A History of Medical Informatics

Edited by

Bruce I. Blum
The Johns Hopkins University

Karen Duncan
Health Information Systems

ACM Press
New York, New York

1990

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts ▪ Menlo Park, California
New York ▪ Don Mills, Ontario ▪ Wokingham, England
Amsterdam ▪ Bonn ▪ Sydney ▪ Singapore
Tokyo ▪ Madrid ▪ San Juan

11/25/87
F270

How DENDRAL was conceived and born.

Joshua Lederberg
Rockefeller University
New York, N.Y.

ACM Symposium on the History of Medical Informatics
National Library of Medicine November 5, 1987

As agreed with your organizers, this will be a somewhat personal history. They have given me permission to recall how I came to work with Ed Feigenbaum on DENDRAL, an exemplar of expert systems and of modelling problem-solving behavior. My recollections are based on a modest effort of historiography, but not a definitive survey of and search for all relevant documents. On the other hand, they will give more of the flow of ideas and events as they happened than is customary in published papers in scientific journals -- accounts so dry that Medawar lugubriously called them fraudulent {43}; cf. Merton & Zuckerman {44, 45, 61}. These authors point out that the standard scientific publication is narrowmindedly devoted to the context of justification. The DENDRAL effort (along with much of medical informatics) is dedicated to discovery: should we use a different standard for its history?

This is a first effort at historical research and informed consensus on the origins of DENDRAL; and we all understand the limitations of a personal account -- especially about what others were thinking at a given moment. Built into the phenomenon of history, as soon as enough time has passed to enable some detached judgment, the evidence becomes frail, and we become vulnerable to the myths we create. Understanding all of these limitations, I will no longer qualify every remark: it should be implicit that each is "to the best of my recollection / or/ as best as can be inferred from the fragmentary documentary record".

I will assume you are generally familiar with DENDRAL, and will concentrate mainly on material not found in the published papers, especially as there is a comprehensive synopsis of its postnatal productions {41}. My story will focus on the period up to the recognition that what we were working on was a knowledge-based system (ca. 1971).

As computer science is not my primary profession, my relationship to it has been more episodic; and I can more readily isolate how I came to take some part in it, at Stanford from 1962 - 1978, mainly in very close collaboration with Ed Feigenbaum, Bruce Buchanan, and a host of others. My central scientific commitments have been to molecular genetics, starting when I was a 20-year old medical student in 1945 {38}. At Columbia and then at Yale, I worked on the genetics of bacteria, a specialty which converged with the function of DNA as genetic information. My first academic appointment was at the University of Wisconsin from 1947 - 1958; then I went to Stanford in 1959 to take part in the reconstruction of its School

of Medicine (formerly in San Francisco) at the Palo Alto campus. My intended role was to found a new Department of Genetics; I had no plan to be working with computers. Fate dictated otherwise: I met Ed Feigenbaum in 1963. Then, promptly after he moved from Berkeley to Stanford faculty in early 1965, we initiated the collaboration that became the DENDRAL project.

These were hardly random events: I go back a few years to pick up the relevant premonitory strands.

Figure 1
Conceptual and Experiential Threads
Leading to the Dendral Project [JL view]

1) 1937-43. Leibniz dream
Logic & Axiomatic Method -- studies in Columbia College

2) 1941, 53, 62. Computer hardware: desultory exposures.

3) 1947 ff. Information-theoretic formulations in genetics

4) 1953 ff. Introspections about the history of bacterial genetics.

5) 1960. Instrumentation development for Mars exploration: NASA

6) 1955, 59, 61, 63. Meet Minsky, Djerassi, McCarthy, Feigenbaum

(In every biographic-historic account in science, one seeks an interplay of personality, ideas, institutional setting, and other externalities.)

---

1) Starting in grade school, I had fantasies that echo Leibniz' dream (see {13}) of a "universal calculus" for the "alphabet of human thought", that all of knowledge might be so systematized that every fact could be tagged with a code. Cf. Mortimer Adler's Propaedia {1}.

New York City in the late 1930's offered wonderful encouragement to self-improvement through education, in my own case at Stuyvesant High School and at the New York Public Library. There I was fascinated with the Dewey Decimal System, which was so helpful in locating the books: if I could but memorize that, it would be proxy for mastery of all the knowledge it classified. In those days, taxonomy dominated biological teaching too. (I will not detain you now with the perils of misplaced confidence in low- dimensional, or insight-free knowledge. They need to be remembered when we try to extract "knowledge" from an expert, measure how much we have, and so forth.)

Although I was committed from a very early age to a career in experimental biology, while in college I was eager to have some understanding of the epistemological roots of science, and enrolled in several courses in logic and scientific method. At Columbia, I was fortunate to have some personal exposure to members of the philosophy department: Ernest Nagel, Justus Buchler, and James Gutmann. With their help, I read George Boole and Whitehead & Russell {58}, and tried to follow J.H. Woodger in his Axiomatic Method in Biology {59} -- an effort to express what was then known of genetics and embryology in the formalisms of relational calculus. Our factual knowledge was sparse enough; but apart from that, I wondered if we really understood our assertions when they were expressed in the jargon of empirical biochemistry. Axiomatic reformulations of biology are only just now returning to the scene {3, 54, 57, 47}. They make the intellectual demand of coping both with the formal logic and the molecular biology.

That background gives some flavor of the retrospection about method that has entered my thought sometimes during, often after, my experimental research projects. That would precondition me to look to AI as a way of expressing my philosophy of scientific method, a perspective eloquently stated by Lindley Darden {66}.

2) My first encounter with a "computer" was in 1941, in a lab for high school students sponsored by IBM {23}. My own instrument was a microscope; but one of my fellows was making innovative improvements on a punch card sorter/tabulator. It was an impressive manifestation of an electro- mechanical automaton, one that could certainly calculate more reliably than I could. It looked like fun. After the war, there was some publicity about the electronic machines, which I read at the level of Scientific American or Science. But my own next step was the IBM 602A, on which I practised in Fred Gruenberger's course at Wisconsin, in 1953, in order to get some concept of programming, albeit on a plugboard! One could do statistics on this machine, as did some of my colleagues in applied genetics; but I had no comparable excuse to play with it.

3) That postwar period also saw the elaboration of information theoretic formulations of genetics. We were starting to say that genes encoded the information needed to specify protein structure. {14, 51} This style of thought and expression became more explicit in the period after 1953 {25} with the recognition of the implications of the Watson-Crick molecular structure of DNA {22}. It would be backward for anyone in my field to ignore this way of looking at the biological world. Then, Marvin Minsky came to see me in 1955 at Wisconsin at the behest of some mutual friend to discuss automata. I am sure I had already heard of some of his own work.

4) My own laboratory research was a very mixed bag of theoretical formulation and empirical encounter. I had been extraordinarily lucky on several occasions - but I didn't want to be a hostage to chance: should there not be a more systematic strategy of problem formulation? And if one could do that, problem-solving might be a throwaway. Serious questions about the rational direction of science were invoked around an examination of why genetic recombination in bacteria had not been explored 40 years earlier. {24, 60}.

5) Starting with the observation of Sputnik, and a conversation with J.B.S. Haldane in Calcutta in November 1957, {67} I had set out to assure that fundamental biological science

was properly represented in the programs of space research that were just emerging. I had met him on a date that was both a lunar eclipse and the 40th anniversary of the Soviet October Revolution (almost precisely 30 years ago). He taunted me with the prospect that we might see a red star (a thermonuclear explosion) on the moon during the eclipse. At best, that was a striking metaphor for the danger that scientific interests would be totally submerged by the international military and propaganda competition. They have never gained first priority; they might have been totally excluded. My own efforts were merely advisory and critical until 1960, after NASA had organized a Life Sciences Research Office and asked me to establish an instrumentation laboratory at Stanford. With Elliott Levinthal's able technical direction of the lab, we became actively involved in the conceptual design of approaches to test for life on Mars, at such time as there might be a mission. I know most of my colleagues thought that would be well into the 21st Century, as we were a decade short of the lunar landings. But the possibility of finding another branch of evolution was of such compelling scientific interest, the stake was worth odds I knew were very long.

Both the internal activities of the Instrumentation Research Laboratory, and design discussions with the engineering managers of spaceflight missions (principally at Caltech's JPL) brought us into intimate conversation with technology of automation, process control, communications and computer management. Furthermore, mass spectrometry soon emerged as a technology of choice for chemical analysis. It has enormous sensitivity, selectivity, and independence of prior bias as to the molecular species expected {33}. As we shall see, it also offered some special opportunities and challenges in computation.

In 1961, I was also invited to serve on a PSAC panel on the management of scientific information. Our report {50} gives modest support to the implications of computer technology, along with "reproducing and microphotographing equipment" for information storage and retrieval. However, I had become acquainted with Eugene Garfield, the inventor of Current Contents, and had helped him set up a trial run of the Science Citation Index in the field of genetics {36, 19}. That experience (with its overtones of the classification of knowledge for purposes of retrieval) was an early success in the use of computers in support of scientific research.

By now, I concluded that I would have to learn much more about computers, at a hands-on level. The opportunity was engendered by the evangelistic efforts of Al Bowker and George Forsythe to establish an intellectual and technical base for and broaden interest in computers throughout the Stanford campus. In company with the development of a new division, then department, of Computer Science, and of a campuswide computer center, elementary programming courses were organized. I enrolled in the BALGOL (Burroughs Algol) course given by Bob Oakford, over the summer, 1962. This had much of the flavor of a course in English for fresh immigrants, the class having a very broad distribution of age and of academic status, specialty and sophistication.

I quickly succumbed to the hacker syndrome, (and have suffered episodic relapses over the last 25 years). This was reinforced by the relentless rectitude of the machine in rejecting my errors - always so obvious in retrospect. "Next time, next time I will master the **** system!" Well, I did shortly become reasonably proficient (eventually, in a range from assembly to higher level languages) mostly out of determination not to be made a fool. In

those days, we had a B220 - which would match a fairly feeble PC today - as the first campus machine. Its operating system would accept decks of punched cards in serial batch mode, with output either from the printer or new punched cards. The usual turn- around time was about 12 hours. If you got to the computer room around midnight, you might get another pass by 2 A.M. The democracy and night-owl ambience of the batch system was a social mixer for several enthusiasts from wide-ranging disciplines. (I particularly recall Tony Hearn, who was starting his symbolic algebra system, REDUCE, on the IBM 7090). The impedance of a one-pass per day turnaround certainly did filter out all but the most enthusiastic. You also spent a lot of energy trying to simulate the machine in your own thought, in contrast to the casual, experimental mode -- "Let's see if this works" -- of today's interactive systems. This mode has unquestioned advantages; but it may weaken programming as a teaching discipline for logical rigor (except insofar as pure, unremitting failure teaches mainly discouragement.)

Our first applications included some that are pertinent to medical informatics, but not to DENDRAL, in areas of genetic epidemiology {6}, including a contract to produce the childspacing report on the 1960 census. Bob Tucker was instrumental in sustaining our effectiveness and sanity through that experience. When we discovered that "children" of some mothers were delivered at 3 month intervals, I again learned the familar GIGO lesson, and a healthy skepticism for mass data repositories. Massive numbers do not take the place of quality controls on individual data entries. Some other inquiries, e.g. of intercorrelations of season of birth and birthweight with postnatal outcomes, taught us the difficulties of removing all the confounding factors. The usual "socio-economic status indicators" do not begin to exhaust the vagaries of stratification of human behavior.

1962 also marked the recruitment of John McCarthy to Stanford. We met around the computer room, soon discovered we had a common friend in Marvin Minsky. I had read Marvin's article on steps toward artificial intelligence in the January 1961, special issue on computers of the Proceedings of the Institute of Radio Engineers {46}, the first issue I received as a newly enrolled member (having joined at the urging of Lloyd Berkner, chair of the Space Science Board). That article and McCarthy's intellect and excitement gave me a sense of tangibility of the possibility of engaging in AI research. When he showed me his new DEC PDP-1 and its interactive CRT displays (viz., Spacewar) I reached the conviction that "computers were going to change the whole style of scientific investigation". This was not going to happen with card deck data entry.

We soon conspired in various projects to try to enhance the interface of computers with medical science. The most ambitious of these was an effort to attract Marvin Minsky to join the faculty of Stanford Medical School; but unhappily for us he decided to stay at M.I.T. We also began to talk about bringing interactive computing, via time-sharing, to Stanford, along the lines of Project MAC, which John had helped to design at M.I.T. These discussions ultimately led to ACME and SUMEX, the first community-access time-shared systems at Stanford, as we discovered that the NIH was able to fund research resources for health research through its Biotechnology Resources Branch. McCarthy's PDP-1 also led us to emulate it as a laboratory interface computer, and our IRL signed on as one of the test sites for the new LINC (laboratory instrumentation computer) whose development NIH was sponsoring. Lee Hundley and Nick Veizades provided the indispensable hardware engineering expertise to enable us to master this marvelous new machine. The LINC was, of course, the

forerunner of the DEC PDP-8, and in turn of the PC revolution.

Meanwhile, the IRL was getting more actively involved in mass spectrometry. Carl Djerassi had come to the Chemistry Department in late 1959, and we had developed a close personal and professional association around his academic research as well as his continued research direction of the Syntex Corporation. (Upon the company's relocation from Mexico City to new laboratories at Stanford Industrial Park in 1961, he asked me to advise on its establishment of the Syntex Institute of Molecular Biology.) He was an accomplished mass spectrometrist; and of course I leaned very heavily on him for the elaboration of this technology for space applications. Conversely, he knew nothing about computers, and I was eager to find helpful applications in the zone of our common interest. The IRL began to work on using the LINC to manage the formidable data management problems of real-time gas-chromatography mass- spectrometry {52}. One central problem was the efficient translation of mass numbers to molecular formulas.

As I reexamine that arithmetic play, it reveals some premonitions of the later work. So I will expand on it beyond the intrinsic worth of the solutions {29}.

The mass spectrometer is an instrument that converts molecules of a sample material into ions that are accelerated and measured one by one. Further, by a combination of magnetic and electrostatic fields, each ion can be sorted by its mass number. For the initial discussion, we will consider only the molecular ion, ignoring further processes of fragmentation. At low resolution, we take atomic masses as integers (H =1; C = 12; N = 14; O = 16; etc.) If we find a mass number of 14, this might be composed of H(14), C + H(2), or N. H(14) is a monstrosity: we have valence rules (H ~ 1; C ~ 4; N ~ 3; O ~2) that limit how many atoms can be bonded to a given atom. The ambiguity already seen at m = 14 is of course greatly multiplied in real cases, like m = 3675, a number which reflects the bounds of current instrumentation. Our first problem is to calculate all the compositional isomers consistent with a given mass number. At this level, it is a knapsack, or change-making problem: finding all the ways coins of different denomination can be combined to add up to a given sum. In non-negative integers, this is a diophantine equation, viz. we seek all the solutions, (i.e. compositions in h, c, n, o) of:

$$h + 12c + 14n + 16o + \ldots = m.$$

The brute force approach is a set of nested iterations,
for (h = 1; h <= Z; h++); for (n = 1; n <= Z; n++) .... m' = h + 14n + ...
and test the m' sums for a match to m. One simplification is to augment m, m" = m + k == 0 mod 12. We then eliminate c and find solutions in h, n, o that satisfy (h+k) +14n + 16o == 0 mod 12. I would be interested to learn of deeper analytic approaches to the problem. For online computation, one thinks of constraining Z, at least by the mass still unassigned in each loop, to reasonable bounds. It transpired that the valence considerations also set constraints on possible values of h; and other tricks allowed still further pruning of the tree generated by the nest, greatly shortening the computation.

Prior aides to mass spectrometrists had been published tables (embracing 570 pages in print) that reported the compositions sorted by m, from about 1 to 500, with n and o no greater than

6 {4}. A full set of tables for m up to 1000 would take about 10,000 pages of fine print.

In reality, the masses of individual nuclides are not integers (subject to the so-called nuclide packing fraction), and we have
H = 1.0078252
C = 12. (by definition)
N = 14.003074
O = 15.994915

With a high resolution mass spectrometer, a given ion might be reported as 718.374 +/- .006 . Hundreds of compositions would match 718 in integers. One should use the fractional mass (.374) as equally important information in limiting the search. We no longer have an equation in integers, owing to the instrumental error. Nevertheless, various arithmetic tricks were devised that took account of valence rules, plausibility of composition, the negative and positive packing fractions of O and N, and the abnormal proportional discrepancy of H, to keep the search down to a manageable scope. For paper and pencil work (in 1964) this was embodied in a handbook of some 50 pages, in which one could quickly look up the "mass defect" of numbers classified by residues modulo 12. {26} Even that small book was later {35} obsoleted by an algorithm that depended on a one-page table with just 72 non-zero entries, and a few arithmetic steps easily done on a 4-function hand calculator. This algorithm has served well in the data system built for a GC-MS (gas-chromatograph/mass-spectrometer) designed around a MAT-711 MS {62}, and the LINC computer. It has evidently been independently rediscovered in China, {63}. By now, however, most machines are coupled with data processors that are oblivious to such economies. (And mass spectrometrists no longer give much thought to the arithmetic of this problem.)

The main point is self-evident: contextual information could be incorporated early into the combinatorics, and reduce a blind generate-and-test search by very large factors.

We turn now to the larger frame of chemical analysis. Molecular ions are important targets for mass spectrometry; in the ideal case they can give unambiguous compositional formulas. Of course, they tell nothing of the topological connectivity of the constituent atoms. To illustrate with a trivial case, C(2) H(6) O has a mass of 46.041866 but this does not distinguish methyl ether (CH3-O-CH3) from ethanol (CH3-CH2-OH), a medically significant matter! Within the mass spectrometer, however, the molecular ion also breaks up into a set of fragments (according to reasonably well understood rules). The spectrum is the array of these fragments, revealed by their mass numbers. It is often an absolutely distinctive fingerprint, diagnostic of a specific structural isomer (as the molecular ion mass number is of the composition).

The elementary problem of inferring composition from molecular mass now well-solved, could we take the next step: model the chemist's inferential procedure in finding the structure from the spectrum?

How to represent organic molecular structures in graphs, and then their dissection into subgraph fragments, as occurs in the mass spectrometer, became my task for 1963-64. Emile Zuckerkandl, an associate of Linus Pauling, also visited my lab. during this interval. We

started some of the first statistical studies on amino acid sequences of proteins, looking for hints of non-random regularities within sequences, and unsuspected evolutionary relationships among different ones. This is a substantial industry today {40}; there were not enough published data in 1963 to offer more than a few tantalizing hints.

6) All this was then the ideological context of my meeting Ed Feigenbaum on April 6, 1963. This was a Saturday meeting that Karl Pribram had organized at the Center for Advanced Studies in Behavioral Sciences on computer models of thought. John McCarthy, Ken Colby and several others were also present. I told Ed how I was groping for ways to represent chemical structures; he was already on the lookout for problem areas in science to which to bring his background on mechanized problem-solving. We stayed in good contact: I have a signed copy of "Computers and Thought" {15} dated 1/17/64.

During 1964, I completed the preliminary graph-theoretical work on representation of organic molecular structures. {30, 32, 28}. That had entailed going back to the elementary graph theory of the 19th century for canonical forms of tree structures {21}. Fortunately, George Polya had done some important work on generating functions in 1936 {49}, and was most generous in his advice about that older literature. When it came to cyclic graphs, I had a particularly entertaining time, almost at the level of recreational mathematics {31, 34}.

See Figure 1: Cyclic Graphs {31}

For a century after the conception of organic molecules as ensembles of connected atoms subject to structural isomerism (Berzelius, 1831: {48}; Crum Brown, Butlerov and Kekule in the 1860's: {20}) no more than desultory attention was given to the formal mathematics of their representation as graphs, to the potentialities of a connection between Hamilton circuits, convex polytopes and organic molecules {53}. It is hard to account for such an egregious lapse, one possibly another candidate for the label of a "postmature discovery" {60}. The topology was perhaps too elementary to engage the interest of serious mathematicians -- but there are still intractable problems in the enumeration of cyclic graphs (after automorphisms!). Related issues, like the notorious map-coloring problem, illustrate the still primitive state of analytical approaches to the taxonomy of graphs. Cayley {12} made a stab (a fallacious one) at the enumeration of the hydrocarbons; this was improved upon by Henze & Blair in the 1930's {5}. In the mid-1960's, Balaban and his colleagues in Romania began their extensive investigations independently of the work at Stanford {2}.

Chemistry has then developed a taxonomy of its own structures that has no coherent mathematical theme. It is full of colorful but trivial names that give no structural information: a few eccentricities like "windowpane" for 4 fused rectangles are a partial exception. A formidable burden in learning chemistry is the enormous amount of rote memorization that is entailed in associating names like butane, cholestane, cytosine, melezitose, xanthopterin -- there are tens of thousands of these -- with graphic representations. One may think of these as the passwords for admission to the secret society; they do deter many a student, and they may also impair a critical analytical perspective about organic chemistry. These pictures also have formal names, but the nomenclatural handbook that gives the rules for their translation occupies a thick book, mostly the idiosyncratic cases.

Dendral-64 is a set of reports to NASA {30, 32, 28} that outlines an approach to formal representation of chemical graph structures, and a generator of all possible ones. Acyclic structures (trees) were readily tractable. Cyclic ones can be dealt with, mainly with the help of a few tricks that rely upon an empirical enumeration of the underlying vertex graphs – this is feasible within the bounds of practical chemistry -- which is analytically unsatisfying. It helped to learn about Hamilton Circuits of graphs (paths that touch each node just once) {27}, since the enumeration of these, and the elimination of automorphisms are greatly simplified. When it came to the implementation of DENDRAL for (typical) organic molecules with imbedded rings, Harold Brown, Larry Hjelmeland and Larry Masinter provided the group-theoretic general mathematical solutions to these perplexities {10, 8, 7}. A few molecules have been constructed precisely because they defy some constraints of topological simplicity -- e.g. topological planarity, namely their connection graphs cannot be drawn on the plane without bonds crossing; as exceptions they make history and can be dealt with as such. {31, 55}

The DENDRAL generator was then designed so that only one canonical form of a possible automorphic proliferation is issued, greatly pruning the space of candidate graphs. This was the essential prerequisite for an AI program that could manage the generator and confront it with information derived from the mass spectrum. But I had no idea how one would go about translating these structural concepts into a computer program, nor whether this would be computationally feasible with available hardware. Even more telling, I had only second-hand access to the field of AI and barely knew how to relate these conjectures to the systematic approaches that were emerging {15}. It was fortunate indeed that Ed Feigenbaum came to Stanford just at this time: we promptly got together again and organized the collaboration that became the DENDRAL project.

Ed now deserves equal time in presenting his personal prehistory. Some of his oral history has appeared in McCorduck's book {42}. In addition, I have a few of his own words, excerpted from an electronic message:

Date: Thu 8 Mar 84 00:22:01-PST
From: Edward Feigenbaum <FEIGENBAUM@SUMEX-AIM.ARPA>
To: lederberg@SUMEX-AIM.ARPA
Subject: our history {*referring to some private notes*}

Josh, all of what you have written accords with my memory of things we discussed in 1965 as we quickly got to know each other better.

Your mention of mass spectral analysis as a problem domain in which we should work came as an answer to a question I posed you. I had decided that I wanted to work on constructing models of EMPIRICAL INDUCTION IN SCIENCE, within the methodology that I had learned from Newell and Simon, i.e. work on a concrete task domain, not in the abstract. So I needed a concrete task domain. You said you knew of one that contained the essence of the empirical induction problem, that you had been working on it for a while, you even had a computational algorithm underlying it (which immediately made me think: aha, legal move generator as in chess-playing programs). ALL of this conversation (embryonic research

planning) took place AFTER I arrived at Stanford Jan 1, 1965, but I remember that I would not have sought you out for advice on the aforementioned puzzlement had I not met you earlier [April 1963] and learned of your interest in machine models of thinking. Recall: there were very very few people to talk to about machine models of thinking at Stanford in early 1965.

We didn't just "bump into each other" as in "lucky accident". You weren't at the [April 63] meeting by "lucky accident". I didn't decide to work with you on the mass spec analysis problem because it was of general intellectual stimulation. You had a definite interest in AI and I had a definite interest in hypothesis- formation/theory-formation. (Incidentally, do you remember how we went round and round on whether to deign to call what DENDRAL did "theory formation"? We decided on "hypothesis formation" to distinguish the case of one spectrum being explained by one (or a few) structures. We reserved the use of the term "theory formation" for a later date, for a more general approach, and decided to use it in describing Meta-DENDRAL (many spectra--> rule set).

P.S. Some things do appear to be "lucky accidents". It would appear to be a genuinely lucky accident that I chose to go to college at Carnegie Tech (an accident of Westinghouse scholarships and my family's financial condition), and a lucky accident that I met Herb Simon through Jim March, and that Herb paid attention to me, and that the Logic Theory program was invented while I was still a Carnegie Tech undergrad and that I was taking a seminar from Herb at the time of its invention. One level deeper: I was an ACTIVE RECEPTOR SITE re the idea of a computer. I had never even heard of an electronic digital computer before Herb handed me an IBM 701 manual, but... I had been entranced by mechanical calculator machines in high school and before. My father was an office manager/ accountant and owned a giant,heavy Monroe or Marchant calculator. I became an expert on its use. I even remember dragging it with me miles on the bus to Weehawken High School, heavy as it was, just to show off my skill with this marvelous technology that no other kid in the high school knew anything about. So when Herb gave me that manual, he was projecting me five or six orders of magnitude into a territory I was already fascinated with. It was also very fortunate that my introduction to the electronic computer was via the computer as general symbol manipulator (Herb never mentioned that it was anything BUT that) and that my introduction to programming was via IPL 1 and 2. (I might add that such a sophisticated early view, given to me by Herb and Al Newell, has taken away most of the awe from later developments; everything else has seemed to be "merely" extensions of the great inventions and discoveries of the 1956-59 period) "

END OF MESSAGE ---------

It is now Spring 1965, and our project is concretely launched. We began to think of and label it "Heuristic Dendral" to mark it as a refinement of the "Dendral Algorithm" for generating all the feasible structures. Ed and Richard Watson circulated a bulletin {16} "An initial problem statement for a machine induction research project" to graduate students in Computer Science; but it was to take a few years of slow accretion to organize a cadre of collaborators. One of our first, Research Associate Georgia Sutherland did a fabulous job on the formidable task of converting the concepts of DENDRAL-64 into a LISP program, interleaving its production with that of a baby: an early prototype of telecommuting. Her first report was issued February 1967 {56}: we finally had a working program with which we could all experiment

with heuristics and other measures to bring its performance to practically useful levels. The choice of LISP was originally mandated by the good match of its data structures to trees, to the sparse connection tables of chemical structures. But the memory and bit crunching requirements were of course monumental -- it's a wonder we got as far as we did with the hardware of the time. I used to remark, in arguments with ideologues, that in the last analysis it was the programming environment of INTERLISP that was its key advantage.

We were fortunate to have continued support from NASA and from DARPA to continue these explorations. We had quickly found that the campus IBM 7090 had too little memory to support our LISP programs; and we were eager to move to more interactive systems for program development. In 1966, our DARPA sponsorhip gave us access to the Q-32 time-sharing system at System Development Corporation (Santa Monica) with a 100 baud teletype interface. My first experience with remote, time-shared hacking was a happy vision of future improvements. Then, John McCarthy acquired a DEC PDP-6, and we approached something closer to the modern era. Bruce Buchanan joined our group, and we had great benefit from his philosophical perspective, patience, insight and administrative acumen, not to mention a lot of hard work in implementation of the software. We gained more and more collaborators, including the explicit involvement of Carl Djerassi and his associates as founts of authentic chemical expertise {64}. As our reports began to appear in refereed chemistry journals, we eventually bolstered our confidence that we were contributing to the scientific domain, as well as to system- building -- a point about which some of my colleagues had been skeptical. Broader access to these computer applications became possible with the help of the NIH-supported computer resources: ACME, a general time-sharing system for the Stanford Medical School, and SUMEX-AIM, a national resource to support research in artifical intelligence in medicine {11}. The history of SUMEX would take us into many lessons about the social organization of cooperative intelligence. However, as this account is now moving into a time of documented history and numerous publications {41}, I omit many details.

Largely owing to the contributions of Carl Djerassi and his colleagues in natural products chemistry, the program was crammed with chemical information. It was becoming an effective assistant in the analysis of spectra and other analytical information. Buchanan recoded DENDRAL's knowledge of mass spectrometry, originally embodied in a collection of LISP procedures, into a table of explicit rules separated from the internal operations of the system. This redesign to facilitate augmenting, validating and editing the informational (i.e. rule) base, was a paradigm shift later to become the standard for expert systems. Balky resistance of the program to input of new ideas remained the limiting factor in its elaboration. At every weekly group meeting, a dozen new ideas would come up: but we knew that each one would take weeks to implement in tested software code, just to test it. Natural intelligence still enjoys a flexibility of hierarchical planning yet to be achieved in machine emulations {17}.

Throughout this time, we would ask ourselves the nagging question: was the growing pragmatic success of DENDRAL in solving chemical problems teaching us anything about artificial intelligence? Had we simply crafted a special case, accumulating a hefty store of chemical knowledge from several experts? We did see the need for -- and Bruce Buchanan made a stab at -- a self- learning system, whereby META-DENDRAL could induce its own rules (as the chemist does) by introspecting about concrete data inputs of mass-spectral

fragmentation of molecules of known structure. This showed real promise {10}, but was impeded by the insufficiency of computer horsepower needed when DENDRAL itself had to be invoked repetitively to test every new rule candidate induced.

We never got a grip on one idea that I hope to return to someday. DENDRAL is remarkably neatly structured (as implied by its name) as a generator of trees of candidate structures {39}.These can easily number in the billions or more, in practical cases: the efficiency of the program depends on the pruning of impossible or implausible cases, as early as possible; preferably large branches at a fell swoop. The order of application of the shears can have a large effect. To give a stupidly trivial example, if N (nitrogen) is absent, we don't generate molecules that may contain N, then retrospectively eliminate each of those twigs. We gave some forethought towards optimizing the sequence of shears; but we know this will be case-specific, sometimes in ways we have dificulty predicting. We should build in recurrent introspection about the shearing sequence, make that a specific planning objective, and experiment with it from time to time. These considerations (I called it Theta-DENDRAL for reasons not recalled) would have broad generalizability to rule-based systems: the sequence of invocation of rules is often totally inaccessible to the user, and rarely if ever (as far as I know) is it dynamically regulated.

We did do some work on the interesting tradeoffs between storing memory of all partially completed branches, vs. regenerating them as needed. Finally, we had many discussions of the desirability of learning to read expertise from the world's published books, to bypass the oral tradition. The ultimate fantasy was to attach a high-order DENDRAL directly to a mass spectrometer, learning directly from Nature.

I wish I had the documentation, but I have an image of a conversation when I was pressing Ed about the limitations of DENDRAL as general intelligence: he responded with the illumination that I may paraphrase: "That's exactly the point! Knowledge, not tricks or metaphysical insight, is what makes the program effective -- and that itself is an insight of general import." That is why I remark, we were trying to invent AI, and in the process discovered an expert system. This shift of paradigm, "that Knowledge IS Power" was explicated in our 1971 paper {17}, and has been the banner of the knowledge-based-system movement within AI research from that moment. Cf. Alan Newell's comments, {65}.

Shortly thereafter, Bruce Buchanan and Ted Shortliffe initiated the MYCIN project {9}. As Alan Newell remarked in his preface to {9}, MYCIN had no pretensions to deep theoretical structure of chemistry, none to outdoing the experts, but only to conveying that expertise as advisory to the general practitioner in optimizing the prescription of antibiotics. Their coding of MYCIN gave a fresh start to the design of rule-based systems that could be readily transported to other applications.

The published documentation after this time is quite rich, and I will refer to that for further historical development. Time now for the numerous morals of the story.

Most problematical is the public utility of private autobiography. But biography remains very popular, albeit the main lesson may be the very idiosyncrasy of personal history and character. Worse than no history would be a false conception of it, that it has rigorous rules. As my tale

shows, chance does play an enormous role in bringing together people, ideas, situation in a productive way. Were we lucky? Who knows what the alternatives might have been?

One lesson of personality should be brought out, especially when the media enjoy characterizing the scientific enterprise as rapacious competition and selfishness. The fraternity that came out of the DENDRAL effort was a high in my life experience, matching the gratifications of scientific excitement and (perhaps belated) recognition. One is not always so lucky in one's colleagues; but we should not glamorize and confuse the pathology as the standard.

The project also dramatized the values of electronic communication in project management. Although we certainly met informally from time to time, most of our serious communication (be it a few yards down the hall) was by electronic mail. In this way, innumerable proposals and drafts could be posted on common bulletin boards, and subjected to consensual review, often through scores of cycles of reiteration. Distance was no consideration, courtesy of the ARPANET, and communication could be sustained during momentary travel, and collaboration continued when participants moved. (This ms. will of course be shared between the Rockefeller and Stanford.) Such draft texts, program modules and outputs needed critical scrutiny of a kind that is only possible when one has a copy of the file to work on from one's own terminal. I went so far as to characterize this mode of communication "The New Literacy", and I meant it {37}. Databases should not be thought of as static, final repositories but as bulletin boards, subjected to dynamic critical attention by the entire knowledgeable community.

Stanford University, in the 1960's, was a fortunate place to be for the pursuit of scientific innovation, and equally for a highly interdisciplinary program. Computer science, medical science, chemistry, were all in a surge of rapid expansion and new opportunity. If there were no specific facilitations for these kinds of interactions, nor were there rigid impediments. There were potential problems of disciplinary homes for the degrees sought by graduate students; but in the event we never found any students who looked for a degree in what might have been a difficult hybrid of say genetics, chemistry and informatics. The graduates in the project were able to justify themselves by the standards of the major department. The laissez-faire philosophy of the institution worked admirably, so long as we were able to secure funding. While we had the usual share of crises, we should look back in awe at the forbearance of the three agencies, NASA, DARPA and NIH who did make significant risk investments in a novel venture. Needless to say, all of the senior professors were also staking their credibility in the process. There is no guarantee that untenured faculty would have been able to feel so secure.

The greatest hurdle in efforts to replicate the experience would be to find experts willing and positioned to be able to forego continued immediate productivity in their own fields, for the sake of longer term ends in system building. Students and fellows may be intimidated by the demands of working across disciplines, and some were concerned that there would be a limited market in say artificial intelligence in molecular biology. Their prudence may be pragmatically justified. The process of knowledge extraction is unbelievably arduous: as always, 90% of the effort must go into debugging and validation. The process can give the expert an opportunity for critical self-reflection about the foundations of the scientific domain.

Some of the return on investment of DENDRAL was in its motivating a fresh study of the conceptual structure of organic chemistry, apart from its actual application in computer programs. This is to be commended in problem choice in other areas of application, scientific or otherwise.

The choice of organic chemistry and mass spectrometry as an object domain was a matter of careful reflection. It was rich in experimental data, and in a conceptual framework of mechanism that lent itself to model construction on the computer; I had no taste for purely statistical correlations. I might have preferred molecular genetics as more germane, and closer to my own experience. But in 1965 I did not feel it had ripened sufficiently to allow a secure theoretical framework for the necessary deductive tests of candidate hypotheses. (By 1975 it had, and this perception was the root of the followon MOLGEN project {18}). In his 1961 review, Minsky had been rather critical of generate-and-test paradigms: "for any problem worthy of the name, the search through all possibilities will be too inefficient for practical use." He had chess playing in mind with 10^120 possible move paths. It is true that equally intractable problems, like protein folding, are known in chemistry and other natural sciences. These are also difficult for human intelligence. The heuristics we have evolved biologically tend instead to relate to real world faculties like speech and image recognition. Nevertheless, solution spaces of 10^6 to 10^12 candidates are both interesting and feasible challenges to computation, and many are of scientific or technological consequence. Our particular problem in chemical analysis is one of exhaustive elimination, to find ALL solutions that match the spectral data set. Further measures may then be needed for a final disambiguation. Theorem-proving is a reasonably good analogy. Our chemical heuristics are second order: to find efficient ways of rigorously pruning the search tree, though it can be helpful to find a single approximate solution from the most plausible genera of chemical structures (e.g. rings limited to 5 or 6 nodes) and examine ways in which it can be altered and give the successfully matching spectrum. Whatever heuristics are used, no search branches can be discarded without the rationale being transparent to the chemist. Unlike chess or image understanding, chemistry does have an intrinsic mathematical structure that permits its move generator to heed the constraints of the data, so that efficiency is more readily achievable. And we have criteria, both for a formally correct candidate (a graph in canonical form), and to know when it is a solution, i.e. the test generates a spectrum that matches the data. We played against Nature. In chess (and in war), you have to play against another "expert".

Other areas of natural science deserve a fresh reconnaissance to inspire a reexamination of their conceptual structure. Biology, in particular, will soon suffocate in the sheer bulk of knowledge about DNA and protein structures, and the complex interactions of the causal chains they initiate, unless new epistemological machinery can be invented. Our education of physicians and scientists must also place more stress on the skills needed to acquire new knowledge as needed than on rote memorization that will promptly be obsolete {68}.

Finally, I would remark that I have never viewed research on artificial intelligence as having much bearing on how the human brain functions: there are too many differences in architecture and in levels of complexity, connectivity, and programmability. Nor do I see how neurobiology has contributed very much to AI. At the highest level of problem-solving routines, expert systems do of course exploit human experience. Lindley Darden's discussion of "the history of science as compiled hindsight" {66} eloquently captures my own

perspectives. My interest in AI has little to do with my background as a biologist, a great deal with curiosity about complex systems that follow rules of their own, and which have great potentialities in preserving the fruits of human labor, of sharing hardwon traditions with the entire community. In that sense, the knowledge-based-system on the computer is above all a remarkable social device, the ultimate form of publication.

Acknowledgment: It is impossible to give fair and sufficient credit to the many graduate students, collaborators, and programmers who made this effort possible. Where possible, they have been named in {41} and its bibliography. We are also indebted to the agency sponsors, primarily in DARPA, NIH and NASA whose fiscal support made the work possible. They often made us work hard on our proposals to justify that support; but it clearly was a gamble that demanded a gift of confidence of unusual measure.

BIBLIOGRAPHY

In addition to the references tabulated, the following archival sources would be instrumental for further historical research:

The Stanford University Computer Science Department: A.I. Lab Reports (microfiches available from Scientific DataLink, Comtex Sci. Corp.)
Most of these are indexed in {41}.

Annual Reports to NIH on the DENDRAL and SUMEX projects
Archived at Stanford and Rockefeller Universities

Annual Reports to NASA on the S.U., Department of Genetics Instrumentation Research Laboratory
Archived as above, and inventoried by the US NTIS

Peer critiques, NIH study sections and councils
Available, in principle, from NIH under the FoI Act.
This should be a rich resource in tracing contemporary reactions.

Buchanan, Feigenbaum, Lederberg, notes, correspondence and electronic mail files. To be archived, as above.

REFERENCES

1. Adler, M.J., 1985
Propaedia (Outline of Knowledge).
Encyclopaedia Britannica, 15 ed. Chicago.

2. Balaban, A.T. (ed. 1976).
Chemical applications of graph theory.
389 p. New York: Academic Press

3. Balzer, W. and Dawe, C. M. 1986. Structure and comparison of genetic theories: (I) Classical Genetics. Brit. J. Phil. Sci. 37, 1 (Mar.), 55-69.

4. Beynon, J. H. and Williams, A. E. 1963. Mass and Abundance Tables for Use in Mass Spectrometry. Elsevier Publ. Co., New York.

5. Blair, C. M. and Henze, H. R. 1931. The number of structurally isomeric alcohols of the methanol series. J. Amer. Chem. 53, 3042-3046.

6. Bodmer W.F., and Lederberg J. 1967.
Census data for studies of genetic demography.
Proc. III Int. Congress of Human Genetics
John Hopkins Press, Baltimore. p 459-471.

7. Brown, H., and Masinter, L.M., 1974
Algorithm for the construction of the graphs of organic molecules.
Discrete Mathematics, 8:227-244.

8. Brown, H., Hjelmeland, L., and Masinter, L.M., 1974
Constructive graph labelling using double cosets.
Discrete Mathematics, 7:1-30.

9. Buchanan, B.G. & Shortliffe, E.H., 1984
Rule-Based Expert Systems.
748 pp. Addison-Wesley: Reading, Mass.

10. Buchanan, B.G., D.H. Smith, W.C. White, R. Gritter, E.A. Feigenbaum, J. Lederberg and C. Djerassi, 1976.
Applications of artificial intelligence for chemical inference.
XXII. Automatic rule formation in mass spectrometry by means of the meta-DENDRAL program.
J. Am. Chem. Soc. 98:6168-6178.

11. Carhart, R.E., S.M. Johnson, D.H. Smith, B.G. Buchanan, R.G. Dromey and J. Lederberg, 1975.
Networking and a collaborative research community: a case study using the DENDRAL programs. pp. 192-217 in Peter Lykos (ed.) Computer Networking and Chemistry, ACS Symposium Series, No. 19. American Chemical Society:

Washington.

12. Cayley, A. 1874. On the mathematical theory of isomers. Phil. Mag. 47, 444-446.

13. Cohen, M. R. and Nagel, E. 1934. An Introduction to Logic and Scientific Method. Harcourt, Brace and Company, New York.

14. Ephrussi, B., Leopold, U., Watson, J. D., and Weigle, J. J. 1953. Terminology in bacterial genetics. Nature 171: 701.

15. Feigenbaum, E.A. & Feldman, J., (eds. 1963)
Computers and Thought.
535 pp. McGraw-Hill: New York

16. Feigenbaum, E.A. & Watson, R., 1965.
An initial problem statement for a machine induction research project.
Stanford AI Memo # 40.

17. Feigenbaum, E.A., Buchanan, B. G. and Lederberg, J. 1971.
On generality and problem solving: a case using the DENDRAL program in Machine Intelligence 6, (B. Meltzer and D. Michie, eds.), Edinburgh University Press, p.165-190.

18. Friedland, P. and Kedes, L. 1985. Discovering the secrets of DNA. Comm. ACM 28:1164-1186.

19. Garfield, E. 1979 Citation Indexing -- Its Theory and Application in Science, Technology, and Humanities.
ISI Press, Philadelphia, pp. 274.

20. Gould, R. F. (ed. 1966). Kekule Centennial. In Advances in Chemistry Series 61. American Chemical Society, Washington.

21. Jordan, C. 1869. Sur les assemblages de lignes.
J. fuer die reine und angewandte Math. 70,185 (1869)

22. Judson, H.J., 1979.
The Eighth Day of Creation: makers of the revolution in biology. 686 pp. New York: Simon & Schuster

23. Kinney, H. 1979. The year of the gifted children. Think (IBM) 45: 12-17.

24. Lederberg J. 1955. Genetics and microbiology.
In Symposium on Perspectives and Horizons in Microbiology.
Rutgers Univ. Press, pp. 24-39.

25. Lederberg J. 1958. A view of genetics.
Les Prix Nobel 170-189.

26. Lederberg J. 1964. Computation of Molecular Formulas for Mass Spectrometry. Holden-Day, Inc., San Francisco.

27. Lederberg J. 1965.
Hamilton circuits of convex trivalent polyhedra (up to 18 vertices).
Am. Math Monthly 74, 522-527.

28. Lederberg J. 1965. Systematics of organic molecules, graph topology and Hamilton circuits.
A general outline of the DENDRAL system.
NASA CR-68899. STAR No. N66-14075.

29. Lederberg J. 1966.
Online computation of molecular formulas from mass number.
NASA CR-95977; Accession # x68-18613.

30. Lederberg, J. 1964. DENDRAL-64. A system for computer construction, enumeration & notation of organic molecules as tree structures and cyclic graphs.
Part I. Notational algorithm for tree structures.
NASA CR-57029. STAR No. N65-13158.

31. Lederberg, J. 1965. Topological mapping of organic molecules.
Proc. Nat. Acad. Sci. U.S. 53, 134-139.

32. Lederberg, J. 1965. DENDRAL-64. Part II. Topology of cyclic graphs.
NASA CR-68898. STAR No. N66-14074.

33. Lederberg, J. 1965. Signs of Life: criterion system of exobiology.
Nature 207, 9-13.

34. Lederberg, J. 1969. Topology of Molecules.
In The Mathematical Sciences (COSRIMS). MIT Press, p. 37-51.

35. Lederberg, J. 1972.

Rapid calculation of molecular formulas from mass values.
J. Chem. Ed. 49, 613.

36. Lederberg, J. 1977.
Foreword to Essays of an Information Scientist by E. Garfield
ISI Press, PA, 1977, pages xi-xi

37. Lederberg, J. 1978.
Digital communications and the conduct of science: The new literacy.
Proc. of the IEEE 1978, 66(11):1314-1319.

38. Lederberg, J. 1986. Forty Years of Genetic Recombination in Bacteria. A Fortieth Anniversary Reminiscence.
Nature, 327:627-628.

39. Lederberg, J., Sutherland, G. L., Buchanan, B. G., Feigenbaum, E. A., A. V. Robertson, Duffield, A. M. and Djerassi, C. 1969.
Applications of artificial intelligence for chemical inference.
I. The number of possible organic compounds. Acyclic structures containing C, H, O, and N.
J. Am. Chem. Soc. 91, 2973-76.

40. Lewin, R. 1984. National networks for molecular biologists.
Science 223, 1379-1380.

41. Lindsay, R.K., B.G. Buchanan, E. A. Feigenbaum and J. Lederberg, 1980. Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project. 194 pp. McGraw-Hill Book Co.,
Now available from Wm. Kaufmann, Inc., 95 1st St., Los Altos CA

42. McCorduck, P. 1979. Machines Who Think. W. H. Freeman and Company, San Francisco.

43. Medawar, P. B. 1964. Is the scientific paper fraudulent?
Saturday Rev. (1 Aug.), 42-43.

44. Merton, R.K., 1968. See pp. 4-7.
Social Theory and Social Structure. 702 pp. New York: Free Press.

45. Merton, R.K., 1977.
The sociology of science: an episodic memoir. In
Merton, R. K and Gaston, J. (eds. 1977). The Sociology of Science in Europe. Southern Illinois University Press, Carbondale , Ill. 383. pp.
See esp. pp. 119-120.

46. Minsky, M., 1961
Steps toward artificial intelligence.
Proc. Inst. Radio Eng., 49:8-30.

47. Models for Biomedical Research. 1985.
National Academy Press, Washington, D. C. pp. 180.

48. Partington, J. R. 1964. A History of Chemistry, vol. IV.
Macmillan & Co. Ltd, London.

49. Polya, G. 1938. Kombinatorische Anzahlbestimmungen fur Gruppen, Grraphen und Chemische Verbindungen. Acta Math. 68, 145-254.

50. Pres. Sci. Adv. Comm., 1963.
Science, Government, and Information.
Panel on Scientific Information Report. (A. Weinberg, chmn.)
52 pp. USGPO: Washington

51. Quastler, H. (ed. 1953). Essays on the Use of Information Theory in Biology. University of Illinois Press, Urbana. pp. 273.

52. Reynolds, W.E., Bacon, V. A., Bridges, J. C., Coburn, T. C., Lederberg, J., Levinthal, E. C., Steed, E. and Tucker, R. B.
A computer operated mass spectrometer system.
Analytical Chem. 42, 1122-1129.

53. Rouvray, D.H., 1971. Graph Theory in Chemistry.
R.I.C. Reviews. 4:173-195.

54. Savageau, M. A. 1976. Biochemical Systems Analysis. A Study of Function and Design in Molecular Biology. Addison-Wesley Publ. Co., Massachusetts, pp. 379.

55. Simmons, H. E. and Maggio, J. E. 1981. Synthesis of the first topologically non-planar molecule. Tetrahedron Letters 22, 287-290.

56. Sutherland, G., 1967 DENDRAL - a computer program for generating and filtering chemical structures. 24 pp. + 10 (appendix.)

Stanford Artificial Intelligence Memo # 49.

57. Thomas, R. (ed. 1979). Lecture Notes in Biomathematics. Kinetic Logic: A Boolean Approach to the Analysis of Complex Regulatory Systems. Springer-Verlag, Berlin and New York.

58. Whitehead, A. N. and Russell, B. 1950. Principia Mathematica, 3 vols. 2nd ed. Cambridge at the University Press, London.

59. Woodger, J. H. 1937. The Axiomatic Method in Biology. Cambridge at the University Press, London.

60. Zuckerman, H. A. & Lederberg, J. 1986. Forty Years of Genetic Recombination in Bacteria. Postmature Scientific Discovery? Nature, 327:629-631.

61. Zuckerman, H.A. & Merton, R.K., 1971. Patterns of Evaluation in Science. Institutionalization, structure and functions of the referee system. Minerva 9:66-100.

---- Additional references:

62. Rindfleisch, T.C., Smith, D.H., Yeager, W.J., Achenbach, M.W. and Wegmann, A. 1980. "Mass Spectrometer Data Acquisition and Processing Systems," in Chapter 3 of "Biomedical Applications of Mass Spectrometry, First Supplementary Volume," G.R. Waller and O.C. Dermer, Eds., John Wiley and Sons, Inc., New York, pp. 55-77.

63. Huang. Zhi-heng and Wu, Wei-qin 1986 MDAL: a novel algorithm for the computation of empirical formulas in high-resolution mass spectrometry. Biomed. Env. Mass. Spec. 13: 187-191.

64. Djerassi, C., 1988. Organic Chemistry: A view through steroid glasses, in Modern Organic Chemistry: Scientific and Historic Perspectives, J. I. Seeman, ed. Washington: American Chemical Society.

65. Newell, A. 1983 Intellectual issues in the history of artificial intelligence. 187-227 in Machlup, F. and Mansfield,U., The study of information: interdisciplinary messages. 743 pp. New York: Wiley-Interscience.

66. Darden, Lindley 1987 Viewing the history of science as compiled hindsight. AI Magazine 8(Summer): 33-41.

67. Lederberg,J. 1987. Sputnik 1957-1987 The Scientist, Oct. 5, 1987.

68. Lederberg, J. 1986 Computers in Science, Communication and Education AAMC Symp. on Medical Informatics, held March 7, 1985. pp. 62-70. "Medical Education in information age". AAMC, Washington, 1986

.nf
Appendix:

This memo is an excellent snapshot of the status of AI research as seen in 1965, and will therefore be appended in its entirety.

.na
.nh
.nf
.tl 'STANFORD ARTIFICIAL INTELLIGENCE PROJECT''April 5,1965'
Memo No. 30

.ce3
AN INITIAL PROBLEM STATEMENT FOR A
MACHINE INDUCTION RESEARCH PROJECT

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

.fi
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 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).

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 sugggested 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 O - 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. The 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 O 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 asa 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 proceed 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 its 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 structures 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

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 test 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 $20 < n < 46$.

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 * 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.

* E.G.:

.ig

---

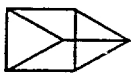
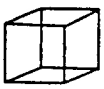
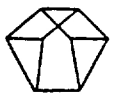
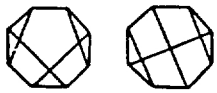
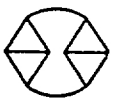
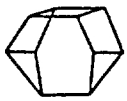
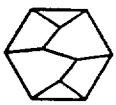
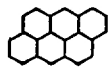
*Ed if you can put that into ASCII, congratulations. Maybe you can blow it up with SCRIPT.

..

Mappings of cyclic chemical structures

onto trivalent graphs - convex polyhedra

from Lederberg (1965)

VERTICES	FIGURE		POLYGON FORM(S)		EXAMPLE	
0		CIRCLE				BENZENE
2A		BICYCLANE				NAPHTHALENE
4A		TETRAHEDRON				TRIPHENYLENE
4B		TRICYCLANE				ANTHRACENE
6A		PRISM				PYRENE
8A		CUBE				CUBANE
8B		BI-PENTAGON				BENZOPYRENE
8C		BI-TETRAHEDRON				PERYLENE
10A		BI-HEXAGON				BENZOPERYLENE
10B						DIBENZO-CHRYSENE
10C						DIBENZO-PYRENE
10D						
10E		PENTAGONAL PRISM				ETHANEDIYLIDENE-CYCLOPENTA-PENTALENE

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 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 bombardment 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". Proc. Nat. Acad. Sci. 53:134-139, 1965. (Available from either author of this memo).

APPENDIX A