Applications of Artificial Intelligence for Chemical Inference. I. The Number of Possible Organic Compounds. Acyclic Structures Containing C, H, O, and N[1]

J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and Carl Djerassi

Contribution from the Departments of Genetics, Computer Science, and Chemistry, Stanford University, Stanford, California 94305. *Received November 18, 1968*

Abstract: The use of the computer program DENDRAL in constructing the total number of possible acyclic structures of C, H, N, and O is described. Those structures containing either chemical absurdities or undesired functional groups are not constructed if these substructures are explicitly listed. Conversely, if it is desired to restrict the output to any functional group(s) then this can be accomplished. Examples of the linear notation used are given. Semilog plots of total numbers of isomers *vs.* carbon content for selected compositions summarize the results. Some broader implications of the program are discussed which forms the basis for the computer-aided interpretation of mass spectra to be reported[2] in subsequent articles from our laboratories.

Chemists have sensed ever since the theory of structural isomerism was conceived that the number of organic compounds possible was astronomical. In retrospect, therefore, it is surprising that there have been so few attempts to find mathematical procedures for evaluating the number of isomers of a given molecular formula. Such enumerations would be of universal interest in defining the boundaries, scope, and limits of the subject. One specific use of lists of possible isomers is in the computerized inference of chemical structures from mass spectra.[2] Formal attempts to devise an algorithm yielding the number of acyclic alkanes for a given carbon content began in 1875 with Cayley,[3] but it was not until 1931 that Henze and Blair solved this problem.[4] They found it necessary to derive first the number of isomeric alkyl groups.[5] The few other references on the application of topology and combinatorial analysis to chemical problems were listed recently by Balaban.[6]

The first general procedure for enumerating the isomers of any given elemental composition was recently devised by Lederberg.[7] The key to the solution, seen from the viewpoint of topological graph theory (the atoms and bonds of a chemical structure forming the nodes and edges, respectively, of the graph) had been foreshadowed for monofunctional acyclic structures, by Henze and Blair.[4,5] It is that any chemical structure, considered as a tree-graph, has a unique centroid. This centroid is either a bond that evenly divides the tree into two parts with equal numbers of atoms (neglecting hydrogen), or a single atom from which each branch carries less than half the atoms. The unique centroid is then the starting point for a canonical mapping of the tree, following rules that arrange the constituent radicals in systematic sequence. These canons of precedence establish priorities between radicals in terms of, say, the relative number of atoms in each (disregarding hydrogen), heteroatom content, unsaturation present, etc., along lines similar to, but more compactly axiomatized than, the Cahn–Ingold–Prelog[8] absolute configuration conventions. In this way, the atomic connectivity can be conveyed in a linear notational form (*i.e.*, written on one line), and the format itself contains the information needed to rank the linear formulas for a set of isomers in a canonical dictionary order. Examples of this linear notation are given in Table I, which shows ten topologically possible linear isomers of $C_4H_9NO_3$ (threonine).

Table I. Ten Topologically Possible Linear Isomers of Threonine Generated by DENDRAL[a]

```
3050.   N...  O.CH3   O.CH3   CH2.CH=O ,
3075.   C.... CH3   CH3   O.NH2   *COOH ,
3100.   C.... CH3   NH2   O.OH   C.=CH3 O ,
3125.   C.... CH3   OH   CH2.OH   *CONH2 ,
3150.   C.... CH3   OH   O.NH2   O.CH=CH2 ,
3175.   C.... CH3   CH2.OH   O.CH3   N=O ,
3200.   C.... NH2   OH   CH=O   CH..CH3 OH ,
3225.   C.... OH   OH   CH=CH2   N..CH3 OH ,
3250.   C.... OH   OH   CH2.OH   C.=CH3 NH ,
3275.   C.... OH   C2H5   CH2.OH   N=O ;
```

[a] The conventions used by DENDRAL are as follows: period denotes a single bond, *COOH and *CONH₂ are obvious abbreviations, = denotes a double bond, hydrogens and spaces are included for readability. The total output amounted to 3294 topologically possible structures of which only ten are reproduced. Each molecule shown is represented as three or four radicals jointed to a central atom. Entry 3125 corresponds to 2,3-dihydroxy-2-methylpropionamide.

The rules or canons of precedence for writing the linear notation, and for assigning to each isomer of a given composition its unique position in the dictionary list, can then be used to generate such an exhaustive, nonredundant list. It is easy to write down the

[1] This research was financially supported by the Advanced Research Projects Agency of the Office of the Secretary of Defense (Grant SD-183), the National Aeronautics and Space Administration (Grant NGR-05-020-004), and the National Institutes of Health (Grants GM-11309 and AM-04257). The award of a Fulbright Travel Grant (to A. V. R.) from the Australian-American Educational Foundation is gratefully acknowledged.

[2] A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *J. Am. Chem. Soc.*, 91, 2977 (1969).

[3] A. Cayley, *Ber.*, 8, 1056 (1875).

[4] H. R. Henze and C. M. Blair, *J. Am. Chem. Soc.*, 53, 3077 (1931).

[5] H. R. Henze and C. M. Blair, *ibid.*, 53, 3042 (1931).

[6] A. T. Balaban, *Rev. Chim., Acad. Rep. Populaire Roumaine*, 12, 875 (1967).

[7] J. Lederberg, "Topology of Molecules, in The Mathematical Sciences," The MIT Press, Cambridge, Mass., 1969, p 37.

[8] R. S. Cahn, C. K. Ingold, and V. Prelog, *Experientia*, 12, 81 (1956); *Angew. Chem. Intern. Ed. Engl.*, 5, 385 (1966).

structure having the lowest ranking in the hierarchy of priorities (canonical hierarchy). Given this first member of the dictionary list of isomers for that composition, the other members are generated in proper order by permutation of structural subunits according to the rules. The linear notation and the fact that the use of the canons involves an ordered sequence of binary decisions means that the system is very well adapted for computer use, and of course it was intentionally designed with that end in view.

The program developed is called DENDRAL[9] (for *Dend*ritic *Alg*orithm). It is written in the list-processing programming language LISP. It requires 40,000 or more words of memory, depending on the number of atoms in the composition and the speed with which one wants to see answers. Many options are available to the chemist at the teletype console; for instance, he can revise the program's theory of chemical instability (see BADLIST, below), he can restrict structure generation to molecules of a specified class (see GOODLIST, below) or he can monitor the structure-generation process through a dialog with the program. Programming details are available.[10]

We present here some of the results obtained with the DENDRAL program, as applied to a range of atomic compositions for C, H, O, N. The output was confined, for now, to acyclic structural isomers; rings, pentavalent N, and geometrical and optical isomerism were excluded from consideration. The program will handle such structural features but with correspondingly greater demands on computer time and memory.

In using the DENDRAL program, anything that can be done to truncate the potential list of isomers as early as possible results in much improved running efficiency. With the perspective of the practical organic chemist in mind, we have deliberately restricted the DENDRAL output here by excluding structures containing functional groups that are as yet unknown, or unstable, or less favored tautomeric forms. This is achieved with BADLIST, which is simply an input list of structural fragments to be excluded during that run. The program then prunes the DENDRAL tree of isomers at every branch for which a BADLIST entry appears. Conversely, if it is desired to restrict the output to a list of isomers which all contain a given structural group(s), such group(s) can be placed on an input feature called GOODLIST. In this event the program ignores *all* other structural possibilities. The numerical consequences of using BADLIST and GOODLIST are illustrated below.

Table II. BADLIST Used in the Generation of Table III

C=C—O—H	H—O—C—O—H	N—N—O
C=C—N—H	H—O—C—N—H	N—O—N
C≡C—O—H	H—N—C—N—H | H	N—O—O
C≡C—N—H		O—N—O
N≡N—N	H—C—N=O	O—O—O
N=N—O	H—O—C=N	H—O—C—O— ‖ O
N=N—H	O—O N—N—N	H—O—C—N— ‖ O

With GOODLIST empty and with the BADLIST shown in Table II, the output summarized in Table III was generated. Selection of the BADLIST entries was arbitrary, with the idea of truncating the DENDRAL output to those isomers with functional groups that are reasonably stable, well known, and represent only the predominant form of any tautomer. Any atomic and bonding arrangement not explicitly on BADLIST is included among the output structures provided it conforms to the usual valence rules. For example, the first entry in Table II excludes enols, but since C=C—O—C is not listed, enol ethers are not excluded. The entry O—N—O, with two single O—N bonds, does not exclude nitrites, O—N=O. The BADLIST selected here for acyclic aliphatic structures is inappropriate for heterocyclic compounds, where some of the entries in Table II do exist as subunits of stable structures. The generality and flexibility of the DENDRAL program ensures that all tastes can be accommodated. We happen to have excluded peroxides (Table II), but anyone with such an explosive interest, for example, could be rapidly and selectively satisfied by running DENDRAL with O—O on GOODLIST.

Table III was generated by Heuristic DENDRAL on the PDP-6 time-sharing system at the Stanford Artificial Intelligence Laboratory. Under time-sharing conditions, the program generates 100 isomers of any of these compositions in approximately 1–4 min of machine time (including system overhead time apportioned to all users). Depending upon many factors, including the number of other users of the system at the run time, 4 min of machine time will require from 4 to 12 min or more of time at the teletype console. And the amount of machine time required for 100 isomers will also vary depending upon the amount of core memory allocated, the amount of past work saved in the program's dictionary, and the heteroatom content of the composition. Thus these estimates are rough indicators at best.

Practical considerations determined the number of entries in this table. We arbitrarily decided to exclude compositions which we estimated would have many more than 3000 isomers (with a few exceptions).

We have found that simple semilog plots of the number of carbon atoms *vs.* the number of possible acyclic isomers yield nearly straight lines as shown in Figure 1. The semilog plots of Table III before completion were used to predict some additional entries, as a check on the extrapolation method we are advocating. In each case the predictions were accurate within 2%, often much less. For example, with seven entries in the $C_nH_{2n+2}O_2$ section the C_8 and C_9 entries were predicted to be 1000 and 2700 isomers: the actual numbers were 990 and 2688. With only four entries in the $C_nH_{2n+3}NO$ section the C_5 and C_6 entries were predicted to be 135 and 375 isomers; the actual numbers were 137 and 376. The graphs reproduced in Figure 1 are not exact straight lines, since as was pointed out by Perry[11] for the alkanes, the numbers are not in strict geometric progression. It may be of some interest to mathematical chemists to find theo-

[9] Labels capitalized in this way are program MODULES.

[10] B. G. Buchanan and G. L. Sutherland, Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry, Memo No. 62, Stanford Artificial Intelligence Project, July 1967.

[11] D. Perry, *J. Am. Chem. Soc.*, **54**, 2918 (1932).

Table III. Numbers of Possible Acyclic Isomers of Selected Compositions

Section	Comp	Number of carbon atoms[a] 1	2	3	4	5	6	7	8	9	10	11	12
A	C_nH_{2n+2}	1	1	1	2	3	5	9	18	35	75	159	355
	C_nH_{2n}		1	1	3	5	13	27	66	153	377	915	2315
	C_nH_{2n-2}		1	2	4	9	23	55	152	375	1048	2877	
	C_nH_{2n-4}			0	2	6	21	59	195	563	1823		
	C_nH_{2n-6}				1	4	15	45	182	629	2270		
	C_nH_{2n-8}					0	5	21	110	511	2113	8057	
	C_nH_{2n-10}						1	8	45	262	1304		
	C_nH_{2n-12}							0	9	77	532		
	C_nH_{2n-14}								1	13	135		
	C_nH_{2n-16}								0	0	17		
	C_nH_{2n-18}										1		
B	$C_nH_{2n+2}O$	1	2	3	7	14	32	72	171	405	989	2460	6123
	$C_nH_{2n}O$	1	1	4	11	33	91	254	698	1936	5296		
	$C_nH_{2n-2}O$		1	4	15	47	156	492	1544				
	$C_nH_{2n-4}O$		0	2	7	32	566	2687					
C	$C_nH_{2n+2}O_2$	0	2	6	18	48	133	359	990	2688			
	$C_nH_{2n}O_2$	1	3	8	32	110	380	1233	4030				
D	$C_nH_{2n+3}N$	1	2	4	8	17	39	89	211	507	1238	3057	
	$C_nH_{2n+1}N$	1	2	5	14	40	111	304	845	2322			
E	$C_nH_{2n+4}N_2$	1	5	11	34	84	235	623	1724				
	$C_nH_{2n+2}N_2$	2	8	24	78	241	751	2334					
F	$C_nH_{2n+3}NO$	2	6	18	50	137	365	995	2727				
	$C_nH_{2n+1}NO$	2	9	31	105	350	1116	3574					
G	$C_nH_{2n+3}NO_2$	2	9	43	160	533	1756	5617					
	$C_nH_{2n+1}NO_2$	3	17	83	362	1430							
H	$C_nH_{2n+3}NO_3$	0	7	56	288	1313							
	$C_nH_{2n+1}NO_3$	3	17	130	751	3740							

[a] No special significance should be attached to the exact numbers shown in this table. As noted elsewhere, these numbers will change appreciably when relevant BADLIST or GOODLIST entries are added or deleted to suit the interests of individual chemists. Also, these tables carry no guarantee of accuracy since there is always a possibility of undetected hardware or software (programming) errors. Running several examples a second time with the same results has given us a relatively high degree of confidence in the machine itself. And we are quite confident that any errors still in the program would have only a small effect on the results.

retical justification for the curves presented here, and for similar curves for the other sections.

The rate at which the number of possible acyclic isomers increases with added heteroatoms and unsaturations is very startling. For instance there are three isomers of C_4H_8 but 362 isomers of $C_4H_9NO_2$. Many other examples are evident from Figure 1 and Table III.

Since Liebig coined the idea of functional groups it has been natural to organize organic chemistry by such a classification. The example shown in Table IV demonstrates numerically the incisive power of this concept as a logical device to truncate the list of possible isomers of a given composition. There exist 4030 isomers of $C_8H_{16}O_2$, using the constraints of Table II, and probably two or three times that number without these constraints. When a functional group of particular interest is placed on GOODLIST, however, the program generates a much smaller list of isomers (Table IV) each of which contains only that functional group. The numbers of this table were computed by the program without regard for considerations of stability, *i.e.*, with BADLIST empty. Thus, with BADLIST (Table II) filtering in addition to GOODLIST selectivity, even fewer structures will be considered in many cases. These figures show the importance of being able to define the functional groups actually present in a molecule by a simple chemical reaction or by spectroscopic methods as one can then eliminate approximately 90% of the possible structures. The practical consequences of such an approach will become evident from a series of papers to be published from our laboratories.

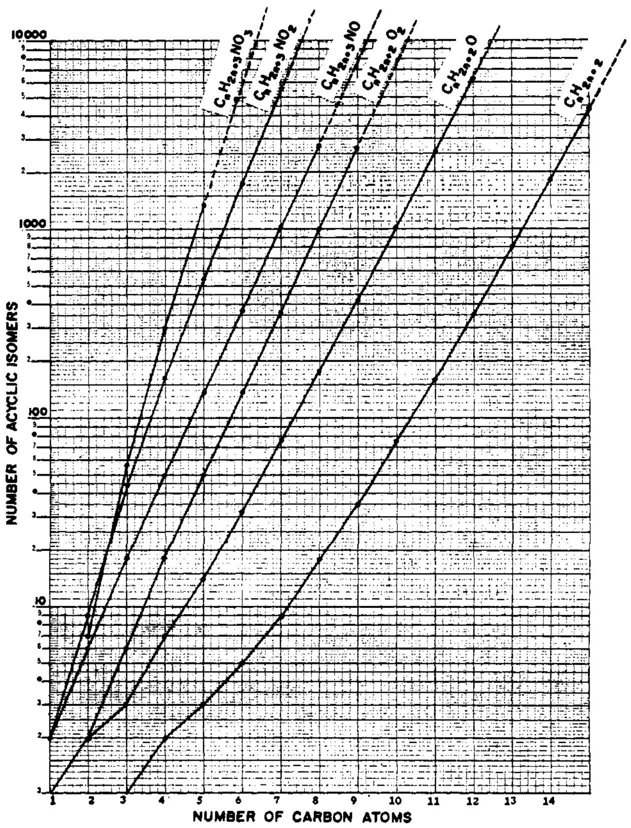
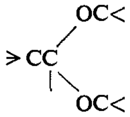
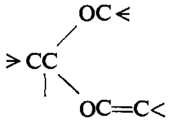
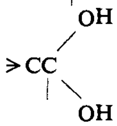


Figure 1. Graphs of relationships between number of acyclic isomers and carbon content for selected compositions.

The use of BADLIST has a practical advantage when the generation of substantial numbers of isomers is required.

Table IV. The Number of Aliphatic Isomers of $C_8H_{16}O_2$ Separated by Functional Groups

No.	Functional group name	No. of isomers of $C_8H_{16}O_2$	Contained subgraph(s)
1	Acid	39	-COOH
2	Ester	105	-COO-
3	Keto ether and aldehyde ether	329	>COC< and -CO-
4	Hydroxy ketone and hydroxyaldehyde	458	>COH and >CCO-
5	Diether (excluding enol ether)	183	$(>COC<)_2$
6	Hydroxy ether	783	>COC< and >COH
7	Enol and ether	305	>COC< and >C=COH
8	Hydroxy enol ether	497	>COH and >C=COC<
9	Unconjugated acetal	102	
10	Conjugated acetal	46	
11	Acyloin enol ether	48	>COC=COC<
12	*gem*-Diol	262	
13	Diol (excluding *gem*-diol and enol)	32	$(>CCOH)_2$
14	Unconjugated peroxide	197	>COOC<
15	Unconjugated hydroperoxide	306	>COOH

Thus with Table II in use there exist 751 isomers of the composition $C_4H_9NO_3$ (Table III, section H). However, with BADLIST inoperative this list amounts to 3294 isomers. Hence by forbidding the generation of isomers containing unwanted or chemically absurd structures BADLIST substantially limits the output level with a corresponding saving on computer time.

It is interesting to note (Table III, section A) that 15 acyclic isomers of benzene exist. Twelve of these are unbranched. In addition the isoprene rule restricts the 1823 possible acyclic isomers of $C_{10}H_{16}$ to 52 structures having two isoprene units linked head to tail with three unsaturations in all possible arrangements.

Of the threonine isomers 385 were selected as arbitrarily interesting for a literature search.[12] This disclosed that only 59 of the isomers had been reported. Similarly, a less detailed investigation showed that only 63 of the 1823 acyclic isomers of $C_{10}H_{16}$ had been described.

DENDRAL has been used extensively in computer interpretation of mass spectra and for details the reader is referred to the following paper.[2]

[12] The literature search was confined to the formula indexes of *Chemical Abstracts* (Vol. **14–65**) and Beilstein and was performed by Mr. Don Greeley.