" and "G", respectively.

ELEMENTARY EXAMPLES OF DENDRAL CODES: TREE STRUCTURES

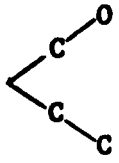
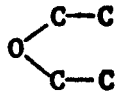
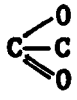
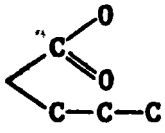
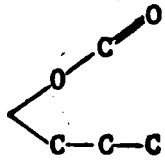
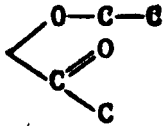
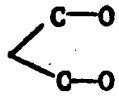
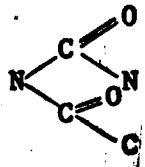
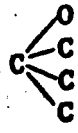
COMPOUND	SKELETON	VERNACULAR DENDRAL	PROPER DENDRAL
Ethanol		C.C.O 2.O 0.2 C..O C	C..C O
Methanol		C.O	.C O
Propanol		3.0 0.3 C..2 O .C.O 2	.2 C.O
Ethyl Ether		2.O.2 C..C 0.2 O..22	O./2
Acetic Acid		C.Y O.G O:C..OC	.C Y
Butyric Acid		C..2Y 3.Y O.C:.03 O.G.2	.3 Y
Propyl Formate		C:.00.3 3.O.C:O O:C.O.3 .O.C:O 3	.3 O.C:O
Ethyl Acetate		2.O.G O..2G	.GO.2
Glycol		O.C.C.O C..OC.O .C.OC.O	/C.O
Acetyl Urea		N..C:.ON G N..C.:NO G N.C:.ON.G	N..G C.:NO
t-Butanol		O.C.//C C.C../OC	C.//.CO

Table 2

CANONS OF DENDRAL ORDER

Hierarchy of Vector Valuation in Decreasing Order of Significance

The DENDRAL-VALUE of a radical consists of its

COUNT

Rings by number of rings[1]

Other atoms (except H)

COMPOSITION of radical

Rings[1] by valuation of ring (see Part II)

Composition, Vertex Group, Path List, Vertex List, Substituent Locations

Other atoms by atomic number (S,P,O,N,C)

UNSATURATIONS (afferent link included; ring paths excluded)

APICAL NODE

Ring Value[1]

Degree: number of efferent radicals[2]

Composition: e.g. (S,P,O,N,C)

Afferent link: (⋮, :, .)

APPENDANT RADICALS

(vectors in canonical order)[2]

Enantiomerism around apex (DL, D, L, unspecified)

---

1 Fixed at 0 for linear DENDRAL and in mapping linear paths on ring.

2 Fixed at degree = 1 (one efferent radical) in mapping linear paths on ring.

Each line is a separate cell of the vector. Fixed items in any comparison may be ignored for that valuation.

Table 3

A SUGGESTED CHARACTER SET FOR ELECTRONIC COMPUTATION

INTEGERS:

Signify strings of C.C.C etc.

Locations of ring substituents

Initialize vertex list

SPECIAL CHARACTERS:

		General Significance	
	Read as	In tree structures[1]	In ring definitions[2]
'	QUOTE		Define literal; path abbrev'n.
( )	START AND CLOSE	Delimit reference	Delimit definition
/	SLASH	Repeat radicals	
,	COMMA		Separator
.	DOT	Single bond	Unspecified dissymmetry
=	DOUBLE	Olefinic bond	Racemic vertex
$	SIGN	Triple bond	Spiro fusion
*	STAR	Conjugated	Aromatic
+	PLUS	Dissymmetry	Dissymmetry
Δ	SPACE	Separate primary radicals	
-	MINUS	Dissymmetry	Dissymmetry; ( )

---

[1] Because of the limited number of characters, the precise meaning depends on the context, as defined.

[2] Path codes resemble those for branches of trees.

Table 3, continued.

LETTERS:

A		N	N
B	Br	O	O
C	C	P	P
D		Q	Special bond initial
E	expunge	R	special node initial
F	F	S	S
G	$-CO-CH_2-$	T	
H	H	U	Unsaturation
I	I	V	Cis-U
J		W	Trans-U
K		X	Variable as defined
L	Cl	Y	-COO-
M		Z	Benzene ring

Table 4

EXTENSION OF NOTATION FOR LINEAR FORMS

NOTATIONAL FORMS WITH ABBREVIATIONS

The following shortcuts have been proven helpful, giving more compact notations. They have no bearing on the logical, ordering operations for which the expanded form is indicated, but are mere abbreviations for the strings they replace.

1. / Repeat. When one or more dots calls upon the same radical as the previous dot, the repeat can be signified by using / in place of the trailing . , and suppressing the redundant string. Thus, ethyl ether O..C.C C.C becomes O./C.C and trimethyl formate C...O.C O.C O.C becomes C.//O.C .

1a. The same rule may also be used for the principal partitions. Thus .C.C.O C.C.O (1,4-butanediol) becomes /C.C.O .

Where the symmetry axis cuts a double or triple bond, the comma is followed by the bond symbol. :C.C.:OO C.C.:OO (maleic acid) becomes /:C.C.:OO .

2. Y Carbonyl. The strings .C.:OO and .C:.OO. (-COOH and -COO-) may be replaced by .Y and .Y. . Do not confuse with the reverse .O.C:.OX . The string C.:OOX.A (X(A)COOH) is replaced by X..AY . Thus, glycine, acetic, glycolic and glyoxylic acids are C..NY, C.Y, C..OY, C:,OY, respectively.

3, G Aceto. The strings of .C:.OC and .C:.OC. ($-COCH_3$ and $-COCH_2-$) are replaced by .G and .G. . The acronyms Y, then G. take precedence over section 5. Thus, maleic acid is :C.Y not :2.:OO and -COCOOH .C:.OC.:OO , becomes .C:.OY not .G.:OO . Pyruvic acid is .YG ; acetoynl .C.C:.OC becomes .C.G .

4. Integer, n. Alkyl. Strings of two or more carbons, C.C , C.C.C , C.C.C.C etc., are designated 2 , 3 , 4 , n etc. 1 is however not used in place of C.

An ambiguity may arise with the integers 22 to 29 , 32 to 39 ,etc. which might generate erroneous formulas if confused with accidental pairs of digits, like thirty-two vs. three , two . These large integers are therefore written out in the form 20.2 for 22 , 30.9 for 39 etc. Caution: distinguish the letter o (oxygen) from the cipher 0 , which has no other application in linear DENDRAL.

5. Space. To facilitate visual interpretation, the radicals of the primary partitition only are separated by a space. In fact, spaces have no syntactic significance and are ignored by the interpreter program.

6. * Conjugation. n* designates a string of conjugated carbon atoms initialized by C: . Thus 4* signifies C:C.C:C. ; 5* C:C.C:C.C: and C.4* is available for C.C:C.C:C. . Note that the * is an unterminated unary operator.

* is also used to indicated aromatic or conjugated character of paths and vertices of rings, as detailed in Part II.

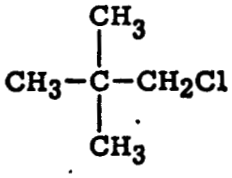
NAS SURVEY OF
CHEMICAL NOTATION SYSTEMS
TEST EXAMPLES AS CODED ON PUNCH CARDS

1A	C•//•C C•L
1B	•C•/C 2•L
1C	3 2•L
2A	/C=C
2B	•3=7 C=6•Y
2C	•C$C C•O
3	(2*A-4/) X(1)C•O
4	(2*A-4/) X(1)O
5	•2 C•N
6	Z•N
7	N•/3
8	N•//3
9	C••/O 2
10	C•//O 2
11	/V3•O•C
12	/W3•O•C
17	•3 C•=OS
18	•3 C=•OS
19	C=••O O•2 C•G
20	C=••O O•2 C=C••CO
21	ZO••NY
22	ZM••NY
23	ZP••NY
24	(4*B4-4,1,4,1)
25	(4*B4-4,2,4)
27	(-6)
28	(-=6)
29	(-=2=4)
30	Z
31	(2*A-4/)
32	(2*A-•4,4)
33	(2A-4•4)
34	(-=2=3)
35	(2*A-•4,•C=2)
36	(4*B4-4,•C,4)
37	(6C6(**)*4,2,,3,2) X(V5)C
38	(6C6-C=3=,2,,3,2) X(V1/V5)C
39	(2*A-•N•2=C,4)
40	(2*A-•N=3,4)
41	(2*A-N•3,•4)
42	(2*A-•2=2,4) X(1=/4)O
43	(2*A-•C=3,4) X(3=/4)O
44	(2A2-=3=C/) X(2=/6)O
45	(6C-2,C'4,2,C)
46	(10E-,C'9/)
47	(10E-,C'9/)
48	N••///B 2
49	N••//L 2
50	(*-N=5) X(1)L
51	QQ2 C••/•O Y C•Y•R•- N••//R•+C
52	QQ2 Z•N•R•+ ZP••C S=/•OO•R•-
53	((X))QPO
54	QCH /C•N•Q•• R24•••/O O•/H L
55	QQ2 QCH (*-N•5)/X(2,1)Q•• R26•/R•+ S=/•/OO•R•-
64	ZP••S•=/OO O•ZP•C
65	ZP••O•C S=/•O ZP•O•C
66	ZP••O•C S=/•O ZP•C•O
67	((2A2-=3=C/)(Z)) X(=3/7,1) N•O N•O•S=/•OZ
68	((2*A-4/)(Z)) X(7,3,1)N•/C S•=/OO N•Z

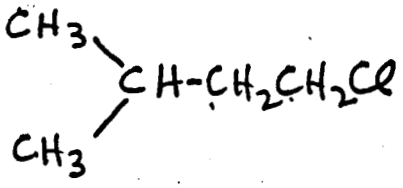
Table 5

The Committee on Modern Methods of Handling Chemical Information, NAS Publication 1150, for its Survey of Chemical Notation Systems,[1] prepared a list of test questions reproduced herewith. DENDRAL codes (some provisional) for these items are presented herewith.

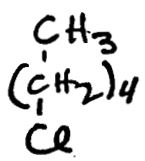
1a



1b



1c



b. Unsaturations in chains, e.g., conjugated double bonds, double bonds, triple bonds, etc. How?

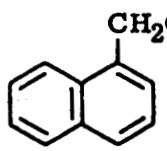
(2) a $CH_2=CH-CH=CH_2$

b $CH_3(CH_2)_5CH=CH(CH_2)_2CH=CH(CH_2)_5\ CO_2H$

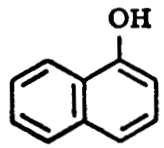
c $HC\equiv C-CH_2-OH$

c. Alcohols and phenols:

(3)



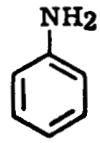
(4)



d. Alkyl amines and aryl amines:

(5) $CH_3CH_2CH_2-NH_2$

(6)



e. Primary, secondary and tertiary amines:

(7) $(CH_3CH_2CH_2)_2-NH$

(8) $(CH_3CH_2CH_2)_3N$

f. Primary, secondary and tertiary alcohols:

(9) $(CH_3CH_2)_2CHOH$

(10) $(CH_3CH_2)_3COH$

g. Cis-trans isomers:

(11) $CH_3-O-CH_2CH_2CH$

$CH_3-O-CH_2CH_2CH$

(12) $CH_3-O-CH_2CH_2CH$

$HCCH_2CH_2-O-CH_3$

h. d, l, dl, meso and unresolved forms; D, L.

(13)

i. alpha, beta forms: steroids substituents

(14) (15) (16)

j. Tautomers:

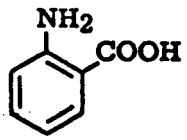
(17) $CH_3CH_2CH_2-\overset{\overset{S}{\|}}{C}-OH$

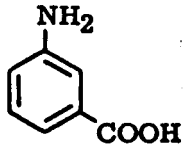
(18) $CH_3CH_2CH_2-\overset{\overset{O}{\|}}{C}-SH$

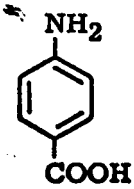
(19) $CH_3-\overset{\overset{O}{\|}}{C}-CH_2-\overset{\overset{O}{\|}}{C}-O-CH_2CH_3$

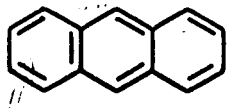
(20) $CH_3-\overset{OH}{C}=CH-\overset{\overset{O}{\|}}{C}-O-CH_2CH_3$

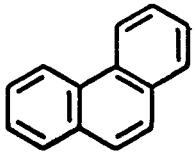
k. Ring position isomers:

(21) 

(22) 

(23) 

(24) 

(25) 

(26)

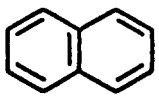
l. Unsaturations in rings:

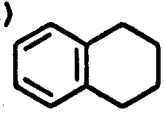
(27) 

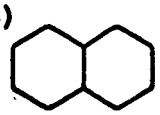
(28) 

(29) 

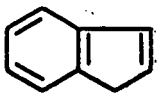
(30) 

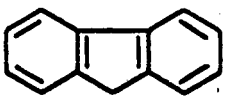
(31) 

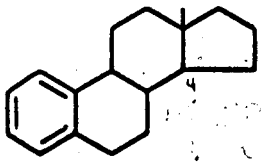
(32) 

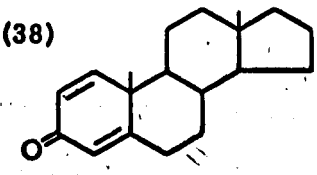
(33) 

(34) 

(35) 

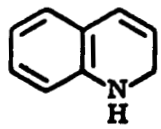
(36) 

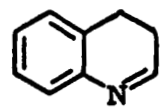
(37) 

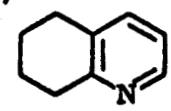
(38) 

Do you define "aromaticity" and if so, how?

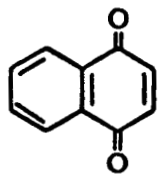
m. Heterocyclics with reduced positions:

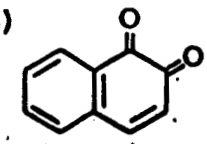
(39) 

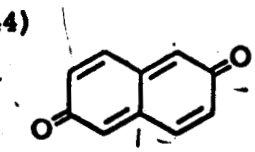
(40) 

(41) 

n. Quinones:

(42) 

(43) 

(44) 

o. Bridged rings:

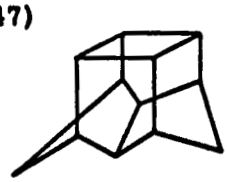
(45) 

Do you define the term "bridge" and if so, how?

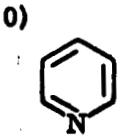
p. Caged rings:

(46) 

q. Three dimensional structures:

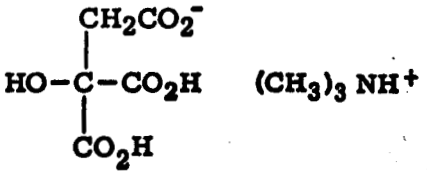
(47) 

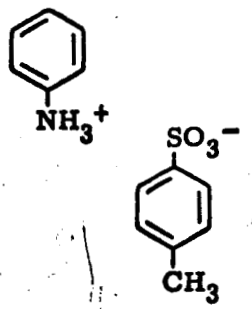
r. Onium compounds; addition compounds:

(48) $(C_2H_5)_4N\ Br$ (49) $(C_2H_5)_3N \cdot HCl$ (50) 

HCl

s. Organic salts; one ion organic; both ions organic:

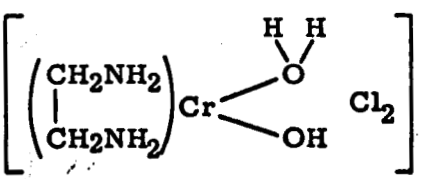
(51) 

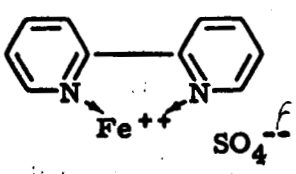
(52) 

t. Polymers:

(53)

u. Chelate compounds:

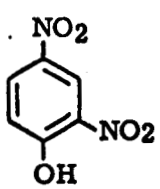
(54) 

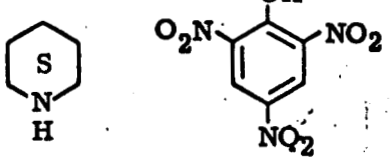
(55) 

v. Structure partially known, partially unknown:

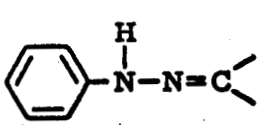
(56)

w. Can you program a search with your code to distinguish nitrophenols from picrate salts and if so, how?

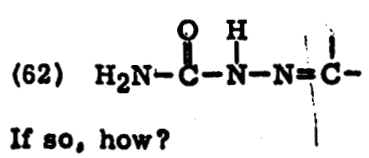
(57) 

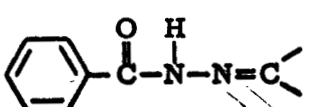
(58) 

x. Can you make a search to distinguish between: How?

(59) $H_2N-N=C$ (60) $H_2N-C=NH$ and (61) 

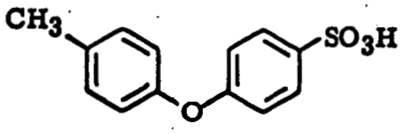
y. Can your code distinguish between:

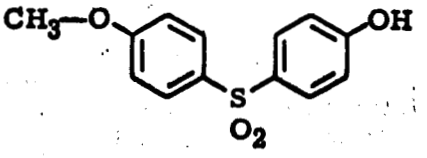
(62) 

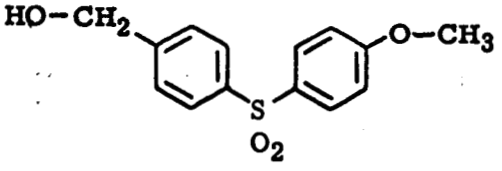
(63) 

If so, how?

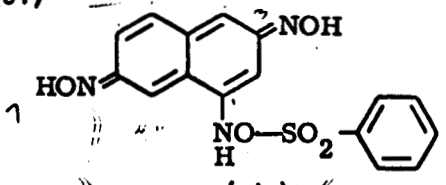
z. How would your code distinguish:

(64) 

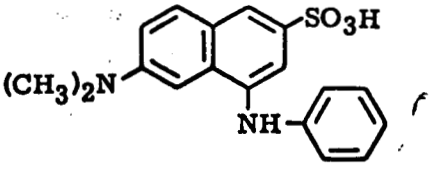
(65) 

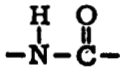
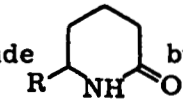
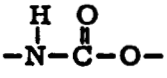
and (66) 

aa. How could you distinguish oximes carrying $-SO_2-$ groups elsewhere in the structure from sulfonic acids with N elsewhere?

(67) 

(sic)

(68) 

bb. How would you make a generic search that would yield all structures with the linkage  that would include  but not  though the person requesting the search were thinking of amides?

cc. Can your code or notation adequately separate —

(1) All fused ring compounds containing three rings, at least one of which is a 5-membered ring, and containing at least one sulfur atom?

(2) All fused ring systems containing two nitrogens in one ring?

(3) All α or β or γ -keto esters?

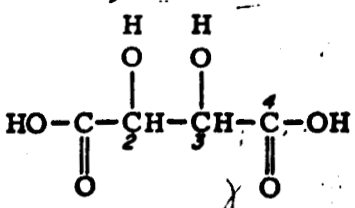
(4) All aryl esters of heterocyclic acids?

(5) All tertiary carbinols in which the tertiary carbon is substituted with an alkyl, an aryl, and a heterocyclic group?

(6) All N,N-dialkylbenzamides?

(7) Sodium salts of all phenols?

For comparison, Mr. Hayward supplied the following notations for several compounds by the Wiswesser, IUPAC, and Hayward systems: (DENDRAL notations also shown)

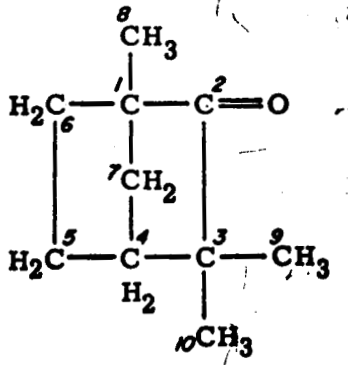


IUPAC: C_4.X,1,4.Q,2,3

Wiswesser: QVYQYQVQ

Hayward: CVQCQCQCVQ

DENDRAL: /C..OY

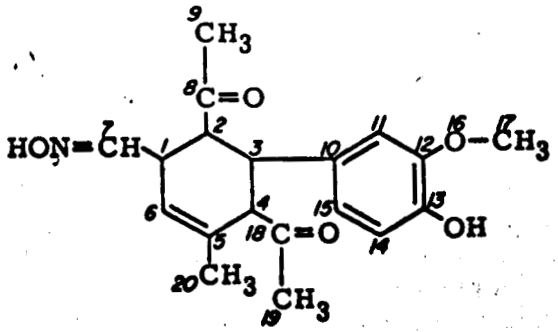


IUPAC: $A5_2$,1-3.C,1,4,4.EQ,5.U2

Wiswesser: $(5^15/cV)bdd$

Hayward: 6L*L(1)MLVLM2L(1)LL

DENDRAL: (2A-2/C)X(V1/2/2,1:)CO



IUPAC=B6.[A6.C_2,1,3.C,4.E,4.EQ, 7,9.ENQ,6.2].QC,3.Q,4

Wiswesser=(6:)ac:NQdVfVeRcOldQ

Hayward=6L(C(=NQ))L(CVM)L(@6RRR (OM)RQRR)L(CVM)LM:L

DENDRAL: ((-:6)Z)X(1,2/4,5,3)CGC:N.O Z(4,3)OO,C