

- 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. Considerable effort has been expended on producing a standard relational database as the appropriate data structure to underlie the OPAL IDS. The PROTEGE system described in the core ONCOCIN section was built upon this relational database.

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. Based on this experience, we continue to explore ways to modify 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 Research in Progress*

The major thrusts in speech input and generalized knowledge acquisition are described in the core research description of ONCOCIN. We will describe here our research in complex therapy planning and its spin-offs in temporal representations and summarization of patient records.

### *C.2.1 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.

A major spin-off from our ONYX work is a program that can summarize temporal trends in patient visits during chemotherapy and produce a summary of the patient's course using both data from the flowsheet and an *underlying model of bone marrow physiology*. This work has led to major improvements in the temporal representation and in the integrate of mathematical and symbolic models. This work is part of Michael Kahn's Ph.D. thesis.

Summarization is defined as the task of combining multiple observations or features into a more general statement and abstraction as the task of selecting a subset of available features considered most relevant to answering a particular question. Both tasks require a model of the underlying system that encodes extensive knowledge about the entities and relationships that cause the system behavior and result in the observations. In the setting of a dynamic system, the model must be capable of representing temporal relationships between entities.

This work proposes that the combination of mathematical and symbolic techniques can be used to construct useful summaries of complex time-ordered data. In particular, mathematical models are used to capture the knowledge about the physiological processes that are responsible for the patient's clinical findings. Model parameters represent physiological concepts that are clinically relevant for medical problem solving. Prior to any patient-specific observations, the model parameters are set to population-based estimates. Standard curve-fitting techniques using a Bayesian updating scheme adjust model parameters to new observations. As more patient-specific observations are obtained, the set of estimated model parameters move further away from the population estimates. Symbolic models are used to augment the mathematical model parameter and state estimates. As the patient's clinical course evolves, the symbolic model captures the concurrent contexts that affect the interpretation of the physiological model results. For example, a heart rate of 120 is considered abnormally high in the context of a resting person but may be inappropriately low in the context of a treadmill stress test. A key feature of the combined mathematical and symbolic approach is that the physiological model changes over time as additional data are obtained and the symbolic model modifies the interpretation of these model changes in light of the clinical contexts present when the data was observed.

The methodology for combining mathematical and symbolic models emphasizes four main elements in summarizing complex time-ordered data:

1. A mechanistically-motivated model (in medicine, a physiological model) forms the basis for converting raw observations into more meaningful concepts. However, the interpretation of these concepts requires additional knowledge such as the contextual information contained in a symbolic model.

2. The initial model is based on general knowledge since no specific observations are available to alter the initial impression. New observations will change the initial model by incorporating the new information. The collection of altered models captures state changes that have evolved over time.
3. Differences in key model features or states form the basis for selective abstraction and effective summarization. A method for determining which features are pertinent to a user question or sufficiently "interesting" to warrant inclusion into a summarization requires additional domain-specific reasoning.
4. The construction of a concise and useful summarization requires the use of additional contextual and domain-specific information so that the generated summary text conforms to the user's expectations and requirements.

These principles form the basis for a computer program designed to summarize the clinical course of individual patients receiving experimental cancer chemotherapy. In this setting, patients are often receiving more than one treatment that have overlapping schedules and durations of action. Thus our temporal model requires the representation and the reasoning with multiple, simultaneous contexts to ensure the proper interpretation of a given observation or model estimate. ONCOCIN uses a specialized structure called the temporal network to represent treatment contexts used in temporal queries into a time-oriented patient record. We have extended the temporal network concept to create a symbolic model of the patient's clinical course over time. This structure permits the representation of multiple, concurrent contexts over time and therefore can capture the complex temporal nature of our patient's clinical course. For the proper interpretation of the mathematical model output, the temporal network provides the set of contexts that existed when the observation and model estimates were obtained. In addition, the interpretation task requires complex context-sensitive reasoning. For example, the interpretation of a model parameter may be different if two contexts were present concurrently than if either context was present alone. The temporal network provides a mechanism for altering the available reasoning methods based on the set of current contexts. In this use of the temporal network, reasoning methods are associated with each context. When a context is present, a temporal network node representing that context is created and the reasoning methods are made available to the interpretation process. A temporal network node may also withdraw methods made available by other temporal network nodes. In this manner, a general rule or method can be suspended if it is not appropriate in particular context.

We believe that the combination of mathematical models along with specialized symbolic structures results in more representational and inferencing power than either method alone. Well established mathematical techniques convert observations into underlying system concepts while symbolic techniques interpret the mathematical results using additional domain knowledge. Although some of these features could possibly be represented using either mathematical or symbolic techniques alone, we believe the resulting "pure" models would be too complex to be useful. In combining the two approaches, concepts best modeled in a physiological paradigm can be expressed within the mathematical model while concepts best modeled symbolically can be represented within the temporal network.

#### *C.2.2 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. We have sent copies of our system to the University of Pittsburgh, and will distribute it to the National Library of Medicine. We have developed a user manual, description of sample interaction, reference card, and graphical flowchart to help with training in the use of ONCOCIN.

*D. Publications Since January, 1987*

1. Walton, J.D., Musen, M.A., Combs, D., Lane, C.D., Shortliffe, E.H., and Fagan, L.M. Graphical access to medical expert systems: III. Design of a knowledge-acquisition environment. Memo KSL-85-30. Methods of Information in Medicine, 26:78-88, 1987.
2. 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.), pp. 15-35 (Chapter 2), Norwood, NJ: Ablex Publishing Company, 1988.
3. 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. Computers in Biomedical Research, 20:279-303, 1987.
4. 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). International Journal of Man-Machine Studies 26(1):105-121, 1987.
5. 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. Also to appear in Frontiers of Medical Information Sciences, Praeger Publishing, 1988.
6. Langlotz, C.P., Shortliffe, E.H., and Fagan, L.M. A methodology for generating computer-based explanations of decision-theoretic advice. Technical report, KSL-86-57, July 1987. To appear in Medical Decision Making.
7. Langlotz, C.P. and Shortliffe, E.H. An analysis of logic and decision-theoretic methods for planning under uncertainty. Report KSL-87-17, 1987.
8. Shortliffe, E.H. Computer programs to support clinical decision making. Memo KSL-87-30. Journal of the American Medical Association, 258:61-67, 1987.
9. Rennels, G.D. and Shortliffe, E.H. Advanced computing for medicine (Memo KSL-87-33). Scientific American, pp. 154-161, October 1987.
10. Langlotz, C.P. Advice generation in an axiomatically-based expert system. Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care, pp. 49-55, Washington, D.C. 1987.

11. Shortliffe, E.H. and Hubbard, S.M. Information systems in oncology. Memo KSL-87-66, November 1987. To appear as a chapter in Cancer: Principles and Practice of Oncology (V.T. DeVita, S. Hellman, and S.A. Rosenberg, eds.), 1988.
12. Musen, M.A. Generation of Model-Based Knowledge-Acquisition Tools for Clinical-Trial Advice Systems. KSL-88-06. Doctoral dissertation, Medical Information Sciences Program, Medical Computer Science Group, Stanford University, January 1988.
13. Wulfman, C.E., Isaacs, E.A., Webber, B.L., and Fagan, L.M. Integration discontinuity: Interfacing users and systems. Memo KSL-88-12, February 1988. Proceedings of Architectures for Intelligent Interfaces: Elements and Prototypes, pp. 57-68, Monterey CA, March 29-April 1, 1988.
14. Musen, M.A. Conceptual models of interactive knowledge-acquisition tools. Report KSL-88-16, March 1988.
15. Musen, M.A. Generation of knowledge-acquisition tools from clinical-trial models. Report KSL-88-26, March 1988. To be published in The Proceedings of Medical Informatics, Europe 1988, Oslo, Norway, August 1988.

#### *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 AAI-84 in Austin, Texas, IJCAI-85 in Los Angeles, AAI-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 MacNeill-Lehrer Report." Most recently it appeared in a special on Artificial Intelligence for TV Ontario (Canadian PBS station) and "Physician's Journal Update" on the Lifetime Cable Network. Due to the frequent requests for ONCOCIN demonstrations, we have produced a videotape about the ONCOCIN research which includes demonstrations ONCOCIN and OPAL. The tape has also been shown to both national and international researchers in biomedical computing. We are producing a tape that describes our speech research.

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*. The EMYCIN experience is now well disseminated via commercial products. We may be able to demonstrate MYCIN from the 2020 after the 2060 is turned off, but this is not a major concern in the transition effort.

### 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 three 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. Several recent papers have been written to describe collaborations between students in our training program and the group at the University of Pittsburgh and Carnegie-Mellon University.

### *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. As described above, we will eventually be moving our software development to a combination of Lisp machines and Mac II's. These will need to be easily networked together while still providing the communications resources of the DEC-20.

## **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 generalize the reasoning and interaction components of the ONCOCIN system for other applications.*
2. *Extensions and generalizations of the OPAL and PROTEGE knowledge acquisition experiments*
3. *Extensions to the strategic planning framework including research on temporal representations, mathematical modeling and summarization*
4. *To continue testing of the workstation version of ONCOCIN.*

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

All of our research takes place in the context of the SUMEX resource. The 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, with a gradual transition to Mac II's. The word processing and communication requirements will be met by the Mac II's plus the excellent network services provided by SUMEX.

*C. Requirements for Additional Computing Resources*

Most of our CPU needs are met with our current equipment. However, our requirements for high bandwidth communication facilities are increased as we have an increasingly distributed environment. We would appreciate support that provides equivalent resources to what we are used to on the DEC-20 including mail support, and file servers. We will continue to need access to the international networks that we now use for much of our communication with colleagues. For example, Mark Musen (who recently finished his Ph.D., and will be an Assistant Professor in our department when he returns) is currently in Holland for 9 months. During this period, we have been in constant electronic communication with him over a set of networks. This has been a significant asset to our research project.

*D. Recommendations for Future Community and Resource Development*

The continuation of network support and file service are our major needs. Help with software selection and implementation for our Mac II's is also an important requirement to maintain research continuity as the DEC-20 is disconnected. Maintaining the excellent mail service and access to the international networks is also essential for our research.

## IV.A.5. 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, and (b) use protein structure determination as a test problem for experiments with large scale constraint satisfaction, an area of increasing interest to artificial intelligence. 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 computational complexity, but by formulating it as search problem, it should be possible to utilize many of the knowledge-based techniques developed in AI for dealing with large search spaces.

#### 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*

Atomic level determination of the structure of cytochrome b562. During the past year we greatly expanded the functionality of PROTEAN towards obtaining meaningful biochemical results. In the first half of the year the primary outcome of this effort was the atomic-level assembly of the protein cytochrome b562 using simulated NMR constraints from the known crystal structure. The strategy for assembling cytochrome involved functionality at the atomic as well as the more abstract solid level, and demonstrated one possible application of our general approach, which we call *heuristic refinement*.

The heuristic refinement method falls within an *exclusion paradigm* for interpreting protein structure. In this paradigm all the atoms in the protein are initially assumed to be everywhere with respect to some arbitrary coordinate system. As constraints are introduced these initially infinite *accessible volumes* are gradually reduced until the remaining accessible volumes are small enough so that a representative set of structures may be enumerated.

The exclusion paradigm has not been tried previously because it appears to be computationally intractable for any reasonably sized protein. For this reason most current techniques (non AI-based) use an *adjustment paradigm* to optimize a single structure towards a global minimum. These techniques all share difficulties common to optimization, namely lack of convergence for large numbers of variables or for starting structures far from the global minimum. In addition, for reasons of computation time, they are not able to provide a representative sample of all structures compatible with the constraints.

The exclusion paradigm, by systematically eliminating structures incompatible with the constraints, has the advantage that all, or at least a representative set, of structures compatible with the constraints are retained. In order to deal with the combinatorics the heuristic refinement method formulates the problem as one of search, which is the fundamental paradigm of artificial intelligence. In particular, the problem is formulated as a geometric constraint satisfaction problem (GCSP), which can be solved by backtrack search if the size of the atomic accessible volumes can be made small enough.

The complexity of solving a GCSP depends on the number of objects and the number of possible locations that each object can have in space. Therefore, the heuristic refinement method utilizes four techniques to reduce these two numbers to the point that backtrack search may be applied. The effect of each of these techniques is to utilize knowledge of protein structure and of techniques for solving GCSPs to prune the search tree:

1. Problem decomposition. The protein is broken into logical subparts such as secondary structures and sidechains. Each of these is treated as a separate GCSP, which is partially solved. Solutions to the subproblems either determine constraints at more abstract levels or are combined into larger subproblems.
2. Problem abstraction. Groups of locally highly constrained atoms are represented as single abstract "solid" level objects. Solutions to atomic level GCSPs are used to create abstract "solid" level constraints, resulting in a solid level GCSP with far fewer objects. Solutions at the solid level are then expanded to the atomic level, resulting in fewer possible locations for the atoms than would have been present if the initial solid level processing had not occurred. Almost all the work reported in previous reports was at this abstract solid level, but it is the current refinement to the atomic level that is of most interest to biochemists.
3. Local satisfaction of constraints. In each local GCSP filtering operations called network consistency algorithms are applied, allowing the accessible volumes to be reduced by looking at constraints pairwise (or to higher order) rather than all at once. These operations, which were initially developed for computer vision, are becoming more popular in AI because of their utility in many forms of constraint satisfaction problems.
4. Heuristic control. At each point in the problem solving there are many possible constraint satisfaction operations that may be applied. If there is no bias then the order of operations should not matter, but the efficiency may vary greatly. Much of the previously reported work in PROTEAN went towards developing the BB1 framework for using heuristics to opportunistically determine the best operation to apply at any given time. The BB1 framework has been used to control the solid abstract level of problem solving, but it has not yet been integrated with the atomic level functionality, which is currently manually controlled.

These principals were used to determine the structure of cytochrome b562. The program was given as input the primary and secondary structure as obtained from the known crystal structure. It was also given a set of 729 distance ranges designed to simulate the expected type of data from NMR. Using these data, as well as knowledge of protein structure, the heuristic refinement method produced a set of 10 atomic level structures, with an average root mean square deviation from the crystal structure of 4.1 angstroms when all backbone alpha-carbons were included, and 2.8 angstroms when random-coil segments were excluded. These deviations were on the order of those found by other methods. Because we systematically excluded structures we could be confident that we had obtained an upper bound on the set of structures compatible with the constraints. These structures could then be adjusted by any of the current techniques (although we didn't do that in this case).

The cytochrome results were obtained over a period of about three to four months, using the previously-described geometry system and BB1 to obtain the solid level results, and LISP and C code to produce the atomic level results. Intermediate results were written to files, and individual programs were written by three different people.

Removing bias. During the latter part of this year we have concentrated on improving the functionality used to obtain the cytochrome results.

In the heuristic refinement method an important tradeoff is efficiency versus the amount of bias introduced by the abstractions. A major source of bias is the assumption that secondary structures are in their ideal configuration. For example, in the cytochrome and all previously reported results alpha helices were modeled as cylinders. This approximation meant that some legal atomic level solutions are prematurely eliminated at the solid level. In order to reduce this bias we have looked at actual examples of helices in the crystal structure database, and have used these to form more realistic constraints at the solid level. This relaxation of the ideal helix assumption has allowed us to obtain more accurate solid level accessible volumes, but at the cost of less constraining power. In later versions of the program we plan to allow the user to specify how much ideality to assume.

Atomic level constraint satisfaction operations. As part of his Ph.D. work Bruce Duncan has extended many of the solid level constraint satisfaction operations to the atomic level. He has developed programs, written in C, which allow atomic locations to be described at different levels of resolution, from a coarse grid initially to a finer grid as refinement proceeds. This approach has been used to reconstruct the crystal structure of a small protein called crambin given exact distances less than five angstroms. In this case no solid level preprocessing was done, and the entire process took about six weeks on three Suns running in parallel. Bruce is currently working to develop more intelligent control before trying his system on a more realistic simulated data set.

Probabilistic operations at the atomic level. Russ Altman, as part of his Ph.D. thesis, has been working to develop an alternative method for representing the accessible volumes of atoms. Currently, we represent accessible volumes as lists of xyz locations sampled on a regular grid at a specified resolution. In Russ's approach accessible volumes are represented as probability density functions described by a mean and covariance matrix. Constraints are also represented by a mean and variance, and a Kalman filter, which is basically another form of Bayes Formula, is used to determine the reduction in accessible volume. This approach has been used to reconstruct the amino acid tyrosine, and is being extended to handle larger numbers of atoms.

Development of an integrated software environment and general strategies. All our

current results were obtained with more or less independently developed software modules. No one person is able to run PROTEAN by himself. For this reason we have designed a framework that allows these modules to communicate via common files organized as a distributed object-oriented knowledge base. Parts of this framework are currently being implemented by a part time programmer. The framework will allow the existing solid level functionality to be integrated with the emerging atomic level, and will allow different assumptions, representations and algorithms to be experimented with.

Porting of heuristic control to TI Explorer in Common Lisp. The reasoning component in BB1 is currently written in InterLisp and runs on Xerox D-machines. Since InterLisp and D-machines are becoming obsolete the code has now been partially ported to the TI Explorer in Common Lisp. Network software has been written that allows the Explorer to control the geometry and graphics programs over the network. Once the port has been completed additional heuristic control experiments will be performed.

Secondary structure prediction. An important pre-processing step for PROTEAN is the determination of secondary structure from NMR. John Brugge, for his masters thesis, designed a program called ABC which emulates expert knowledge in assigning secondary structure from NMR. The program is written in BB1 and uses symbolic patterns in the NMR to predict structure. Validation studies showed that the program did about as well as experts.

Peak Assignment. When an adequate amount of purified protein is available, the routine determination of protein structure in solution by nuclear magnetic resonance would require efficient methods for: 1) collecting data, 2) extracting structural constraints from the data, and 3) generating protein structures which satisfy these constraints. Many two-dimensional nuclear magnetic resonance (2D-NMR) experiments now exist to carry out step 1), and a number of computer methods have been presented to deal with step 3). However, step 2) is a severe bottleneck.

The main difficulty in that step lies in the assignment problem in which each peak from 2D-NMR spectra must be matched to a pair of specific atoms from specific residues. Typically, hundreds of resonances must be identified, a tedious, slow, and error prone task. It involves keeping track of a large number of peaks, each of which initially has a large number of possible assignments. This complexity, however, contrasts with the relatively small number of principles and pattern matching techniques that are actually needed to assign the peaks in a 2D-NMR protein spectra.

Craig Cornelius, Guido Haymann-Haber, and Olivier Lichtarge have developed a prototype program called PEAKS that uses constraint satisfaction methods to apply patterns expected from amino acid residues to actual spectra. The program has been tested on simulated data sets for small proteins of 20 to 30 residues. The side chains of most residues can be correctly identified and sometimes assigned. The program can also deal with slightly incomplete spectra with degenerate or missing peaks.

#### 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.* Proceedings of the AAI, pp 399-404, 1987.

3. Altman, R., Duncan, B., Brinkley, J., Buchanan, B., Jardetzky, O.: *Determination of the spatial distribution of protein structure using solution data*, Proceedings of the Alfred Benzon Symposium 26: NMR Spectroscopy and Drug Development, Copenhagen, 1987.
4. 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.
5. 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.
6. Brinkley, J.F., Altman, R.B., Duncan, B.S., Buchanan, B.G., and Jardetzky, O.: *The heuristic refinement method for the derivation of protein solution structures: Validation on cytochrome-b562*, KSL Technical Report 88-03, 1988.
7. Brugge, J.A., Buchanan, B.G., and Jardetzky, O.: *Toward automating the process of determining polypeptide secondary structure from <sup>1</sup>H NMR data*, To be published in J. Computational Chemistry, 1988.
8. 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.
9. Duncan, B., Buchanan, B., Hayes-Roth, B., Lichtarge, O., Altman, R., Brinkley, J., Hewett, M., Cornelius, C., and Jardetzky, O.: *PROTEAN: A new method for deriving solution structures of proteins*, Bull. Mag. Res., 8:111-119, 1987.
10. Garvey, Alan, Cornelius, Craig, and Hayes-Roth, Barbara: *Computational Costs versus Benefits of Control Reasoning*. Report KSL 87-11, Department of Computer Science.
11. Hayes-Roth, B.: *The Blackboard Architecture: A General Framework for Problem Solving?* Report HPP-83-30, Department of Computer Science, Stanford University, 1983.
12. 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.
13. Hayes-Roth, B.: *A Blackboard Architecture for Control*. Artificial Intelligence 26:251-321, 1985.
14. Hayes-Roth, B. and Hewett, M.: *Learning Control Heuristics in BB1*. Report HPP-85-2, Department of Computer Science, 1985.
15. 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.

16. 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.
17. Lichtarge, Olivier: *Structure determination of proteins in solution by NMR*. Ph.D. Thesis, Stanford University, November, 1986.
18. 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*, *Proteins: Structure, Function and Genetics*, 2:340-358, 1987.

#### *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: 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  
(direct and indirect)

Current award period and amount: 10/1/87 - 9/30/88 \$1,467,300  
(direct and indirect)

PROTEAN component is \$72,956, or 5.0 % 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 biochemical goal is to build a system that can automatically interpret NMR and other types of constraints, in order to produce the family of structures compatible with the constraints. In so doing we have encountered and will continue to encounter important and interesting artificial intelligence problems in large scale, knowledge-based constraint satisfaction.

The major steps towards these goals are as follows, more or less in order of expected order, although the validations will proceed in parallel with the addition of new features. Several of these steps have been worked on within the past year (see Highlights of Research Progress).

*A.1. Develop an integrated system.* In order for the heuristic refinement method to be widely used it must be integrated into a computer system that contains the required functionality, is reasonably efficient, is semi-automatically controlled, relatively easy to use, and extensible. Using our previous work as a basis we will continue to work towards this goal. The required substeps in order to achieve the goal of an integrated system are:

- Develop improved functionality. This is our current major focus and includes efforts to relax the ideal helix assumption, to utilize proven techniques from other structure determination methods (such as the use of distance bound smoothing algorithms as a pre-processing step, and the use of adjustment procedures as a post-processing step), to develop better methods of representing the spatial locations of both three-dimensional (non-oriented) objects, and six-dimensional (oriented) objects, and to develop more efficient constraint satisfaction algorithms. Progress has been made in several of these areas over the past year, especially by the Ph.D. students as they develop their dissertations.
- Develop system architecture. During this year we have also designed a basic architecture for a distributed system that can be extended as new modules and representations are developed. Much of the programmers effort this year will go towards implementing this system and towards incorporating some of the advances made by the Ph.D. students.
- Develop control strategies and knowledge base. As the integrated functionality is developed it will be made available either as remote servers accessible over the network (as with our current geometry system), or as loadable modules. BB1, or another suitable object-oriented expert system shell, will be used to create the control modules. These modules will use heuristics to intelligently pick the most efficient constraint satisfaction operation(s) to perform next. Progress towards one such control module has been made with the porting of the BB1 system to the TI Explorer.
- Develop user interface and improve graphics. To be useful the system should be relatively easy to run by a non-programmer biochemist. We have not yet begun to explore this issue, but other AI in Medicine projects at Stanford and elsewhere have worked in this area. Examples at Stanford are Oncocin and its derivatives, and Apple's HyperCard as an interface to an object-oriented belief network. Once the programmer has created a rudimentary and extensible system he will begin to look at this issue.

*A.2. Extend the system.* As the required functionality, framework and control

strategies become integrated a version of the system will be made available (at least at Stanford) for analyzing real data. At the same time we will use the integrated system as the basis for extending its capabilities. Developing these new capabilities will require solution to some important AI problems dealing with object representation and reasoning over time as well as space. These extensions include:

- Studying larger proteins and protein complexes. Since our system is hierarchical there is in theory no reason why we can't move all the way up the anatomic spectrum from atoms to organs.
- Add new types of constraints, including global constraints such as volume and surface, and probabilistic criteria. The current work of Russ Altman should be very useful in this regard.
- Automate the earlier interpretation of NMR data. We have made good beginnings in this area with the ABC program for secondary structure prediction and the peak detection programs. We will also continue our development of the PEAKS system for interpretation of 2-D NMR spectra, to be integrated with ABC and the rest of the PROTEAN system.
- Analyze structure. The purpose of determining structure is to understand function, and a component of this is analyzing the structures produced. We have already developed several modules for analyzing structure, such as programs for determining root mean square deviations and for calculating distance distributions. Russ Altman has also done some work in automatically extracting features of interest from a protein, and will continue this as part of his dissertation.
- Study protein dynamics. In some cases proteins in solution may be moving, a fact which greatly complicates the analysis of NMR data. So far we have ignored this issue because it is hard enough to solve the problem with stationary structures, but as we gain experience we will begin to consider these issues. Many problems dealing with incompatible constraints will arise as we look at this problem.

*A.3. Validate on known crystal structures.* As new functionality is developed, it will be verified by attempting to reconstruct proteins whose structures are known from crystallography. In this way we will at least know that our methods work, even if there is no gold standard for solution data.

*A.4. Use on actual NMR and theoretical constraints.* As the system is developed and verified on crystal data it will be made available to biochemists. Only those capabilities that have been verified will be supported. Funds for actually using the system on real data will be expected to be provided in the context of specific research projects.

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

We will continue to use the SUMEX facilities for integrating and expanding the PROTEAN system. The 2060 (or whatever replaces it) will primarily be a focus for integration and communication among machines and humans. Most of the actual program development will be on workstations running C or Common Lisp, communicating via the excellent local area network maintained by SUMEX. The communication and clearinghouse capability provided by SUMEX and the SUMEX staff is essential for continued success of such a large and evolving distributed project.

*C. Need for other computing resources*

At this time our computational resources are almost adequate as long as we can continue to use the SUMEX-supported workstations and local area network. We could always use faster machines since many of our computations are very numerically intensive, but our major need is for personnel and systems support. Since PROTEAN is evolving into a distributed problem-solving system we actively support and require extensive networking and remote procedure access capabilities. We believe that the requirements of PROTEAN make it an ideal test problem for many of the distributed systems issues being considered by the SUMEX core research staff.

## IV.A.6. RADIX and PENGUIN Projects

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

**The PENGUIN Project: Applying Database and Knowledge base  
Technology to Medical Instrumentation**

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

**Thierry Barsalou, M.D.  
Medical Computer Science  
Stanford University**

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

### I. SUMMARY OF RESEARCH PROGRAM

The RADIX research has been phased out during this year. Although proposals to NLM and NCHSR for continued research support were approved, they have not been funded. Some closeout funding enabled the project to be brought to an orderly conclusion, and publications continue to appear. The experience gained here and under the predecessor project, RX, continues to influence basic research directions. The problem of automated knowledge acquisition remains an important area of research, and we expect that the foundations laid here will influence work of others as well.

Some of the basic research has influenced the KBMS/KSYS project. Specifically problems encountered in the management and structuring of large quantities of knowledge are now being addressed, primarily under DARPA sponsorship. A subproject, on knowledge validation is being supported by IBM and will draw directly on the data and knowledge bases established for RADIX in its initial phases.

The KBMS/KSYS work, in turn, is leading to another important medical application: PENGUIN. This project is currently mainly supported by KBMS, but a substantial research grant has been approved by NLM, and given a high priority, so that we are confident that we will be able to soon move forward intensely in this project.

We will, below, provide two distinct sections. The first one will provide the final report of achievements under RADIX and the second will outline the current state and plans for PENGUIN.

### RADIX PROJECT GOALS AND PROGRESS

### A. RADIX -- Technical Goals

The objectives of the RADIX project were 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. Research was performed on the first two objectives.

*RADIX Discovery Module:* Computerized clinical databases and automated medical records systems are starting to contain valuable long-term records of medical practice and experience. We have utilized the data collected by 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.

A 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 first phase of the RADIX computer program helps to examine the time-oriented ARAMIS database displays information which helps recognize possibly causal relationships. The important concept is here the mapping of base information to higher level abstractions which are meaningful to physicians. Instances of data, when mapped, can display a patient's course in a compact and meaningful fashion.

The intent was also that knowledge acquisition from physician-researchers can then be brought to a level where the relationships are meaningful. Instances of such relationships were acquired in the RX projects, but that work was not yet based on the more general concepts envisaged for RADIX.

*RADIX Summarization Module:* The management of inpatients and outpatients is often complicated by the size and disorganization of patient charts. Computerized patient records are becoming increasingly available. While computerization of records renders data, it worsens the problem of information overload. The ability to automatically create patient summaries represents 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 sub-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. We have reported on the results of this work.

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

As a test bed for system development, our focus of attention were 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.

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. RADIX -- Highlights of Research Progress*

#### *C.1 April 1987 to September 1988*

Our primary accomplishments in this period have been the following:

1. Implementation and demonstration of a second generation of the automated summarization program.
2. Application of algorithms for transforming the Internist knowledge base into standard Bayes form.
3. Publication of papers on automated summarization, and presentation of results at medical conferences.
4. 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 completed a second generation of our prototype automated summarization program. This work is described in DeZegher-Geets, 1987, noted in the publications section. It 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 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 used within the system to represent uncertainty, called evoking strengths and frequencies, are poorly defined. This makes it difficult to tune the method used by the program to assign likelihoods to diseases (the scoring scheme) and makes it difficult to transport knowledge contained in the program to other medical diagnostic systems.

In collaboration with R. Miller and D. Heckerman we have transformed the values in the INTERNIST KB into a probabilistic form. Various combinations of multiple regressions were performed on the evoking strengths, frequencies, probabilities of disease,  $p(D)$ , and probabilities of manifestation,  $p(M)$ , versus the likelihood ratios  $L(D|M)$  and  $L(D|\text{not } M)$ . This process yielded some interesting and unexpected results. For example, the multiple regression of evoking strength AND  $p(M)$  vs.  $L(D|M)$  showed an  $r$ -squared of 0.84, significantly better than the  $r$ -squared value for evoking strength vs.  $L(D|M)$  alone. Also, the transformation from frequency,  $p(M)$ , and  $p(D)$  into  $L(D|\text{not } M)$  revealed a correlation coefficient of 0.58. These results suggest a low cost method for converting the knowledge in INTERNIST-1 to a probabilistic form. In particular, assessments of  $p(D)$  and  $p(M)$  (only about 4500 numbers) can be used in conjunction with evoking strengths and frequencies in the KB (about 40,000 numbers) to construct likelihood ratios.

#### *C.1.3 Publication of papers on automated summarization, and presentation of results at medical conferences*

We have submitted and had accepted several papers, as noted in the section on publications. Some of these were presented at medical conferences.

#### *C.1.4 Training Post-Doctoral researchers, participants in RADIX, in methods of medical artificial intelligence research*

We have been training three post-doctoral researchers on the project during the earlier part of the current reporting year; Andrew G. Freeman, M.D., Isabelle de Zegher-Geets, M.D., and Donald Rucker, M.D.. Andrew Freeman has developed the Internist transformation algorithms. Isabelle de Zegher-Geets completed a thesis on Automated Summarization as part of Stanford's Medical Information Sciences program.

### **PENGUIN PROJECT GOALS AND PROGRESS**

#### *A. PENGUIN Project -- Rationale*

Databases and expert systems share a common goal -- generating useful information for action -- but accomplish their tasks separately, using different principles. It is clear, however, that future information systems will require both the problem-solving capabilities of expert systems (ESs) and the data-handling capabilities of database

management systems (DBMSs). Indeed, combining database and expert system technologies into expert database systems (EDSs) is an emerging research area. One can define an EDS as "a system for developing applications requiring knowledge-directed processing of shared information". From a perspective of developing advanced biomedical information systems, this definition conveys two precise scenarios: (1) enhancing DBMSs with structuring and manipulation tools that take more semantics into account; (2) allowing ESs to access and to handle efficiently information stored in database(s).

The object-oriented paradigm has gained much attention in recent years. In the database field, object-oriented DBMSs have emerged. The concept of an entity is also widely used by database design tools. In the field of artificial intelligence, frames are a well-known knowledge representation scheme. Although frames were conceived separately from the object paradigm, the two are in fact consistent with each other. In this research project, called PENGUIN, we investigate the hypothesis that the object-oriented approach can also serve as a unifying scheme for developing EDSs.

We have the opportunity to explore this hypothesis in a practical biomedical environment. Fluorescence Activated Cell Sorting (FACS) is emerging as a major source of information for biomedical research and clinical practice. Currently, achieving good FACS performance requires analyzing and integrating various complex data and knowledge sources. PENGUIN is thus aimed at developing methods for bringing together expert system and database technologies in an integrated advice system that could fulfill the information needs of FACS investigators. Central to this work is a very close collaboration between researchers in the Medical Information Sciences Program and the Departments of Computer Science and Genetics.

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

FACS has already demonstrated significant promise in clinical as well as research-oriented areas. As indicated earlier, basic studies identifying lymphocyte subpopulations have now been translated into clinical applications including monitoring of AIDS patients. Similar FACS applications can now be found in such diverse specialties as Hematology, Oncology, Infectious Diseases and even Gynecology-Obstetrics where current studies are evaluating FACS analysis of maternal blood sample as a potential non-invasive method for prenatal diagnoses.

The major block to proliferation of these kinds of valuable studies involves the requirement for greater skills in reagent selection and FACS machine operation than are typically available in basic and clinical research settings.

Thus developing sophisticated computer tools that provide a new level of automatic analysis and control for FACS and greatly facilitates FACS use by reducing the need for on-site human expertise should significantly improve the potential for using this versatile methodology in biomedical research and clinical practice.

#### *C. PENGUIN -- Highlights of Research Progress*

Our current effort focuses on developing a computer-based advisory system to assist the FACS investigator in designing experiment protocols. As mentioned above, this system will combine database and artificial intelligence methodologies in an integrated framework, as to provide 1) several levels of abstraction for information management and retrieval from a relational database system, 2) access to and integration of the results of past similar experiments as additional units of information through an interface with existing databases of FACS data and 3) inference

capabilities coupled to the database for high-level interactions with scientists during the design process. Although this work is motivated by a specific application, the formal design of the system will be domain-independent; it is then our belief that ideas, principles and programs developed in this process will be applicable to other medical and non-medical areas.

In the past year, we have developed the concept of an object-based interface on top of a relational database system and implemented an initial prototype of the interface. We have drawn an analogy between the notions of object and database view. Using this analogy, we have defined three components in the object interface:

1. The object *generator* maps relations into object templates where each template can be a complex combination of join (combining two relations through shared attributes) and projection (restricting the set of attributes of a relation) operations on the base relations. In addition, an object network groups together related templates, thereby identifying different object views of the same database. The whole process is knowledge-driven, using the semantics of the database structure. We define the object schema as the set of object networks constructed over a given database. Like the data schema for a relational database, the object schema represents the domain-specific information needed to gain access to PENGUIN's objects; this information enables us to combine well-organized, regular tabular structures -- the relations -- into complex, heterogeneous entities -- the objects.
2. The object *instantiator* provides nonprocedural access to the actual object instances. First, a declarative query (e.g., Select instances of template x where attribute y < 0.5) specifies the template of interest. Combining the database-access function (stored in the template), and the specific selection criteria, PENGUIN automatically generates the relational query and transmits it to the DBMS, which in turn transmits back the set of matching relational tuples. In addition to performing the database-access function, the object template specifies the structure and linkage of the data elements within the object. This information is necessary for the tuples to be correctly assembled into the desired instances. Those instances are then made available to the expert system directly, or to the user through a graphic interface.
3. The object *decomposer* implements the inverse function; that is, it maps the object instances back to the base relations. This component is invoked when changes to some object instances (e.g., deletion of an instance, update of some attributes) need to be made persistent at the database level. An object instance is generated by collapsing (potentially) many tuples from several relations. By the same token, one update operation on an object may result in a number of update operations that need to be performed on the base relations.

Preliminary results of PENGUIN's implementation indicate that such an object-based architecture provides (1) efficient access to and manipulation of complex units of biomedical information while preserving the advantages associated with persistent storage of data in relational format, and (2) a domain-independent, bidirectional communication channel between relational database systems and expert systems.

In parallel to developing the object interface, we are currently engaged in a knowledge acquisition process to define the structure of the database and of the knowledge base. Through interviews with our experts in the Genetics department, we are eliciting the structure of the various components (genetic, histological, and

serological information) of the FACS reagents database. Finally, we are exploring the use of hypertext (more specifically, the HyperCard program for the Macintosh) as a possible framework for developing the user interface component of PENGUIN.

*D. Publications of the RADIX and PENGUIN projects*

1. Barsalou, Thierry and Gio Wiederhold: "Automating a Cell Counter"; to be published in *The International Journal of Artificial Intelligence in Engineering*, Computational Mechanics Publ., UK, 1988.
2. Barsalou, Thierry and Gio Wiederhold: "Applying a Semantic Model to an Immunology Database"; in W.W. Stead (editor), *Proceedings of the Eleventh Symposium on Computer Applications in Medical Care*, pages 871-877, IEEE Computer Society Press, Washington, D.C., November 1987.
3. Barsalou, Thierry, W.A. Moore, L.A. Herzenberg, and G. Wiederhold: "A Database System to Facilitate the Design of FACS Experiment Protocols" (abstract); *Cytometry*, Vol.97, August 1987.
4. Barsalou, T. and G. Wiederhold. *Knowledge-based mapping of relations into objects*. To appear in the *International Journal of AI in Engineering*, 1988.
5. Barsalou, T. *An object-based architecture for biomedical expert database systems*. Submitted to the IEEE Twelfth Symposium on Computer Applications in Medical Care, 1988.
6. Blum, Robert L, *Computers and Artificial Intelligence in Clinical Medicine: Current Systems and Future Prospects*, Computer News for Physicians, October, 1987.
7. Blum, R. L., and Walker, M.G.: *LISP as an Environment for Software Design: Powerful and Perspicuous*. In *Proceedings of the Tenth Annual Symposium on Computer Applications in Medical Care*, pages 326-331. IEEE Computer Society, October, 1986.
8. Blum, R. L., and Walker, M.G.: *Automated Medical Discovery from Clinical Databases: An Overview of the RADIX Project*. In *Proceedings of the Fifth Toyobo Biotechnology Foundation Symposium: Artificial Intelligence in Medicine*. The Toyobo Foundation, Tokyo, Japan, August, 1986.
9. Blum, R. L., and Walker, M. G.: *Automated Medical Discovery from Clinical Databases: an Overview of the RADIX Project*. In *Proceedings of the First International Conference on Artificial Intelligence and Its Impacts in Biology and Medicine*, pages 59-83. Groupement Scientifique pour le Developpement de l'Intelligence Artificielle en Languedoc-Roussillon, Montpellier, France, September, 1986.
10. Blum, Robert L. and Gio C.M. Wiederhold: *Studying Hypotheses on a Time-Oriented Clinical Database: An Overview of the RX Project*. In J.A. Reggia and S. Thurim: 'Computer Assisted Medical Decision-Making'; Springer Verlag, 1985, pp.245-253.
11. Blum, R.L.: *Two Stage Regression: Application to a Time-Oriented Clinical Database*. Knowledge Systems Laboratory Technical Report. 1985.

12. Blum, R.L.: *Modeling and encoding clinical causal relationships*. Proceedings of SCAMC, Baltimore, MD, October, 1983.
13. Blum, R.L.: *Representation of empirically derived causal relationships*. IJCAI, Karlsruhe, West Germany, August, 1983.
14. Blum, R.L.: *Machine representation of clinical causal relationships*. MEDINFO 83, Amsterdam, August, 1983.
15. Blum, R.L.: *Clinical decision making aboard the Starship Enterprise*. Chairman's paper, Session on Artificial Intelligence and Clinical Decision Making, AAMSI, San Francisco, May, 1983.
16. Blum, R.L. and Wiederhold, G.: *Studying hypotheses on a time-oriented database: An overview of the RX project*. Proc. Sixth SCAMC, IEEE, Washington D.C., October, 1982.
17. Blum, R.L.: *Induction of causal relationships from a time-oriented clinical database: An overview of the RX project*. Proc. AAAI, Pittsburgh, August, 1982.
18. Blum, R.L.: *Automated induction of causal relationships from a time-oriented clinical database: The RX project*. Proc. AMIA San Francisco, 1982.
19. Blum, R.L.: *Discovery and Representation of Causal Relationships from a Large Time-oriented Clinical Database: The RX Project*. In D.A.B. Lindberg and P.L. Reichertz (Eds.), LECTURE NOTES IN MEDICAL INFORMATICS, Springer-Verlag, 1982.
20. Blum, R.L.: *Discovery, confirmation, and incorporation of causal relationships from a large time-oriented clinical database: The RX project*. Computers and Biomed. Res. 15(2):164-187, April, 1982.
21. Blum, R.L.: *Discovery and representation of causal relationships from a large time-oriented clinical database: The RX project* (Ph.D. thesis). Computer Science and Biostatistics, Stanford University, 1982.
22. Blum, R.L.: *Displaying clinical data from a time-oriented database*. Computers in Biol. and Med. 11(4):197-210, 1981.
23. Blum, R.L.: *Automating the study of clinical hypotheses on a time-oriented database: The RX project*. Proc. MEDINFO 80, Tokyo, October, 1980, pp. 456-460. (Also STAN-CS-79-816)
24. Blum, R.L. and Wiederhold, G.: *Inferring knowledge from clinical data banks utilizing techniques from artificial intelligence*. Proc. Second SCAMC, IEEE, Washington, D.C., November, 1978.
25. DeLaPaz, R., Hanson, W., Bernstein, R., and Walker, M.G. "Tissue Characterization of MRI Using Fuzzy C-means Cluster Analysis" In preparation for Radiology.
26. DeLaPaz, R., Hanson, W., Bernstein, R., and Walker, M.G. "Clinical Application of Automated MRI Tissue Classification" In preparation for Magnetic Resonance in Medicine.
27. DeLaPaz, R., Bernstein, R., Chang, P., Hanson, W., and Walker, M.G.