D. Requirements for Additional Computing Resources

With the addition of two new D-machines for this work, our computing needs will be adequately met in the coming 1-2 years at least.

The D-machine's large address space permits development of the large programs that complex computer-aided instruction requires. Graphics enable us to develop new methods for presenting material to naive users. We also plan to use the D-machine as a reliable, constant "load-average" machine, for running experiments with physicians and students. The development of GUIDON2 on the D-machine will demonstrate the feasibility of running intelligent consultation or tutoring systems on small, affordable machines in physicians' offices, schools, and other remote sites.

E. Recommendations for Future Community and Resource Development

As we shift our development of systems to personal Lisp machines, such as the Dolphin, it becomes more difficult to access these programs remotely for access from our homes (so that we may work conveniently during the evenings and weekends) and from remote sites for collaboration and demonstration. This problem will be partly ameliorated by "dial-up" (modem) access to these machines, but the use of bitmapped displays requiring a high bandwidth makes the phone lines inadequate for our purposes. Further technological development of networks, probably involving access over cables, will be necessary.

As computer resources become more distributed, the need for a central machine does not diminish. Programs and knowledge bases continue to be shared, requiring high-speed network connections among computers and file servers. SUMEX-AIM's role will shift slightly over the next few years to accommodate these needs, but its identity as a central resource will only change in kind, not importance. Moreover, sophisticated printing devices, such as the Xerox RAVEN, must necessarily be shared, again using a network. Maintenance of this network and its shared devices will become a key activity for the SUMEX staff. Thus, while computing resources will be provided by the "outboard engines" of personal machines, the community will remain intricately linked and dependent on common, but peripheral, resources.

From this perspective, future resource development should focus on improving the capabilities of networks, file servers, and attached devices to respond to individual requests. Multi-processing becomes a necessity in such an environment, so a request can be honored while the user returns to continue his programming or editing.

6.1.2. MOLGEN Project

MOLGEN - Applications of Artificial Intelligence to Molecular Biology: Research in Theory Formation, Testing, and Modification

Prof. E. Feigenbaum and Dr. P. Friedland
Department of Computer Science
Stanford University

Prof. Charles Yanofsky
Department of Biology
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The MOLGEN project has focused on research into the applications of symbolic computation and inference to the field of molecular biology. This has taken the specific form of systems which provide assistance to the experimental scientist in various tasks, the most important of which have been the design of complex experiment plans and the analysis of nucleic acid sequences. Our current research concentrates on scientific discovery within the subdomain of regulatory genetics. We desire to explore the methodologies scientists use to modify, extend, and test theories of genetic regulation, and then emulate that process within a computational system.

Theory or model formation is a fundamental part of scientific research. Scientists both use and form such models dynamically. They are used to predict results (and therefore to suggest experiments to test the model) and also to explain experimental results. Models are extended and revised both as a result of logical conclusions from existing premises and as a result of new experimental evidence.

Theory formation is a difficult cognitive task, and one in which there is substantial scope for intelligent computational assistance. Our research is toward building a system which can form theories to explain experimental evidence, can interact with a scientist to help to suggest experiments to discriminate among competing hypotheses, and can then revise and extend the growing model based upon the results of the experiments.

The MOLGEN project has continuing computer science goals of exploring issues of knowledge representation, problem-solving, discovery, and planning within a real and complex domain. The project operates in a framework of collaboration between the Heuristic Programming Project (HPP) in the Computer Science Department and various domain experts in the departments of Biochemistry, Medicine, and Biology. It draws from the experience of several other projects in the HPP which deal with applications of artificial intelligence to medicine, organic chemistry, and engineering.

B. Medical Relevance and Collaboration

The field of molecular biology is nearing the point where the results of current research will have immediate and important application to the pharmaceutical and chemical industries. Already, clinical testing has begun with synthetic interferon and human growth hormone produced by recombinant DNA technology. Governmental reports estimate that there are more than 200 new and established industrial firms already undertaking product development using these new genetic tools.

The programs being developed in the MOLGEN project have already proven useful and important to a considerable number of molecular biologists. Currently several dozen researchers in various laboratories at Stanford (Prof. Paul Berg's, Prof. Stanley Cohen's, Prof. Laurence Kedes', Prof. Douglas Brutlag's, Prof. Henry Kaplan's, and Prof. Douglas Wallace's) and over 400 others throughout the country have used MOLGEN programs over the SUMEX-AIM facility. We have exported some of our programs to users outside the range of our computer network (University of Geneva [Switzerland], Imperial Cancer Research Fund [England], and European Molecular Biology Institute [Heidelberg] are examples). The pioneering work on SUMEX has led to the establishment of a separate NIH-supported facility, BIONET, to serve the academic molecular biology research community with MOLGEN-like software. BIONET is now serving many of the computational needs of over 1000 academic molecular biologists in the United States.

C. Highlights of Research Progress

C.1 Accomplishments

The current year has seen the completion of our initial study of the Yanofsky project on genetic regulation in the *trp* operon. In addition we have tested several models of qualitative simulation of biological systems and begun our design of a theory discovery system. Finally, a new application program for DNA sequence analysis was developed by one of our research collaborators. The highlights of this work are summarized in several categories below.

C.1.1 The Scientific Process of Theory Formation, Modification, and Testing

The first goal of our work in scientific theory discovery was to extensively study an existing example of the process. Professor Charles Yanofsky's work in elucidating the structure and function of regulation in the trp operon of E. coli provided us with an excellent subject that spanned twelve years of research, dozens of collaborators, and almost one hundred research papers.

We have conducted extensive interviews with Professor Yanofsky and many of his former students and collaborators. We have examined most of the relevant research papers. We believe we now have a good understanding of the three major classes of knowledge that were important in the discovery of the theory of regulation in the trp operon: knowledge about the relevant biological objects, knowledge about the techniques used to elicit new information, and discovery heuristics used to build new models.

In addition, we have developed an initial model for the inference mechanisms used during the discovery process. This model includes at least four different types of reasoning: data-driven, theory-driven, analogy to closely-related biological systems, and analogy to other systems (railroad engines and tracks, for example).

C.1.2 Knowledge-Based Simulation of the Trp Operon

The first major programming task of our project was to build a knowledge base representing the initial state of knowledge about the tryptophan operon system at the beginning of the Yanofsky research. This initial knowledge base contains information relevant to genetic regulation in general and to the trp operon system in particular. The information relates both to structure, i.e. the physical characteristics of the biological objects, and to function, i.e. the operational characteristics of the biological objects. In addition, the procedural knowledge needed to relate structure to function plays an important part in the knowledge base.

The goal was to have a knowledge base that can be used "actively" to simulate the result of various possible changes in the underlying regulatory model. For example, a common experimental method for studying a biological system is to introduce a mutation which destroys the functionality of some piece of the system. The regulatory knowledge base should be able to simulate and describe the results of such a "deletion mutation."

As a first experiment, we built the knowledge base using the Unit System (developed under previous MOLGEN work). We were able to successfully model most of the important processes of Jacob-Monod repression, the initial model of genetic regulation used in the Yanofsky research.

C.1.3 A Model for Theory Discovery

In parallel with our work on knowledge base construction, we designed an initial architecture for theory proposal, extension, and correction. In human scientists we have observed at least four major types of reasoning during the cognitive process. The first is data-driven reasoning when the major goal is to explain individual experimental results. The second is theory-driven reasoning which occurs when a partial theory or model drives its own extension. The third type of reasoning involves looking at closely related biological systems (e.g, noticing a similar behavior in the his operon system). The final type of reasoning relates to more distant analogies; thinking of DNA polymerase moving along a nucleotide sequence as similar to a railroad engine moving along a set of tracks. Our discovery system architecture embraces all of these reasoning types within a blackboard-style hybrid architecture.

In addition, we have fit our overall model of simulation and discovery into a framework of research on machine learning. This framework involves interacting performance and learning elements. The performance element, here the knowledge-based system for qualitative simulation of regulatory genetics, is asked to explain observations from the real world. The learning element, here the discovery architecture described above, is able to evaluate the explanations and "tune" the performance element by changing its model (or theory) of the world.

C.1.1.4 Simultaneous alignment of DNA sequences--MULTAN

Previously, MOLGEN researchers have developed numerous programs to aid in the symbolic analysis of DNA sequences. During the last year Dr. William Bains (a postdoctoral scholar in Professor Kedes' laboratory), completed a program called MULTAN which allows the facile alignment of three or more DNA sequences. This was a major unsolved problem in sequence analysis and the program is now undergoing final testing on the BIONET resource. In the future, we expect that BIONET will support development of application-oriented programs of this type, while MOLGEN and SUMEX will focus on research-oriented systems with major AI goals.

C.2 Research in Progress

We have two major goals over the next several months. The first is to convert and enhance our knowledge-based simulation model within the KEE tool from IntelliCorp, Inc. KEE will be a significant improvement over the Unit System in three areas: speed, functionality, and support. IntelliCorp is providing KEE for use in our research without charge. Studies have indicated that using KEE will unable us to produce a reasonable prototype of our discovery system in about half the time or using the Unit System. Our second goal is to more formally define the learning element of our discovery system and to build a first test system that operates upon the simulation system knowledge base.

D. Publications

1. Bach, R., Friedland, P., Brutlag, D. and Kedes, L.: *MAXIMIZE, a DNA sequencing strategy advisor.* Nucleic Acids Res. 10(1):295-304, January, 1982.

2. Bach, R., Friedland, P., and Iwasaki, Y.: *Intelligent computational assistance for experiment design.* Nucleic Acids Res. 12(1):11-29, January, 1984.

3. Brutlag, D., Clayton, J., Friedland, P. and Kedes, L.: *SEQ: A nucleotide sequence analysis and recombination system.* Nucleic Acids Res. 10(1):279-294, January, 1982.

4. Clayton, J. and Kedes, L.: *GEL, a DNA sequencing project management system.* Nucleic Acids Res. 10(1):305-321, January, 1982.

5. Feitelson, J. and Stefik, M.J.: *A case study of the reasoning in a genetics experiment.* Heuristic Programming Project Report HPP-77-18 (working paper), May, 1977.

6. Friedland, P.: *Knowledge-based experiment design in molecular genetics.* Proc. Sixth IJCAI, August, 1979, pp. 285-287.

7. Friedland P.: *Knowledge-based experiment design in molecular genetics.* Stanford Computer Science Report STAN-CS-79-760 (Ph.D. thesis), December, 1979.

8. Friedland, P., Kedes, L. and Brutlag D.: *MOLGEN--Applications of symbolic computation and artificial intelligence to molecular biology.* Proc. Battelle Conference on Genetic Engineering, April, 1981.

9. Friedland, P.: *Acquisition of procedural knowledge from domain experts.* Proc. Seventh IJCAI, August, 1981, pp. 856-861.

10. Friedland, P., Kedes, L., Brutlag, D., Iwasaki, Y. and Bach R.: *GENESIS, a knowledge-based genetic engineering simulation system for representation of genetic data and experiment planning.* Nucleic Acids Res. 10(1):323-340, January, 1982.

11. Friedland, P., and Kedes, L.: *Discovering the secrets of DNA.* (To appear in a joint issue of Communications of the ACM and IEEE/Computer, October, 1985).

12. Friedland, P. and Iwasaki Y.: *The concept and implementation of skeletal plans.* (To appear in Journal of Automated Reasoning, Vol. 1, No. 2, 1985).

13. Friedland, P., Armstrong, P., and Kehler, T.: *The role of computers in biotechnology.* BIO\TECHNOLOGY 565-575, September, 1983.

14. Iwasaki, Y. and Friedland, P.: *SPEX: A second-generation experiment design system.* Proc. of Second National Conference on Artificial Intelligence, August, 1982, pp. 341-344.

15. Martin, N., Friedland, P., King, J. and Stefik M.J.: *Knowledge base management for experiment planning in molecular genetics.* Proc. Fifth IJCAI, August, 1977, pp. 882-887.

16. Meyers, S. and Friedland, P.: *Knowledge-based simulation of regulatory genetics in bacteriophage Lambda.* Nucleic Acids Res. 12(1):1-9, January, 1984.

17. Stefik, M. and Friedland, P.: *Machine inference for molecular genetics: Methods and applications.* Proc. of NCC, June, 1978.

18. Stefik, M.J. and Martin N.: *A review of knowledge based problem solving as a basis for a genetics experiment designing system.* Stanford Computer Science Report STAN-CS-77-596, March, 1977.

19. Stefik, M.: *Inferring DNA structures from segmentation data: A case study.* Artificial Intelligence 11:85-114, December, 1977.

20. Stefik, M.: *An examination of a frame-structured representation system.* Proc. Sixth IJCAI, August, 1979, pp. 844-852.

21. Stefik, M.: *Planning with constraints.* Stanford Computer Science Report STAN-CS-80-784 (Ph.D. thesis), March, 1980.

E. Funding Support

The MOLGEN grant is titled: MOLGEN: Applications of Artificial Intelligence to Molecular Biology: Research in Theory Formation, Testing, and Modification. It is NSF Grant MCS-8310236. Current Principal Investigators are Edward A. Feigenbaum Professor of Computer Science and Charles Yanofsky, Professor of Biology. MOLGEN is currently funded from 11/84 to 10/85 at $131,621 including indirect costs as the first year of a three year grant.

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

SUMEX-AIM continues to provide the bulk of our computing resources. The facility has not only provided excellent support for our programming efforts but has served as a major communication link among members of the project. Systems available on SUMEX-AIM such as INTERLISP, TV-EDIT, and BULLETIN BOARD have made possible the project's programming, documentation and communication efforts. The interactive environment of the facility is especially important in this type of project development.

We strongly approve of the network-oriented approach to a programming environment that SUMEX has begun to evolve into. The ability to utilize LISP workstations for intensive computing while still communicate with all of the other SUMEX resources has been very valuable to our work. We see a satisfactory mode of operation where most programming takes place on the workstations and most electronic communications, information sharing, and document preparation takes place within the mature TOPS-20 environment. The evolution of SUMEX has alleviated most of our previous problems with resource loading and file space. Our current workstations are not quite fast nor sophisticated enough, but we are encouraged by the progress that has been made.

We have taken advantage of the collective expertise on medically-oriented knowledge-based systems of the other SUMEX-AIM projects. In addition to especially close ties with other projects at Stanford, we have greatly benefited by interaction with other projects at yearly meetings and through exchange of working papers and ideas over the system.

The ability for instant communication with a large number of experts in this field has been a determining factor in the success of the MOLGEN project. It has made possible the near instantaneous dissemination of MOLGEN systems to a host of experimental users in laboratories across the country. The wide-ranging input from these users has greatly improved the general utility of our project.

We find it very difficult to find fault with any aspect of the SUMEX resource management. It has made it easy for us to expand our user group, to give demonstrations (through the 20/20 adjunct system as well as the LISP workstations), and to disseminate software to non-SUMEX users overseas.

III. RESEARCH PLANS

A. Project Goals And Plans

Our current work has the following major goals:

1. Use the knowledge base to explain observations that are indeed explainable without changes to the current model. For example, "I have observed a mutation that causes constitutive (uncontrolled) production of tryptophan. How can that be explained within the Jacob-Monod model?" This process will be accomplished by some combination of forward simulation and backward rule-chaining.

2. Begin to recognize when observations are "interesting." Interesting here has one of the following broad meanings:

   a. A seeming direct contradiction to the existing theory.

   b. A statistically rare occurrence (one that is understandable by the current theory, but should not occur very often).

   c. A dramatic confirmation of the existing model.

   d. An observation currently unpredictable by the current model because the model is either not detailed enough or incomplete. The observation in this case must have a relation to the model because an important object of the model is involved or it relates to an effect predicted by the model.

3. Build a mechanism for postulating extensions or corrections to the current theory: a contrained regulatory theory generator. The overall approach to this mechanism is perhaps the most interesting problem in our work. In discussions with other computer scientists, the notion of "or" reasoning where the theory construction process consists of hierarchical refinement of abstract ideas into more detailed ones, and "and" reasoning where the theory is built up in little pieces at many different levels simultaneously has emerged. We see strong evidence for both types of reasoning within Yanofsky's project. In fact, as stated above, the global model of Yanofsky's laboratory is a hybrid one. Individual graduate students performed "and" tasks--filling in details of seemingly unrelated pieces of the model. Yanofsky was the master "or" reasoner, slowly building a hierarchical model of the new regulatory mechanism. It is in this area of our research where the greatest discussion with AI colleagues is needed and which may produce the most significant AI benefits.

4. Build a mechanism for evaluating alternative theories. This would include rating the theories based on plausibility, selectability, completeness, significance, and so on. We hope the evaluation process produces information useful in discriminating among the possible theories.

5. Test the entire structure on the evolving trp operon regulatory system. Experiment with different initial knowledge bases to see how the discovery process is altered by the availability of new techniques, analogous systems, etc.

B. Justification and Requirements for Continued SUMEX Use

The MOLGEN project depends heavily on the SUMEX facility. We have already developed several useful tools on the facility and are continuing research toward applying the methods of artificial intelligence to the field of molecular biology. The community of potential users, is growing nearly exponentially as researchers from most of the biomedical-medical fields become interested in the technology of recombinant DNA. We believe the MOLGEN work is already important to this growing community and will continue to be important. The evidence for this is an already large list of pilot exo-MOLGEN users on SUMEX.

We support with great enthusiasm the acquisition of satellite computers for technology transfer and hope that the SUMEX staff continues to develop and support these systems. One of the oft-mentioned problems of artificial intelligence research is exactly the problem of taking prototypical systems and applying them to real problems. SUMEX gives the MOLGEN project a chance to conquer that problem and potentially supply scientific computing resources to a national audience of biomedical-medical research scientists.

6.1.3. ONCOCIN Project

ONCOCIN Project

Edward H. Shortliffe, M.D., Ph.D.
Departments of Medicine and Computer Science
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The ONCOCIN Project is one of many Stanford research programs devoted to the development of knowledge-based expert systems for application to medicine and the allied sciences. The central issue in this work has been to develop a program that can provide advice similar in quality to that given by human experts, and to insure that the system is easy to use and acceptable to physicians. The work seeks to improve the interactive process, both for the developer of a knowledge-based system, and for the intended end user. In addition, we have emphasized clinical implementation of the developing tool so that we can ascertain the effectiveness of the program's interactive capabilities when it is used by physicians who are caring for patients and are uninvolved in the computer-based research activity.

B. Medical Relevance and Collaboration

The lessons learned in building prior production rule systems have allowed us to create a large oncology protocol management system much more rapidly than was the case when we started to build MYCIN. We introduced ONCOCIN for use by Stanford oncologists in May 1981. This would not have been possible without the active collaboration of Stanford oncologists who helped with the construction of the knowledge base and also kept project computer scientists aware of the psychological and logistical issues related to the operation of a busy outpatient clinic.

C. Highlights of Research Progress

C.1 Background and Overview of Accomplishments

The ONCOCIN Project is a large interdisciplinary effort that has involved over 35 individuals since the project's inception in July 1979. With the work currently in its sixth year, we summarize here the milestones that have occurred in the research to date:

- *Year 1:* The project began with two programmers (Carli Scott and Miriam Bischoff), a Clinical Specialist (Dr. Bruce Campbell) and students under the direction of Dr. Shortliffe and Dr. Charlotte Jacobs from the Division of Oncology. During the first year of this research (1979-1980), we developed a prototype of the ONCOCIN consultation system, drawing from programs and capabilities developed for the EMYCIN system-building project. During that year, we also undertook a detailed analysis of the day-to-day activities of the Stanford Oncology Clinic in order to determine how to introduce ONCOCIN with minimal disruption of an operation which is already running smoothly. We also spent much of our time in the first year giving careful consideration to the most appropriate mode of interaction with physicians in order to optimize the chances for ONCOCIN to become a useful and accepted tool in this specialized clinical environment.

- *Year 2:* The following year (1980-1981) we completed the development of a special interface program that responds to commands from a customized keypad. We also encoded the rules for one more chemotherapy protocol (oat cell carcinoma of the lung) and updated the Hodgkin's Disease protocols when new versions were released late in 1980; these exercises demonstrated the generality and flexibility of the representation scheme we had devised. Software protocols were developed for achieving communication between the interface program and the reasoning program, and we coordinated the printing routines needed to produce hard copy flow sheets, patient summaries, and encounter sheets. Finally, lines were installed in the Stanford Oncology Day Care Center, and, beginning in May 1981, eight fellows in oncology began using the system three mornings per week for management of their patients enrolled in lymphoma chemotherapy protocols.

- *Year 3:* During our third year (1981 - 1982) the results of our early experience with physician users guided both our basic and applied work. We designed and began to collect data for three formal studies to evaluate the impact of ONCOCIN in the clinic. This latter task required special software development to generate special flow sheets and to maintain the records needed for the data analysis. Towards the end of 1982 we also began new research into a *critiquing model* for ONCOCIN that involves "hypothesis assessment" rather than formal advice giving. Finally, in 1982 we began to develop a query system to allow system builders as well as end users to examine the growing complex knowledge base of the program.

- *Year 4:* Our fourth year (1982-1983) saw the departure of Carli Scott, a key figure in the initial design and implementation of ONCOCIN, the promotion of Miriam Bischoff to Chief Programmer, and the arrival of Christopher Lane as our second scientific programmer. At this time we began exploring the possibility of running ONCOCIN on a single-user professional workstation and experimented with different options for data-entry using a "mouse" pointing device. Christopher Lane became an expert on the Xerox workstations that we are using. In addition, since ONCOCIN had grown to such a large program with many different facets, we spent much of our fourth year documenting the system. During that year we also modified the clinic system based upon feedback from the physician-users, made some modifications to the rules for Hodgkin's disease based upon changes to the protocols, and completed several evaluation studies.

- *Year 5:* The project's fifth year (1983-1984) was characterized by growth in the size of our staff (three new full-time staff members and a new oncologist joined the group). The increased size resulted from a DRR grant that permitted us to begin a major effort to rewrite ONCOCIN to run on professional workstations. Dr. Robert Carlson, who had been our Clinical Specialist for the previous two years, was replaced by Dr. Joel Bernstein, while Dr. Carlson assumed a position with the nearby Northern California Oncology Group; this appointment permitted him to continue his affiliation both with Stanford and with our research group. In August of 1983, Larry Fagan joined the project to take over the duties of the ONCOCIN Project Director while also becoming the Co-Director of the newly formed Medical Information Sciences Program. Dr. Fagan continues to be in charge of the day-to-day efforts of our research. An additional programmer, Jay Ferguson, joined the group in the fall to assist with the effort required to transfer ONCOCIN from SUMEX to the 1108 workstation. A fourth programmer, Joan Differding, joined the staff to work on our protocol acquisition effort (OPAL).

- *Year 6:* During our sixth year (1984-1985) we have further increased the size of our programming staff to help in the major workstation conversion effort. The ONCOCIN and OPAL efforts were greatly facilitated by a successful application for an equipment grant from Xerox Corporation. With a total of 15 Xerox LISP machines now available for our group's research, all full time programmers have dedicated machines, as do several of the senior graduate students working on the project. Christopher Lane took on full-time responsibility for the integration and maintenance of the group's equipment and associated software. Two of our programming staff moved on to jobs in industry (Bischoff and Ferguson) and three new programmers (David Combs, Cliff Wulfman, and Samson Tu) were hired to fill the void created by their departure and by the reassignment of Christopher Lane.

With daily coordination by the project's data manager, Janice Rohn, the DEC-20 version of ONCOCIN continues to be used on a limited basis in the Stanford Oncology Clinic. The continued dependence on this time-shared computer, however, has prevented us from using ONCOCIN in in many clinical problem areas (other than the lymphomas where clinics are held three mornings per week, and breast cancer where clinic is held one day per week) because of our inability to assure the system's availability with reasonable response time. It is this latter point that has accounted for our decision not to spend a great deal of time developing new protocols to run on the DEC-20 ONCOCIN prototype. Instead we have pressed our effort to adapt ONCOCIN to run on professional workstations which can eventually be dedicated to full time clinic use. We envision these workstations as the model for eventual dissemination of this kind of technology.

In addition to funding from DRR for the workstation conversion effort, we have support from the National Library of Medicine that supports our more basic research activities regarding biomedical knowledge representation, knowledge acquisition, therapy planning, and explanation as it relates to the ONCOCIN task domain. A grant from the NLM to study the therapy planning process was received, and this work (led by Dr. Fagan) is in its second year. This research is investigating how to represent the therapy planning strategies used to decide treatment for patients on the oat cell carcinoma protocol who run into serious problems requiring consultation with the protocol study chairman. Dr. Branimar Sikic, a faculty member from the Stanford University Department of Medicine, and the Study Chairman for the oat cell protocol, is collaborating on this project.

C.2 Research in Progress

The major efforts of the ONCOCIN project over the last year have fallen into three major categories: (1) conversion of ONCOCIN to run on workstations, (2) development of a knowledge acquisition interface (OPAL) for entering new protocols, and (3) research on modeling the strategic therapy selection process (ONYX). Efforts are also in progress to evaluate the system, to document the results of the research, and to disseminate the technology to sites beyond Stanford. We summarize these ongoing research efforts below.

C.2.1 Transfer of the ONCOCIN system from the DEC-20 to the Xerox 1108

In an effort to improve the efficiency of the reimplemented system (and thereby to improve its response time and make it more acceptable to physicians), we have undertaken a substantial system redesign while transferring it to the new machines. An additional commitment in time and programming effort has resulted, but we are confident that the resulting system will be a substantial improvement over the prototype. There have been several aspects to the system's reimplementation during the current year:

- *Reorganization and recoding of existing programs for improved efficiency.* In last year's report, we discussed our first steps in reorganizing the program. A further analysis during the year suggested that we should consider a redesign of the program to take advantage of our experience with the existing program and to respond to advances in artificial intelligence representation methods since ONCOCIN was first designed. In addition, our work during the year on new methods for entering knowledge into the system suggested corresponding improvements in the ways to represent oncologic knowledge in the system (see paper by Musen, et al. for more details on the redesign of the ONCOCIN system).

- *Redesign of the reasoning component.* As a major part of the redesign of the system, we decided to concentrate on methods that would allow for a more efficient search of the knowledge base during the running of a case. We have implemented and are currently debugging a reasoning program that uses a discrimination network to process the cancer protocols. This network allows for a compact representation of information that overlaps elements of multiple protocols, but does not require the program to consider and then disregard information related to protocols that are irrelevant to a particular patient.

- *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 (see the paper by Kahn, et al. for a more detailed description of the temporal network).

- *Extensions to the user interface.* The user interface has been extended so that it can read patient data files of the type that are created by the original ONCOCIN system. This will allow us to transfer currently active patients to the new version of the ONCOCIN system. A detailed description of the user interface is available in the paper by Lane, et al.

- *Connecting the components of the ONCOCIN system.* The reasoning component, user interface, and knowledge acquisition program (described below) have been developed as separate programs. In the final version of the system, the knowledge acquisition program must be able to automatically translate from the graphical input forms into the knowledge base. The reasoner and user interface components are independent programs that run in parallel while communicating with each other. Each of these connections between components has been tested on a limited basis and will continue to be exercised during the next several months.

- *Knowledge engineering tools.* The challenge of coordinating a large software development project, with multiple programmers working in parallel, has necessitated the development of specialized tools to facilitate the process of system construction and maintenance. One area of particular concern has been the need for tools to assist with knowledge base maintenance (see paper by Tsuji and Shortliffe for a discussion of our initial work in this area).

- *System support for the reorganization.* The LISP language that we used to build the first version of ONCOCIN does not explicitly support basic knowledge manipulation techniques (viz. 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 provides the reliability, speed, size, or special memory-manipulation techniques that are needed for our project.

We have accordingly developed a "minimal" object-oriented system to meet these specifications. The object system is currently in use by each component of the new version of ONCOCIN and in the software used to connect the components. In addition, several student projects are now able to use this programming environment.

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

A major effort in this grant year has been the development of software (termed the OPAL system) that will permit physicians who are not computer programmers to enter protocol information into a structured set of forms on a graphical display. Most early expert systems required tedious (and occasionally erroneous) entry of the system's medical knowledge. Each segment of knowledge was transferred from physician to programmer and then entered into the program by the computer expert. Although many programs allowed for specification of a structure within which to organize the information, only minimal attempts were made to define a description that would be generic enough to provide a basis for a series of related knowledge bases in one medical area.

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: multiple protocols (that may be related to each other), experimental research arms in each protocol, drug combinations, individual drugs, and drug modifications. Using the graphically-oriented workstations, this information is presented to the user as computer-generated forms that appear on the screen. As the protocol is described, new forms are added to the computer display to allow for the specification of the special cases that make the protocols so complicated.

Although this design appears to be organized specifically for cancer treatment plans, we believe that the technique 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 could accept *any* possible problem specification.

Using this program we have entered several versions of a small cell lung cancer protocol, and a complicated lymphoma protocol with several different therapies. We are currently implementing the changes suggested by entering these protocols.

C.2.3 Strategic Therapy Planning (ONYX)

As mentioned above, we have begun a new research project to study the therapy planning process, and how strategies which are used to plan therapy in difficult cases might be represented on a computer. This project, which we call the ONYX project, has as its goals: to conduct basic research into the possible representations of the therapy planning process; to develop a computer program to represent this process; and eventually to interface the planning program with ONCOCIN. The project members (Fagan, Tu, Langlotz, and Williams) have spent many hours meeting with Dr. Sikic trying to understand how he plans therapy for patients whose special clinical situation precludes following the standard therapeutic plan described in the protocol document. In March of last year, the group spent two days at Xerox Palo Alto Research Center (PARC), working with Mark Stefik, Daniel Bobrow and Sanjay Mittal of PARC on possible representations for the knowledge structures and how such a program might run using the LOOPS knowledge programming system. A prototype version of this program is currently being tested. The prototype program has been designed as two components: the strategic planning program and the qualitative simulation builder. The strategic planning program is capable of turning the patient's medical data and knowledge of the intent of the protocol into a small number of plausible protocol modifications for the current point in time, and conditional modifications for the near future. Another component of the system is capable of building simulation models using the graphical abilities of the 1108 workstation. The first test of this component is the construction of a model of the effects of chemotherapy drugs on the bone marrow of the patient. During the next year of research this type of qualitative simulation model will be integrated into the strategic planning program.

C.2.4 Evaluations of ONCOCIN's performance

We have completed our first three formal studies of ONCOCIN's DEC-20 version (see papers by Kent et al. and Hickam et al. for results of two of these; written reports on the third is in preparation). Lessons learned in these initial studies have led to revisions both in the design of ONCOCIN and in our plans for evaluation studies of the 1108 version of the system when it is implemented at non-Stanford sites in later years.

C.2.5 Documentation

We have developed a videotape that discusses and demonstrates our research on the workstation version of our system. This tape has been shown at national meetings and has been extensively distributed to researchers internationally who have shown an interest in our work. The publication list that accompanies this report further documents the design decisions we have made in developing the new version of ONCOCIN.

C.2.6 Dissemination

In anticipation of completion of the workstation version of ONCOCIN, we are beginning to plan for an experiment in which we will install ONCOCIN workstations in private oncology offices in San Jose and Fresno. An application proposing this work is current under review.

D. Publications Since January 1984

1. (*) Buchanan, B.G. and Shortliffe, E.H.: *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, MA., 1984. [book]

2. (*) Clancey, W.J. and Shortliffe, E.H.: *Readings in Medical Artificial Intelligence: The First Decade.* Addison-Wesley, Reading, MA., 1984. [book]

3. Clancey, W.J. and Shortliffe, E.H.: *Strategies for medical knowledge engineering: Lessons from the first decade.* To appear in the Proceedings of the AAMSI Congress 85, San Francisco, CA., May 1985.

4. Differding, J.C.: *The OPAL interface: General Overview.* Working paper. August 1984.

5. (*) Fagan, L.: *New Directions for Expert Systems: Examples from the ONCOCIN Project.* To appear in the Proceedings of AAMSI Congress 85, San Francisco, CA., May 1985.

6. (*) Hickam, D.H., Shortliffe, E.H., Bischoff, M.B., Scott, A.C., Jacobs, C.D.: *A study of the treatment advice of a computer-based cancer chemotherapy protocol advisor.* Submitted for publication, May 1985.

7. (*) Kahn, M.G., Ferguson, J., Shortliffe, E.H., Fagan, L.: *An approach for structuring temporal information in the ONCOCIN system.* To appear in the

Proceedings of the Symposium on Computer Applications in Medical Care, Baltimore, MD., November 1985.

8. (*) Kent, D.L., Shortliffe, E.H., Carlson, R.W., Bischoff, M.B., Jacobs, C.D.: *Improvements in data collection through physician use of a computer-based chemotherapy treatment consultant.* Submitted for publication, March 1985.

9. (*) Lane, C.D., Differding, J.C., Shortliffe, E.H.: *Design of a graphic interface for a medical expert system.* (Memo KSL-85-15). Working paper.

10. (*) Langlotz, C., Fagan, L., Tu, S., Williams, J., Sikic, B.: *ONYX: An architecture for planning in uncertain environments.* To appear in the Proceedings of International Joint Conference on Artificial Intelligence, Los Angeles, CA., August 1985.

11. (*) Langlotz, C.P. and Shortliffe, E.H.: *Adapting a consultation system to critique user plans.* In Developments in Expert Systems, (M. Coombs, ed.), pp. 77-94, London: Academic Press, 1984.

12. (*) Musen, M., Langlotz, C., Fagan, L., Shortliffe, E.H.: *Rationale for knowledge base redesign in a medical advice system.* To appear in the Proceedings of AAMSI Congress 85, San Francisco, CA., May 1985.

13. Shortliffe, E.H.: *The science of biomedical computing.*Medical Informatics, Vol.9, Nos. 3/4, 185-193 (1984).

14. (*) Shortliffe, E.H.:*Reasoning methods in medical consultation systems: artificial intelligence approaches* (tutorial). Computer Programs in Biomedicine 18:5-14 (1984).

15. Shortliffe, E. H.: *Explanation capabilities for medical consultation systems* (tutorial). Proceedings of AAMSI Congress 84 (D. Lindberg and M. Collen, Eds.), pp. 193-197, San Francisco, May 1984.

16. Shortliffe, E.H. and Fagan, L.M.: *Artificial intelligence: the expert systems approach to medical consultation.* Proceedings of the 6th Annual International Symposium on Computers in Critical Care and Pulmonary Medicine, Heidelberg, Germany, June 1984.

17. (*) Shortliffe, E.H.: *Update on ONCOCIN: A chemotherapy advisor for clinical oncology.* Proceedings of the Symposium on Computer Applications in Medical Care, November 1984.

18. (*) Tsuji, S. and Shortliffe, E.H.: *Graphics for knowledge engineers: a window on knowledge base management* (Memo KSL-85-11). Submitted for publication, April 1985.

E. Funding Support

Grant Title: "Studies in the Dissemination of Consultation Systems"
Principal Investigator: Edward H. Shortliffe
Agency: Biomedical Research Technology Program, Division of Research Resources
ID Number: RR 01613
Term: July 1983 to June 1986
Total award: $624,455

Current award: (7/84-6/85): $222,511 (Direct costs)

Grant Title: "Therapy-planning strategies for consultation by computer"
Principal Investigator: Edward H. Shortliffe
Agency: National Library of Medicine
ID Number: LM-04136
Term: August 1983 to July 1986
Total award: $211,851
Current award: (8/84-7/85) $69,875 (Direct costs)

Grant Title: "Postdoctoral Training in Medical Information Science"
Principal Investigator: Edward H. Shortliffe
Agency: National Library of Medicine
ID Number: 1 T32 LM07033
Term: July 1, 1984 - June 30, 1989
Total award: $903,718
Current award: (7/84-6/85) $79,059 (Direct costs)

Grant Title: Explanation of Computer-Assisted Therapy Plans"
Principal Investigator: Lawrence M. Fagan
Agency: National Library of Medicine (New Investigator Grant)
ID Number: 1 R23 LM04316
Term: February 1985-January 1988
Total award: $107,441
Current award: (2/85-1/86) $37,500 (Direct Costs)

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 1986, renewable until June 1988
Total award: $150,000 ($50,000 annually).

Grant Title: Information structure and use in knowledge-based expert systems
Principal Investigator: Bruce G. Buchanan
Co-Principal Investigator: Edward H. Shortliffe
Agency: National Science Foundation - IST83-12148
Term: March 1, 1984 - February 28, 1987
Total award: $330,000 (includes indirects)

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 (both the prototype system running in the clinic and the newer work transferring the system to professional workstations). We also demonstrated our developing workstation code in the Xerox exhibit in the trade show associated with AAAI-84 in Austin, Texas. 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*, whereas ONCOCIN has been featured on the "NBC Nightly News", the PBS

"Health Notes" series, and "The MacNeil-Lehrer Report." Due to the frequent requests for ONCOCIN demonstrations, we have produced a videotape about the ONCOCIN research which includes demonstrations of our the 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.

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 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 start out 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 on 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 that 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 exchange 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 on opposite coasts of the country.

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 hard 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. *To transfer the oncology prototype from its current research computer to a professional workstation that provides a model for cost-effective dissemination of clinical consultation systems.* To meet this specific aim we will 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 coming year, we will:

    - Complete our OPAL protocol entry system

    - Continue entry of additional protocols, hopefully at the rate of one protocol/month (including testing)

    - Place a version of the OPAL protocol entry system into the clinic for use by physicians as a graphical reference guide to the protocols.

3. *To introduce ONCOCIN gradually for ongoing use so that by mid-1986 two professional workstations will be available in the oncology clinic to assist in the management of cancer patients.* During the next year, we will:

    - Implement the first workstation-based ONCOCIN system for use by physicians in the oncology clinic by the end of the calendar year 1985, adding a second workstation within a few months thereafter

    - Continue to operate the DEC-2020 version to maintain continuity of support in the clinic setting until the workstation version is fully operational.

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 is shifting 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, and the ONCOCIN prototype currently used in the clinic depends upon proximity to the DEC 2020 which enables us to use a 9600 baud interface.

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 Dandelion lisp machines that we have purchased or have been donated by XEROX corporation. We also have special needs for more computing power for our ONYX therapy planning research, and have been able to share an upgraded Dandelion loaned by SUMEX for this work.

C. Requirements for Additional Computing Resources

The acquisition of the DEC 2020 by SUMEX was crucial to the growth of our research work. It has insured 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 (Dandelions) 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 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 D-machines have provided a reliable, constant "load-average" machine for running experiments with physicians and doing development work. The development of ONCOCIN on the Dandelion 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 Dandelion 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.

6.1.4. 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 [34].

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 speed up the determination.

C. Highlights of Progress

We have constructed a prototype of such a program, called PROTEAN, designed on the blackboard model [16], [26]. It is implemented in BB1 [27], a framework system for building blackboard systems that control their own problem-solving behavior [28](see discussion of BB1 above). We have coupled the reasoning program with an IRIS graphics terminal (shared with SUMEX) which displays protein structures at different levels of detail. This provides a visual understanding of how the program is behaving, which is essential for this problem.

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 partitions the problem of positioning all of a protein's constituent structures within a global coordinate system into sub-problems of positioning 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. It also

introduces additional constraints in the global characteristics of internally constrained partial solutions. For example, the conformations of partial protein solutions constrain their composability with other partial solutions.

2. The problem-solver attempts to solve sub-problems and coordinate solutions at multiple levels of abstraction, where lower levels of abstraction partition solution elements with finer granularity. For example, PROTEAN operates at three levels of abstraction. At the "Solid" level, it positions elements of the protein's secondary structure: alpha-helices, beta-sheets, and random coils. At the "Blob" level, it positions elements of the protein's primary structure of amino acids: peptide units and side-chains. At the "Atom" level, it positions the protein's individual atoms. Partial solutions at higher levels of abstraction reduce the combinatorics of search at lower levels. Conversely, tightly constrained partial solutions at lower levels introduce new constraints on higher-level solutions.

3. The problem-solver forbears hypothesizing specific partial solutions for a sub-problem in favor of preserving 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 only as constrained as the data are constraining. It also accommodates incompatible constraints by permitting disjunctive sub-families. For PROTEAN, disjunctive sub-volumes imply that the associated structure lies within any one of the sub-volumes or, if the structure is mobile, that it may move from one sub-volume to another.

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, successively restricting the spatial volumes within which they may lie. Independent application of different constraints finesses the problem of integrating qualitatively different kinds of constraints by simply integrating their results. In addition, successive restriction of the family of solutions obviates guessing which specific solutions within a family are likely to be consistent with subsequently applied constraints and the otherwise inevitable back-tracking.

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. Toleration of overlapping partial solutions is another accommodation of incomplete or incompatible constraints and potentially dynamic solutions. For PROTEAN, overlapping volumes for two protein 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 (see section 3.2.1). It also

provides a foundation for the problem-solver's explanation of problem-solving activities and intermediate partial solutions (see section 3.2.2) and for its learning of new control heuristics (see section 5.5).

The current version of PROTEAN has six knowledge sources that demonstrate the reasoning techniques described above. These knowledge sources develop partial solutions that position multiple helices at the Solid level and refine those helices at the Blob level. Proposed work will introduce knowledge sources that operate on other protein structures at the Solid level, as well as knowledge sources that apply the reasoning techniques at the Blob and Atom levels. We also will investigate emergent constraints entailed in reliable partial solutions, composition of partial solutions into complete solutions, and intelligent control.

D. Relevant Publications

1. Erman, L.D., Hayes-Roth, B., Lesser, V.R., Reddy, D.R.:*The HEARSAY-II Speech Understanding System: Integrating Knowledge to Resolve Uncertainty.* ACM Computing Surveys 12(2):213-254, June, 1980.

2. Hayes-Roth, B.: *The Blackboard Architecture: A General Framework for Problem Solving?* Report HPP-83-30, Department of Computer Science, Stanford University, 1983.

3. Hayes-Roth, B.: *BB1: 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.

4. Hayes-Roth, B.:*A Blackboard Architecture for Control.* Artificial Intelligence In Press, 1985.

5. Hayes-Roth, B. and Hewett, M.: *Learning Control Heuristics in BB1.* Report HPP-85-2, Department of Computer Science, 1985.

6. 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.

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

Total Amount: $100,000

Dates: Nov 1, 1984/Oct 31 1986

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

Robert Langridge was visiting at Stanford last year, and informal discussions with him and his group have continued in this year.

C. Critique of Resource Management

The SUMEX staff has continued to be most cooperative in getting this project started. Without their persistence, we would not have been able to obtain Ethernet software for the IRIS graphics terminal from Xerox.

III. RESEARCH PLANS

A. Goals & Plans

Our long-range goal is to build an automatic interpretation system similar to CRYSALIS (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 six knowledge sources that demonstrate the reasoning techniques described above. These knowledge sources develop partial solutions that position multiple helices at the Solid level and refine those helices at the Blob level. The proposed research would expand PROTEAN to include knowledge sources that:

1. construct partial solutions combining helices, beta sheets, and random coils at the Solid level;

2. merge highly constrained partial solutions at the Solid level;

3. refine Solid level solutions in terms of the relative positions of constituent peptide units and side chains at the Blob level;

4. further restrict the relative locations of peptide units and side chains relative to one another at the Blob level;

5. propagate emergent constraints at the Blob level back up to the Solid level to further restrict the relative positions of superordinate helices, beta sheets, and random coils;

6. refine Blob level solutions at the Atom level;

7. further restrict the relative locations of atoms relative to one another;

8. propagate emergent constraints at the Atom level back up to the Blob level to further restrict the relative positions of superordinate peptide units and side chains.

The research will also develop a set of control knowledge sources to guide PROTEAN's application of constraints to identify the family of legal protein conformations as efficiently as possible. And we expect to improve the graphics interface to provide more functionality and options for viewing partial structures.

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 the Blackboard model and knowledge structures most flexibly and quickly.

C. Need for other computing resources

In this stage of development we need more computer cycles and hope to have access to additional D-machines. We expect to upgrade the Silicon Graphics IRIS terminal to a workstation for more efficiency in the subprograms doing computational geometry.

6.1.5. 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 long-range objectives of our project, called RADIX (formerly RX), are 1) to increase the validity of medical knowledge derived from large time-oriented databases containing routine, non-randomized clinical data, 2) to provide knowledgeable assistance to a research investigator in studying medical hypotheses on large databases, 3) to fully automate the process of hypothesis generation and exploratory confirmation. For system development we have used a subset of the ARAMIS database.

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