

Thus, in addition to providing an information resource about protocols, the use of a graphically-oriented program provided a way to learn about the software style and hardware used in the workstation version of ONCOCIN.

We discontinued the mainframe version of ONCOCIN, and began using the workstation version exclusively. The performance of the mainframe version of ONCOCIN was documented in two evaluation papers that appeared in clinical journals (see Hickam and Kent's papers).

We continued our basic research in the design of advanced therapy-planning programs: the ONYX project. We developed a model for planning which includes techniques from the fields of artificial intelligence, simulation, and decision analysis. Artificial intelligence techniques are used to create a small number of possible plans given the ideal therapy and the patient's past treatment history. Simulation techniques and decision analysis are used to examine and order the most promising plans. Our goal is to allow ONCOCIN to give advice in a wider range of situations; in particular, the system should be able to recommend plans for patients who have an unusual response to chemotherapy.

During this year, Stephen Rappaport, M.D. joined us as a programmer on the therapy planning research. Clinical expertise for ONCOCIN was provided by Richard Lenon, M.D. and Robert Carlson, M.D.

- *Year 8:* This year (1986-87) concentrated on two diverse tasks: 1) scaling up the use of the workstation version of ONCOCIN in the clinic, and 2) generalization of each of the components. The latter task is described in the core research sections of this report(see page 19).

In 1986, we placed the workstation version of ONCOCIN into the Oncology Day Care clinic. This version is a completely different program from the version of ONCOCIN that ran on the DECsystem 20--using protocols entered through the OPAL program, with a new graphical data entry interface, and a revised knowledge representation and reasoning component. One of the Oncology Clinical Fellows (Andy Zelenetz) became responsible for verifying how well our design goals for ONCOCIN had been accomplished. His suggestions have included the addition of key protocols and the ability to have the program used as a data management tool if the complete treatment protocol had not yet been entered into the system. Both of these suggestions were carried out during this year, and the program has achieved wider use in the clinic setting. In addition, laser-printed flowsheets and progress notes have been added to the clinic system.

The process of entering a large number of treatment protocols in a short period of time led to other research topics including: design of an automated system for producing meaningful test cases for each knowledge base, modification of the design and access methods for the time-oriented database, and the development of methods for graphically viewing multiple protocols that are combined into one large knowledge base. These research efforts will continue into the next year. In addition, some of the treatment regimens developed for the original mainframe version are still in use and can be transferred to the new version of ONCOCIN. The process of converting this knowledge will also be undertaken in the next year. As the knowledge base grows, additional mechanisms will be needed for the incremental update and retraction of protocols. Additional changes in the reasoning and interface components of the system are described below.

A new research project related to ONCOCIN was started this last year. We are exploring the use of continuous speech recognition as an alternate entry method for communicating with ONCOCIN. This project requires the connection of speech recognition equipment produced by Speech Systems, Inc. of Tarzana to the ONCOCIN interface module. Christopher Lane has already developed a prototype network connection and command interpreter between the speech module (running on a Sun with special hardware added) and the Xerox 1186 computer that runs ONCOCIN. Clifford Wulfman has designed a series of modifications to the ONCOCIN user interface to allow for verbal commands. Graduate student Danielle Fafchamps has helped to design experiments to elicit how clinicians would like to phrase their requests to ONCOCIN.

Janice Rohn is creating a new version of the Librarian program which facilitates the physician's initial communication with the ONCOCIN system (based on the original version by Cliff Wulfman). We continue to collaborate with Andy Zelenetz, Richard Lenon, Robert Carlson, and Charlotte Jacobs on the design and implementation of ONCOCIN in the clinic. Stephen Rappaport has started a residency program to continue his medical education.

## *C.2 Research in Progress*

Our research in the ONCOCIN project over the last year comprised three major categories: (1) conversion of ONCOCIN to the workstation version, (2) development of a knowledge acquisition interface (OPAL) for entering new protocols, and (3) modeling of the strategic therapy selection process (ONYX). We are now able to explore ways to test the system beyond the Stanford environment.

A summary of our current research endeavors follows.

### *C.2.1 Transfer of the ONCOCIN system from the DEC-20 to the Xerox 1100 Series machines*

During the process of converting to the workstation version of ONCOCIN, we redesigned segments of the program. We have completed the major portion of that work, and our experience with the new version has suggested additional areas for improving the reasoning techniques and knowledge representation of ONCOCIN.

- *Redesign of the reasoning component.* A major impetus for the redesign of the system was to develop more efficient methods to search the knowledge base during the running of a case. We have implemented a reasoning program that uses a discrimination network to process the cancer protocols. This network provides for a compact representation of information which is common to many protocols but does not require the program to consider and then disregard information related to protocols that are irrelevant to a particular patient. We continue to improve portions of the reasoning component that are associated with reasoning over time; e.g., modeling the appropriate timing for ordering tests and identifying the information which needs to be gathered before the next clinic visit. In general, we are concentrating on improving the representation of the knowledge regarding sequences of therapy actions specified by the protocol.

Our experience with adding a large number of protocols has led to the evaluation of the design of the internal structure of the knowledge base (e.g., the way we describe the relationships between chemotherapies, drugs, and treatment visits). We will continue to improve the method for traversing

the plan structure in the knowledge base, and consider alternative arrangements for representing the structure of chemotherapy plans. Currently, the knowledge base of treatment guidelines and the patient database are separated. We propose to tie these two structures closer together. Additional work is anticipated on turning ONCOCIN into a critiquing system, where the physician enters their therapy and ONCOCIN provides suggestions about possible alternatives to the entered therapy. Although we have concentrated our review of the ONCOCIN design primarily on the data provided by additional protocols, we know that non-cancer therapy problems may also raise similar issues. The E-ONCOCIN effort is designed to produce a domain-independent therapy planning system that includes the lessons learned from our oncology research. Samson Tu is primarily responsible for continued improvement of the reasoning component of ONCOCIN.

- *Development of a temporal network.* The ability to represent temporal information is a key element of programs that must reason about treatment protocols. The earlier version of the ONCOCIN system did not have an explicit structure for reasoning about time-oriented events. We are experimenting with different configurations of the temporal network, and with the syntax for querying the network. We are also adapting this network so that it can interface with the ONYX therapy-planning systems. This research on temporal reasoning is part of Michael Kahn's Ph.D. thesis. Michael is a student in the Medical Information Sciences Program at University of California at San Francisco.
- *Extensions to the user interface.* We continue to experiment with various configurations of the user interface. Many of the changes have been in response to requests for a more flexible data management environment. We are occasionally faced with data that becomes available corresponding to a time before the current visit. This can happen if a laboratory result is delayed, or a patient's electronic flowsheet is started in the middle of the treatment. We have added the ability to create new columns of data, and are designing the changes to the temporal processing components of ONCOCIN to allow for data that is inserted out of order. We have also extended the flowsheet to allow for patient specific parameters (e.g., special test results or symptoms) that the physician wishes to follow over time. The flowsheet layouts have been modified to create protocol specific flowsheets, e.g., lymphoma flowsheets have a different configuration than lung cancer flowsheets. The basic structure of the interface has been modified to use object-oriented methods, which allows for more flexible interaction between different components of the flowsheet and the operations performed on the flowsheet.

A continuing area of research concerns how to guide the user to the most appropriate items to enter (based on the needs of the reasoning program) without disrupting the fixed layout of the flowsheet. The mainframe version of ONCOCIN modified the order of items on the flowsheet to extract necessary information from the user. In the workstation version, we have developed a guidance mechanism which alerts the user to items that are needed by the reasoning program. The user is not required to deviate from a preferred order of entry nor required to respond to a question for which no current answer is available. Cliff Wulfman is primarily responsible for improvements to the user interface of ONCOCIN.

- *System support for the reorganization.* The LISP language, which we used to

build the first version of ONCOCIN, does not explicitly support basic knowledge manipulation techniques (such as message passing, inheritance techniques, or other object-oriented programming structures). These facilities are available in some commercial products, but none of the existing commercial implementations provide the reliability, speed, size, or special memory-manipulation techniques that are needed for our project. We have therefore developed a "minimal" object-oriented system to meet our specifications. The object system is currently in use by each component of the new version of ONCOCIN and in the software used to connect these components. In addition, all ONCOCIN student projects are now based on this programming environment. Christopher Lane created and is responsible for modifications to the object-oriented system.

### *C.2.2 Interactive Entry of Chemotherapy Protocols by Oncologists (OPAL)*

A major effort in this grant year has been the continued development and testing of software (the OPAL system) that will permit physicians who are not computer programmers to enter protocol information on a structured set of forms presented on a graphics display. Most expert systems require tedious entry of the system's knowledge. In many other medical expert systems, each segment of knowledge is transferred from the physician to the programmer, who then enters the knowledge into the expert system. We have taken advantage of the generally well-structured nature of cancer treatment plans to design a knowledge entry program that can be used directly by clinicians. The structure of cancer treatment plans includes:

- choosing among multiple protocols (that may be related to each other);
- describing experimental research arms in each protocol;
- specifying individual drugs and drug combinations;
- setting the drug dosage level;
- and modifying either the choice of drugs or their dosage.

Using the graphics-oriented workstations, this information is presented to the user as computer-generated forms which appear on the screen. After the user fills in the blanks on the forms, the program generates the rules used to drive the reasoning process. As the user describes more detailed aspects of the protocol, new forms are added to the computer display; these allow the user to specify the special cases that make the protocols so complicated. Although the user is unaware of the creation of the knowledge base from the interaction with OPAL, a complex set of translations are taking place. The user's entries are mapped into an intermediate data structure (IDS) that is common for all protocols. From the IDS, a translation program generates rules for creating and modifying treatment, and integrates them with the existing ONCOCIN knowledge base. Improving the design of the IDS and the rule translation programs will be a major research effort of this year.

Although the "forms" were specifically designed for cancer treatment plans, the techniques used to organize data can be extended to other clinical trials, and eventually to other structured decision tasks. The key factor is to exploit the regularities in the structure of the task (e.g., this interface has an extensive notion of how chemotherapy regimens are constructed) rather than to try to build a knowledge-entry system that can accept *any* possible problem specification. The OPAL program is based upon a domain-independent forms creation package designed and implemented by David Combs. This program will provide the basis for our extension of OPAL to other application areas.

We have now entered thirty-five protocols covering many different organ systems and styles of protocol design (increased from 6 in last year's annual report). Based on this experience, we are modifying OPAL to increase the percentage of the protocol that can be entered directly by our clinical collaborators. One direction in which we have extended the OPAL program is in providing a graphical interface of nodes and arcs to specify the procedural knowledge about the order of treatments and important decision points within the treatments. This work is described in several papers by Musen.

### *C.2.3 Strategic Therapy Planning (ONYX)*

As mentioned above, we have continued our research project (ONYX) to study the therapy-planning process and to determine how clinical strategies are used to plan therapy in unusual situations. Our goals for ONYX are: (1) to conduct basic research into the possible representations of the therapy-planning process, (2) to develop a computer program to represent this process, and (3) eventually to interface the planning program with ONCOCIN. We have worked with our clinical collaborators to determine how to create therapy plans for patients whose special clinical situation preclude following the standard therapeutic plan described in the protocol document.

The prototype program design has four components: (1) to review the patient's past record and recognize emerging problems, (2) to formulate a small number of revised therapy plans based on existing problems, (3) to determine the results of the generated plans by using simulation, and (4) to weight the results of the simulation and rank order the plans by performing decision analysis. This model is described in the papers by Langlotz.

We have built an expert system based on decision analytic techniques as part of the solution to the fourth step of the ONYX planning problem. The program carries out a dialogue with the user concerning the particular treatment choices to be compared, potential problems with the treatments, and the patient-specific utilities corresponding to the possible outcomes. A decision tree is automatically created, displayed on the screen, and solved. The solution is presented to the user, and is compatible with an explanation program for decision trees being developed as part of the Ph.D. research of Curtis Langlotz.

### *C.2.4 Documentation*

In 1986, we videotaped a lecture and demonstration of the ONCOCIN and OPAL systems at the XEROX Palo Alto Research Center. This videotape is available for loan from our offices. Our previous videotapes have been shown at scientific meetings and have been distributed to many researchers in other countries. The publications described below further document our recent work on ONCOCIN.

### *C.2.5 Dissemination*

We are planning experimental installation of ONCOCIN workstations in private oncology offices in San Jose and San Francisco. An application proposing this project is currently under review.

### *D. Publications Since January, 1986*

1. Musen, M.A., Rohn, J.A., Fagan, L.M., and Shortliffe, E.H. Knowledge engineering for a clinical trial advice system: Uncovering errors in protocol specification (Memo KSL-85-51). Proceedings of AAMSI Congress 86 (A. Levy and B. Williams, eds.), pp. 24-27, Anaheim, 8-10 May 1986.
2. Langlotz, C.P., Fagan, L.M., and Shortliffe, E.H. Overcoming limitations of

- artificial intelligence planning techniques. Memo KSL-85-52. Proceedings of AAMSI Congress 86 (A. Levy and B. Williams, eds.), pp. 92-96, Anaheim, 8-10 May 1986.
3. Musen, M.A., Fagan, L.M., and Shortliffe, E.H. Graphical specification of procedural knowledge for an expert system. Memo KSL-85-53. Presented at the Second IEEE Computer Society Workshop on Visual Languages, pp. 167-178, Dallas, TX, June 1986. Reprinted in Expert Systems: The User Interface (J. Hendler, ed.). Norwood, NJ: Ablex Publishing Company, 1987.
  4. Langlotz, C.P., Fagan, L.M., Tu, S.W., Sikic, B.I., and Shortliffe, E.H. A therapy planning architecture that combines decision theory and artificial intelligence techniques. KSL-85-55. Submitted for publication, November 1986.
  5. Combs, D.M., Musen, M.A., Fagan, L.M., and Shortliffe, E.H. Graphical entry of procedural and inferential knowledge. Memo KSL-85-56. Proceedings of AAMSI Congress 86 (A. Levy and B. Williams, eds.), pp. 298-302, Anaheim, 8-10 May 1986.
  6. Lane, C.D., Frisse, M.E., Fagan, L.M., and Shortliffe, E.H. Object-oriented graphics in medical interface design. Memo KSL-85-58. Proceedings of AAMSI Congress 86 (A. Levy and B. Williams, eds.), pp. 293-297, Anaheim, 8-10 May 1986.
  7. Musen, M.A., Fagan, L.M., Combs, D.M., and Shortliffe, E.H. Facilitating knowledge entry for an oncology therapy advisor using a model of the application area. Memo KSL-86-1. Proceedings of MEDINFO-86, pp. 46-50, Washington, D.C., October 1986.
  8. Langlotz, C.P., Fagan, L.M., Tu, S.W., Sikic, B.I., and Shortliffe, E.H. Combining artificial intelligence and decision analysis for automated therapy planning assistance. Memo KSL-86-3. Proceedings of MEDINFO-86, pp. 794-798, Washington, D.C., October 1986.
  9. Kahn, M.G., Fagan, L.M., and Shortliffe, E.H. Context-specific interpretation of patient records for a therapy advice system. Memo KSL-86-4. Proceedings of MEDINFO-86, pp. 175-179, Washington, D.C., October 1986.
  10. Musen, M.A., Fagan, L.M., Combs, D.M., and Shortliffe, E.H. Use of a domain model to drive an interactive knowledge-editing tool. Memo KSL-86-24. To appear in the International Journal of Man-Machine Studies, 1987.
  11. Langlotz, C.P., Shortliffe, E.H., and Fagan, L.M. Using decision theory to justify heuristics. Memo KSL-86-26. Proceedings of AAAI-86, pp. 215-219, Philadelphia, August 1986.
  12. Shortliffe, E.H. Artificial Intelligence in Management Decisions: ONCOCIN. Memo KSL-86-39. Proceedings of a Conference on Medical Information Sciences, University of Texas Health Sciences Center at San Antonio, July 1985. To appear in Frontiers of Medical Information Sciences, Praeger Publishing, 1986.
  13. Lane, C. The Ozone (O<sub>3</sub>) Reference Manual. KSL-86-40, July 1986.

14. Musen, M.A., Combs, D.M., Walton, J.D., Shortliffe, E.H., and Fagan, L.M. OPAL: Toward the computer-aided design of oncology advice systems. Memo KSL-86-49. Proceedings of the Tenth Annual Symposium on Computer Applications in Medical Care, pp. 43-52, Washington, D.C., October 1986. Reprinted in Topics in Medical Artificial Intelligence (P.L. Miller, ed.), New York: Springer-Verlag, 1987.
15. Shortliffe, E.H. Medical expert systems: Knowledge tools for physicians. Memo KSL-86-52. Special issue on Medical Informatics, West. J. Med. 145:830-839, 1986.
16. Shortliffe, E.H. Medical expert systems research at Stanford University. Memo KSL-86-53. Presented at the Twentieth IBM Computer Science Symposium, Shizuoka, Japan, October 1986.
17. Langlotz, C.P., Shortliffe, E.H., and Fagan, L.M. A methodology for computer-based explanation of decision analysis. Working paper, KSL-86-57, November 1986.
18. Shortliffe, E.H. Computers in support of clinical decision making. Memo KSL-87-25, 1986. To appear in Lippincott's forthcoming Textbook of Internal Medicine (W.N. Kelley, ed.).
19. Langlotz, C.P. and Shortliffe, E.H. The relationship between decision theory and default reasoning. Working paper KSL-87-17, 1987.
20. Shortliffe, E.H. Computer programs to support clinical decision making. Memo KSL-87-30. To appear in JAMA, July 1987.

#### *E. Funding Support*

Grant Title: "Therapy-planning strategies for consultation by computer"  
 Principal Investigator: Edward H. Shortliffe  
 Project Management: Lawrence M. Fagan  
 Agency: National Library of Medicine  
 ID Number: LM-04136  
 Term: April 1987 to March 1990  
 Total award: \$380,123

Grant Title: "Knowledge Management for Clinical Trial Advice Systems"  
 Principal Investigator: Edward H. Shortliffe  
 Project Management: Lawrence M. Fagan  
 Agency: National Library of Medicine  
 ID Number: 1 R01 LM04420-01  
 Term: September 1985 through August 1988  
 Total award: \$314,707

Grant Title: Postdoctoral Training in Medical Information Science  
 Principal Investigator: Edward H. Shortliffe  
 Project Management: Edward H. Shortliffe  
 Agency: National Library of Medicine  
 ID Number: 1 T32 LM07033  
 Term: July 1, 1984 - June 30, 1989  
 Total award: \$903,718

Grant Title: Henry J. Kaiser Faculty Scholar in General Internal Medicine

Principal Investigator: Edward H. Shortliffe  
Agency: Henry J. Kaiser Family Foundation  
Term: July 1983 to June 1988  
Total award: \$250,000 (\$50,000 annually).

Grant Title: Explanation of Computer-assisted therapy plans  
Principal Investigator: Lawrence M. Fagan  
Agency: National Institutes of Health  
ID Number: 1 R23 LM04316  
Term: 2/1985-1/1988  
Total award: \$107,441

## II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

### *A. Medical Collaborations and Program Dissemination via SUMEX*

A great deal of interest in ONCOCIN has been shown by the medical, computer science, and lay communities. We are frequently asked to demonstrate the program to Stanford visitors. We also demonstrated our developing workstation code in the Xerox exhibit in the trade show associated with AAAI-84 in Austin, Texas, IJCAI-85 in Los Angeles, AAAI-86 in Philadelphia, and Medinfo 86. Physicians have generally been enthusiastic about ONCOCIN's potential. The interest of the lay community is reflected in the frequent requests for magazine interviews and television coverage of the work. Articles about MYCIN and ONCOCIN have appeared in such diverse publications as *Time* and *Fortune*, and ONCOCIN has been featured on the "NBC Nightly News," the PBS "Health Notes" series, and "The MacNeil-Lehrer Report." Most recently it appeared in a special on Artificial Intelligence for TV Ontario (Canadian PBS station). Due to the frequent requests for ONCOCIN demonstrations, we have produced a videotape about the ONCOCIN research which includes demonstrations of our professional workstation research projects and the 2020-based clinic system. The tape has been shown at several national meetings, including the 1984 Workshop on Artificial Intelligence in Medicine, the 1984 meeting of the Society for Medical Decision Making, and the 1985 meeting of the Society for Research and Education in Primary Care Internal Medicine. The tape has also been shown to both national and international researchers in biomedical computing. We have also completed an updated tape.

Our group also continues to oversee the MYCIN program (not an active research project since 1978) and the EMYCIN program. Both systems continue to be in demand as demonstrations of expert systems technology. MYCIN has been demonstrated via networks at both national and international meetings in the past, and several medical school and computer science teachers continue to use the program in their computer science or medical computing courses. Researchers who visit our laboratory often begin their introduction by experimenting with the MYCIN/EMYCIN systems. We also have made the MYCIN program available to researchers around the world who access SUMEX using the GUEST account. EMYCIN has been made available to interested researchers developing expert systems who access SUMEX via the CONSULT account. One such consultation system for psychopharmacological treatment of depression, called Blue-Box (developed by two French medical students, Benoit Mulsant and David Servan-Schreiber), was reported in July of 1983 in *Computers and Biomedical Research*.

### *B. Sharing and Interaction with Other SUMEX-AIM Projects*

The community created on the SUMEX resource has other benefits which go beyond actual shared computing. Because we are able to experiment with other developing systems, such as INTERNIST/CADUCEUS, and because we frequently interact with

other workers (at AIM Workshops or at other meetings), many of us have found the scientific exchange and stimulation to be heightened. Several of us have visited workers at other sites, sometimes for extended periods, in order to pursue further issues which have arisen through SUMEX- or workshop-based interactions. In this regard, the ability to exchange messages with other workers, both on SUMEX and at other sites, has been crucial to rapid and efficient dissemination of ideas. Certainly it is unusual for a small community of researchers with similar scholarly interests to have at their disposal such powerful and efficient communication mechanisms, even among those researchers on opposite coasts of the country.

During this past two years, we have had extensive interactions with Randy Miller at Pittsburgh. Via floppy disks and SUMEX, we have experimented with several versions of the QMR program. The interaction was very much facilitated by the availability of SUMEX for communication and data transmission.

### *C. Critique of Resource Management*

Our community of researchers has been extremely fortunate to work on a facility that has continued to maintain the high standards that we have praised in the past. The staff members are always helpful and friendly, and work as diligently to please the SUMEX community as to please themselves. As a result, the computer is as accessible and easy-to-use as they can make it. More importantly, it is a reliable and convenient research tool. We extend special thanks to Tom Rindfleisch for maintaining such high professional standards. As our computing needs grow, we have increased our dependence on special SUMEX skills such as networking and communication protocols.

## III. RESEARCH PLANS

### *A. Project Goals and Plans*

In the coming year, there are several areas in which we expect to expend our efforts on the ONCOCIN System:

1. *Development of a workstation model for cost-effective dissemination of clinical consultation systems.* To meet this specific aim we will continue the basic and applied programming efforts (ONCOCIN, OPAL, and ONYX) described earlier in this report.
2. *To encode and implement for use by ONCOCIN the commonly used chemotherapy protocols from our oncology clinic.* In the upcoming year, we will:
  - Extend the OPAL protocol entry system
  - Continue entry of additional protocols at the rate of one protocol/month (including testing)
3. *To continue testing of the workstation version of ONCOCIN.*
4. *To generalize the reasoning and interaction components of the ONCOCIN system for other applications.*

### *B. Justification and Requirements for Continued SUMEX Use*

All the work we are doing (ONCOCIN plus continued use of the original MYCIN program) continues to be dependent on daily use of the SUMEX resource. Although much of the ONCOCIN work has shifted to Xerox workstations, the SUMEX 2060 and

the 2020 continue to be key elements in our research plan. The programs all make assumptions regarding the computing environment in which they operate.

In addition, we have long appreciated the benefits of GUEST and network access to the programs we are developing. SUMEX greatly enhances our ability to obtain feedback from interested physicians and computer scientists around the country. Network access has also permitted high quality formal demonstrations of our work both from around the United States and from sites abroad (e.g., Finland, Japan, Sweden, Switzerland).

The main development of our project will continue to take place on LISP machines which we have purchased or which have been donated by the XEROX Corporation.

### *C. Requirements for Additional Computing Resources*

The acquisition of the DEC 2020 by SUMEX was crucial to the growth of our research work. It ensured high quality demonstrations and has enabled us to develop a system (ONCOCIN) for real-world use in a clinical setting. As we have begun to develop systems that are potentially useful as stand-alone packages (i.e., an exportable ONCOCIN), the addition of personal workstations has provided particularly valuable new resources. We have made a commitment to the smaller Interlisp-D machines ("D-machines") produced by Xerox, and our work will increasingly transfer to them over the next several years. Our current funding supports our effort to implement ONCOCIN on workstations in the Stanford oncology clinic (and eventually to move the program to non-Stanford environments), but we will simultaneously continue to require access to Interlisp on upgraded workstations for extremely CPU-intensive tasks. Although our dependence on SUMEX for workstations has decreased due to a recent gift from XEROX, our requirements for network support of the machines has drastically increased. Individual machines do not provide sufficient space to store all of the software used in our project, nor to provide backup or long-term storage of work in progress. It is the networks, file storage devices, protocol converters, and other parts of the SUMEX network that hold our project together. In addition, with a research group of about 20 people, we are taking advantage of file sharing, electronic mail, and other information coordinating activities provided by the DEC 2060. We hope that with systems support and research by SUMEX staff, we will be able to gradually move away from a need for the central coordinating machine over the next five years.

The acquisition of the DEC 2060, coupled with our increasing use of workstations, has greatly helped with the problems in SUMEX response time that we had described in previous annual reports. We are extremely grateful for access both to the central machine and to the research workstations on which we are currently building the new ONCOCIN prototype. The D-machine's greater address space is permitting development of the large knowledge base that ONCOCIN requires. The graphics capability of the workstations has also enabled us to develop new methods for presenting material to naive users. In addition, the workstations have provided a reliable, constant "load-average" machine for running experiments with physicians and for development work. The development of ONCOCIN on the D-machine will demonstrate the feasibility of running intelligent consultation systems on small, affordable machines in physicians' offices and other remote sites.

### *D. Recommendations for Future Community and Resource Development*

SUMEX is providing an excellent research environment and we are delighted with the help that SUMEX staff have provided implementing enhanced system features on the 2060 and on the workstations. We feel that we have a highly acceptable research environment in which to undertake our work. Workstation availability is becoming increasingly crucial to our research, and we have found over the past year that workstation access is at a premium. The SUMEX staff has been very helpful and understanding about our needs for workstation access, allowing us D-machine use

wherever possible, and providing us with systems-level support when needed. We look forward to the arrival of additional advanced workstations and the development of a more distributed computing environment through SUMEX-AIM.

*E. Responses to Questions Regarding Resource Future*

1. "What do you think the role of the SUMEX-AIM resource should be for the period after 7/86, e.g., continue like it is, discontinue support of the central machine, act as a communications crossroads, develop software for user community workstations, etc.?"

We believe that the trend towards distributed computing that characterized the early 1980's will continue during the second half of the decade. Although we have begun this process by moving much of our research activity to LISP machines, the SUMEX DEC-20 continues to be a major source of support for all communication, collaboration, and administrative functions. It also continues to provide a quality LISP environment for rapid prototyping, student projects in the early stages before workstations are made available, and for demonstrating system features to people at a distance. These latter functions are still not well handled by distributed machines, and we believe that a logical role for the resource in the future is to develop software and communications techniques that will allow us to further decrease our dependence on the large central machine.

2. "Will you require continued access to the SUMEX-AIM 2060 and if so, for how long?"

As indicated above, our needs could still be met with a gradual phaseout of the 2060 over the next 3-5 years, provided that current services such as file handling and backup, mail, document preparation, and advanced network support are available from other machines (e.g., SAFE file server plus the Medical Computer Science file server). This implies maintenance of an ARPANET connection, connections to other campus machines, and facilities for linking together the heterogeneous collection of computing equipment upon which our research group depends. SUMEX would need to concentrate on providing software support for networks and systems software for workstations if it were to provide the same level of service we now experience while moving to a fully distributed environment.

3. "What would be the effect of imposing fees for using SUMEX resources (computing and communications) if NIH were to require this?"

Since all our research is NIH-supported, we see nothing but administrative headaches without benefits if there were to be a move to require fee-for-service billing for access to shared SUMEX resources. The net effect would simply be a transfer of funds from one arm of NIH to another (assuming that the agencies that currently fund our work could supplement our grants to cover SUMEX charges), and there would be a simultaneous restraining effect on the research environment. The current scheme permits experimentation and flexibility in use that would be severely inhibited if all access incurred an incremental charge.

4. "Do you have plans to move your work to another machine workstation and if so, when and to what kind of system?"

As mentioned above, and described in greater detail in our annual report, we are making a major effort to move much of our research activity to LISP machines (currently Xerox 1108's, 1186's and HP-9836's). Our familiarity with this technology, and our commitment to it, have resulted solely from the foresight of the SUMEX resource in anticipating the technology and providing for it at the time of their last renewal. However, for the reasons mentioned above, we continue to depend upon the central communication node for many aspects of our activities and could effectively adapt to its demise only if the phaseout were gradual and accompanied by improved support for a totally distributed computing environment.

## IV.A.4. PROTEAN Project

### PROTEAN Project

Oleg Jardetzky  
Nuclear Magnetic Resonance Lab, School of Medicine  
Stanford University

Bruce Buchanan, Ph.D.  
Computer Science Department  
Stanford University

### I. SUMMARY OF RESEARCH PROGRAM

#### *A. Project Rationale*

The goals of this project are related both to biochemistry and artificial intelligence: (a) use existing AI methods to aid in the determination of the 3-dimensional structure of proteins in solution (not from x-ray crystallography proteins), and (b) use protein structure determination as a test problem for experiments with the AI problem solving structure known as the Blackboard Model. Empirical data from nuclear magnetic resonance (NMR) and other sources may provide enough constraints on structural descriptions to allow protein chemists to bypass the laborious methods of crystallizing a protein and using X-ray crystallography to determine its structure. This problem exhibits considerable complexity, yet there is reason to believe that AI programs can be written that reason much as experts do to resolve these difficulties [12].

#### *B. Medical Relevance*

The molecular structure of proteins is essential for understanding many problems of medicine at the molecular level, such as the mechanisms of drug action. Using NMR data from proteins in solution will allow the study of proteins whose structure cannot be determined with other techniques, and will decrease the time needed for the determination.

#### *C. Highlights of Progress*

During the past year, we have expanded our initial prototype program, called PROTEAN, designed on the blackboard model. It is implemented in BB1 (discussed in the Core AI Research section of this report), a framework system for building blackboard systems that control their own problem-solving behavior.

The reasoning component of PROTEAN directs the actions of the Geometry System (GS), a set of programs that performs the computationally intensive task of positioning portions of a molecule with respect to each other in three dimensions. The GS runs in the UNIX environment on a Silicon Graphics IRIS 3020 graphics workstation, which provides computing performance comparable to a VAX 11/780 for our task. The reasoning program (in Lisp in BB1) is coupled to the GS by a local area computer network, maintained by SUMEX.

Pictures of the results of GS computations are displayed on the graphics screen of the IRIS workstation, using a locally developed program called DISPLAY to draw the evolving protein structures at several levels of detail. The DISPLAY program can be used to view structures generated by the GS either under the direct control of the user or as directed by the reasoning system running in BB1. MIDAS and MMS are two

other molecular modeling and display systems to manipulate protein structures, particularly those obtained from crystallographic techniques as found in the Protein Data Bank. The ability to observe structures in three dimensions is essential to understanding the behavior of the PROTEAN's reasoning and geometry systems and provides essential insights on the problem solving process.

PROTEAN embodies the following experimental techniques for coping with the complexities of constraint satisfaction:

1. The problem-solver partitions each problem into a network of loosely-coupled sub-problems. PROTEAN first positions individual pieces of structures and their immediate neighbors within local coordinate systems. It subsequently composes the most constrained partial solutions developed for these sub-problems in a complete solution for the entire protein. This partitioning and composition technique reduces the combinatorics of search.
2. The problem-solver attempts to solve sub-problems and coordinate solutions at multiple levels of abstraction. For example, PROTEAN operates at two levels of abstraction. At the "Solid" level, it positions elements of the protein's secondary structure: alpha-helices, beta-sheets, and coils. At the "Atom" level, it positions the protein's individual atoms. Partial solutions at the solid level reduce the combinatorics of search at the lower level. Conversely, tightly constrained partial solutions at the lower level introduce new constraints on solid level solutions.
3. The problem-solver preserves the "family" of solutions consistent with all constraints applied thus far. For example, in positioning a helix within a partial solution, PROTEAN does not attempt to identify a unique spatial position for the helix. Instead, it identifies the entire spatial volume within which the helix might lie, given the constraints applied thus far. Preserving the family of legal solutions accommodates problems with incomplete constraints; the solution is constrained only as the data indicate. It also accommodates incompatible constraints by permitting disjunctive sub-families, which may be necessary for flexible proteins.
4. The problem-solver applies constraints one at a time, successively restricting the family of solutions hypothesized for different sub-problems. PROTEAN successively applies constraints on the positions of protein structures, restricting spatial volumes within which they may lie. This allows the different kinds of constraints to be applied by integrating their effects on a family of solutions.
5. The problem-solver tolerates overlapping solutions for different sub-problems. For example, in identifying the volume within which structure-a might lie in partial solution 1, PROTEAN may include part of the volume identified for structure-b. Overlapping volumes for two structures indicate either: (a) that the two structures actually occupy disjoint sub-volumes that cannot be distinguished within the larger, overlapping volumes identified for them because the constraints are incomplete; or (b) that the two structures are mobile and alternately occupy the shared volume.
6. The problem-solver reasons explicitly about control of its own problem-solving actions: which sub-problems it will attack, which partial solutions it will expand, and which constraints it will apply. Control reasoning guides the problem-solver to perform actions that minimize computation, while maximizing progress toward a complete solution. It also provides a foundation for the problem-solver's explanation of problem-solving

activities and intermediate partial solutions and for its learning of new control heuristics.

Multiple blackboards in PROTEAN allow several sets of knowledge to be used. A biochemical knowledge base stores information about proteins and secondary structures, amino acids, and atoms. A concept blackboard describes a concept hierarchy of natural types, object types, role types, contexts, constraint types, and problem solving methods. The ACCORD language blackboard explicitly represents the actions that can be taken in the language for arrangement assembly problems. The problem blackboard describes the protein to be solved and all experimental data observed for the molecule. Finally, the evolving solution of the protein structure is built on a third solution blackboard.

PROTEAN determines the structure of a protein by assembling the protein from components at several levels of detail. Initially, the major secondary structures of the protein are positioned relative to each other by considering them as solid structures, ignoring the side chains of the amino acids and representing constraints with respect to atoms of the protein backbone. This *solid level* approximation is sufficient to determine the overall shape of the molecule, but leaves details of the structure indistinct. Second, an *atomic level* representation of the protein including side chains is used with more precise distance, bond length, and bond angle constraints to remove chemically infeasible structures generated at the solid level. The atomic level description allows a more detailed description of the structure, at the cost of larger numbers of components to consider and increased computation time.

The reasoning component of PROTEAN includes domain and control knowledge sources for the assembly of a protein. Each domain knowledge source directs a small portion of the construction of the molecule. These knowledge sources develop partial solutions that position alpha helices, beta strands, and coils at the *solid level* and refine the resulting state families using all available distance constraints. Control knowledge sources determine which of the possible assembly actions is the best to perform at each stage of the problem solving.

We have built a first extension to PROTEAN that assembles a protein at the level of the atomic backbone. The facilities available include programs to manipulate protein data bank files and generate test data automatically, use atomic level constraints to prune solid level solutions, generate example instances of the protein backbone from the solid level structures, and generate candidate structures for unstructured coil segments of a protein. Work is in progress to combine the atomic level of assembly with the solid level to provide additional constraints at the more abstract level of assembly.

The PROTEAN system has been used to construct a complete solution at the solid level of detail for the Lac-repressor headpiece, a protein with fifty-one amino acids consisting of four coil sections and three alpha helices. In this work, the constraints were determined experimentally from NMR studies.

In addition to the Lac-repressor headpiece protein, we have applied PROTEAN to sperm whale myoglobin, T4 lysozyme, and cytochrome B. Each of these latter proteins has a known crystal structure. In each case, we extracted features of the protein structure and distance constraints from the crystal structure to build data sets for PROTEAN. We then applied the PROTEAN system to the resulting data sets to determine the behavior of the system with different kinds of input.

To determine the correctness and capabilities of the PROTEAN method, we have applied PROTEAN to sperm whale myoglobin, a molecule whose crystal structure is known. In this test, we used distance constraints that would be measured as NOEs, overall size information, and the interaction between the heme group and the amino acids. We also systematically explored the dependence of the precision and accuracy of

the solutions on the quality of the input data available. In all cases, the solutions obtained from PROTEAN enclose the actual structure of the molecule, with the best results coming from data that includes many short range constraints.

We have also defined representations for structures such as the heme group in myoglobin and other cofactors that can be used in constraint satisfaction operations to further restrict the positions of the secondary structures in the protein.

The PROTEAN system takes the secondary structure as input. For molecules in solution, the extent of the helical, sheet, and unstructured coil segments of a protein is derived largely from NMR data between backbone and side chain hydrogen atoms. We have developed a knowledge-based system called ABC that uses heuristic knowledge and NMR data to automate this important step in protein structure determination. ABC is implemented using the BBI blackboard architecture. In addition to solving the secondary structure classification problem, ABC provides a flexible and extensible framework for experimenting with identification methods for secondary structures as well as for data interpretation and pattern recognition techniques.

Work is proceeding on several aspects of the protein structure problem, including assembly of several partial arrangements and integration of these pieces of solution into larger structures, using atomic level volume exclusion of atoms and information on sidechain packing to produce more precise atomic level solutions, and developing more appropriate representations for unstructured coil sections of proteins.

#### D. Relevant Publications

1. Altman, R. and Jardetzky, O.: *New strategies for the determination of macromolecular structures in solution*. Journal of Biochemistry (Tokyo), Vol. 100, No. 6, p. 1403-1423, 1986.
2. Altman, R. and Buchanan, B.G.: *Partial Compilation of Control Knowledge*. To appear in Proceedings of the AAAI 1987.
3. Brinkley, J., Cornelius, C., Altman, R., Hayes-Roth, B., Lichtarge, O., Duncan, B., Buchanan, B.G., Jardetzky, O.: *Application of Constraint Satisfaction Techniques to the Determination of Protein Tertiary Structure*. Report KSL-86-28, Department of Computer Science, 1986.
4. Brinkley, James F., Buchanan, Bruce G., Altman, Russ B., Duncan, Bruce S., Cornelius, Craig W.: *A Heuristic Refinement Method for Spatial Constraint Satisfaction Problems*. Report KSL 87-05, Department of Computer Science.
5. Buchanan, B.G., Hayes-Roth, B., Lichtarge, O., Altman, A., Brinkley, J., Hewett, M., Cornelius, C., Duncan, B., Jardetzky, O.: *The Heuristic Refinement Method for Deriving Solution Structures of Proteins*. Report KSL-85-41. October 1985.
6. Garvey, Alan, Cornelius, Craig, and Hayes-Roth, Barbara: *Computational Costs versus Benefits of Control Reasoning*. Report KSL 87-11, Department of Computer Science.
7. Hayes-Roth, B.: *The Blackboard Architecture: A General Framework for Problem Solving?* Report HPP-83-30, Department of Computer Science, Stanford University, 1983.
8. Hayes-Roth, B.: *BBI: An Environment for Building Blackboard Systems that Control, Explain, and Learn about their own Behavior*. Report HPP-84-16, Department of Computer Science, Stanford University, 1984.

9. Hayes-Roth, B.: *A Blackboard Architecture for Control*. Artificial Intelligence 26:251-321, 1985.
10. Hayes-Roth, B. and Hewett, M.: *Learning Control Heuristics in BBI*. Report HPP-85-2, Department of Computer Science, 1985.
11. Hayes-Roth, B., Buchanan, B.G., Lichtarge, O., Hewett, M., Altman, R., Brinkley, J., Cornelius, C., Duncan, B., and Jardetzky, O.: *PROTEAN: Deriving protein structure from constraints*. Proceedings of the AAAI, 1986, p. 904-909.
12. Jardetzky, O.: *A Method for the Definition of the Solution Structure of Proteins from NMR and Other Physical Measurements: The LAC-Repressor Headpiece*. Proceedings of the International Conference on the Frontiers of Biochemistry and Molecular Biology, Alma Alta, June 17-24, 1984, October, 1984.
13. Lichtarge, Olivier: *Structure determination of proteins in solution by NMR*. Ph.D. Thesis, Stanford University, November, 1986.
14. Lichtarge, Olivier, Cornelius, Craig W., Buchanan, Bruce G., Jardetzky, Oleg: *Validation of the First Step of the Heuristic Refinement Method for the Derivation of Solution Structures of Proteins from NMR Data.*, April 1987. Submitted to Proteins: Structure, Function, and Genetics.

#### *E. Funding Support*

Title: Interpretation of NMR Data from Proteins Using AI Methods

PI's: Oleg Jardetzky and Bruce G. Buchanan

Agency: National Science Foundation

Grant identification number: DMB-8402348

Total Award Period and Amount: 2/1/87 - 9/30/89 \$120,000  
(includes direct and indirect costs)

Current award period and amount: 2/1/87 - 9/30/89 \$120,000  
(includes direct and indirect costs)

The following grants and contracts each provide partial funding for PROTEAN personnel.

Title: Modeling Exper Control

PI: Bruce G. Buchanan

Agency: Office of Naval Research

Grant Identification Number: ONR N00014-86-K-0652

Total award period and amount: 6/1/85 - 5/31/85, \$96,879  
(direct and indirect)

Current award period and amount: 6/1/85 - 5/31/85, \$96,879  
 (direct and indirect)  
 PROTEAN component is \$48,440 (direct & indirect) or 50% of grant

Title: Research on Blackboard Problem-Solving Systems

PI's: Edward A. Feigenbaum and Bruce G. Buchanan

Agency: Boeing Computer Services Corporation

Grant identification number: W-271799

Total award period and amount: 8/1/86 - 7/31/87, \$245,432  
 (direct and indirect)

Current award period and amount: 8/1/86 - 7/31/87, \$245,432  
 (direct and indirect)  
 PROTEAN component is \$12,730 (direct & indirect) or 5% of grant

Title: Knowledge-Based Systems Research

PI: Edward A. Feigenbaum

Agency: Defense Advanced Projects Research Agency

Grant identification number: N00039-86-0033

Total award period and amount: 10/1/85 - 9/30/88 \$4,130,230 (in negotiation)  
 (direct and indirect)

Current award period and amount: 10/1/86 - 9/30/87 \$1,549,539  
 (direct and indirect)  
 PROTEAN component is \$29031, or 1.9 % of grant total

## II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

### *A. Medical Collaborations*

Several members of Prof. Jardetzky's research group are involved in this research.

### *B. Interactions with other SUMEX-AIM projects*

We are occasionally in contact with researchers at Robert Langridge's laboratory at the University of San Francisco.

### *C. Critique of Resource Management*

The SUMEX staff has continued to be most cooperative in supporting PROTEAN research. The SUMEX computer facility is well maintained and managed for effective support of our work. The computer network and Lisp workstations are supported very effectively by the SUMEX staff.

### III. RESEARCH PLANS

#### *A. Goals & Plans*

Our long-range goal is to build an automatic interpretation system similar to CRYVALIS (which worked with x-ray crystallography data). In the shorter term, we are building interactive programs that aid in the interpretation of NMR data on small proteins. The current version of PROTEAN has domain and control knowledge sources that implement the reasoning techniques described above to build a solution using a dynamically created strategic plan. These knowledge sources develop partial solutions that position multiple alpha helices, coils, and beta structures at the Solid level and refine those helices using distance, surface, and volume constraints.

PROTEAN also includes programs that use atomic level representations of the amino acid backbone and side chains. These routines use more precise atomic level distance constraints to prune the solutions obtained by the more abstract solid level geometry computations. Programs are also available to find acceptable backbone segments for unstructured coil segments between alpha helices and beta structures.

The proposed research would expand PROTEAN to include knowledge sources that:

1. merge highly constrained partial solutions at the Solid level.
2. propagate emergent constraints at the atomic level back up to the solid level to further restrict the relative positions of superordinate helices, beta sheets, and coils.
3. further restrict the relative locations of atoms relative to one another.
4. select instances of structures to be used as starting points for other kinds of refinement procedures, such as the solution of the Bloch equations, which define the NMR spectrum that can possibly arise from a given structure. These equations provide a very strong test of the correctness of our method, as well as providing an additional constraint on proposed structures.
5. develop efficient and effective control strategies for the solution of intermediate and large molecules.
6. reason about mobility of structures when the data indicate that mobility is possible.

We have built an effective strategy for automatically determining the families of solid level solutions for small proteins, such as the Lac-repressor headpiece. We will extend the current work to develop control strategies to guide PROTEAN's constraint satisfaction in medium and large protein to identify the family of legal protein conformations as efficiently as possible.

#### *B. Justification for continued SUMEX use*

We will continue to use SUMEX for developing parts of the program before integrating them with the whole system. We are using Interlisp to implement PROTEAN within the Blackboard model flexibly and quickly. In addition, the local area network that SUMEX maintains is crucial to the communications between our reasoning system in BB1, running on Xerox Lisp machines, and our geometry programs and display systems, running on the IRIS 3020 workstation.

*C. Need for other computing resources*

At this time our computational resources are almost adequate. However, access to Lisp machines for program development is often a limiting factor in our ability to continue the research. In addition, faster computation of the operations of the GS would be facilitated by a special-purpose array processor or an additional workstation for computing.

## IV.A.5. RADIX Project

### **The RADIX Project: Deriving Medical Knowledge from Time-Oriented Clinical Databases**

**Robert L. Blum, M.D., Ph.D.**  
Department of Computer Science  
Stanford University

**Gio C. M. Wiederhold, Ph.D.**  
Departments of Computer Science and Medicine  
Stanford University

#### **I. SUMMARY OF RESEARCH PROGRAM**

##### *A. Technical Goals - Introduction*

Medical and Computer Science Goals -- The objectives of the RADIX project are 1) Discovery: to provide knowledgeable assistance to a research investigator in studying medical hypotheses on large databases, and to automate the process of hypothesis generation and exploratory confirmation, 2) Summarization: to develop a program and set of techniques for automated summarization of patient records, and 3) Peer Review: to develop a program to assist physician reviewers examine case databases for medical peer review and quality assurance. For system development we have used a subset of the ARAMIS database. We will first describe our work on discovery, followed by summarization and peer review.

##### **RADIX Discovery Module**

Computerized clinical databases and automated medical records systems have been under development throughout the world for at least a decade. Among the earliest of these endeavors was the ARAMIS Project, (American Rheumatism Association Medical Information System) under development since 1969 in the Stanford Department of Medicine. ARAMIS contains records of over 17,000 patients with a variety of rheumatologic diagnoses. Over 62,000 patient visits have been recorded, accounting for 50,000 patient-years of observation. The ARAMIS Project has now been generalized to include databases for many chronic diseases other than arthritis.

The fundamental objective of the ARAMIS Project and many other clinical database projects is to use the data that have been gathered by clinical observation in order to study the evolution and medical management of chronic diseases. Unfortunately, the process of reliably deriving knowledge has proven to be exceedingly difficult. Numerous problems arise stemming from the complexity of disease, therapy, and outcome definitions, from the complexity of causal relationships, from errors introduced by bias, and from frequently missing and outlying data. A major objective of the RADIX Project is to explore the utility of symbolic computational methods and knowledge-based techniques at solving some of these problems.

The RADIX computer program is designed to examine a time-oriented clinical database such as ARAMIS and to produce a set of (possibly) causal relationships. The algorithm exploits three properties of causal relationships: time precedence, correlation, and nonspuriousness. First, a Discovery Module uses lagged, nonparametric correlations to generate an ordered list of tentative relationships. Second, a Study Module uses a

knowledge base (KB) of medicine and statistics to try to establish nonspuriousness by controlling for known confounders.

The principal innovations of RADIX are the Study Module and the KB. The Study Module takes a causal hypothesis obtained from the Discovery Module and produces a comprehensive study design, using knowledge from the KB. The study design is then executed by an on-line statistical package, and the results are automatically incorporated into the KB. Each new causal relationship is incorporated as a machine-readable record specifying its intensity, distribution across patients, functional form, clinical setting, validity, and evidence. In determining the confounders of a new hypothesis the Study Module uses previously "learned" causal relationships.

In creating a study design the Study Module follows accepted principles of epidemiological research. It determines study feasibility and study design: cross-sectional versus longitudinal. It uses the KB to determine the confounders of a given hypothesis, and it selects methods for controlling their influence: elimination of patient records, elimination of confounding time intervals, or statistical control. The Study Module then determines an appropriate statistical method, using knowledge stored as production rules. Most studies have used a longitudinal design involving a multiple regression model applied to individual patient records. Results across patients are combined using weights based on the precision of the estimated regression coefficient for each patient.

More recently, we have undertaken a new component to the RADIX program: a knowledge-based discovery module. The goal of the knowledge-based discovery module is to overcome some of the limitations of the original, statistics-based, RX discovery module. In creating disease hypotheses, researchers make extensive use of notions of causation, mechanism of action, tempo, and quantitative sufficiency, as well as detailed knowledge of pathophysiology. We are seeking to automate this process of hypothesis formation by replicating selected discoveries in rheumatology using data from the ARAMIS database.

### **RADIX Summarization Module**

The management of inpatients and outpatients is often complicated by the size and disorganization of patient charts. The current paper chart is ill-suited to serve as the major means of communication among health care providers. In recognition of this problem, computerized patient records are becoming increasingly available. While computerization of records at least renders them legible and available, it does not solve the problem of information overload. The ability to automatically create patient summaries would represent a useful adjunct to a patient record for rapid review of a case, for clinical decision making and patient monitoring, and for surveillance of quality of care. The goal of the RADIX summarization program is to infer a summary of a patient's clinical history from lengthy on-line medical records.

The RADIX summarization program is a knowledge-based system which produces intelligent summaries from a time-oriented data base of Systemic Lupus Erythematosus patients. Medical concepts in the system are represented by three entities of increasing complexity: abnormal primary attributes, abnormal states and diseases. Abnormal states and diseases are derived from the abnormal primary attributes by the Reasoner using a combination of model-driven and data-driven algorithms. Uncertainty associated with the derived states is handled with a Bayesian approach supplemented by boolean predicates, using likelihood ratios obtained from a transformation of the INTERNIST knowledge base. After summarizing the data, the system generates interactive, graphical displays with optional explanation windows.

The prototypes we have implemented have shown that intelligent summarization of medical records is feasible and that interactive graphical display is of great help in

conveying complex medical information. However, the system is still under development and has not been formally evaluated. There is much work remaining to be done in the process of creating a complete, clinically useful summary. The knowledge base must be tested and enlarged, the temporal aspect of the reasoning must be improved and more sophisticated displays must be developed. Finally, although our program currently works only with the ARAMIS data base, we hope to extend it and produce a General Summarization System that could be interfaced with any time-oriented medical data base. This general system would include other data base dictionaries and would allow the user to enter medical knowledge tailored to his data base.

### **RADIX Peer Review Program**

We have begun design of a program to assist physician reviewers with medical peer review and quality assurance. This work builds on the Summarization module, and extends it with a new Screening module. The Summarization module, described above, will allow a reviewer to rapidly scan a detailed, longitudinal record. It will summarize major events in the record by displaying them as labels on a time line. The new Screening module will take as input a reviewer's specification of rules of practice that he is interested in checking in the records. The module will transform these rules into an internal form in which they will be matched against the patient records. The output will be a set of episodes in the patient record in which apparent violations of the rules of practice have occurred. The reviewer will then be able to interactively examine each of these episodes using the Summarization module to determine whether a violation was substantiated by the context in which the medical decision was made.

#### *B. Medical Relevance and Collaboration*

As a test bed for system development, our focus of attention has been on the records of patients with systemic lupus erythematosus (SLE) contained in the Stanford portion of the ARAMIS Data Bank. SLE is a chronic rheumatologic disease with a broad spectrum of manifestations. Occasionally the disease can cause profound renal failure and lead to an early death. With many perplexing diagnostic and therapeutic dilemmas, it is a disease of considerable medical interest.

In the future we anticipate possible collaborations with other project users of the TOD System such as the National Stroke Data Bank, the Northern California Oncology Group, and the Stanford Divisions of Oncology and of Radiation Therapy.

We believe that this research project is broadly applicable to the entire gamut of chronic diseases that constitute the bulk of morbidity and mortality in the United States. Consider five major diagnostic categories responsible for approximately two thirds of the two million deaths per year in the United States: myocardial infarction, stroke, cancer, hypertension, and diabetes. Therapy for each of these diagnoses is fraught with controversy concerning the balance of benefits versus costs.

1. Myocardial Infarction: Indications for and efficacy of coronary artery bypass graft vs. medical management alone. Indications for long-term antiarrhythmics ... long-term anticoagulants. Benefits of cholesterol-lowering diets, exercise, and so forth.
2. Stroke: Efficacy of long-term anti-platelet agents, long-term anticoagulation. Indications for revascularization.
3. Cancer: Relative efficacy of radiation therapy, chemotherapy, surgical excision - singly or in combination. Optimal frequency of screening procedures. Prophylactic therapy.

4. Hypertension: Indications for therapy. Efficacy versus adverse effects of chronic antihypertensive drugs. Role of various diagnostic tests such as renal arteriography in work-up.
5. Diabetes: Influence of insulin administration on microvascular complications. Role of oral hypoglycemics.

Despite the expenditure of billions of dollars over recent years for randomized controlled trials (RCT's) designed to answer these and other questions, answers have been slow in coming. RCT's are expensive in terms of funds and personnel. The therapeutic questions in clinical medicine are too numerous for each to be addressed by its own series of RCT's.

On the other hand, the data regularly gathered in patient records in the course of the normal performance of health care delivery are a rich and largely underutilized resource. The ease of accessibility and manipulation of these data afforded by computerized clinical databases holds out the possibility of a major new resource for acquiring knowledge on the evolution and therapy of chronic diseases.

The goal of the research that we are pursuing on SUMEX is to increase the reliability of knowledge derived from clinical data banks with the hope of providing a new tool for augmenting knowledge of diseases and therapies as a supplement to knowledge derived from formal prospective clinical trials. Furthermore, the incorporation of knowledge from both clinical data banks and other sources into a uniform knowledge base should increase the ease of access by individual clinicians to this knowledge and thereby facilitate both the practice of medicine as well as the investigation of human disease processes.

The medical relevance of the automated summarization program is readily apparent. A practicing physician or medical researcher, faced with a patient chart, often with dozens of visits and scores of attributes, rarely has time to read the entire chart. He (or she) would like a succinct summary of the important events in that patient's record to assist his decision making. The use of computerized medical records improves the quality of information but does not solve the problem of information overload. For this reason, it would be useful to have the ability to automatically summarize patient records into meaningful clinical events.

### *C. Highlights of Research Progress*

#### *C.1 April 1986 to April 1987*

Our primary accomplishments in this period have been the following:

- 1) Design and implementation of a second generation of the automated summarization program.
- 2) Design and implementation of a bit-mapped display program for chronic patient data.
- 3) Development of algorithms for transforming the Internist knowledge base into standard Bayes forma.
- 4) Design of a Peer Review program based on the Summarization program.
- 5) Publication of papers on automated discovery and automated summarization, and presentation of results at medical conferences.
- 6) Training post-doctoral researchers, participants in RADIX, in methods of medical artificial intelligence research.

### *C.1.1 Design and implementation of a second generation of the prototype automated summarization program*

We have designed and implemented a second generation of our prototype automated summarization program. This work is described in Dezegher-Geets, 1987, noted in the publications section. The current program improves upon a prototype implemented by Downs (Downs 1986); the knowledge base has been substantially enlarged, the inference mechanisms refined and enhanced for temporal reasoning, and the graphical display capability has been expanded. The summarization program produces intelligent summaries from a time-oriented data base of Systemic Lupus Erythematosus patients. Medical concepts in the system are represented by three entities of increasing complexity: abnormal primary attributes, abnormal states and diseases. Abnormal states and diseases are derived from the abnormal primary attributes by the Reasoner using a combination of model-driven and data-driven algorithms. Uncertainty associated with the derived states is handled with a Bayesian approach supplemented by boolean predicates, using likelihood ratios obtained from a transformation of the INTERNIST knowledge base. After summarizing the data, the system generates interactive, graphical displays with optional explanation windows.

### *C.1.2 Design and implementation of a bit-mapped display program for chronic patient data*

The new display program provides graphic, synoptic, intelligent displays of chronic patient data. The goals of our implementation are:

- 1) Provide a good approximation of what each user actually wants and needs to see, without excess data.
- 2) Provide "intelligent" grouping of attributes based on knowledge of groups of related attributes, for example related to organ system, differential diagnoses, manifestations, and evidence.
- 3) Provide "intelligent" selection of attributes by prioritizing and selecting attributes by their clinical importance for the patient.
- 4) Provide interactive, editable displays, with choices available immediately through menus for the common displays.

The architecture is designed so that the Display Module sits "on top" of the AI components. It is designed to interact with a separate knowledge base or "expert system". The Display is separated from the knowledge base specifically to make it transportable and generalizable.

The knowledge based component contains knowledge of diseases, disease hierarchies, causal relations, equivalence relationships (e.g. proteinuria is part of Nephrotic syndrome), and so on. The display module has information that such relationships exist in medicine, and when to request specific information from the knowledge base. The Display module's knowledge of general medical concepts that are relevant for display includes the severity, belief, import, differential of a manifestation, complications of a disease, manifestations, organ system or user-specified attribute groupings, causal relationships, and equivalence relationships.

### *C.1.3 Development of algorithms for transforming the Internist knowledge base into standard Bayes form*

INTERNIST-1 is an expert system for diagnosis across a broad spectrum of disease. Over twenty man-years of effort have gone into the construction of its knowledge base which contains relationships between approximately 600 diseases and 4,000 manifestations of disease. A major limitation of INTERNIST-1 is that the quantities