Dendral and Meta-Dendral:

Their Applications Dimension

Bruce G. Buchanan

Computer Science Department, Stanford University, Stanford CA 94305, U.S.A.

Edward A. Feigenbaum

Computer Science Department, Stanford University, Stanford CA 94305, U.S.A.

1. Introduction

The DENDRAL and Meta-DENDRAL programs are products of a large, interdisciplinary group of Stanford University scientists concerned with many and highly varied aspects of the mechanization of scientific reasoning and the formalization of scientific knowledge for this purpose. An early motivation for our work was to explore the power of existing AI methods, such as heuristic search, for reasoning in difficult scientific problems [7]. Another concern has been to exploit the AI methodology to understand better some fundamental questions in the philosophy of science, for example the processes by which explanatory hypotheses are discovered or judged adequate [18]. From the start, the project has had an applications dimension [9, 10, 27]. It has sought to develop "expert level" agents to assist in the solution of problems in their discipline that require complex symbolic reasoning. The applications dimension is the focus of this paper.

In order to achieve high performance, the DENDRAL programs incorporate large amounts of knowledge about the area of science to which they are applied, structure elucidation in organic chemistry. A "smart assistant" for a chemist needs to be able to perform many tasks as well as an expert, but need not necessarily understand the domain at the same theoretical level as the expert. The over-all structure elucidation task is described below (Section 2) followed by a description of the role of the DENDRAL programs within that framework (Section 3). The Meta-DENDRAL programs (Section 4) use a weaker body of knowledge about the domain of mass spectrometry because their task is to formulate rules of mass spectrometry by induction from empirical data. A strong model of the domain would bias the rules unnecessarily.

1.1. Historical perspective

The DENDRAL project began in 1965. Then, as now, we were concerned with the conceptual problems of designing and writing symbol manipulation programs that used substantial bodies of domain-specific scientific knowledge. In contrast, this was a time in the history of AI in which most laboratories were working on general problem solving methods, e.g., in 1965 work on resolution theorem proving was in its prime.

The programs have followed an evolutionary progression. Initial concepts were translated into a working program: the program was tested and improved by confronting simple test cases; and finally a production version of the program including user interaction facilities was released for real applications. This intertwining of short-term pragmatic goals and long-term development of new AI science is an important theme throughout our research. The results presented here have been produced by DENDRAL programs at various stages of development.

2. The General Nature of the Applications Tasks

2.1. Structure elucidation

The application of chemical knowledge to elucidation of molecular structures is fundamental to understanding important problems of biology and medicine. Areas in which we and our collaborators maintain active interest include: (a) identification of natural products isolated from terrestrial or marine sources, particularly those products which demonstrate biological activity or which are key intermediates in biosynthetic pathways; (b) verification of the identity of new synthetic materials; (c) identification of drugs and their metabolites in clinical studies; and (d) detection of metabolic disorders of genetic, developmental, toxic or infectious origins by identification of organic constituents excreted in abnormal quantities in human body fluids.

In most circumstances, especially in the areas of interest summarized above, chemists are faced with structural problems where direct examination of the structure by X-ray crystallography is not possible. In these circumstances they must resort to structure elucidation based on data obtained from a variety of physical, chemical and spectroscopic methods.

This kind of structure elucidation involves a sequence of steps that is roughly approximated by the following scenario. An unknown structure is isolated from some source. The source of the sample and the isolation procedures employed already provide some clues as to the chemical constitution of the compound. A variety of chemical, physical and spectroscopic data are collected on the sample. Interpretation of these data yields structural hypotheses in the form of functional groups or more complex molecular fragments. Assembling these fragments into complete structures provides a set of candidate structures for the unknown. These candidates are examined and experiments are designed to differentiate among them. The experiments, usually collecting additional spectroscopic data and

executing sequences of chemical reactions, result in new structural information which serves to reduce the set of candidate structures. Eventually enough information is inferred from experimental data to constrain the candidates to the correct structure.

As long as time permits and the number of unknown structures is small, a manual approach will usually be successful, as it has been in the past. However, the manual approach is amenable to a high degree of computer assistance, which is increasingly necessary for both practical and scientific reasons. One needs only examine current regulatory activities in fields related to chemistry, or the rate at which new compounds are discovered or synthesized to gain a feeling for the practical need for rapid identification of new structures. More important, however, is the contribution such computer assistance can make to scientific creativity in structure elucidation in particular, and chemistry in general, by providing new tools to aid scientists in hypothesis formation. The automated approaches discussed in this paper provide a systematic procedure for verifying hypotheses about chemical structure and ensuring that no plausible alternatives have been overlooked.

2.2. Structure elucidation with constraints from mass spectrometry

The Heuristic DENDRAL Program is designed to help 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 analytic technique. It is not ordinarily the only analytic technique used by chemists, but is one of a broad array, including nuclear magnetic resonance (NMR), infrared (IR), ultraviolet (UV), and "wet chemistry" analyses. Mass spectrometry is particularly useful when the quantity of the sample to be identified is very small, for it 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 histogram 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 form the 100–300 data points in the mass spectrum. The data are highly redundant because molecules fragment along different pathways. Thus two different masses may or may not include atoms from the same part of the molecules. In short, the theory of mass spectrometry is too incomplete to allow unambiguous reconstruction of the structure from overlapping fragments.

Throughout this paper we will use the following terms to describe the actions of molecules in the mass spectrometer:

(1) Fragmentation—the breaking of a connected graph (molecule) into fragments by breaking one or more edges (bonds) within the graph.

(2) Atom migration—the detachment of nodes (atoms) from one fragment and their reattachment to a second fragment. This process alters the masses of both fragments.

(3) Mass spectral process (or processes)—a fragmentation followed by zero or more atom migrations.

2.3. Structure elucidation with constraints from other data

Other analytic techniques are commonly used in conjunction with, or instead of, mass spectrometry. Some rudimentary capabilities exist in the DENDRAL programs to interpret proton NMR and Carbon 13 (^{13}C) NMR spectra. For the most part, however, interpretation of other spectroscopic and chemical data has been left to the chemist. The programs still need to be able to integrate the chemist's partial knowledge into the generation of structural alternatives.

3. Heuristic DENDRAL as an Intelligent Assistant

3.1. Method

Heuristic DENDRAL is organized as a Plan—Generate—Test sequence. This is not necessarily the same method used by chemists, but it is easily understood by them. It complements their methods by providing such a meticulous search through the space of molecular structures that the chemist is virtually guaranteed that any candidate structure which fails to appear on the final list of plausible structures has been rejected for explicitly stated chemical reasons.

The three main parts of the program are discussed below, starting with the generator because of its fundamental importance.

3.1.1. *The generator*

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 [1–4]. This algorithm specifies a systematic enumeration 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. Because graphs with cycles presented special problems,[1] initial work was limited to chemical structures without rings (with the exception of [21]).

The number of chemical graphs for molecular formulas of interest to chemists can be extremely large. Thus it is essential to constrain structure generation to only *plausible* molecular structures. The CONGEN program [44],[2] is the DEN-

[1] The symmetries of cyclic graphs prevented prospective avoidance of duplicates during generation. Brown, Hjelmeland and Masinter solved these problems in both theory and practice [31, 36].

[2] Named for constrained generator.

DRAL hypothesis generator now in use. It accepts problem statements of (a) the number of atoms of each type in the molecule and (b) constraints on the correct hypothesis, in order to generate all chemical graphs that fit the stated constraints. These problem statements may come from a chemist interpreting his own experimental data or from a spectrometric data analysis program.

The purpose of CONGEN is to assist the chemist in determining the chemical structure of an unknown compound by (1) allowing him to specify certain types of structural information about the compound which he has determined from any source (e.g., spectoscopy, chemical degradation, method of isolation, etc.) and (2) generating an exhaustive and non-redundant list of structures that are consistent with the information. The generation is a stepwise process, and the program allows interaction at every stage: based upon partial results the chemist may be reminded of additional information which he can specify, thus limiting further the number of structural possibilities.

CONGEN breaks the problem down into several types of subproblems, for example: (i) hydrogen atoms are omitted; (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 (or higher) 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 [49].

At the heart of CONGEN are two algorithms whose validity has been mathematically proven and whose computer implementation has been well tested. The structure generation algorithm [31, 36, 39, 40] is designed to determine all topologically unique ways of assembling a given set of atoms, each with an associated valence, into molecular structures. The atoms may be chemical atoms with standard chemical valences, or they may be names representing molecular fragments ("superatoms") of any desired complexity, where the valence corresponds to the total number of bonding sites available within the superatom. Because the structure generation algorithm can produce only structures in which the superatoms appear as single nodes (we refer to these as intermediate structures), a second procedure, the imbedding algorithm [36, 44] is needed to expand the superatoms to their full chemical identities.

A substantial amount of effort has been devoted to modifying these two basic procedures, particularly the structure generation algorithm, to accept a variety of other structural information (constraints), using it to prune the list of structural possibilities. Current capabilities include specification of good and bad substructural features, good and bad ring sizes, proton distributions and connectivities of isoprene units [49]. Usually, the chemist has additional information (if only some general rules about chemical stability) of which the program has little knowledge but which can be used to limit the number of structural possibilities. For example, he may know that the chemical procedures used to isolate the compound would change organic acids to esters and thus the program need not consider structures with unchanged acid groups. Also, he is given the facility to impart this knowledge interactively to the program.

To make CONGEN easy to use by research chemists, the program has been provided with an interactive "front end." This interface contains EDITSTRUC, an interactive structure editor, DRAW, a teletype-oriented structure display program [58], and the CONGEN "executive" program which ties together the individual subprograms, such as subprograms for defining superatoms and substructures, creating and editing lists of constraints or superatoms, and saving and restoring superatoms, constraints and structures from secondary storage (disc). The resulting system, for which comprehensive user-level documentation has been prepared, is running on the SUMEX computing facility at Stanford and is available nationwide over the TYMNET network [46]. The use of CONGEN by chemists doing structure elucidation is discussed in Section 3.4.

3.1.2. *The Planning Programs*

Although CONGEN is designed to be useful as a stand-alone package some assistance can also be given with the task of inferring constraints for the generator. This is done by *planning* programs that analyze instrument data and infer constraints (see [10, 22, 28]).

The DENDRAL Planner uses a large amount of knowledge of mass spectrometry to infer constraints. For example, it may infer that the unknown molecule is probably a ketone but definitely not a methylketone. Planning information like this is put on the generator's lists of good and bad structural features. Planning has been limited almost entirely to mass spectrometry, but the same techniques can be used with other data sources as well.

The DENDRAL Planner [28], allows for cooperative (man-machine) problem solving in the interpretation of mass spectra. It uses the chemist's relevant knowledge of mass spectrometry and applies it systematically to the spectrum of an unknown. That is, using the chemist's definitions of the structural skeleton of the molecule and the relevant fragmentation rules, the program does the bookkeeping of associating peaks with fragments and the combinatorics of finding consistent ways of placing substituents around the skeleton.

The output from the DENDRAL Planner is a list of structure descriptions with as much detail filled in as the data and defined fragmentations will allow. Because there are limits to the degree of refinement allowed by mass spectrometry alone, sets of atoms are assigned to sets of skeletal nodes. Thus the task of fleshing out the plan—specifying possible structures assigned to specific skeletal nodes—is left to CONGEN.

3.1.3. *The Testing and Ranking Programs*

The programs MSPRUNE [61] and MSRANK [59] use a large amount of knowledge of mass spectrometry to make 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 [10, 59, 61].

MSPRUNE works with (a) a list of candidate structures from CONGEN, and (b) the mass spectrum of the unknown molecule. It uses a fairly simple theory of mass spectrometry to predict commonly expected fragmentations for each candidate structure. Predictions which deviate greatly from the observed spectrum are considered *prima facie* evidence of incorrectness; the corresponding structures are pruned from the list. MSRANK then uses more subtle rules of mass spectrometry to rank the remaining structures according to the number of predicted peaks found (and not found) in the observed data, weighted by measures of importance of the processes producing those peaks.

3.2. Research Results

The Heuristic DENDRAL effort has shown that it is possible to write a computer program that equals the performance of experts in some limited areas of science. Published papers on the program's analysis of aliphatic ketones, amines, ethers, alcohols, thiols and thioethers [15, 19, 20, 22] make the point that although the program does not know more than an expert (and in fact knows far less), it performs well because of its systematic search through the space of possibilities and its systematic use of what it does know. A paper on the program's analysis of estrogenic steroids makes the point that the program can solve structure elucidation problems for complex organic molecules [28] of current biological interest. 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 some problems [32]. With mixtures, the program succeeds, and people fail, because of the magnitude of the task of correlating data points with each possible fragmentation of each possible component of the mixture. Several articles based on results from CONGEN demonstrate its power and utility for solving current research problems of medical and biochemical importance [42, 48, 50, 53, 62, 58].

3.3. Human Engineering

A successful applications program must demonstrate *competence*, as the previous section emphasized. However, it is also necessary to design the programs to achieve *acceptability*, by the scientists for whom the AI system is written. That is, without proper attention to human engineering, and similar issues, a complex applications program will not be widely used. Besides making the I/O language easy for the user to understand, it is also important to make the scope and limitations of the problem solving methods known to the user as much as possible [60].

The features designed into DENDRAL programs to make them easier and more pleasant to use include graphical drawings of chemical structures [58], a stylized, but easily understood language of expressing and editing chemical constraints [44], on-line help facilities [60], depth-first problem solving to produce some solutions quickly, estimators of problem size and (at any time) amount of work remaining. Documentation and user manuals are written at many levels of detail. And one of our staff is almost always available for consultation by phone or message [46].

3.4. Applications of CONGEN to Chemical Problems

Many persons have used DENDRAL programs (mostly CONGEN) in an experimental mode. Some chemists have used programs on the SUMEX machine, others have requested help by mail, and a few have imported programs to their own computers.

Copies of programs have been distributed to chemists requesting them. However, we have strongly suggested that persons access the local versions by TYMNET to minimize the number of different versions we maintain and to avoid the need for rewriting the INTERLISP code for another machine.

Users do not always tell us about the problems they solve using the DENDRAL programs. To some extent this is one sign of a successful application. The list below thus represents only a sampling of the chemical problems to which the programs have been applied. CONGEN is most used, although other DENDRAL subprograms have been used occasionally.

Since the SUMEX computer is available over the TYMNET network, it is possible for scientists in many parts of the world to access the DENDRAL programs on SUMEX directly. Many scientists interested in using DENDRAL programs in their own work are not located near a network access point, however. These chemists use the mail to send details of their structure elucidation problem to a DENDRAL Project collaborator at Stanford.

DENDRAL programs have been used to aid in structure determination problems of the following kinds:

terpenoid natural products from plant and marine animal sources
marine sterols
organic acids in human urine and other body fluids
photochemical rearrangement products
impurities in manufactured chemicals
conjugates of pesticides with sugars and amino acids
antibiotics
metabolites of microorganisms
insect hormones and pheremones

CONGEN was also applied to published structure elucidation problems by students in Professor Djerassi's class on spectroscopic techniques to check the

accuracy and completeness of the published solutions. For several cases, the program found structures which were plausible alternatives to the published structures (based on a problem constraints that appeared in the article). This kind of information thus serves as a valuable check on conclusions drawn from experimental data.

4. Meta-DENDRAL

Because of the difficulty of extracting domain-specific rules from experts for use by DENDRAL, a more efficient means of transferring knowledge into the program was sought. Two alternatives to "handcrafting" each new knowledge base have been explored: 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 critical.

One of the stumbling blocks with programs for the interactive transfer of knowledge is that for some areas of chemistry there are no experts with enough specific knowledge to make a high performance problem solving program (see [16]). It is desirable 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. For these reasons an effort to build an automatic rule formation program (called Meta-DENDRAL) was initiated.

The DENDRAL programs are structured to read their task-specific knowledge from tables of production rules and execute the rules in new situations, under rather elaborate control structures. The Meta-DENDRAL programs have been constructed to aid in building the knowledge base, i.e., the tables of rules.

4.1. The Task

The present Meta-DENDRAL program [51, 63] interactively helps chemists determine the dependence of mass spectrometric fragmentation on substructural features, under the hypothesis that molecular fragmentations are related to topological graph structural features of molecules. Our goal is to have the program suggest qualitative explanations of the characteristic fragmentations and rearrangements among a set of molecules. We do not now attempt to rationalize all peaks nor find quantitative assessments of the extent to which various processes contribute to peak intensities.

The program emulates many of the reasoning processes of manual approaches to rule discovery. It reasons symbolically, using a modest amount of chemical knowledge. It decides which data points are important and looks for fragmentation processes that will explain them. It attempts to form general rules by correlating plausible fragmentation processes with substructural features of the molecules. Then, as a chemist does, the program tests and modifies the rules.

Each I/O pair for Meta-DENDRAL is: (INPUT) a chemical sample with uniform molecular structure (abbreviated to "a structure"): (OUTPUT) one X–Y point from the histogram of fragment masses and relative abundances of fragments (often referred to as one peak in the mass spectrum).

Since the spectrum of each structure contains 100 to 300 different data points, each structure appears in many I/O pairs. Thus, the program must look for several generating principles, or processes, that operate on a structure to produce many data points. In addition, the data are not guaranteed correct because these are empirical data which may contain noise or contributions from impurities in the original sample. As a result, the program does not attempt to explain every I/O pair. It does, however, choose which data points to explain on the basis of criteria given by the chemist as part of the imposed model of mass-spectrometry.

Rules of mass spectrometry actually used by chemists are often expressed as what AI scientists would call production rules. These rules (when executed by a program) constitute a simulation of the fragmentation and atom migration processes that occur inside the instrument. The left-hand side is a description of the graph structure of some relevant piece of the molecule. The right-hand side is a list of processes which occur: specifically, bond cleavages and atom migrations. For example, one simple rule is

$$\text{(R1)}\quad \text{N—C—C—C} \rightarrow \text{N—C*C—C}$$

where the asterisk indicates breaking the bond at that position and recording the mass of the fragment to the left of the asterisk. (No migration of atoms between fragments is predicted by this rule.)

Although the vocabulary for describing individual atoms in subgraphs is small and the grammar of subgraphs is simple, the size of the subgraph search space is large. In addition to the connectivity of the subgraph, each atom in the subgraph may have up to four (dependent) attributes specified: (a) Atom type (e.g., carbon), (b) Number of connected neighbors (other than hydrogen), (c) Number of hydrogen neighbors, and (d) Number of doubly-bonded neighbors. The size of the space to consider, for example, for subgraphs containing 6 atoms, each with any of (say) 20 attribute-value specifications, is 20^6 possible subgraphs.

The language of processes (right-hand sides of rules) is also simple but can describe many combinations of actions: one or more bonds from the left-hand side may break and zero or more atoms may migrate between fragments.

4.2. Method

The rule formation process for Meta-DENDRAL is a three-stage sequence similar to the plan–generate–test sequence used in Heuristic DENDRAL. In Meta-DENDRAL, the generator (RULEGEN), described in section 4.2.2 below, 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 very large. The model

of mass spectrometry in the program is highly 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 (as before, 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 [63].

The planning part of the program (INTSUM), described in 4.2.1, collects and summarizes the evidential support. The testing part (RULEMOD), described in 4.2.3, 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. These three major components are discussed briefly in the following subsections.

4.2.1. *Interpret Data as Evidence for Processes*

The INTSUM program [33] (named for data interpretation and summary) interprets spectral data of known compounds in terms of possible fragmentations and atom migrations. For each molecule in a given set, INTSUM first produces the plausible processes which might occur, i.e., breaks and combinations of breaks, with and without atom migrations. These processes are associated with specific bonds in a portion of molecular structure, or skeleton, that is chosen because it is common to the molecules in the given set. Then INTSUM examines the spectra of the molecules looking for evidence (spectral peaks) for each process.

Notice that the association of processes with data points may be ambiguous. For instance, in the molecule $CH_3—CH_2—CH_2—NH—CH_2—CH_3$ a spectral peak at mass 29 may be attributed to a process which breaks either the second bond from the left or one which breaks the second bond from the right, both producing $CH_3—CH_2$ fragments.

4.2.2. *Generate Candidate Rules*

After the data have been interpreted by INTSUM, control passes to a heuristic search program known as RULEGEN [51], for rule generation. RULEGEN creates general rules by selecting "important" features of the molecular structure around the site of the fragmentations proposed by INTSUM. These important features are combined to form a subgraph description of the local environment surrounding the broken bonds. Each subgraph considered becomes the left hand side of a candidate rule whose right hand side is INTSUM's proposed process. Essentially RULEGEN searches (within the constraints) through a space of these subgraph descriptions looking for successively more specific subgraphs that are supported by successively "better" sets of evidence.

Conceptually, the program begins with the most general candidate rule, X*X (where X is any unspecified atom and where the asterisk is used to indicate the broken bond, with the detected fragment written to the left of the asterisk). Since the most useful rules lie somewhere between the overly-general candidate, X*X, and the overly-specific complete molecular structure descriptions (with specified bonds breaking), the program generates refined descriptions by successively specifying additional features. This is a coarse search; for efficiency reasons RULEGEN sometimes adds features to several nodes at a time, without considering the intermediate subgraphs.

The program systematically adds features (attribute-value pairs) to subgraphs, starting with the subgraph X*X, and always making each successor more specific than its parent. (Recall that each node can be described with any or all of the following attributes: atom type, number of non-hydrogen neighbors, number of hydrogen neighbors, and number of doubly bonded neighbors.) Working outward, the program assigns one attribute at a time to all atoms that are the same number of atoms away from the breaking bond. Each of the four attributes is considered in turn, and each attribute *value* for which there is supporting evidence generates a new successor. Although different values for the same attribute may be assigned to each atom at a given distance from the breaking bond, the coarseness of the search prevents examination of subgraphs in which this attribute is totally unimportant on *some* of these atoms.

4.2.3. *Refine and Test the Rules*

The last phase of Meta-DENDRAL (called RULEMOD) [51] evaluates the plausible rules generated by RULEGEN and modifies them by making them more general or more specific. In contrast to RULEGEN, RULEMOD considers negative evidence (incorrect predictions) of rules in order to increase the accuracy of the rule's applications within the training set. While RULEGEN performs a coarse search of the rule space for reasons of efficiency, RULEMOD performs a localized, fine search to refine the rules.

RULEMOD will typically output a set of 5 to 10 rules covering substantially the same training data points as the input RULEGEN set of approximately 25 to 100 rules, but with fewer incorrect predictions. This program is written as a set of five tasks, corresponding to the five points below.

Selecting a Subset of Important Rules. The local evaluation in RULEGEN has ignored negative evidence and has not discovered that different RULEGEN pathways may yield rules which are different but explain many of the same data points. Thus there is often a high degree of overlap in those rules and they may make many incorrect predictions. The initial selection removes most of the redundancy in the rule set.

Merging Rules. For any subset of rules which explain many of the same data points, the program attempts to find a slightly more general rule that (a) includes all the evidence covered by the overlapping rules and (b) does not bring in extra negative evidence. If it can find such a rule, the overlapping rules are replaced by the single compact rule.

Deleting Negative Evidence by Making Rules More Specific. RULEMOD tries to add attribute-value specifications to atoms in each rule in order to delete some negative evidence while keeping all of the positive evidence. This involves local search of the possible additions to the subgraph descriptions that were not considered by RULEGEN. Because of the coarseness of RULEGEN's search, some ways of refining rules are not tried, except by RULEMOD.

Making Rules More General. RULEGEN often forms rules that are more specific than they need to be. Thus RULEMOD seeks a more general form that covers the same (and perhaps new) data points without introducing new negative evidence.

Selecting the Final Rule Set. The selection procedure applied at the beginning of RULEMOD is applied again at the very end of RULEMOD in order to remove redundancies that might have been introduced during generalization and specialization.

4.3. Meta-DENDRAL Results

One measure of the proficiency of Meta-DENDRAL is the ability of the corresponding performance program to predict correct spectra of new molecules using the learned rules. One of the DENDRAL performance programs ranks a list of plausible hypotheses (candidate molecules) according to the similarity of their predictions (predicted spectra) to observed data. The rank of the correct hypothesis (i.e. the molecule actually associated with the observed spectrum) provides a quantitative measure of the "discriminatory power" of the rule set.

The Meta-DENDRAL program has successfully rediscovered known, published rules of mass spectrometry for two classes of molecules. More importantly, it has discovered new rules for three closely related families of structures for which rules had not previously been reported. Meta-DENDRAL's rules for these classes have been published in the chemistry literature [51]. Evaluations of all five sets of rules are discussed in that publication.

Recently Meta-DENDRAL has been adapted to a second spectroscopic technique, 13C-nuclear magnetic resonance (13C-NMR) spectroscopy [62, 64]. This new version provides the opportunity to direct the induction machinery of Meta-DENDRAL under a model of 13C-NMR spectroscopy. It generates rules which associate the resonance frequency of a carbon atom in a magnetic field with the local structural environment of the atom. 13C-NMR rules have been generated and used in a candidate molecule ranking program similar to the one described above. 13C-NMR rules formulated by the program for two classes of structures have been successfully used to identify the spectra of additional molecules (of the same classes, but outside the set of training data used in generating the rules).

The quality of rules produced by Meta-DENDRAL has been assessed by

(a) obtaining agreement from mass spectroscopists that they are reasonable explanations of the training data and provide acceptable predictions of new data, and

(b) testing them as discriminators of structures outside the training set.

The question of agreement on previously characterized sets of molecules is relatively easy, since the chemist only needs to compare the program's rules and predictions against published rules and spectra. Agreement has been high on test sets of amines, estrogenic steroids, and aromatic acids. On new data, however, the chemist is forced into spot checks. For example, analyses of some individual androstane spectra from the literature were used as spot checks on the program's analysis of the collections of androstane spectra.

The discrimination test is to determine how well a set of rules allows discrimination of known structures from alternatives on the basis of comparing predicted and actual spectra. For example, given a list of structures (S1, . . ., S*n*) and the mass spectrum for structure S1, can the rules *predict* a spectrum for S1 which matches the *given* spectrum (for S1) better than spectra *predicted* for S2–S*n* match the given spectrum. When this test is repeated for each available spectrum for structures S1–S*n*, the discriminatory power of the rules is determined. The program has found rules with high discriminatory power [51], but much work remains before we standardize on what we consider an optimum mix of generality and discriminatory power in rules.

4.3.1. *Transfer to Applications Problems*

The INTSUM program has begun to receive attention from chemists outside the Stanford community, but so far there have been only inquiries about outside use of the rest of Meta-DENDRAL. INTSUM provides careful assistance in associating plausible explanations with data points, within the chemist's own definition of "plausible". This can save a person many hours, even weeks, of looking at the data under various assumptions about fragmentation patterns.

The uses of INTSUM have been to investigate the mass spectral fragmentations of progesterones [54, 55], marine sterols and antibiotics [in progress].

5. Problems

The science of AI suffers from the absence of satellite engineering firms that can map research programs into marketable products. We have sought alternatives to developing CONGEN ourselves into a program that is widely available and have concluded that the time is not yet ripe for a transfer of responsibility. In the future we hope for two major developments to facilitate dissemination of large

AI programs: (a) off-the-shelf, small (and preferably cheap) computers that run advanced symbol manipulating languages, especially INTERLISP, and (b) software firms that specialize in rewriting AI applications programs to industrial specifications.

While the software is almost too complex to export, our research-oriented computer facility has too little capacity for import. Support of an extensive body of outside users means that resources (people as well as computers) must be diverted from the research goals of the project.

At considerable cost in money and talent, it has been possible to export the programs to Edinburgh.[3] But such extensive and expensive collaborations for technology transfer are almost never done in AI. Even when the software is rewritten for export, there are too few "computational chemists" trained to manage and maintain the programs at local sites.

6. Computers and Languages

The DENDRAL programs are coded largely in INTERLISP and run on the DEC KI-10 system under the TENEX operating system at the SUMEX computer resource at Stanford. Parts of CONGEN are written in FORTRAN and SAIL including some I/O packages and graph manipulation packages. We are currently studying the question of rewriting CONGEN in a less flexible language in order to run the program on a variety of machines with less power and memory. Peripheral programs for data acquisition, data filtering, library search and plotting exist for chemists to use on a DEC PDP 11/45 system, but are coupled to the AI programs only by file transfer.

7. Conclusion

CONGEN has attracted a moderately large following of chemists who consult it for help with structure elucidation problems. INTSUM, too, is used occasionally by persons collecting and codifying a large number of mass spectra.

With the exceptions just noted, the DENDRAL and Meta-DENDRAL programs are not used outside the Stanford University community and thus they represent only a successful *demonstration* of scientific capability. These programs are among the first AI programs to do even this. The achievement is significant in that the task domain was not "smoothed" or "tailored" to fit existing AI techniques. On the contrary, the intrinsic complexity of structure elucidation problems guided the AI research to problems of knowledge acquisition and management that might otherwise have been ignored.

The DENDRAL publications in major chemical journals have introduced to chemists the term "artificial intelligence" along with AI concepts and methods. The large number of publications in the chemistry literature also indicates substantial and continued interest in DENDRAL programs and applications.

[3] R. Carhart is working with Professor Donald Michie's group to bring up a version of CONGEN there.

ACKNOWLEDGMENTS

The individuals who, in addition to the authors, are collectively responsible for most of the AI concepts and code are:

Raymond Carhart, Carl Djerassi, Joshua Lederberg and Dennis Smith.
Harold Brown, Allan Delfino, Geoff Dromey, Alan Duffield, Larry Masinter, Tom Mitchell, James Nourse, N. S. Sridharan, Georgia Sutherland, Tomas Varkony, and William White.

Other contributors to the DENDRAL project have been:

M. Achenbach, C. Van Antwerp, A. Buchs, L. Creary, L. Dunham, H. Eggert, R. Engelmore, F. Fisher, N. Gray, R. Gritter, S. Hammerum, L. Hjelmeland, S. Johnson, J. Konopelski, K. Morrill, T. Rindfleisch, A. Robertson, G. Schroll, G. Schwenzer, Y. Sheikh, M. Stefik, A. Wegmann, W. Yeager, and A. Yeo.

A large number of individuals have worked on programs for data collection and filtering from the mass spectrometer, as well as on operation and maintenance of the instruments themselves. We are particularly indebted to Tom Rindfleisch for overseeing this necessary part of the DENDRAL project.

In its early years DENDRAL research was sponsored by NASA and ARPA. More recently DENDRAL has been sponsored by the NIH (Grant RR-00612). The project depends upon the SUMEX computing facility located at Stanford University for computing support. SUMEX is sponsored by the NIH (Grant RR-00785) as a national resource for applications of artificial intelligence to medicine and biology.

SELECTED REFERENCES IN CHRONOLOGICAL ORDER

Included below are publications written for chemists, as well as selected papers for computer scientists. For this article, in which applications of AI are stressed, the most noteworthy publications are the 25 papers in the series "Applications of Artificial Intelligence for Chemical Inference." (Reference [14] is the first paper in that series.)

1. Lederberg, J., DENDRAL-64—a system for computer construction, enumeration and notation of organic molecules as tree structures and cyclic graphs (technical reports to NASA, also available from the author and summarized in [12]). (1a) Part I. Notational algorithm for tree structures (1964), CR.57029; (1b) Part II. Topology of cyclic graphs (1965), CR.68898; (1c) Part III. Complete chemical graphs; embedding rings in trees (1969).
2. Lederberg, J., *Computation of Molecular Formulas for Mass Spectrometry*, Holden-Day, Inc. (1964).
3. Lederberg, J., Topological mapping of organic Molecules, *Proc. Nat. Acad. Sci.* **53** 1 (1965).
4. Lederberg, J., Systematics of organic molecules, graph topology and Hamilton circuits. A general outline of the DENDRAL system. NASA CR-48899 (1965).
5. Lederberg, J., Hamilton circuits of convex trivalent polyhedra (up to 18 vertices), *Am. Math. Monthly* **74** 5 (1967).
6. Sutherland, G. L., DENDRAL—A computer program for generating and filtering chemical structures, Stanford Heuristic Programming Project Memo HPP-67-1 (February 1967).
7. Lederberg, J., and Feigenbaum, E. A., Mechanization of inductive inference in organic chemistry, in Kleinmuntz, B. (Ed.), *Formal Representations for Human Judgment*, New York: Wiley (1968).

8. Lederberg, J., Online computation of molecular formulas from mass number, NASA CR-94977 (1968).
9. Feigenbaum, E. A. and Buchanan, B. G., Heuristic DENDRAL: a program for generating explanatory hypotheses in organic chemistry, in Kinariwala, B. J., and Kuo, F. F. (Eds.), *Proceedings, Hawaii International Conference on System Sciences*, University of Hawaii Press (1968).
10. Buchanan, B. G., Sutherland, G. L. and Feigenbaum, E. A., Heuristic DENDRAL: a program for generating explanatory hypotheses in organic chemistry, in Meltzer, B. and Michie, D. (Eds), *Machine Intelligence* **4**, Edinburgh: Edinburgh University Press (1969).
11. Feigenbaum, E. A., Artificial intelligence: themes in the second decade, in *Final Supplement to Proceedings of the IFIP*68 *International Congress*, Edinburgh (August 1968).
12. Lederberg, J., Topology of Molecules, in *The Mathematical Sciences—A Collection of Essays*, Edited by the National Research Council's Committee on Support of Research in the Mathematical Sciences (COSRIMS), Cambridge, Mass.: The M.I.T. Press (1969).
13. Sutherland, G., Heuristic DENDRAL: a family of LISP programs, Stanford Heuristic Programming Project Memo HPP-69-1 (March 1969).
14. Lederberg, J., Sutherland, G. L., Buchanan, B. G., Feigenbaum, E. A., Robertson, A. V., Duffield, A. M. and Djerassi, C., Applications of artificial intelligence for chemical inference I. The number of possible organic compounds: acyclic structures containing C, H, O and N., *Journal of the American Chemical Society* **91** (1969) 2973.
15. Duffield, A. M., Robertson, A. V., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A. and Lederberg, J., Application of artificial intelligence for chemical inference. II. Interpretation of low resolution mass spectra of ketones, *Journal of the American Chemical Society* **91** (1969) 11.
16. Buchanan, B. G., Sutherland, G. L. and Feigenbaum, E. A., Toward an understanding of information processes of scientific inference in the context of organic chemistry, in Meltzer, B. and Michie, D. (Eds), *Machine Intelligence* **5** Edinburgh: Edinburgh University Press (1970).
17. Lederberg, J., Sutherland, G. L., Buchanan, B. G. and Feigenbaum, E. A., A heuristic program for solving a scientific inference problem: summary of motivation and implementation, in Banerji, R. and Mesarovic, M. D. (Eds.), *Theoretical Approaches to Non-Numerical Problem Solving*, New York: Springer-Verlag (1970).
18. Churchman, C. W. and Buchanan, B. G., On the design of inductive systems: some philosophical problems, *British Journal for the Philosophy of Science* **20** (1969) 311.
19. Schroll, G., Duffield, A. M., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A. and Lederberg, J., Application of artificial intelligence for chemical inference. III. Aliphatic ethers diagnosed by their low resolution mass spectra and NMR data, *Journal of the American Chemical Society* **91** (1969) 7440.
20. Buchs, A., Duffield, A. M., Schroll, G., Djerassi, C., Delfino, A. B., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A. and Lederberg, J., Applications of artificial intelligence for chemical inference. IV. Saturated amines diagnosed by their low resolution mass spectra and nuclear magnetic resonance spectra, *Journal of the American Chemical Society* **92** (1970) 6831.
21. Sheikh, Y. M., Buchs, A., Delfino, A. B., Schroll, G., Duffield, A. M., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A. and Lederberg, J., Applications of artificial intelligence for chemical inference. V. An approach to the computer generation of cyclic structures. Differentiation between all the possible isomeric ketones of composition C6H10O, *Organic Mass Spectrometry* **4** (1970) 493.
22. Buchs, A., Delfino, A. B., Duffield, A. M., Djerassi, C., Buchanan, B. G., Feigenbaum, E. A. and Lederberg, J., Applications of artificial intelligence for chemical inference. VI. Approach to a general method of interpreting low resolution mass spectra with a computer, *Helvetica Chimica Acta* **53** (1970) 1394.
23. Feigenbaum, E. A., Buchanan, B. G. and Lederberg, J., On generality and problem solving: A case study using the DENDRAL program, in Meltzer, B. and Michie, D. (Eds.), *Machine Intelligence* **6** Edinburgh: Edinburgh University Press (1971).
24. Buchs, A., Delfino, A. B., Djerassi, C., Duffield, A. M., Buchanan, B. G., Feigenbaum, E. A., Lederberg, J., Schroll, G. and Sutherland, G. L., The application of artificial intelligence in the interpretation of low-resolution mass spectra, *Advances in Mass Spectrometry* **5** (1971) 314.
25. Buchanan, B. G. and Lederberg, J., The heuristic DENDRAL program for explaining empirical data, in *Proceedings of the IFIP Congress* **71** Ljubljana, Yugoslavia (1971).
26. Buchanan, B. G., Feigenbaum, E. A. and Lederberg, J., A heuristic programming study of theory formation in science, in *Proceedings of the Second International Joint Conference on Artificial Intelligence*, Imperial College, London (September 1971).
27. Buchanan, B. G., Duffield, A. M. and Robertson, A. V., An application of artificial intelligence to the interpretation of mass spectra, *Mass Spectrometry Techniques and Applications*, in Milne, G. W. A. (Ed.), New York: Wiley (1971) p. 121.
28. Smith, D. H., Buchanan, B. G., Engelmore, R. S., Duffield, A. M., Yeo, A., Feigenbaum, E. A., Lederberg, J. and Djerassi, C., Applications of artificial intelligence for chemical inference VIII. An approach to the computer interpretation of the high resolution mass spectra of complex molecules. Structure elucidation of estrogencic steroids, *Journal of the American Chemical Society* **94** (1972) 5962.
29. Buchanan, B. G., Feigenbaum, E. A. and Sridharan, N. S., Heuristic theory formation: data interpretation and rule formation, in Meltzer, B. and Michie, D. (Eds.), *Machine Intelligence* **7** Edinburgh: Edinburgh University Press (1972).
30. Lederberg, J., Rapid calculation of molecular formulas from mass values, *Journal of Chemical Education* **49** (1972) 613.
31. Brown, H., Masinter, L. and Hjelmeland, L., Constructive graph labeling using double cosets, *Discrete Mathematics* **7** (1974) 1.
32. Smith, D. H., Buchanan, B. G., Engelmore, R. S., Adlercreutz, H. and Djerassi, C., Applications of artificial intelligence for chemical inference. IX. Analysis of mixtures without prior separation as illustrated for estrogens, *Journal of the American Chemical Society* **95** (1973) 6078.
33. Smith, D. H., Buchanan, B. G., White, W. C., Feigenbaum, E. A., Djerassi, C. and Lederberg, J., Applications of artificial intelligence for chemical inference. X. Intsum: a data interpretation program as applied to the collected mass spectra of estrogenic steroids, *Tetrahedron* **29** (1973) 3117.
34. Buchanan, B. G. and Sridharan, N. S., Rule formation on non-homogeneous classes of objects, in *Proceedings of the Third International Joint Conference on Artificial Intelligence*, Stanford, California (August 1973).
35. Michie, D. and Buchanan, B. G., Current status of the heuristic DENDRAL program for applying artificial intelligence to the interpretation of mass spectra, in Carrington, R. A. G. (Ed.), *Computers for Spectroscopy*, London: Adam Hilger (1973).
36. Brown, H. and Masinter, L., An algorithm for the construction of the graphs of organic molecules, *Discrete Mathematics*, **8** (1974) 227.
37. Smith, D. H., Masinter, L. M. and Sridharan, N. S., Heuristic DENDRAL: analysis of molecular structure, in Wipke, W. T., Heller, S., Feldmann, R. and Hyde, E. (Eds.), *Computer Representation and Manipulation of Chemical Information*, New York: Wiley (1974), p. 287.
38. Carhart, R. and Djerassi, C., Applications of artificial intelligence for chemical inference. XI. The analysis of C13 NMR data for structure elucidation of acyclic amines, *Journal of the Chemical Society (Perkin II)* (1973) 1753.
39. Masinter, L., Sridharan, N. S., Carhart, R. and Smith, D. H., Application of artificial intelligence for chemical inference XII: exhaustive generation of cyclic and acyclic isomers, *Journal of the American Chemical Society* **96** (1974) 7702.

40. Masinter, L., Sridharan, N. S., Carhart, R. and Smith, D. H., Applications of artificial intelligence for chemical inference. XIII. Labeling of objects having symmetry, *Journal of the American Chemical Society* **96** (1974) 7714.
41. Dromey, R. G., Buchanan, B. G., Lederberg, J. and Djerassi, C., Applications of artificial intelligence for chemical inference. XIV. A general method for predicting molecular ions in mass spectra, *Journal of Organic Chemistry* **40** (1975) 770.
42. Smith, D. H., Applications of artificial intelligence for chemical inference. XV. Constructive graph labeling applied to chemical problems. Chlorinated hydrocarbons, *Analytical Chemistry* **47** (1975) 1176.
43. Carhart, R. E., Smith, D. H., Brown, H. and Sridharan, N. S., Applications of artificial intelligence for chemical inference. XVI. Computer generation of vertex graphs and ring systems, *Journal of Chemical Information and Computer Science* **15** (1975), 124.
44. Carhart, R. E., Smith, D. H., Brown, H. and Djerassi, C., Applications of artificial intelligence for chemical inference. XVII. An approach to computer-assisted elucidation of molecular structure, *Journal of the American Chemical Society* **97** (1975) 5755.
45. Buchanan, B. G., Applications of artificial intelligence to scientific reasoning, in *Proceedings of Second USA–Japan Computer Conference*, American Federation of Information Processing Societies Press, (1975).
46. Carhart, R. E., Johnson, S. M., Smith, D. H., Buchanan, B. G., Dromey, R. G. and Lederberg, J., Networking and a collaborative research community: a case study using the DENDRAL program, in Lykos, P. (Ed.), *Computer Networking and Chemistry*, Washington, D.C.: American Chemical Society (1975), p. 192.
47. Smith, D. H., Applications of artificial intelligence for chemical inference. XVII. The scope of structural isomerism, *Journal of Chemical Information and Computer Science* **15** (1975) 203.
48. Smith, D. H., Konopelski, J. P. and Djerassi, C., Applications of artificial intelligence for chemical inference. XIX. Computer generation of ion structures, *Organic Mass Spectrometry* **11** (1976) 86.
49. Carhart, R. E. and Smith, D. H., Applications of artificial intelligence for chemical inference. XX. "intelligent" use of constraints in computer-assisted structure elucidation, *Computers and Chemistry* **1** (1976) 79.
50. Cheer, C., Smith, D. H., Djerassi, C., Tursch, B., Braekman, J. C. and Daloze, D., Applications of artificial intelligence for chemical inference. XXI. Chemical studies of marine invertebrates. XVII. The computer-assisted identification of [+]-palustrol in the marine organism cespitularia sp., aff. subvirdis, *Tetrahedron* **32** (1976) 1807.
51. Buchanan, B. G., Smith, D. H., White, W. C., Gritter, R., Feigenbaum, E. A., Lederberg, J. and Djerassi, C., 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* **96** (1976) 6168.
52. Varkony, T. H., Carhart, R. E. and Smith, D. H., Computer-assisted structure elucidation. Modelling chemical reaction sequences used in molecular structure problems, in Wipke, W. T. (Ed.), *Computer-Assisted Organic Synthesis*, Washington, D.C.: American Chemical Society (1977).
53. Smith, D. H. and Carhart, R. E., Applications of artificial intelligence for chemical inference. XXIV. Structural isomerism of mono- and sesquiterpenoid skeletons, *Tetrahedron* **32** (1976) 2513.
54. Hammerum, S. and Djerassi, C., Mass spectrometry in structural and stereochemical problems. CCXLV. The electron impact induced fragmentation reactions of 17-oxygenated progesterones, *Steroids* **25** (1975) 817.
55. Hammerum, S. and Djerassi, C., Mass spectrometry in structural and stereochemical problems. CCXLIV. The influence of substituents and stereochemistry on the mass spectral fragmentation of progesterone, *Tetrahedron*, **31** (1975) 2391.
56. Dunham, L. L., Henrick, C. A., Smith, D. H. and Djerassi, C., Mass spectrometry in structural and stereochemical problems. CCXLVI. Electron impact induced fragmentation of juvenile hormone analogs, *Organic Mass Spectrometry* **11** (1976) 1120.
57. Dromey, R. G., Stefik, M. J., Rindfleisch, T. and Duffield, A. M., Extraction of mass spectra free of background and neighboring component contributions from gas chromatography/mass spectrometry data, *Analytical Chemistry* **48** (1976) 1368.
58. Carhart, R. E., A model-based approach to the teletype printing of chemical structures, in *Journal of Chemical Information and Computer Sciences* **16** (1976) 82.
59. Varkony, T. H., Carhart, R. E. and Smith, D. H., Computer assisted structure elucidation, ranking of candidate structures, based on comparison between predicted and observed mass spectra, in *Proceedings of the ASMS Meeting*, Washington, D.C. (1977).
60. Buchanan, B. G. and Smith, D. H., Computer assisted chemical reasoning, in Ludena, E. V., Sabelli, N. H. and Wahl A. C. (Eds.), *Computers in Chemical Education and Research*, New York: Plenum Publishing (1977), p. 388.
61. Smith, D. H. and Carhart, R. E., Structure elucidation based on computer analysis of high and low resolution mass spectral data, in Gross, M. L. (Ed.), *Proceedings of the Symposium on Chemical Applications of High Performance Spectrometry*, Washington, D.C.: American Chemical Society (in press).
62. Mitchell, T. M. and Schwenzer, G. M., Applications of artificial intelligence for chemical inference. XXV. A computer program for automated empirical 13C NMR rule formation, *Organic Magnetic Resonance* (forthcoming).
63. Buchanan, B. G. and Mitchell, T. M., Model-directed learning of production rules, in Waterman, D. A. and Hayes-Roth, F. (Eds.), *Pattern-Directed Inference Systems*, New York: Academic Press (forthcoming).
64. Schwenzer, G. M., Computer assisted structure elucidation using automatically acquired 13C NMR rules, in Smith, D. (Ed.), *Computer Assisted structure Elucidation*, ACS Symposium Series, vol. 54:58 (1977).
65. Varkony, T. H., Smith, D. H. and Djerassi, C., Computer-assisted structure manipulation: studies in the biosynthesis of natural products, *Tetrahedron* (in press).