Table I. Results of Molecular Ion Determination for the Unknown Compound, X, whose Mass Spectrum is Presented in Figure 5.

CANDIDATE	RANKING INDEX
263.0	100
307.0	41
299.0	38
295.0	34
281.0	25

The MOLION program is written to operate on either low or high resolution mass spectra. The program has certain limitations which have been summarized in detail previously(11).

MOLION is available on SUMEX. A FORTRAN version, initially for low resolution mass spectra, is being written so that the program can be run on smaller computers and exported to others. However, it will continue to be available via SUMEX so that others can access it easily. MOLION is contained within PLANNER as one of the available methods for detecting candidate molecular ions.

PLANNER(12). The PLANNER program is designed to analyze the mass spectrum of a compound or of a mixture of related compounds. Because there is no ab initio way of relating a mass spectrum of a complex organic molecule to the structure of that molecule, PLANNER requires fragmentation rules for the class of compounds to which the unknown belongs. This is its major limitation. For our example the class was unknown, forcing us to resort to other means of assistance.

Applications and limitations of PLANNER have been discussed extensively(12,13). The program is very powerful in instances where mass spectrometry rules are strong (i.e., general, with few exceptions). In instances where rules are weak or nonexistent, additional work on known structures and spectra may yield useful rules to make PLANNER applicable (see INTSUM and RULEGEN, below). One important feature of PLANNER is its ability to analyze the spectra of mixtures in a systematic and thorough way. Thus, it can be applied to spectra obtained as mixtures when GC/MS data are unavailable or impossible to obtain. PLANNER is available in an interactive version over SUMEX, requiring three kinds of information as input: the high or low resolution mass spectrum, the characteristic skeletal structure for molecules in the specific compound class, and the fragmentation rules for the class. Additional knowledge about the unknown can be used by the program to constrain the structural possibilities.

CONGEN(14,15). Structure problems are usually not solved with mass spectrometry alone. Even when sample size is too limited for obtaining other spectroscopic data, knowledge of chemical isolation and results of derivatization procedures frequently act as powerful

constraints on structural possibilities. Larger amounts of sample permit determination of other spectroscopic data. Taken together, this information allows determination of structural features (substructures) of the molecule and constraints on the plausibility of ways in which the substructures may be assembled. The CONGEN program is capable of providing assistance in solution of such problems.

CONGEN performs the task of construction, or generation, of structural isomers under constraints. The program accepts as input known structural fragments of the molecule ("superatoms") and any remaining atoms (C,N,O,P,...), together with constraints on how they may be assembled. It is based on the exhaustive structure generator(16,17) and extensions(18) which permit a stepwise assembly of structures.

In an interactive session with the program, a user supplies structural information determined by his own analysis of the data (perhaps with the help of the above programs), together with whatever other constraints are available concerning desired and undesired structural features, ring sizes and so forth. The program builds structures in a series of steps, during which a user can interact further with the procedure, for example, to add new constraints. Although very much a developing program, its ability to accept user-inferred constraints from many data sources makes CONGEN a general tool for structure elucidation which we are making available via SUMEX-AIM in its current form.

For the unknown X, the observed fragment ions from the molecular ion (M) at m/e 263 (Figure 5) suggest several structural features when coupled with the knowledge of the chemical derivatization procedures used on this fraction of the urine extract. The ion at m/e 194 represents loss of 69 amu, probably CF3, from fragmentation of a trifluoroacetyl derivative of an amine. This suggests the partial structure 2, Figure 5. The ions at m/e 190 (M-74 amu) and m/e 162 (M-101 amu) suggest the characteristic fragmentation of an n-butyl ester resulting from the second derivatization procedure, formation of the n-butyl esters of free carboxylic acid functions. This suggests the partial structure 1, Figure 5. Taken together, all the above information implies (if no other elements are present) that the empirical formula contains an odd number of nitrogen atoms, at least three oxygen atoms, three fluorine atoms and at least seven carbon atoms. Interestingly, there is only one plausible empirical formula under these constraints, C11H12NO3F3.

Structural fragments ("superatoms") 1 and 2 were supplied to CONGEN, together with the remaining four carbon atoms and three degrees of unsaturation (that is, rings plus multiple bonds). With no additional constraints, 155 structures result. The inclusion of other plausible constraints (e.g., no allenes, acetylenes, cyclopropenes, cyclobutenes) reduces the number of structural candidates to just the two isomeric forms of 3, Figure 5.

This problem represents a simple example of a large class of such problems. Although a chemist could probably reach the same conclusions quickly in this case, in the general case, piecing together potential solutions is not a trivial task.

Although still a developing program, CONGEN is, capable of considerable assistance in a wide variety of structure problems. Some areas of current application are summarized in the subsequent section. It is already proving its value in structure elucidation problems by suggesting solutions with a guarantee that no plausible alternatives have been overlooked.

The program has a great deal of flexibility. Many of the types of constraints normally brought to bear on structure elucidation problems can be expressed. However, some types of constraints cannot be easily expressed (e.g., disjunctions of features and stereo-constraints). Recent work by our group and Wipke's(19) will make it possible to add considerations of stereoisomerism relatively easily (a good example of collaboration via SUMEX). We are depending on a broad user community to help us guide further development of CONGEN.

Programs for Knowledge Acquisition

INTSUM(20) and RULEGEN(21). When the mass spectrometry rules for a given class of compounds are not known, the INTSUM and RULEGEN programs can help a chemist formulate those rules. Essentially, these programs categorize the plausible fragmentations for a class of compounds by looking at the mass spectra of several molecules in the class. All molecules are assumed to belong to one class whose skeletal structure must be specified. Also, the mass spectra and the structures of all the molecules must be given to the program.

INTSUM collects evidence for all possible fragmentations (within user-specified constraints) and summarizes the results. For example, a user may be interested in all fragmentations involving one or two bonds, but not three; aromatic rings may be known to be unfragmented; and the user may be interested only in fragmentations resulting in an ion containing a heteroatom. Under these constraints, the program correlates all peaks in the mass spectra with all possible fragmentations. The summary of results shows the molecules whose spectra display evidence for each particular fragmentation,along with the total (and average) ion current associated with the fragmentation.

The RULEGEN program attempts to explain the regularities found by INTSUM in terms of the underlying structural features around the bonds in question that seem to "direct" the fragmentations. For example, INTSUM will notice significant fragmentation of the two different bonds alpha to the carbonyl group in aliphatic ketones. It is left to RULEGEN to discover that these are both instances of the same fundamental alpha-cleavage process that can be predicted any time a bond is alpha to a carbonyl group.

These programs are part of the so-called Meta-DENDRAL effort, whose general goal is to understand rule formation activities. Both INTSUM and RULEGEN are available as interactive programs on SUMEX, the former being much more highly developed than the latter. Although these programs can be very useful to chemists interested in finding new mass spectrometry rules, they require having the collection of

mass spectra and molecular structure descriptions available in one computer file. Because of this, they have been used mostly by chemists at Stanford.

Applications and Resource Sharing

The DENDRAL programs are being developed to serve a broad community of chemists with structure elucidation problems. Our experience is admittedly limited. In this section we discuss some of the applications, both local and from remote sites, where these programs have proven useful.

CLEANUP and MOLION. These programs are in routine use as part of the Genetics Research Center's GC/LRMS efforts. In addition, MOLION has been incorporated as part of PLANNER. Their generality has proven very useful in applications to a variety of GC/MS problems involving structural studies of urinary metabolites.

PLANNER. The planning program has been used to infer plausible placement of substituents around a skeletal structure for numerous test problems in which the class of the sample was known and the fragmentation rules for the class were known. Those tests have resulted in a program that we believe is general. We have applied this program to unknown mixtures of estrogenic steroids(13). We are preparing to use PLANNER for screening mass spectra of marine sterols to identify quickly those spectra of known compounds and to suggest structures for spectra of new compounds.

CONGEN. CONGEN is being used locally and from remote sites in a wide variety of applications. We have used it for construction of ring systems under constraints(22) and for generation of structures of chlorocarbons(23). We have investigated several monoterpenoid and sesquiterpenoid structure problems to suggest solutions and to ensure that all alternatives had been considered. We are currently investigating the scope of terpenoid isomerism. Two problems relating to unknown photochemical reaction products have been analyzed and results used to suggest further experiments. In most cases we do not know the precise problems under study by remote users, only that they are using the program.

CONGEN will perhaps be the most widely used (by remote users) program of those mentioned above as accessible through SUMEX. This is primarily a result of the wider scope of problems which might benefit from use of the program. However, the need for remote users to have their mass spectral data available at SUMEX for analysis present a significant energy barrier to use of the programs which require these data.

INTSUM and RULEGEN. INTSUM is essentially a production program now, and is being used as such in a variety of applications involving correlations of molecular structures with their respective spectra. Recent or current applications include analysis of the mass spectra of progesterones and related steroids, androstanes, macrolide

antibiotics, insect juvenile hormones and phytoecdycones. These studies serve to develop fragmentation rules which, if of sufficient generality, can in turn be used in PLANNER in the study of unknown compounds.

III. PROBLEMS RELATED TO NETWORKING

During this first year of operation, the SUMEX-AIM facility has encountered a variety of problems arising from its network availability. In most cases, there has been no clear precedent for the handling of these situations, in fact, many problem-areas still reflect the influences of a yet-developing policy. The hope is that this presentation and discussion of problems and their solutions may give foresight to others who contemplate networking or network use. The problems to be discussed can be loosely associated into three classes; those related to the management of the facility, those pertaining to research activities on the system, and those involving psychological barriers to network use.

Managerial problems

"Gatekeeping." The most general problem faced by the organizers of the SUMEX-AIM facility is the question of "gatekeeping." In order to insure a high quality of pertinent research, some kind of refereeing system is needed to assess the value of proposed new projects. The organizers of the facility would seem to be the best source of such judgements; yet, because we are both organizers and members of the SUMEX community, there is a danger that our decisions would unfairly favor local priorities. In order to establish credibility in SUMEX-AIM as a truly national resource, a management system has been instituted that allocates a defined fraction (initially 50%) of the SUMEX resource to external users, under the jurisdiction of an independent national committee (the AIM advisory group). The remaining 50%, allocated for local use, contains a portion for flexible experiments outside of local projects, but on our own responsibility.

Choice of computer and operating system. A second management level problem is the choice of a computer and operating system which optimize the usefulness of the facility for a majority of users, and which encourage intercommunication between remote collaborators. Because SUMEX-AIM is intended to be used primarily for applications of artificial intelligence, and because interactive LISP(24) is a primary language in this type of work, the choice of TENEX(25) as an operating system was dictated somewhat by necessity. TENEX incorporates multiple address spaces, thereby allowing multiple "fork" structure and paging, a design which is necessary to create the large-memory virtual machine required by INTERLISP.

The PDP-10 is a popular machine for interactive computing of all

sorts in university research environments, and thus an added benefit of this choice was expected - the possibility of easily transferring to SUMEX programs developed at other sites. Many of these programs were written not under TENEX but under the 10/50 monitor supplied by the manufacturer. Because a large and useful program library was already available under the 10/50 monitor, one of the design criteria of TENEX was compatibility with such programs; when a 10/50 program is run under TENEX, a special "compatibility package" of routines is invoked to translate 10/50 monitor calls into equivalent TENEX monitor calls. Although the concept is sound, we have found that in practice very few programs written for the 10/50 monitor are able to run under TENEX without extensive modification. Other problems with TENEX include weaknesses in the support of peripheral devices and the lack of a default line-editor. The latter has caused a proliferation of editing programs, and some confusion has resulted because editor conventions vary from program to program. These difficulties have dampened somewhat our initial enthusiasm for the TENEX system.

Nonetheless, TENEX provides some features which are crucial to a comfortable network environment. The standard support programs included with this system facilitate both the sending of messages to other users (either at the same site or at other sites on the ARPA network) and the transfer of data and programs from site to site on that network; also, the ability to "link" two or more terminals allows users to communicate easily and immediately. Both the linking and message facility have been found to be invaluable aids in inter-group communications and in such problems as interactive program debugging. When two terminals are linked, their output streams are merged, thus allowing each terminal to display everything typed at the other terminal. Since only the output stream is affected under these circumstances, it is still possible for each terminal to be used to provide input to separate programs, in addition to being used in a conversational mode.

Resource allocation. As noted above, the computational resources of the SUMEX-AIM facility are apportioned by the AIM advisory group and SUMEX management. Some extensions to the basic TENEX system have been made to reflect this apportioning in the actual use of the facility. Basically, it was recognized that users of the facility are members of groups working on specific projects, and it is among these projects that the facility is apportioned. Disk space and cpu cycles are now distributed among groups instead of among individual users. For example, a user may exceed his individual disk allocation somewhat without any ill effect, so long as the total allocation of his group remains within the limits. Similarly, a Reserve Allocation Scheduler has been added to TENEX which tries to match the administrative cycle distribution over a ninety second time frame. Thus a particular group cannot dominate the machine if a lot of its members are logged in at one time.

It is typical for usage of a facility to peak through the middle hours of the day. Indeed, one of the advantages of having users from around the country is the spreading of the load caused by the difference in time zones. Even so, the facility could offer better service if more people would shift their main usage hours toward

either end of the day. To encourage "soft-scheduling" within groups on the system, SUMEX-AIM publishes a weekly plot of diurnal loading . This plot shows the total number of jobs on the system as well as the number of LISP jobs, since these jobs seem to make the biggest demands of system resources. The result has been an increased awareness by users of system loading and a noticeable increase in the number of users at all hours of the night and early morning.

Protection and system security. Protection for a computer system covers a range of ideas. It means the ability to maintain secrecy - for example, to guarantee the privacy of patient records. It also guarantees integrity by assuring that programs and data are not modified by an unauthorized party.

Questions of protection generally become more interesting and complex as more sharing is involved. Consider the example of a proprietory program which generates layouts given a user's circuit data. The program owner demands assurance that he will be paid whenever his program is used and that copies of the program cannot be made. The user wants guarantees that his data sets cannot be destroyed or copied for a competitor. Yet the user must have access to the program and the program must have access to the data. Unable to support such complicated examples of protection, SUMEX-AIM assumes that sharing takes place between friendly users. This is not to imply that issues of protection and sharing have not appeared. For example, in an effort to improve the human engineering of programs for public use, the capability of recording a session has been built into several of the programs. Studied by the program designers to pinpoint confusing aspects of programs, these recordings serve to improve program design. Since the issue of violation of privacy has been raised, some of these programs now request permission to record a session before doing so. At this time, any guarantee of privacy must be provided by the program designer because TENEX itself does not have the ability to render the protection .

The general design for systems offering "state of the art" protection involves a tolerance for failure; that is, if a potential offender succeeds in breaking through some of the defenses, he still does not place the entire computer system at his mercy. Encrypting of data files provides an additional line of defense. This method is used by at least two calendar or appointment programs on the computer. More general purpose facilities to allow users to encrypt and decrypt any of their files whenever they wish are being developed and will be available soon.

Tenex provides the usual keyword protection at login time and a measure of file protection. Owners of a file may assign a protection number which specifies some combination of READ, WRITE, EXECUTE, or APPEND access to a file for owners, members of a group, or other users. This level of protection is basically enough to prevent accidents and most mischief. System programmer's around the country are aware of a number of TENEX bugs which permit this access to be violated. One user of our system found a way to place himself in a mode where he could modify any file on the system. To date, we have no examples of such activity~ actually having a deleterious effect on SUMEX-AIM.

To make the use of SUMEX-AIM programs easily available on a trial basis for prospective users, a "guest" account system has been established. Since this makes logging into SUMEX-AIM so easy, it has invited some misuse by people using those accounts to play the computer games. A proposed extension to the system now being implemented is a special "guest EXEC" which would extend the protection of the TENEX monitor by allowing guest accounts access to only a more restricted set of programs.

File backup. In order to assure the user maximum protection against loss of valuable work, SUMEX operates a multi-level file backup system. In addition to routine file backup system there are facilities to enable the user to selectively archive his or her disk files. By issuing a simple command to the TENEX executive the user can transmit a message to the operator to copy specified files to magnetic tape. Each such file is copied to two magnetic tapes within 24 hours of issuing the archive command. File retrieval is affected by a similar process. The user also has the alternative option of being able to lodge files in a special backup directory. Files are held in this directory until the next exclusive file dump (see below) at which time they are deleted. In this way the user can remove files from his directory at his own choosing knowing they will be archived by the exclusive dump.

On a system level, an effort is made to maintain file backups such that the maximum possible loss, in the event of a crash fatal to the file system, would amount to no more than one day's work. Once each day all files that have been read or written within the last 48 hours are dumped onto magnetic tape. Files that exist for 48 hours are thus held on two separate tapes. The rotation period for files dumped in this way is 60 days. Once each week a full file dump is made to separate disk storage. Each such dump is kept for two weeks at which time it is replaced by a new file dump. Each month there is a full system dump from disk to magnetic tape. Files can be recovered from the system backup by sending a message to the operator specifying the file name(s) and when the file was last read or written (if such information is available).

Excessive demand for production programs. One of the concepts behind the creation of a shared resource is elimination of the problems which arise when large, complex computer programs are exported. Since, in theory, exportability is no longer a problem, there is greater latitude in choice of a language in which program development can take place. In the case of some of the DENDRAL programs, it was thought that program development should take place in INTERLISP, a language that lends itself well to the artificial intelligence nature of these programs, but does not lead to particularly efficient run-time code.

In order to ascertain the usefulness of these programs and to determine what areas remain in need of work, chemist collaborators are being sought. As these users increase in number and begin to use the programs more frequently, it is almost certain that the inherent slowness of the predominately LISP code will affect the whole system as well as handicap the efficient use of the DENDRAL programs.

Additionally, some of the chemist-users who are finding the programs most useful and who are most enthusiastic about their potential use, are persons working in industry. Although, in one sense, this interest from industry could be interpreted as an indication of the "real-world" usefulness of the programs, it came as rather a surprise to both SUMEX and DENDRAL personnel.

The fact that SUMEX-AIM is funded by NIH as a national resource prohibits the facility from providing a service, at taxpayer's expense, to a private industry. Although there is precedent for a site funded via government grant to charge a fee for service, such an arrangement leads to highly complicated bookkeeping, and is contrary to the essential purpose of SUMEX-AIM; to be a research-oriented rather than service-oriented facility. This leaves the industrial users in the position of being more than willing to pay for the use of the programs, but of having no mechanism whereby they can be charged. Furthermore, the fact that the programs are coded in LISP for a highly specialized environment, almost guarantees the impossibility of export, except to an almost identical computer system.

An intermediate solution that will help to solve the problem of industrial users on SUMEX and will help to alleviate the system loading resulting from heavy usage of LISP coded production programs, is to mount CONGEN on a closely related computer which is operated on a fee for service basis. However, in order to make this program available at a reasonable fee, it has become evident that it will be necessary to recode the LISP sections of the program into a more efficient and easily exportable language.

Research-oriented problems

Community mindedness. Those involved in computer science research at SUMEX face a general problem which is absent or greatly lessened at non-network sites; the problem of community mindedness. The network provides a large and varied set of other researchers and users who have an interest in their work. Although the network-TENEX combination provides new forms of communication with these remote parties, the traditional means of fully describing the use and structure of a complex program, a detailed person-to-person discussion, is not convenient. Comprehensive documentation gains importance in such a situation, and within the DENDRAL project a great deal of time has been needed in the development of program descriptions which are adequate for a diverse audience. Also, in both DENDRAL and MYCIN, effort has been and is being directed toward "human engineering" in program design; to provide the user with commands which assist him in using the programs, in understanding the logic by which the programs reach certain decisions and in communicating questions or comments on the programs' operation to those responsible for development. Such "housekeeping" tasks can often be neglected, yet are quite important in smoothing interaction with the community.

Choice of programming language. High level programming languages which are designed for ease of program development are

frequently poor as production-level languages. This is because developmental languages free the researcher from a raft of programming details, thus allowing him to concentrate upon the central logical issues of the problem, but the automatic handling of these details is seldom optimal. Also, because such languages tend to be specialized for certain computers and operating systems, the exportation of programs can be a serious problem. One solution to these problems is the recoding of research-level programs into more efficient language when fast and exportable versions are needed.

Networking greatly eases the problem of exportability, but can also aggravate the the problem of efficiency. As mentioned in the previous section, the DENDRAL programs, which are undergoing constant development, found a substantial number of production-level users. Because of the inefficiencies of INTERLISP (a 50- to 100-fold improvement in running time is not uncommon when an INTERLISP program is translated into FORTRAN), this use adversely impacted the entire system. Because the DENDRAL programs are quite large and complex, their translation into other languages is impractically tedious. A partial solution to this problem is provided by the TENEX operating system, which allows some interface between programs written in different languages. With such intercommunication, time-consuming segments of an INTERLISP program which are not undergoing active development can be reprogrammed in another more efficient language. The developmental parts of the program are left in INTERLISP, where modifications can easily be made and tested. The CONGEN program uses three languages; INTERLISP, FORTRAN and SAIL(26). The SAIL segment was added when a new feature, whose implementation was fairly straightforward, was included in CONGEN. Since then, the SAIL portion gradually has been taking over some of the more time-consuming tasks. This method allows a balance in the trade-off between ease of program development and efficiency of the final program.

Accumulation of expert knowledge in knowledge-based programs. Just as statistics-based programs need to worry about accumulation of large data bases, knowledge based programs need to worry about the accumulation of large amounts of expertise. The performance of these programs is tied directly to the amount of knowledge they have about the task domain -- in a phrase, knowledge is power. Therefore, one of the goals of artificial intelligence research is to build systems that not only perform as well as an expert but that also can accumulate knowledge from several experts.

Simple accretion of knowledge is possible only when the "facts", or inference rules, that are being added to the program are entirely separate from one another. It is unreasonable to expect a body of knowledge to be so well organized that the facts or rules do not overlap. (If it were so well organized, it is unlikely that an artificial intelligence program would be the best encoding of the problem solver.) One way of dealing with the overlap is to examine the new rules on an individual basis, as they are added to the system in order to remove the overlap. This was the strategy for developing the early DENDRAL programs. However, it is very inefficient and becomes increasingly more difficult as the body of knowledge grows.

The problem of removing conflicts, or potential conflicts, from overlapping rules becomes more acute when more than one expert adds new rules to the knowledge base. Of course, the advantages of allowing several experts to "teach" the system are enormous -- not only is the program's breadth of knowledge potentially greater than that of a single expert, but the rules are more apt to be refined when looked at by several experts. On the other hand, one can expect not only a greater volume of new rules but a higher percentage of conflicts when several experts are adding rules.

Having a computer program that can accumulate knowledge presupposes having an organization of the program and its knowledge base that allows accumulation. If the knowledge is built into the program as sequences of low-level program statements -- as often happens -- then changing the program becomes impossible. Thus current artificial intelligence research stresses the importance of separating problem-solving knowledge from the control structure of the program that uses that knowledge.

Another problem, at a political rather than a programming level, becomes apparent with one accumulation process: how does the program distinguish an expert from a novice? In the MYCIN program we have circumvented the problem by having the program ask the current user for a keyword that would identify him as an expert. It is then a bureaucratic decision as to which users are given that keyword. There is nothing subtle in this solution, and one can imagine far better schemes for accomplishing the same thing. The point here is that not every user should have the privilege of changing rules that experts have added to the system, and that some safeguards must be implemented.

"Human nature" barriers to SUMEX use

Countering disbelief. There is sometimes a tendency among those unfamiliar with the capabilities and limitations of computers and computer programs to express disbelief. This is not disbelief in the sense of worrying that the programs have errors and produce erroneous results. Indeed, the fact that a problem is being done by a computer seems to generate some faith that it might be right, or at least significantly reduces questions about correctness. The disbelief is that programs, which are designed to model, or to emulate, human problem solving will not be capable of useful performance. This, of course, is the classic argument against artificial intelligence - we think in mysterious ways and have such a complex brain that a computer program must be inferior. In some cases, authors of artificial intelligence programs have brought such criticism upon themselves by not stressing limitations, or by making extravagant claims.

In the DENDRAL project, we have tried to counter this type of disbelief in a number of ways. We have tried to stress that our programs are designed to assist, not replace chemists. We have always discussed limitations to give a reasonable perspective on capabilities vs. limitations of a program. Most importantly however, we have

focused on those aspects of problems which are amenable to systematic analysis, i.e., those problems which can be done manually, but only with difficulty and with the consumption of a great deal of time which a chemist could better spend on more productive pursuits. Examples of this would include the application of PLANNER to mixtures where all fragmentations may have to be considered as possible fragments of every molecular ion, the systematic analysis by INTSUM of possible fragmentation processes, the consideration by MOLION of all plausible possibilities, and the structure generation capabilities of CONGEN.

We have also tried to reduce chemists' disbelief by blurring the "outsider-insider" distinction, in particular by having trained chemists work on the programs and make them useful to themselves first. Further, when "outside" chemists are first introduced to the programs, the introduction is done by another chemist who has already thought through and can readily explain many of the chemistry-related problems.

The ultimate way to counter disbelief, however, is to illustrate high levels of performance. If a potential user is aware of the goals (intent) of a program and its limitations, a few examples of results which would be extremely difficult to obtain without the program are very convincing.

The "security" of a local facility. Networking is still a relatively new concept to many people, and there is a resistance to departing from the "traditional" modes of computing. There is a sense of security in having a local computing facility with knowledgeable consultants within walking distance, and in having "hard" forms of input (eg, boxes of computer cards) and output (eg, voluminous listings). These props are difficult to simulate over a network connection - in most cases a user's interaction with the remote site takes place exclusively through a computer terminal - yet the quality of service can match or exceed that of a local facility; programs and large data sets can be entered and stored on secondary storage as can large output files; all types of program and data editing can be done with interactive editing programs; programs can be written in an interactive mode so that small amounts of control information can be input and key results output in "real time" over the terminal; And as noted in a previous section, consultation can be significantly more productive providing that the remote operating system supports the appropriate types of communication possibilities.

There can, of course, be no denying that there are problems in learning to use a distant computer system, be it for program development or for the use of certain programs. Whether or not overcoming these problems to gain access to the special resources which are available, is worth the effort, is a question answerable only by the individuals involved. Fortunately, there will always be those persons who have a pressing problem in need of solution and who are willing to try a new approach; regardless of whether or not they have had prior network experience.

IV. THE SUMEX-AIM FACILITY

The SUMEX-AIM computer facility consists of a Digital Equipment Corporation model KI-10 central processor operating under the TENEX time sharing monitor. It is located at Stanford University Medical Center, Stanford, California.

The system has 256K words (36 bit) of high speed memory; 1.6 million words of swapping storage; 70 million words of disk storage; two 9-track, 800 bpi industry compatible tape units; one dual DEC-tape unit; a line printer; and communications network interfaces providing user terminal access via both TYMNET and ARPANET.

Software support has evolved, and will continue to evolve, based on user research goals and requirements. Major user languages currently include INTERLISP, SAIL, FORTRAN-10, BLISS-10, BASIC and MACRO-10. Major software packages available include OMNIGRAPH, for graphics support of multiple terminal types, and MLAB, for mathematical modeling.

The SUMEX-AIM computer generally is left with no operator in attendance; thereby helping to eliminate some overhead, but also creating some problems. Users who wish to run jobs requiring tapes must make arrangements to mount their own tapes. Likewise, obtaining listings from the line printer can be somewhat difficult since there is no regular schedule for distribution of this output. The solution to these two problems has been to make keys to the machine room available at strategic locations, convenient to all groups of local users. This experiment in basic "resource sharing" has not resulted in any of the major problems one might expect from having a fairly large group of people with hands-on access to a computer.

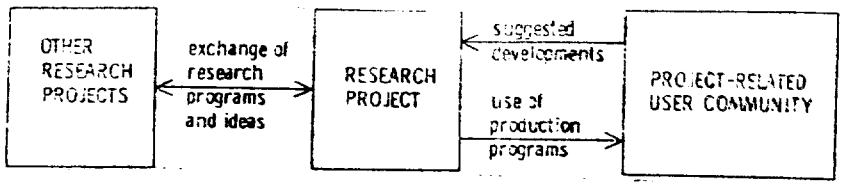

Figure 1. Interactions in the SUMEX-AIM Community

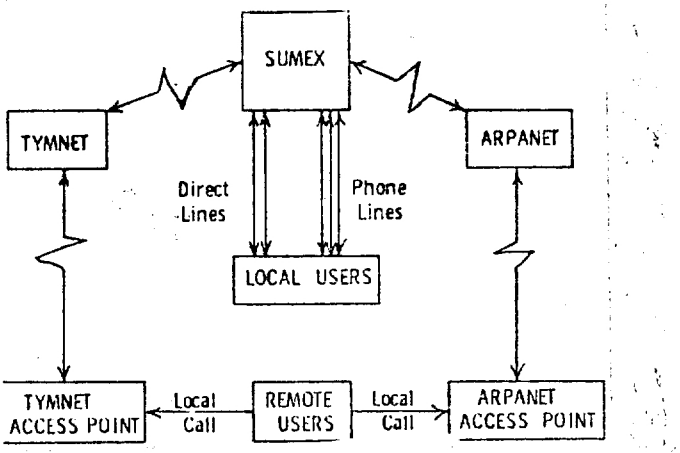

Figure 2. Access to SUMEX-AIM

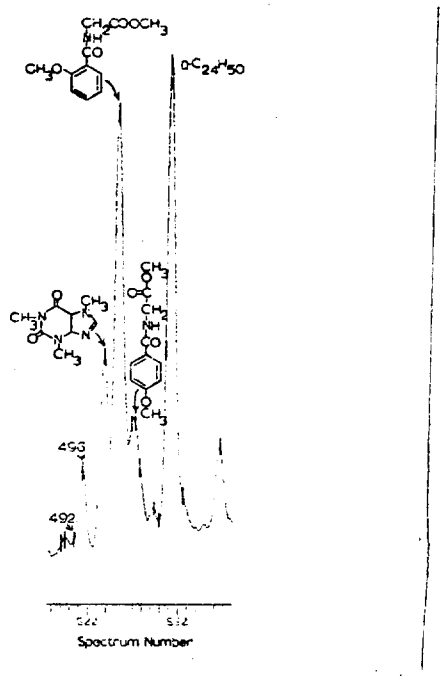

Figure 3. Total Ion Current vs. Spectrum Number in a GC/LRMS Run

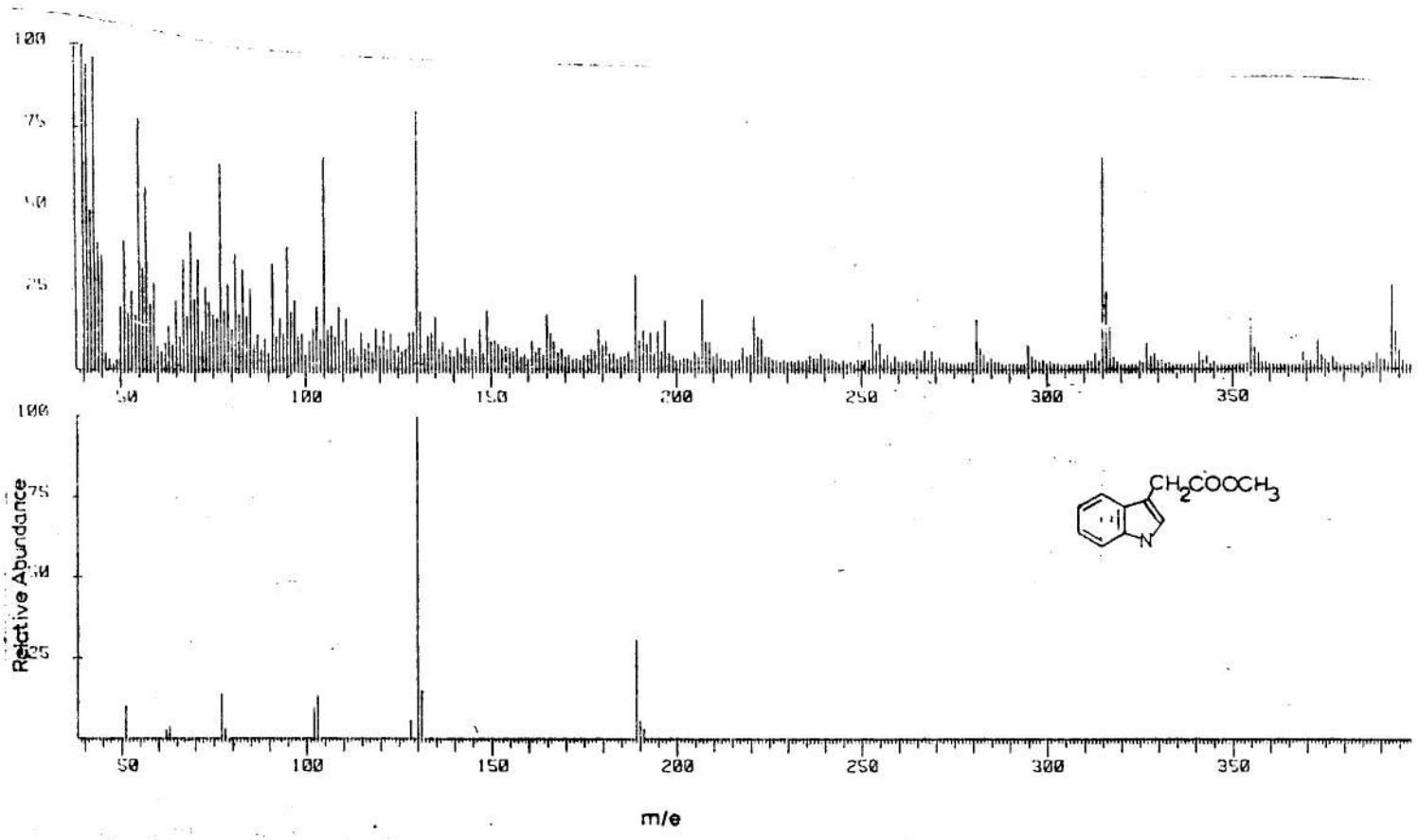

Figure 4. Spectrum number 492 from the GC/LRMS trace shown in Figure 3: (top) Raw data; (bottom) Spectrum output by CLEANUP

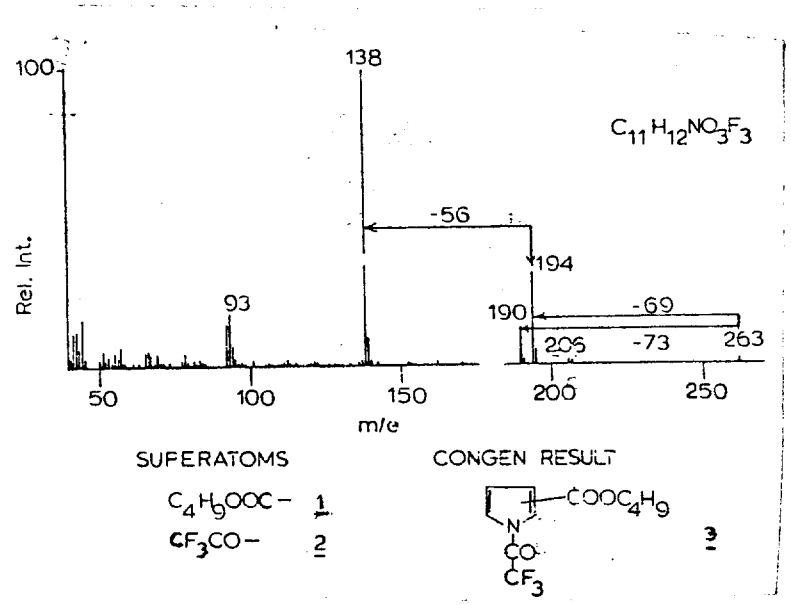

Figure 5. Low-resolution Mass Spectrum of Unknown X. The indicated superatoms were deduced from the spectrum and the chemical history of the sample. Based on these and other constraints, CONGEN obtains the indicated result.

REFERENCES

(1) Gordon, R. M., Datamation(1975), 21(2), 127.

(2) "World List of Crystallographic Computer Programs," Second Edition, D. P. Shoemaker, Bronder-Offset, Rotterdam, 1966.

(3) Professor Joshua Lederberg, Principle Investigator.

(4) Lederberg, J., Sutherland, G. L., Buchanan, B. G., Feigenbaum, E. A., Robertson, A. V., Duffield, A. M. and Djerassi, C., J. Amer. Chem. Soc.(1969), 91, 2973.

(5) Duffield, A. M., Robertson, A. V., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A. and Lederberg, J., J. Amer. Chem. Soc.(1969), 91, 2977.

(6) Buchanan, B. G., Duffield, A. M. and Robertson, A. V., "Mass Spectrometry: Techniques and Applications," G. W. A. Milne, Ed., p. 121, John Wiley and Sons, New York, 1971.

(7) Dromey, R. G., unpublished results, preprint available on request, Dept. of Computer Science, Serra House, Stanford University, Stanford, Calif. 94305.

(8) Biller, J. E. and Biemann, K., Anal. Lett.(1974), 1974, 515.

(9) Several libraries of mass spectral data are available in various forms. The Aldermaston Data Centre (see the "Mass Spectrometry Bulletin") can provide information on the availability of such libraries.

(10) Hertz, H. S., Hites, R. A. and Biemann, K., Anal. Chem.(1971), 43, 681.

(11) Dromey, R. G., Buchanan, B. G., Smith, D. H., Lederberg, J. and Djerassi, C., J. Org. Chem.(1975), 40, 770.

(12) Smith, D. H., Buchanan, B. G., Engelmore, R. S., Duffield, A. M., Yeo, A., Feigenbaum, E. A.,Lederberg, J. and Djerassi, C., J. Amer. Chem. Soc.(1972), 94, 5962.

(13) Smith, D. H., Buchanan, B. G., Engelmore, R. S., Adlerkreutz, H. and Djerassi, C., J. Amer. Chem. Soc.(1973), 95, 6078.

(14) Smith, D. H. and Carhart, R. E., Abstracts, 169th Meeting of the American Chemical Society, Philadelphia, April 6-11, 1975.

(15) Carhart, R. E., Smith, D. H., Brown, H. and Djerassi, C., J. Amer. Chem. Soc., submitted for publication.

(16) Masinter, L. M., Sridharan, N. S., Lederberg, J and Smith, D. H., J. Amer, Chem, Soc.(1974), 96, 7702.

(17) Masinter, L. M., Sridharan, N. S., Carhart, R. E. and Smith, D. H., J. Amer, Chem. Soc.(1974), 96, 7714.

(18) Brown, H., SIAM Journal of Computing, submitted for publication.

(19) Wipke, W. T. and Dyott, T. M., J. Amer. Chem. Soc.(1974), 96, 4825.

(20) Smith, D. H., Buchanan, B. G., White, W. C., Feigenbaum, E. A., Djerassi, C. and Lederberg, J., Tetrahedron(1973), 29, 3117.

(21) Buchanan, B. G., to appear in the Proceedings of NATO Advanced Study Institute on Computer Oriented Learning Processes, 1974, Bonas, France.

(22) Carhart, R. E., Smith, D. H. and Brown, H., J. Chem. Inf. Comp. Sci., in press (May, 1975).

(23) Smith, D. H., Anal. Chem., in press (May 1975).

(24) Teitelman, W., "INTERLISP Reference Manual," Xerox Corp. (Palo Alto Research Center), Palo Alto, Calif., 1974.

(25) Bobrow, D. G., Burchfiel, J. D. and Tomlinson, R. S., Commun. ACM(1972), 15(3), 135.

(26) VanLehn, K. A., "SAIL User Manual," Stanford Artificial Intelligence Laboratory, Stanford, Calif., 1973.

APPENDIX G

Management Committee Membership

The following are the membership lists of the various SUMEX-AIM management committees at the present time:

AIM EXECUTIVE COMMITTEE:

LEDERBERG, Dr. Joshua (LEDERBERG) (Chairman)
Department of Genetics, S331
Stanford University Medical Center
Stanford, California 94305
(415) 497-5801

AMAREL, Dr. Saul (AMAREL)
Department of Computer Science
Rutgers University
New Brunswick, New Jersey 08903
(201) 932-3546

BREWER, Dr. Carl R. (BREWER)
Biotechnology Resources Branch
National Institutes of Health
Building 31, Room 5B25
9000 Rockville Pike
Bethesda, Maryland 20014
(301) 496-5411

LINDBERG, Dr. Donald (LINDBERG) (Adv Grp Member)
605 Lewis Hall
University of Missouri
Columbia, Missouri 65201
(314) 882-6966

AIM ADVISORY GROUP:
==========

LINDBERG, Dr. Donald (LINDBERG) (Chairman)
605 Lewis Hall
University of Missouri
Columbia, Missouri 65201
(314) 882-6966

AMAREL, Dr. Saul (AMAREL)
Department of Computer Science
Rutgers University
New Brunswick, New Jersey 08903
(201) 932-3546

BREWER, Dr. Carl R. (BREWER) (Executive Secretary)
Biotechnology Resources Branch
National Institutes of Health
Building 31, Room 5B25
9000 Rockville Pike
Bethesda, Maryland 20014
(301) 496-5411

BOBROW, Dr. Daniel G. (BOBROW)
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, California 94304
(415) 494-4438

FEIGENBAUM, Dr. Edward (FEIGENBAUM)
Serra House
Department of Computer Science
Stanford University
Stanford, California 94305
(415) 497-4878

FELDMAN, Dr. Jerome (FELDMAN)
Department of Computer Science
University of Rochester
Rochester, New York
(716) 275-5478

LEDERBERG, Dr. Joshua (LEDERBERG) (Ex-officio)
Principal Investigator - SUMEX
Department of Genetics, S331
Stanford University Medical Center
Stanford, California 94305
(415) 497-5801

MILLER, Dr. George (GMILLER)
The Rockefeller University
1230 York Avenue
New York, New York 10021
(212) 360-1801

REDDY, Dr. D.R. (REDDY)
Department of Computer Science

Carnegie-Mellon University
Pittsburgh, Pennsylvania
(412) 621-2600, Ext. 149

SAFIR, Dr. Aran (SAFIR)
Department of Ophthalmology
Mount Sinai School of Medicine
City University of New York
Fifth Avenue and 100th Street
New York, New York 10029
(212) 369-4721

STANFORD COMMUNITY ADVISORY COMMITTEE

LEDERBERG, Dr. Joshua (LEDERBERG) (Chairman)
Department of Genetics, S331
Stanford University Medical Center
Stanford, California 94305
(415) 497-5801

FEIGENBAUM, Dr. Edward (FEIGENBAUM)
Serra House
Department of Computer Science
Stanford University
Stanford, California 94305
(415) 497-4878

GREENES, Robert A., M.D.
Department of Community and
Preventative Medicine, A152
Stanford University Medical Center
Stanford, California 94305
(415) 497-5492

LEVINTHAL, Dr. Elliott C. (LEVINTHAL)
Department of Genetics, S047
Stanford University Medical Center
Stanford, California 94305
(415) 497-5813

APPENDIX H

User Information - General Brochure

SUMEX-AIM

Revised May 1975

The Stanford University Medical Experimental Computer (SUMEX) was established in January, 1974, to provide the first shared national computing facility for medical research. Directed by Dr. Joshua Lederberg, Professor and Chairman of the Department of Genetics, SUMEX is an innovative effort to help biomedical scientists meet today's research requirements and to explore computer applications in many health fields ranging from basic research to bedside care. The project is funded by a grant from the Division of Research Resources of the National Institutes of Health (Biotechnology Resources Branch) for an initial term that expires in July 1978.

At present, SUMEX consists of a powerful PDP-10 computer available to approved users throughout the United States over a computer communication network on a time-shared basis. The project's goals over the next 5 years are: 1) the encouragement of applications of artificial intelligence in medicine (AIM), and 2) the managerial, administrative and technical demonstration of a nationally-shared technological resource for health research.

Such a resource offers scientists both a significant economic advantage in sharing expensive equipment and a greater opportunity to share ideas about their research. This is especially true in computer science, a field whose intellectual and technological complexity has made it difficult to avert the development of relatively isolated research groups. Each group may then tend to pursue its line of investigation with limited convergence on working programs available from others. The SUMEX-AIM project seeks to lower these barriers to scientific cooperation in the field of artificial intelligence applied to health research.

ARTIFICIAL INTELLIGENCE

The term "artificial intelligence" (AI) refers to research efforts aimed at studying and mechanizing information processing tasks that generally have been considered to require human intelligence. The current emphasis in the field is to understand the underlying principles in efficient acquisition and utilization of material knowledge and representation of conceptual abstractions in reasoning, deductive, and problem-solving activities. AI systems are characterized by complex computational processes that are primarily

non-numeric, e.g., graph-searching and symbolic pattern analysis. They involve procedures whose execution is controlled by different types and forms of knowledge about a given task domain, such as models, fragments of "advice", and systems of constraints or heuristic rules. Unlike conventional algorithms commonly based on a well-tailored method for a given task, AI procedures typically use a multiplicity of methods in a highly conditional manner--depending on the specific data in the task and a variety of sources of relevant information. The tangible objective of this approach is the production of computer programs which, using formal and informal knowledge together with mechanized hypothesis formation and problem-solving procedures, will be more general and effective consultative tools for the clinician and medical scientist.

Each authorized project in the SUMEX-AIM community is concerned in some way with the application of these principles to medical problems. This type of "intelligent" assistance by computer program is perhaps best illustrated by the following brief descriptions of some SUMEX-AIM projects.

DENDRAL

The DENDRAL project at Stanford, under the direction of Dr. Lederberg, Professor Edward Feigenbaum, Computer Science, and Professor Carl Djerassi, Chemistry, is aimed at assisting the biochemist in interpreting molecular structures from mass spectral and other chemical information. In cases where the characteristic spectrum of a compound is not catalogued in a library, the DENDRAL programs carry out the rather laborious processes a chemist must go through to interpret the spectrum from "first principles". By symbolically generating "reasonable" candidate structures from hints within the spectrum and a knowledge of organic chemistry and mass spectrometry, the program infers the unknown structure to be the one which best explains the observed spectrum. There is no direct algorithmic path available to determine such a molecular structure from the spectral data--only the inferential process of hypothesis generation and testing within the domain of reasonable solutions defined by a knowledge of organic and physical chemistry.

This process, as implemented in the computer, is a simplified example of the cycle of inductive hypothesis--deductive verification that is often taught as a model of the scientific method (Whether this is a faithful description of contemporary science is arguable, and how it may be implemented in the human brain is unknown. Regardless, these are useful leads rather than absolute preconditions for the pragmatic improvement of mechanized intelligence for more efficient problem-solving.). The elaboration of these approaches with existing hardware and software technologies is the most promising approach to enhancing computer application to the vaguely structured problems that dominate our task domains.

THE RUTGERS PROJECT
COMPUTERS IN BIOMEDICINE

Professor Saul Amarel, a Rutgers University computer scientist, directs several research efforts designed to introduce advanced methods in computer science--particularly in artificial intelligence and interactive data-base systems--into specific areas of biomedical research.

For example, a group of computer scientists led by Professor Casimir Kulikowski is developing computer-based consultation systems for diseases of the eye in collaboration with Dr. Aran Safir, an ophthalmologist from the Mount Sinai School of Medicine. An important development in this area is the establishment of a national network of collaborators for computer diagnosis and treatment of glaucoma. The computer system, which includes an elaborate pathophysiologic model of the disease, is being tested through the SUMEX-AIM network at three eye centers: Mount Sinai Hospital and Medical Center, New York; Washington University, St. Louis; and The Johns Hopkins University, Baltimore. Glaucoma, in one form or another, affects 2% of all people over 40 years of age. It is a disease in which increased pressure within the eye may lead to irreparable optic nerve damage and blindness. The computer-based program has great potential for assisting clinicians and researchers in understanding the disease, diagnosing it more accurately and improving its treatment.

In another project, Professor Charles Schmidt, a social psychologist, is developing a theory of how people arrive at interpretation of the social actions of others. The theory will be tested in situations such as the psychiatric interview and the legal trial. The computer system which currently represents the theory is called "Believer". It includes a large body of statements about people's motivations and actions. The SUMEX-AIM environment provides an excellent medium for collaboration between Dr. Schmidt's group and other researchers around the Country in the development and testing of this computer-based theory.

The Rutgers project includes, in addition, several fundamental studies in artificial intelligence and system design. These provide much of the support needed for the development of complex systems such as the glaucoma consultation and the "Believer" programs.

MYCIN
Computer-based Consultation in Clinical Therapeutics

Dr. Stanley Cohen, Associate Professor and Head of the Division of Clinical Pharmacology at Stanford, directs this research in collaboration with Dr. Stanton Axline and with computer scientists interested in artificial intelligence and medical computing. An evolving computer program developed to assist physician nonspecialists in the selection of therapy for patients with bacterial infections, MYCIN attempts to model the decision processes of medical experts. It consists of three closely integrated components: the Consultation System asks questions, makes conclusions and gives advice; the Explanation System answers questions from the user to justify the program's advice and explain its methods; and the Rule-Acquisition