JL. date?

DRAFT

1

~~Berzelius: Mechanized Induction in Organic Chemistry~~

The fundamental problem of organic chemistry is the topological structure of a molecule. This was first brought into focus by Jons Jakob Berzelius (1779-1848), the Swedish chemist established the occurence of chemical isomers. These are different organic molecules having the same chemical composition or ensemble of atoms; hence they have different structures, i.e. connectivities of the atoms with respect to atom-to-atom bonds. For simplest example, take C_2H_6O, which has the two isomers, dimethyl ether and ethanol (Figure 1). Since the concept of isomerism was ~~Most of the intellectual effort of the important science of organic chemistry is devoted to the discrimination of isomeric structures~~. To determine that the composition of a compound once obtained as a pure sample is, say, C_2H_6O is an essentially mechanical process of quantitative analysis. To assign it to one of the possible isomers is a much more demanding intellectual exercise.

1A

In practical problem solving the chemist uses every possible datum. For example, smell can help him decide between dimethyl ether and ethanol, had he ~~if he did~~ not already recognized that the ether would be much more volatile than its isomeric alcohol. He also has a repertoire of reagents that can help to detect various fragments (called radicals) in the molecule, for example, -OH. More recently a specialized instrument, the mass spectrometer, has been developed which facilitates a unified systematic attack on structural problems. Briefly, a molecule is bom-

At this level of analysis, structure means connectivity, not geometry. In fact, with the help of X-ray diffraction analyses, a great deal can be learned about the actual disposition in space of the atoms in a molecule in the crystalline state. However, the molecules, especially in the liquid or gaseous states, may be undergoing a variety of dynamic transitions - linear, rotational and rocking modes about every chemical bond. Chemical geometry is beyond the scope of the present discussion, but what we know of it could be superimposed upon the topological framework developed hereinbelow.

The preceding paragraph can be summarized: a chemical structure is represented by an undirected graph whose nodes are atoms, whose edges are chemical bonds. While this analogy was recognized almost 100 years ago, it was left to Polya (1937) to make a systematic inquiry based upon it, and this outlook ~~remains virtually unknown to organic chemists today~~ has still to penetrate the teaching of organic chemistry.

barded by an electron beam which sputters off an electron, leaving a positively charged molecule-ion. A fraction of these fragment, giving radical ions of various sizes corresponding to different modes of cleavage, often complicated by further rearrangements and reactions of fragments. Finally, the ensemble of molecule- and radical-ions is resolved by careful acceleration through electrostatic and magnetic fields. The mass spectrum is a paired list of mass numbers and their relative intensities. Mass spectrometers of very high resolution have been built, capable of distinguishing between radicals of different composition but the same integer atomic weight. For example, the radical NH, M= 15.0110 can be distinguished from the radical CH_3, M = 15.0215. This capability is especially useful for determining the formula of the intact molecule. Unless we specify otherwise, we have in mind the more ordinary low resolution mass spectrometer which lumps together species having the same integral mass. However, more precise data are readily accomodated and avidly used by the programs logic.

The stated goal of our program is then an inductive solution of the mass spectrum. That is, a molecular formula and its mass spectrum are given as data. We must induce the structure (hypothesis) that best satisfies the data. Our basic approach to this has been first to furnish the computer with a language in which chemical structure hypotheses can be expressed, then to inter-

rogate chemists and their literature for the rules and techniques they have used in problem solving and attempt to translate these into computer algorithms. In the course of searching for these heuristics, we have in fact discovered a number of algorithms which are much more systematic than the approaches commonly used by chemists in this field.

Under lying the solution of virtually every problem and sub-problem in structural organic chemistry is the potential exhaustion of the list of possible isomers of a given molecule or radical. It is remarkable that while hundreds of thousands of students of elementary organic chemistry are challenged in this way every year, no algorithm for generating and verifying complete lists of isomers has hitherto been presented. Each student is left to work out his own intuitive approach to this problem, which may account for the bafflement with which very many students approach the subject upon their first exposure to it.

The core of ~~[redacted]~~ DENDRAL is ~~the~~ a notation ~~[redacted]~~ used for chemical structures and an algorithm, ~~[redacted]~~ capable of producing all distinct isomers and casting each of them into a canonical representation. This will be outlined in some more detail further on.

One of the principal motives for this investigation has been to provide a utilitarian machine that could in fact be of assistance to chemists working on practical structure problems. Their actual utilization of the machine in problem solving should furnish invaluable information about their own problem solving techniques, and in this way further the development of artificial intelligence and mechanized judgment in this specialized field. It soon became apparent, howedver, that structural organic chemistry is an especially favorable arena for the mechanization of the scientific method. To a degree shared by few other empirical sciences, both the data and the hypotheses can be expressed in fairly simple machinable form. Thus the data of mass spectrometry are simply a list of numbers, while the hypotheses of structural organic chemistry are a list of topological maps, i.e., graphs indicating the connectivity of the component atoms. Redundant hypotheses, that is, isomorphic graphs, can be readily detected by casting them in canonical form. Compare this situation with sciences whose hypothesis statements must be expressed in a natural language! The algebra of chemical maps also gives one confidence that one could compute an exhaustive list of potential hypotheses, each of them at least meaningful, that is, compatible with the data already considered. Most of the permutations of characters or words

that might be used in forming natural language sentences would of course be pure gibberish.

The lowest level of DENDRAL might be called the topologist. This machine considers only the valence rules and elementary graph theory in constructing lists of isomers. It uses two elementary concepts, one,the center of a graph as a point of departure, and two, a recursive procedure for evaluating a radical as a way of specifying the canonical representation of a given molecule. After the center of the map is fixed, being either a bond or an atom of known valence, the radicals pendant on the center must be listed in non-decreasing value. The apical node of each radical is then regarded as a new center and the process continues recursively. A few examples of canonical and non-canonical representations will be help to illustrate this principle. For details please refer to complete outlines already published.( ).

The same approach can be used to make a generator from DENDRAL. From the formula or composition list a bond or a given species of atom is first taken as the central feature and the remaining atoms partitioned in appropriate ways, and these partitions assigned tentatively to the pendant radicals. For each radical then successive allocations are made for the apical node and then partitions are allocated to the pendant subradicals, etc. Table 1 illustrates the computation

of all of the isomers generated by the topologist for the formula $C_3H_7NO_2$, one of whose isomers is the common amino acid, alanine. This exercise is already at the very margin of human capability, barring the possible rediscovery of this algorithm. In practice no intelligent human has the patience to attempt to generate such a list by the intuitive process. The chemist will often ~~then~~ demand redundant information at this point in order to narrow the range of possibilities he is obliged to consider before he will make the effort to produce an exhaustive list. ~~Th~~

The topologist knows only the valence rules as quasi-empirical data, i.e., that four bonds must issue from each carbon atom, three from any nitrogen, two from any oxygen, and but one from hydrogen. With this very limited quota of chemical insight, the topologist produces many structures that would be regarded as absurdities by the experienced chemist, for example no. of the above list. The next stage in the development of DENDRAL is then to impart a certain amount of additional chemical information taken from the real world. IN doing this a definite context is implied, even if this is not immediately overt. There are probably many realms of organic chemistry, i.e. at ultra low temperatures that are beyond our present experience. The implicit context we have in fact adopted is that of the natural product, that is to say, molecular species that

might be reasonably stable at ambient temperatures, and therefore stand some chance of persisting or being isolated from natural sources. However, this rule has been applied rather cautiously and the lists that will be adduced for further illustration still contain a number of items which would be regarded as quite dubious by this criterion. However, the program is quite amenable to adjustment to any given set of facts. Indeed, a certain stage in the program can be switched on to interrogate the chemist to help to find the context in which various rules will be applied or not. At this stage chemical insight is given most explicitly by providing a list of forbidden substructures. Whenever these substructures are encountered during the building of a potential molecule, the generator is adjusted to pass over that entire branch of synthetic possibilities. In order to effectuate this use of a "badlist" a graph matching algorithm has been incorporated into the DENDRAL program. We have followed the line suggested by Sussenguth for this purpose. At best, however, graph matching is an expensive proposition and it soon became necessary to seek ways of economizing on redundant computation. The least important feature, nodal string matching, merely exploits an idiosyncrasy of the DENDRAL program that it is rather easy to detect linear sequences of nodes that might be on a forbidden list of such sequences, for example, -N-N-N or -O-O.

A far greater generality is the use of a dictionary of solved subproblems. As soon as the program has gone a short way towards a solution of any practical problem, DENDRAL would find itself constantly redoing the same subproblems over and over again as it builds radicals on one side of the molecules again after reconstructing the other side. Inorder to avoid the waste involved in this redundancy, the program automatically generates a list of compositions which is consulted whenever a new radical is to be generated. If the composition of the new radical appears in the dictionary, the dictionary contents are simply copied out. If not, the problem is solved and a new dictionary item is entered for further use later. Insofar as the dictionary has already been filtered with respect to BADLIST, a great deal of effort can be saved, and in fact the program would not be practical for molecules of even moderate complexity were it not for this feature. As an example, the dictionary that has been generated in the solution of the alanine problem is given in Table 2. The headings for the dictionary entries are radical compositions expressed in the form U C O , etc. where U stands for double bonds, C for carbon, O for oxygen, etc. (It is convenient in the DENDRAL generator to replace the specification of numbers of hydrogen atoms by an equivalent specification of the number of double bonds in the molecule, represented by U.)

It is also feasible and desirable to give chemical insight into the program by overt manipulation of the dictionary. That is to say, when a given context calls for it, the radicals corresponding to a given composition can be entered directly, usually with the aim of excluding certain idiosyncratic items. This must be done with great care, since the list of larger radicals that may be generated later relies upon the dictionary already established for smaller radicals.

A serious problem encountered in practice is managing the trade-off between the growth of the dictionary and the corresponding loss of scratch space for the list program to maneuver in. If left unchecked the dictionary building can easily reach the point of exhausting available computing room and paralyzing the program. A heuristic management of the dictionary would be a close analog to the human solution to this problem and is being studied at the present time. For example, very large dictionaries could be stored on external memories, and only those segments kept in core needed for the current operations of the program.

These facilities have been built into the DENDRAL generator program in such a way as to leave it in a state of high efficiency. Thus the filters are not applied at the end after the production of a larger redundant list, they are applied at the earliest possible stage in the tree building program. When $C_3H_7NO_2$

is examined by this filtered DENDRAL generator the results of Table 4 are obtained. Each of these is a moderately plausible chemical isomer. No. is the actual structure of alanine. The order of output is the canonical DENDRAL sequence. It may of some interest that three of the structures in Table have apparently not yet been reported in the chemical literature, although they would appear to be reasonable candidates for synthesis by a chemistry graduate student. With even slightly more complex molecules, one should expect to find that only a small minority of the potential structural species are in fact already known to chemical science. Without an algorithmic generator, however, it has not hitherto been possible to make any realistic estimates of the extent of empirical coverage of the theoretical expectations.

It should be perfectly obvious that again with a small increase in complexity the number of possible isomers will grow very quickly and one may have to rely upon a heuristic rather than an exhaustive approach to the generation of hypotheses apt to a given set of data. In particular it might be desirable to use some a priori notions of plausibility in the generator and then to seek ways of adjusting the program so that the parameters for plausibility sequences were already sensitive to qualities in the data themselves. One approach to this uses

goodlist, ~~that is to say~~ an ordered list of preferred substructures. That is to say, we would assign the highest plausibility and therefore ~~would like to see first~~ priority for deductive corroboration of those molecules which contain items in goodlist. In order to accomplish this each goodlist item is regarded as a "super atom" of appropriate valence, and the corresponding subset of atoms from the compositional formula is allocated to the super atom. Thus the very common radical -COOH, the carboxyl radical, is ~~generally the preferred way in which~~ a very common ensemble of a double bond, a carbon atom, and two oxygen atoms ~~should be associated.~~ Insofar as the molecular formula permits, various numbers of these sets of atoms are assigned to carboxyl groups, and the construct *COOH is then regarded as if it were a univalent superatom. Certain housekeeping details must be looked after to be sure of avoiding redundant representations and to reconvert the constructions to canonical form. They will, however, no longer be in canonical sequence, but rather have some implicit order of plausibility in the sequence with which they are put out. When alanine is subjected to such a procedure, the ordering of Figure 5 is obtained. It will be noted that alanine itself is a very early entry in this table.

With these facilities we are now ready to attempt to ~~use~~ apply DENDRAL to explicit data ~~for the first stages of hypothesis.~~ The actual processes in the mass spectrometer are too complicated to be dealt with head-on in the first instance. We therefore

deal with various models of the behavior of the mass spectrometer, the theories of mass spectrometry. ~~To exercise the simpler logical elements of BERZELIUS,~~ we begin with a zero order theory, one which postulates that the mass spectrum is obtained by assigning a uniform intensity to each fragment that can be secured by breaking just one bond in the molecule. We neglect the splitting of bonds affecting only a hydrogen atom. To test the program we do not at first use a real spectrum, but rather the spectrum predicted by this idealized theory for some given isomer.

It is quite characteristic of the scientific method to observe a ~~constant~~ oscillation between experimental observation and theoretical prediction, and then the confrontation of the two.

As before, the predictor is deeply embedded within the DENDRAL generator, so that the structure building tree is truncated at the earliest point that a violation of the theory by the data set is encountered. This leads to a very efficient set of trials, not of completed, but of tentative and partial structures when the program is given a molecular composition and a hypothetical zero-order spectrum. This is illustrated in Table . The essence of the program is to generate all of the partitions at a given level, and then to scan these for compatibility with the mass list of the fragments. There are also some pertinent a priori considerations about the partitioning of molecular compositions, and this

has been used to reorder the primary partitions in the most plausible sequence.

~~These facilities have the advantage of permitting the embedding of a certain amount of judgment with respect to the most likely~~ This sequences of the priority arrangements of hypotheses without losing anyone and without ~~taking the risk of~~ forever, ~~excluding a hypothesis that may eventually prove to be quite viable.~~ We manage the sequence with which hypotheses are tested but still retain the exhaustive and irredundant character of the generator. Owing to imperfect memory and non-standard formats, human judgment rarely succeeds so well at this.

Each of the plausibility operations plainly should and can be related to a statement of context. For example, in setting up the GOODLIST the chemist will be interrogated about the likelihood of certain radicals, and cues for this can also be obtained directly from the mass spectrum. For example, the program is aware that mass number 45 is almost pathognomic for the radical -COOH. Hence, ~~The residue of~~ this superatom will be set to zero in the absence of a signal at that mass. In a high-resolution analysis the occurrence of mass number 44.998 would justify fixing *COOH as non-zero.

The description so far characterizes an operational program ~~whose~~ main features can be routinely demonstrated ~~more or less reliably~~ without special preparation, by remote teletypewriter interactions with the PDP-6 computer at Stanford University.

~~Since the program still operates with idealized spectra and is not quite sophisticated enough to cope generally with real data, it is somewhat difficult to evaluate its~~ DENDRAL has been tested in a number of ways in an attempt performance as a working tool. ~~Berzelius~~ will, of course, vastly outdo the human chemist in such contrived but potentially useful exercises as making an exhaustive ~~xxxx~~ and irredundant list of isomers of a given formula. (Fig. shows this for $C_3H_7NO_2$) In many cases, particularly when an adequate dictionary has been previously built and no further entries are being made, the computer will output its solutions at ~~a rate close to the~~ teletype speed. The program is also slightly faster than the human operator at subgraph-matching, that is, searching a series of molecular structures for the presence of any member of a given list of forbidden embedded subgraphs. It will outdo the human by approximately 100:1, or perhaps better, if accuracy is given due weight in converting structural representations into canonical form and testing for isomorphism.

~~15~~ (14A.00)

Facilities have been provided in the past, but are not available on our present computer system owing to hardware limitations, for providing two-dimensional graphic displays of structural maps as translations of DENDRAL notation. These programs also enabled man-computer interactions where the chemist could manipulate chemical structures to a substantial degree. Where ~~Berzelius~~ DENDRAL begins to be shaky

(14A). A few real spectra have been input, with surprisingly crisp results in view of the known imperfections of the zero order theory of mass spectrometry. Thus alanine was solved (Fig.) with data obtained from our Bendix time-of-flight instrument. Threonine, $C_4H_9NO_3$, elicited the correct structure, and one of the ~~C.... CH3 NH2 CH2.OH COOH~~ (Fig), ~~This is threonine is not available to~~ 2-methyl serine. (Although this compound has been synthesized ~~as for obtaining its mass spectrum, and genesis suggested~~ ( ), its mass spectrum has not been reported. ~~in the chemical literature.~~ However, its ~~the~~ spectrum can be predicted to ~~resemble~~ resemble that of threonine very closely in its qualitative features. Having been ~~Since it is~~ returned as a solution, the spectrum deduced from the zero-order theory must be included in the spectrum input as the data.

data - Math ( ).

14B It is not easy to test the exhaustiveness of the DENDRAL generator without extensive files of known recursive structures. However, it is possible to write combinatorial expressions to count the expected numbers of isomeric alkanes (C_nH_{2n+2}) and alkyl radicals ($-C_nH_{2n+1}$) as shown in table 1. These numbers have been verified through $n = 9$, at which point the LISP program structure became too unwieldy to continue in core memory. The number of isomers is of course vastly greater for compounds containing some N and O atoms.

Heuristic Basis Logic

is, as usual, when confronted with subtle changes of context which the user may often find difficult to ~~express~~ communicate precisely to the program, even when he can ~~communicate~~ do this readily to his fellow scientists. As far as possible we seek to get out of this difficulty by building interrogation subroutines into the program so that the chemist can provide data rather than obliging him to write new program text in the LISP language. ~~At this instant our~~ Present efforts are concentrated on elaborating the theory of mass spectrometry as represented in the predicter sub-program. THis is giving very promising results, the chief limitations being (1) the precise definition of the rules actually used by the chemist and operant in nature, and (2) the translation of these conceptual algorithms into viable program. These two issues are, however, not as independent as might be imagined. It is the clumsiness of the program writing and debugging that impedes rapid testing of the correctness with which a rule has been formulated. In our experience each half hour of conference has generated approximately a man-month of programming effort. ~~It is obvious that~~ despite the simplicity of the DENDRAL notation for chemical structures, we still have a long way to go in the development of a language for the simple expression of other conceptual constructs of organic chemistry, particularly context definitions and reaction mechanisms. Insofar as programs are

also graphs and an effective subroutine may be regarded as a hypothesis that matches its intended functions, the latter being both logically deducible and operationally testable by running the subroutine, program writing may be regarded as an inductive process roughly analogous to the induction of structural formulas as solutions to sets of chemical data. We believe it may be necessary to produce a solution to this meta language puzzle before the implementation of human ideas in computer subroutines can proceed efficiently enough for the rapid and effective transfer of human insights into machine judgment. Nevertheless, by the rather laborious process that we have outlined, the program Berzelius has proceeded to that stage of sophistication where it is at least no longer an occasion of embarrassment to demonstrate it to our scientific colleagues and friends who have no interest whatsoever in computers per se.

~~The~~ DENDRAL ~~and BERZELIUS systems~~ was developed in the LISP 1.5 and 1.6 dialects. The original package was composed by Mr. William White working from the specifications summarized in Table 6 taken from ~~reference~~ Lederberg ( ), and a version of DENDRAL which almost worked was generated on the IBM 7090 with the help of a time-shared editing system run on the PDP-1. In (month, year) the LISP system on System Development Corporation's Q-32 became available to us, and we pursued a vigorous programming effort by remote teletype communication from Stanford to Santa Monica. This proved to be a very powerful and remarkably reliable system and the expenditure of approximate 1 man-year of effort by Mr. White and by Mrs. Georgia Sutherland resulted in the perfection of the program on that computer. In retrospect it is quite obvious that the program simply could never have been written and debugged without the help of the rapid interaction provided by the time-sharing system. We stress "never" advisedly, in the light of our own experience with the human frustrations involved in the typical turnaround times for error detection and error correction under the operating system for the IBM 7090. In November 1966 we moved our operations to LISP 1.5 on the PDP-6 computer installed for the Artificial Intelligence Project at Stanford. Despite the avowed close compatibility of the LISP systems, approximately 3 man-months of effort were required to transfer the program from one dialect to the other.

Somewhere I'd like to work in the point that if we indeed could have easy access to facilities for other kinds of heuristics in a language strictly compatible with our own, we believe we would do very much more experimentation with far-out ideas. It is characteristic of experimental science that whenever a facility is made available, considerable ingenuity is spent in trying to find uses for it, and that this is often an extremely effective approach to the experimental sciences. And finally, I think we ought to have a paragraph or two, not more than that, about our expectations that the development of displays with the structural manipulating facilities that are given in BERZULIUS, and especially by the synthetic chemist, will sufficiently attract a number of working chemists that we can use the system for further extraction of their own heuristics in problem solving in organic chemistry.

As the candidate structures ~~intended to be dealt with~~ become more and more complex we ~~will clearly~~ had have to abandon the idea of exhaustive enumeration of possible structures. Instead we speculate from the data to ~~and appeal in a speculative way to the data to provide~~ cues that offer even a small likelihood of preference for certain kinds of structures as starting points. As we keep examining the problem we do find more and more ways in which such cues can be exploited. For example, an elementary pattern analysis of the period with which mass numbers are represented, e.g., ~~and particularly examination~~ for gaps ~~in the sequence of mass numbers are represented, and particularly exmaination for gaps~~ in the sequence of mass numbers with significant intensity around a period of about 14 (CH_2), can give significant hints about the existence of a number of branch points within the molecule. If these can be limited, the extent of the necessary tree building can be drastically curtailed from first principles. Likewise, an examination of mass numbers approximating half the total molecular weight can lead to some trial hypotheses about the major partition of the molecule, which again can truncate the development. We do not, however, yet have a program sophisticated enough to make a profound reexamination of its own strategy at any level more complicated than the resetting of numerical parameters, a limitation closely related to the meta language challenge mentioned above. In sum, we find that the development of this program has not

encountered very much that is fundamentally new in principle: problem solving in this field has much the same flavor as the solutions already adduced for chess, checkers, theorem proving, etc. One possible advantage of pursuing investigations in artificial intelligence and heuristic programming within this framework is that the practical utility of what has already been produced should ~~suffice to~~ engage ~~the attention of a considerable number of~~ human chemists working on practical problems in a fashion that lends itself to machine observation and emulation of their techniques. This is a hint that the same process ought to be applied to the game of writing programs themselves, to which the same considerations should apply a fortiori.

21 The game of writing programs becomes more and more an experimental science as the complexity of the programs increases. At the limit, the programmer has the insecure hope that his text will (1) run and (2) accomplish the intended goals, that is, his program is a hypothesis that needs deductive elaboration to verify it. This suggests that program-writing ought to be mechanized by a process analogous to the DENDRAL system, and starting with mechanized observations of human techniques of problem-solving.

The pervasive role of analogy in human judgment suggests that much would be gained in artificial intelligence if a large compatible tool kit of successful programs

were available both to the human and the mechanical programmer. ~~Unfortunately, it is rare for two programs written on the same computer system at different times to be sufficiently~~

Unfortunately, artificial intelligencers have a penchant for originality and uniqueness of dialects and there is no easy way in which past successes can be immediately tried out for a new problem. Experimental science, on the other hand, is replete with important advances that resulted from the provocative availability of a new technique waiting to find a use. Indeed, mass spectrometry itself has exactly that history.