STANFORD ARTIFICIAL INTELLIGENCE PROJECT
Memo No. 30

April 5, 1965

AN INITIAL PROBLEM STATEMENT FOR A MACHINE INDUCTION RESEARCH PROJECT

by E. A. Feigenbaum and R. W. Watson

Abstract: A brief description is given of a research project presently getting under way. This project will study induction by machine using organic chemistry as a task area. Topics for graduate student research related to the problem are listed.

The research reported here was supported in part by the Advanced Research Projects Agency of the Office of the Secretary of Defense (SD-183)

AI 30

AN INITIAL PROBLEM STATEMENT FOR A MACHINE INDUCTION RESEARCH PROJECT

by E. A. Feigenbaum and R. W. Watson

We are engaged, in conjunction with Professor Lederberg of the medical school, in a research project which offers possibilities for graduate research, both well defined problems suitable for C.S. 239 projects and not so well defined problems suitable for Ph.D thesis topics. In this memorandum we will define the problem briefly and then outline some suggested projects. If you are interested in any of the projects or topics suggested, or have a topic to suggest related to this project see either of us for further details.

The long range goal of this research is to attempt to come to grips with the problem of induction by machine. That is, how does one build a machine (write a program) which can interact through a suitable interface with its environment and build and improve models of the environment.

The specific task area chosen in which to attack this problem is organic chemistry and in particular, the determination of the structure of organic molecules from mass spectrograph data. The problem presently facing a chemist is roughly the following:

1) A quantity of an organic molecule is supplied to a mass spectrometer.

2) The molecules are bombarded with electrons which break up the molecules into ionized subparts.

3) The mass spectrometer outputs a spectrum (i.e. a distribution of the masses of the subparts).

4) The largest mass in the distribution which occurs in any quantity above a given noise level is that of the parent molecule.

5) By trying various combinations of atoms the chemist finds molecular compositions which have a mass equal to that determined in 4. If the resolution of the mass spectrometer is fine enough the determination of a unique composition is possible.

6) Once the chemical composition, or possible compositions, of the molecule is determined, the chemist uses various heuristics in conjunction with the mass spectrum to determine the structure of the molecule.

The computer science problem is to automate the above process. At the present time we see the project as progressing in the following stages.

Stage 0 - Display of Chemical Structures

Professor Lederberg has developed a linear notation for organic molecular structures. Further, he has devised an algorithm which given a chemical composition as an input will produce as an output all topologically unique organic structures corresponding to this composition. This system is called "Dendral" and exists as an Algol program for the B-5000 written by Larry Tessler.

At the present time many of the structures are not chemically meaningful. Therefore, our first task will be to develop a system which will interact with a chemist and the Dendral system and determine rules for chemically meaningful structures. These rules will be automatically incorporated into a "filter" for the Dendral system.

Presently a program for the PDP-1 exists which accepts a linear Dendral string and displays a chemical graph on the Philco scope. The problem then of Stage 0 is to improve this program and to develop the software for tying it in with Larry Tessler's program through the disc and which will allow us to use LISP on the 7090 from the Philco scopes.

Stage 1 - Chemist at the Philco Keyboard

During Stage 1 we will develop the programming techniques which will allow a dialogue to take place between the chemist and the system for growing the filter on the Dendral output. This system will involve the display of a graph and the chemist's determination of whether or not it is chemically meaningful. The system must then question the chemist to find out what rules the chemist is using for his determinations and accept his answers in a suitable language. In general, the chemist will not be explicitly conscious of the rules he is using and the machine will serve the important function of helping to bring these rules to a precise awareness.

The end result of Stage 1 is that we will have an improved Dendral system and have learned some important and useful computing techniques. An improved Dendral system and associated display should also be of value to those interested in the problems of information retrieval associated with the chemical sciences.

Stage 2 - Mass Spectrograph Analysis

In stage 2 a chemist and a machine interact in real time through the medium of a scope, scope keyboard, typewriter and possibly light pen or tablet. If the machine were used strictly for performing clerical and algorithmic processes the following dialogue would result.

1) The machine would be supplied with the mass spectrum and would display on the scope face a histogram and the chemical composition(s) of the molecule.

2) The chemist using his experience and peripheral information would then input a linear description of a trial structure which would then be displayed on the scope as a chemical graph, or the Dendral system would be invoked to systematically display chemical graphs which correspond to the given composition.

3) The chemist, using his knowledge of likely places for breaks to occur in the above structure when under electron bombardment, would indicate such a break on the graph. The machine would then compute the mass of the subparts and indicate whether or not such a mass exists in the histogram. Or, the chemist would indicate a mass number in the histogram and the machine would indicate whether or not a subgraph exists which has this mass and if it does exist indicate which subgraph it is.

4) The chemist may also want to move various subgraphs from one place to another and then procced as above. The machine will then compute the linear canonical form of these new graphs and possibly change the display to a canonical form. Further, the Dendral system may be invoked to systematically change a given subgraph.

5) The chemist eventually finds a structure which he hypothesizes as capable of yielding the mass spectrum.

What we want is for the machine to be used not only for clerical work, but more importantly to learn from the chemist's behavior and therefore take over much of the analysis on its own. To this end we visualize the following variation of the above dialogue.

Initially the machine would be input the correct structures corresponding to different chemical compositions. The chemist would then proceed to present an example analysis of this structure in conjunction with is mass spectrum; finally concluding with the known result that the structure could have yielded the given mass spectrum. During this process the machine will probe the chemist for the rules leading to his behavior. The machine will incorporate these rules in a data structure which will allow the machine to perform a similar analysis.

The machine will then be given a chemical structure corresponding to a given mass spectrum and will be asked to proceed on a step by step analysis of its own. The machine will report its "reasoning" to the chemist as it proceeds. When the machine makes an incorrect step the chemist will interrupt and a dialogue will take place until the machine can make the correct step.

Finally, when the machine can correctly analyze structures known to correspond to given mass spectrums the system will be given a composition and the Dendral generator will be invoked to systematically present for analysis possible structures. Then a dialogue of the following type will take place. The machine will proceed with an analysis as far as it is able and then the chemist will take over. As the chemist manipulates the graph with machine aid, the machine queries the chemist for the rules governing his behavior and a dialogue takes place.

Eventually the chemist reaches a hypothesis that the given structure could or could not yield the given mass spectrum. The machine then proceeds to analyze the structure on its own to see if it would reach the same hypothesis. If not, a further dialogue takes place until the machine can reach the hypothesis of the chemist.

When the machine seems adequate at this task we proceed to Stage 3.

Stage 3 - Good Initial Guesses as to Chemical Structure

In stage 2 the man and machine proceeded systematically through the structures produced by Dendral. Clearly for any large structures the number of isomers of a given chemical composition could run into the millions. Therefore, the chemist must make a good initial guess as to a possible structure and only rely on the Dendral generator to modify subgraphs. Again the chemist and system interact, with the machine querying the chemist to determine the rules for proposing initial structures. The procedures to be followed will be similar to those of the previous stage.

Stage 4 - Refinement of the System

When stage 3 is completed the system will be a good mass spectrum analyzer. However, the data structures produced during this stage will be complicated, duplicated and in general unlikely to be optimum. Therefore, the program and associated data structures which result from Stage 3 will be carefully analyzed to determine how to write an efficient compact system and to determine which sections contain general chemical knowledge and which contain knowledge of a specialized character, useful mainly for mass spectrograph analysis. The final efficient program which results will form the software for some experiments to be undertaken by a suggested mars probe and the efficient program minus the specialized strucutures will form the basis for a system to be applied to some other chemical tasks such as the synthesis of organic molecules.

The following problems suggest themselves as possible research projects.

1. Display Problems:

In order that the display of the chemical graphs be as useful as possible to the chemist, it should display the graphs in a form as close as possible to that to which the chemist is trained. This task is difficult to do automatically with our present experience. Therefore, one possible approach at this time is to develop a system which automatically displays a graph close to that desired by the chemist and then allows the chemist to manipulate substructures by simple rotations and bond length adjustments. Another possibility is to allow the chemist to "draw" the graph from the keyboard or with a light pen when it is available.

Because of the size limitations of the scope face it will not be possible to display large molecular structures in their entirety. Therefore, it would be useful to have a "window" mechanism which will allow the chemist to study subsections.

Other features are needed which will allow one to save displays, display more than one graph at a time and perform text editing on the linear input. It would also be useful to allow the chemist to build an initial structure and to later make insertions and deletions as well as move a given substructure to another point on the graph.

As the work on the display proceeds feedback from chemists will indicate other useful refinements to the display system.

2. Various programs need to be written which will allow us to use the facilities of the 7090 from the Philco keyboard.

3. Problems relating to Dendral:

Dendral is a system for canonical representation of chemical structures. However, the chemist is usually not trained in this system and would probably find it easier to input a non-canonical linear string. Therefore, it would be of value to have a routine which would convert this string to a canonical one.

Other more abstract problems relating to the Dendral generator are supplied by Professor Lederberg in appendix A.

4. Mass spectrograph analysis problems:

The chemist will want to have a histogram displayed or some display containing equivalent information. The chemist will further want to indicate a given mass number and have the system determine whether or not there is a subgraph with the indicated mass. The work on this problem will lead to abstract research on the searching and comparison of list structures.

It will also be of use to the chemist to be able to indicate a given bond as a likely place for a break to have occurred when under electron bonbardment and have the system determine if the masses of the subparts are in the distribution. The chemist will also want to be able to invoke the Dendral generator to systematically mark and change subgraphs.

5. The Dendral filter growing problem:

As mentioned before, the Dendral generator will generate all topologically unique structures regardless of whether or not they are chemically meaningful. The problem here is to grow, on-line, a filter which will only allow chemically meaningful structures to be displayed. To solve this problem, techniques need to be developed so that the chemist can be questioned for his rules of chemical meaningfulness and so that his responses can be dynamically incorporated in a changing data structure. Because the chemist will not always give correct rules, methods must be introduced to guard against the possibility of incorrect rules permanently entering the system. Persons interested in natural language and the computer or formal languages may be interested in this phase of the work.

6. Advanced mass spectrograph analysis problems:

Related to the problem above will be the development of techniques which allow the rules supplied by the chemist for analyzing structures to be directly introduced into an internal machine structure. This structure will allow the system to perform the same functions as the chemist and report to the chemist the important stages of its analysis. The detailed problems in this area will only become clear as we proceed.

It would seem to us that the problems related to the display are the most suitable for M.S. projects as they are quite well defined. The more challenging problems related to the Dendral system and filter and Stage 2 would seem to be of the greatest interest to those contemplating doctoral research.

References

Lederberg, J., "Dendral - 64 A System for Computer Construction, Enumeration and Notation of Organic Molecules as Free Structures and Cyclic Graphs", Interim Report to the National Aeronautics and Space Administration, December 15, 1964. (Available from either author of this memo).

Lederberg J., "Topological Mapping of Organic Molecules", Proceedings of the National Academy of Sciences, Vol. 53, No. 1, pp 134-139, January 1965. (Available from either author of this memo).

APPENDIX A

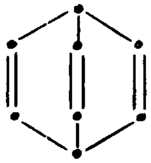
A number of problems in combinatorial graph theory, abstract groups, symmetry, and related subjects have arisen. Some of these would contribute to the elegance and efficiency of the DENDRAL system. Other questions are more abstract and have been suggested by the chemical graphs.

a. Enumeration of cyclic trivalent graphs. This includes the polyhedra. Grace (a former Stanford Mathematics graduate student) has done a possibly vulnerable enumeration up to the 18th order.

b. Efficient tests for isomorphism and reduction to canonical forms.

c. Programming to anticipate symmetries and avoid retrospective elimination of isomorphs.

d. What is the least polyhedron lacking a Hamilton circuit? Now known $20 \leq n \leq 46$.

e. Generalization of the Hamilton circuit (in the sense of mapping a graph on to segments of a circle) to mappings on higher order figures. In DENDRAL-64 the treatment of non-Hamiltonian cyclic graphs (the simplest is ) remains somewhat messy.

f. Heuristic approaches to finding a Hamilton circuit of a graph.

g. Enumeration of graphs with some 4-valent vertices. In DENDRAL-64 this is also somewhat messy, being treated by the collapse of 4-node circuits into 4-valent nodes.

J. Lederberg