Scientific DataLink

The DENDRAL Project: A Short Summary.
Bruce G. Buchanan,
Mar 1977

HPP-77-17 Working Paper

The DENDRAL Project: A Short Summary

Bruce Buchanan
Stanford University
March, 1977

The DENDRAL research project was started in 1965 by Professors J. Lederberg and E.A. Feigenbaum and now includes Professor C. Djerassi in the Chemistry Department and about 20 other persons in Computer Science, Chemistry and Genetics. There are several aspects to the whole project, including research in chemistry and genetics, development of new chemical instrumentation and supporting computer programs, as well as artificial intelligence research. We have had two main computer science goals in this work: to study scientific inference and to aid working scientists. The two programs described below illustrate these concerns.

Heuristic DENDRAL

The Heuristic DENDRAL Program is designed to aid organic chemists determine the molecular structure of unknown compounds. Parts of the program have been highly tuned to work with experimental data from an analytical instrument known as a mass spectrometer. Mass spectrometry is a new and still developing analytical technique. It is not ordinarily the only analytic technique used by chemists, but is one of a broad array of analytic techniques including NMR, IR, UV, and "wet chemistry" analysis. It is particularly useful when the quantity of the sample to be identified is very small, for mass spectrometry requires only micrograms of sample.

A mass spectrometer bombards the chemical sample with electrons causing fragmentations and rearrangements of the molecules. Charged fragments are collected by mass. The data from the instrument, recorded in a bar plot known as a mass spectrum, show the masses of charged fragments plotted against the relative abundance of the fragments at a mass. Although the mass spectrum for each molecule may be nearly unique, it is still a difficult task to infer the molecular structure from the 100-300 data points in the mass spectrum because not only does a spectrum contain "noise peaks" and overlapping peaks originating from many parts of the molecule, but the theory of mass spectrometry is not complete. One way of determining molecular structure of an unknown is to compare its mass spectrum (or

experimental data from any source) to a library of collected data from known compounds. This is a standard solution. But we preferred to approach the problem from a completely different point of view, viz., to reason from the data to the molecular structure that best explains the data. The reason for this choice is twofold. First, we were interested in modelling scientific reasoning. Second, we wanted to develop a program that would be of some assistance to scientists on structure problems that were not already solved and cataloged.

We made several strategy decisions in the course of developing the program, four of which were particularly important in shaping the project. We decided (1) to use a heuristic search paradigm, (2) to use an existing high-level programming language instead of developing our own, (3) to build a high performance system, and (4) to give the system its reasoning power by giving it a large amount of task-specific knowledge. These four points are discussed separately below.

(1) Heuristic Search. The heart of a heuristic search program is a generator of the search space. In a chess playing program, for example, the legal move generator completely defines the space of moves and move sequences. In Heuristic DENDRAL the legal move generator is based on the DENDRAL algorithm developed by J. Lederberg. This algorithm defines a systematic enumerator of molecular structures. It treats molecules as planar graphs and generates successively larger graph structures until all chemical atoms are included in graphs in all possible arrangements. Graphs with cycles presented special problems - their symmetries prevented prospective avoidance of duplicates during generation. Brown, Hjelmeland and Masinter solved these problems in both theory and practice [HPP-72-6, 73-3].

Because the number of chemical graphs can be astronomically large, it is essential to constrain structure generation to only <u>plausible</u> molecular structures. The CONGEN program (named for constrained generator) accepts problem statements of the number of atoms of each type in the molecule and statements of constraints in order to generate all chemical graphs that fit the stated criteria. These problem statements may come from a chemist interpreting his own experimental data or from a program. CONGEN breaks the problem down in many different ways, for example: (i) hydrogen atoms are omitted until the very end; (ii) parts of the graph containing no cycles are generated separately from cyclic parts (and combined at the end); (iii) cycles containing only unnamed nodes are generated before labeling the nodes with names of chemical atoms (e.g., carbon or nitrogen); (iv) cycles containing only three-connected nodes (e.g., nitrogen or tertiary carbon) are generated before mapping two-connected nodes (e.g., oxygen or secondary carbon) onto the edges. At each step several constraints may be applied to limit the number of emerging chemical graphs [HPP-75-11].

CONGEN is designed to be useful in isolation and has been used by several outside chemists in their own work. We have also experimented with adding a planning program at the beginning to set constraints for the generator, and a testing program at the end to filter candidate explanations coming from the generator. These three components are arranged in what we call the Plan-Generate-Test paradigm.

The planning program uses a large amount of knowledge of mass spectrometry to infer structural constraints from the empirical data. For example, it may infer that the unknown molecule is probably a ketone but definitely not a methyl-ketone. This information is put on the generator's lists of good and bad structural features called, naturally, GOODLIST & BADLIST. Planning has been limited almost entirely to mass spectrometry, but the same techniques can be used with other data sources as well.

The testing program also uses a large amount of knowledge of mass spectrometry. Here it is used for the purpose of making testable predictions from each plausible candidate molecule. Predicted data are compared to the data from the unknown compound to throw out some candidates and rank the others.

(2) LISP. We chose to use an existing language, LISP, because we wanted to work on solving the scientific problem instead of building a new problem solving language. LISP was well suited for the development of the program because of its list-handling and general symbol manipulation capabilities and its extreme flexibility. Having completed the program's initial development, it is now reasonable to think of recoding it in a more efficient (and less flexible) language, which we are doing to some extent.

The DENDRAL programs are mostly written in INTERLISP, although some important parts are in FORTRAN and SAIL. They currently run on the dual KI-10 computer at the SUMEX-AIM facility at Stanford. We have begun to explore ways of making the programs more widely available, including reprogramming. At this time, however, the programs are accessible via TYMNET and ARPANET networks but not exportable.

(3) High Performance. From the start, we wanted the program to perform at high levels of competence. We wanted it to have the capability for solving new classes of problems and not just solve a few demonstration problems extremely well [HPP-70-5]. To that end, CONGEN in particular has been developed to accept a wide variety of types of constraints and has been engineered for easy use by chemists. The planning program also has the capability for assimilating knowledge of new classes of compounds. And the prediction programs accept the chemist's definition of the predictive rules to use for ranking candidates.

We have shown that it is possible to write a computer program that equals the performance of experts in some limited classes of problems. Published papers on the program's analysis of aliphatic ketones, amines, ethers, alcohols, thiols and thioethers make the point that the program does not know more than an expert (and in fact knows far less), but it succeeds because of its painstakingly thorough application of the rules it does know. A paper on the program's analysis of estrogenic steroids makes the point that the program is not limited to simple problems. As long as the expert's knowledge of the structural class can be given to the program, the program can patiently examine the data using all of the knowledge, and doing all of the bookkeeping correctly, on problem after problem. Another paper on the analysis of mass spectra of mixtures of estrogenic steroids (without prior separation) establishes the program's ability to do better than experts on occasional problems. With this problem, the program succeeds, and humans fail, because of bookkeeping chores of correlating data points with each possible fragmentation of each possible component of the mixture. Several articles [e.g., HPP-76-6] based on results from CONGEN demonstrate its power and utility for solving problems of medical and biochemical importance.

In its first years, the DENDRAL project focused on fundamental problems in computer science. We were mainly concerned with developing:

a strategy for systematic hypothesis generation;
a symbolic representation for a scientific theory; and
a representation for judgmental rules of the science.

As the interest of working chemists increased, the central focus of the project shifted away from pure AI research to development of a scientific tool. Because few AI programs at that time had ever been applied to practical problems, we saw this as an interesting challenge. Up to that time, we had perhaps demonstrated the feasibility of applying a large AI program to science. But transferring the technology to the working laboratory scientist introduced new problems. Among them were to:

increase the capabilities of the programs;
increase the knowledge base;
increase the programs' speed, reliability and exportability.

(4) Production Rules. We decided to give the program its problem solving power by encoding knowledge of the task domain in flexible tables of rules. In a program that would have to grow, it was not wise to embed the problem solving knowledge deep in the code. Thus we chose as flexible a representation as we could find -- a table of

conditional rules, or productions. We had to rely on expert chemists to perceive errors in the knowledge base and suggest modifications to it. Another reason we wanted the knowledge to be in rule tables instead of LISP procedures is that we felt the production rules were easier for the chemist to understand and easier for us to explain. We were not as strict about moving all of the program's knowledge to rule tables as we might have been. For example, we have written many special-purpose procedures in order to provide chemists with immediate results. But the price for taking those shortcuts was rewriting those parts later.

Meta-DENDRAL

The success of any reasoning program is strongly dependent on the amount of domain-specific knowledge it contains. This is now almost universally accepted in AI, partly because of DENDRAL's success. Because of the difficulty of extracting specific knowledge from experts to put into the program, many years ago we explored the problems of efficiently transferring knowledge into a program. We have looked at two alternatives to "hand-crafting" each new knowledge base: interactive knowledge transfer programs and automatic theory formation programs. In this enterprise the separation of domain-specific knowledge from the computer programs themselves has been a critical component of our success.

One of the stumbling blocks with the interactive knowledge transfer programs is that for some domains there are no experts with enough specific knowledge to make a high performance problem solving program. [See HPP-69-2]. We were looking for ways to avoid forcing an expert to focus on original data in order to codify the rules explaining those data because that is such a time-consuming process. Therefore we began working on an automatic rule formation program (called Meta-DENDRAL) that examines the original data in order to discover for itself the inference rules for that part of the domain.

The problem solving paradigm for Meta-DENDRAL is also the plan-generate-test paradigm used in Heuristic DENDRAL. In this case one part of the program (RULEGEN) generates plausible rules within syntactic and semantic constraints and within desired limits of evidential support. The model used to guide the generation of rules is particularly important since the space of rules is enormous. The model of mass spectrometry in the program is extremely flexible and can be modified by the user to suit his own biases and assumptions about the kinds of rules that are appropriate for the compounds under consideration. The model determines (i) the vocabulary to be used in constructing rules, (ii) the syntax of the rules (the left-hand side of a rule describes a chemical graph, the right-hand side describes a fragmentation and/or rearrangement process to be expected in the mass spectrometer), (iii) some semantic constraints governing the

plausibility of rules. For example, the chemist can use a subset of the terms available for describing chemical graphs and can restrict the number of chemical atoms described in the left-hand sides of rules and can restrict the complexity of processes considered in the right-hand sides. [See HPP-77-6]. The planning part of the program (INTSUM) collects and summarizes the evidential support. The testing part (RULEMOD) looks for counterexamples to rules and makes modifications to the rules in order to increase their generality and simplicity and to decrease the total number of rules.

Meta-DENDRAL successfully formulated rules of mass spectrometry that were new to the science. These rules, along with a discussion of the methodology, were published in the scientific literature [HPP-76-4]. The program was tested to see if it could rediscover the rules of mass spectrometry for two classes of compounds (amines and estrogenic steroids) for which mass spectrometry rules were well known. Then we tried forming rules for three classes of chemical compounds whose mass spectrometry was not as well known (mono-, di-, and tri-ketoandrostanes). The program produced three sets of rules that explained much of the significant data for these classes. The time for manual rule formation for these data was estimated to be several months. The Meta-DENDRAL programs have bee extended to form rules from data collected in a 13C NMR spectrometer as well [HPP-77-4].

Selected Publications

HPP-69-2 Bruce G. Buchanan, G.L. Sutherland, E.A. Feigenbaum, "Toward an Understanding of Information Processes of Scientific Inference in the Context of Organic Chemistry," (September 1969). Also in Machine Intelligence-5, Edinburgh Univ., Press, (1970).

HPP-70-5 Edward A. Feigenbaum, Bruce G. Buchanan, Joshua Lederberg, "On Generality and Problem Solving: a Case Study Using the DENDRAL Program," (August 1970). Also in Machine Intelligence-6, Edinburgh Univ. Press, 1971.

HPP-72-6 H. Brown, L. Masinter, and L. Hjelmeland, "Constructive Graph Labeling Using Double Cosets," Discrete Mathematics 7:1, (1974).

HPP-73-3 Harold Brown and Larry Masinter, "An Algorithm for the Construction of the Graphs of Organic Molecules", Discrete Mathematics 8:227, (1974).

HPP-77-17 Working Paper

HPP-75-11 R.E. Carhart, D.H. Smith, H. Brown and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference XVII. An Approach to Computer-Assisted Elucidation of Molecular Structure," Journal of the American Chemical Society, 97:5755, (1975).

HPP-76-4 B.G. Buchanan, D.H. Smith W.C. White, R.J. Gritter, E.A. Feigenbaum, J. Lederberg, and Carl Djerassi, "Applications of Artificial Intelligence for Chemical Inference XXII. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program," Journal of the American Chemical Society, 98:6168, (1976).

HPP-76-6 D.H. Smith, and R.E. Carhart, "Applications of Artificial Intelligence for Chemical Inference XXIV. Structural Isomerism of Mono and Sesquiterpenoid Skeletons 1,2-," Tetrahedron, 32:2513, Pergamon Press (May 1976).

HPP-77-4 T.M. Mitchell, and G.M. Schwenzer, "Applications of Artificial Intelligence for Chemical Inference XXV. A Computer Program for Automated Empirical 13C NMR Rule Formation," (Submitted to JACS, January 1977).

HPP-77-6 Bruce G. Buchanan and Tom Mitchell, "Model-Directed Learning of Production Rules," Submitted to the Proceedings for the Workshop on Pattern-Directed Inference Systems in Hawaii, (February 1977).