

23. Wiederhold, G.: IN D.A.B. Lindberg and P.L. Reichertz (Eds.), *Databases for Health Care*, Lecture Notes in Medical Informatics, Springer-Verlag, 1981.
24. Wiederhold, G.: *Database technology in health care*. J. Medical Systems 5(3):175-196, 1981.

E. Funding Support Status

- 1) Representation and Use of Causal Knowledge for Inference from Databases
Robert L. Blum, M.D., Ph.D.: Principal Investigator
Total award: \$89,597 (direct + indirect)
Term: March 15, 1984 through March 14, 1986
- 2) Deriving Knowledge from Clinical Databases
Gio C. M. Wiederhold, Ph.D.: Principal Investigator
National Library of Medicine
Total award: \$291,192 (direct)
Term: May 1, 1984 through November 30, 1986

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

A. Collaborations

During the last year we have completed System Programmer's Manuals and a User's Manual as steps towards making the system available to outside collaborators. We have had preliminary discussions with Drs. Edward Shortliffe and Lawrence Fagan concerning use of components of RADIX in the ONCOCIN project. Once the RADIX program is developed, we would anticipate collaboration with some of the ARAMIS project sites in the further development of a knowledge base pertaining to the chronic arthritides. The ARAMIS Project at the Stanford Center for Information Technology is used by a number of institutions around the country via commercial leased lines to store and process their data. These institutions include the University of California School of Medicine, San Francisco and Los Angeles; The Phoenix Arthritis Center, Phoenix; The University of Cincinnati School of Medicine; The University of Pittsburgh School of Medicine; Kansas University; and The University of Saskatchewan. All of the rheumatologists at these sites have closely collaborated with the development of ARAMIS, and their interest in and use of the RADIX project is anticipated. We hasten to mention that we do not expect SUMEX to support the active use of RADIX as an on-going service to this extensive network of arthritis centers, but we would like to be able to allow the national centers to participate in the development of the arthritis knowledge base and to test that knowledge base on their own clinical data banks.

B. Interactions with Other SUMEX-AIM Projects

Several of the concepts incorporated into the design of the RADIX Project have been inspired by other SUMEX-AIM Projects. The RADIX knowledge base is similar to the Units Package of the MOLGEN PROJECT. The production rule inference mechanism used by us is similar to that in the MYCIN Project.

Several programs developed by the MYCIN group are regularly used by RADIX. These include disk hash file facilities, text editing facilities, and miscellaneous LISP functions.

Regular communication on programming details is facilitated by the on-line mail system.

C. Critique of Resource Management

The DEC System 20 continues to provide acceptable performance, but it is frequently heavily loaded at peak hours.

The SUMEX resource management continues to be accessible and cooperative.

III. RESEARCH PLANS

A. Project Goals and Plans

The overall goal of the RADIX Project is to develop a computerized medical information system capable of accurately extracting medical knowledge pertaining to the therapy and evolution of chronic diseases from a database consisting of a collection of stored patient records.

SHORT-TERM GOALS --

For the last two years we have concentrated more heavily on publishing and presentation of our earlier AI results, on acquisition of a 1700 patient database, on medical studies based on the enlarged database, and on reporting the medical results and statistical techniques arising from our research. This is in concert with the long-term goal of ensuring that the work of the SUMEX / Artificial Intelligence in Medicine community be disseminated and applied in the general medical community.

During the coming two years we will concentrate much more on the artificial intelligence aspects of RADIX. We were successful this year in obtaining funding from the National Library of Medicine and the National Science Foundation to pursue this work. In particular, we will be deeply concerned with the representation of causal, temporal, and quantitative medical knowledge. It has become clear that these types of knowledge are crucial for the RADIX tasks of automated discovery of medical knowledge and the provision of intelligent automated assistance to clinical researchers, in addition to their generally perceived value in other medical expert systems applications.

LONG-RANGE GOALS -- There are two inter-related long-range goals of the RADIX Project: 1) automatic discovery of knowledge in a large time-oriented database and 2) provision of assistance to a clinician who is interested in testing a specific hypothesis. These tasks overlap to the extent that some of the algorithms used for discovery are also used in the process of testing an hypothesis.

We hope to make these algorithms sufficiently robust that they will work over a broad range of hypotheses and over a broad spectrum of data distributions in the patient records.

B. Justification and Requirements for Continued Use of SUMEX

Computerized clinical data banks possess great potential as tools for assessing the efficacy of new diagnostic and therapeutic modalities, for monitoring the quality of health care delivery, and for support of basic medical research. Because of this potential, many clinical data banks have recently been developed throughout the United States. However, once the initial problems of data acquisition, storage, and retrieval have been dealt with, there remains a set of complex problems inherent in the task of accurately inferring

medical knowledge from a collection of observations in patient records. These problems concern the complexity of disease and outcome definitions, the complexity of time relationships, potential biases in compared subsets, and missing and outlying data. The major problem of medical data banking is in the reliable inference of medical knowledge from primary observational data.

We see in the RADIX Project a method of solution to this problem through the utilization of knowledge engineering techniques from artificial intelligence. The RADIX Project, in providing this solution, will provide an important conceptual and technological link to a large community of medical research groups involved in the treatment and study of the chronic arthritides throughout the United States and Canada, who are presently using the ARAMIS Data Bank through the CIT facility via TELENET.

Beyond the arthritis centers which we have mentioned in this report, the TOD (Time-Oriented Data Base) User Group involves a broad range of university and community medical institutions involved in the treatment of cancer, stroke, cardiovascular disease, nephrologic disease, and others. Through the RADIX Project, the opportunity will be provided to foster national collaborations with these research groups and to provide a major arena in which to demonstrate the utility of artificial intelligence to clinical medicine.

C. Recommendations for Resource Development

The on-going acquisition of personal work-station Lisp processors is a very positive step, as these provide an excellent environment for program development, and can serve as a vehicle for providing programs to collaborators at other sites. Continued acquisitions are very desirable.

Another resource that would be highly desirable is a faster and more reliable means for transferring data and programs interactively between SUMEX and the CIT IBM 370. The addition of a reliable local network facility would greatly facilitate our ability to transfer patient files from CIT to SUMEX.

II.A.2. National AIM Projects

The following group of projects is formally approved for access to the AIM aliquot of the SUMEX-AIM resource. Their access is based on review by the AIM Advisory Group and approval by the AIM Executive Committee.

In addition to the progress reports presented here, abstracts for each project and its individual users are submitted on a separate Scientific Subproject Form.

II.A.2.1. CADUCEUS Project

CADUCEUS Project

**J. D. Myers, M.D. and Harry E. Pople, Jr., Ph.D.
University of Pittsburgh
Decision Systems Laboratory
Pittsburgh, Pa., 15261**

I. SUMMARY OF RESEARCH PROGRAM

A. Project rationale

The principal objective of this project is the development of a high-level computer diagnostic program in the broad field of internal medicine as an aid in the solution of complex and complicated diagnostic problems. To be effective, the program must be capable of multiple diagnoses (related or independent) in a given patient.

A major achievement of this research undertaking has been the design of a program called INTERNIST-I, along with an extensive medical knowledge base. This program has been used over the past decade to analyze many hundreds of difficult diagnostic problems in the field of internal medicine. These problem cases have included cases published in medical journals (particularly Case Records of the Massachusetts General Hospital, in the New England Journal of Medicine), CPCs, and unusual problems of patients in our Medical Center. In most instances, but by no means all, INTERNIST-I has performed at the level of the skilled internist, but the experience has high-lighted several areas for improvement.

B. Medical Relevance and Collaboration

The program inherently has direct and substantial medical relevance.

The institution of collaborative studies with other institutions has been deferred pending completion of the programs and knowledge base enhancements required for CADUCEUS. The installation of our own, dedicated VAX computer can be expected to aid considerably any future collaboration.

C. Highlights of Research Progress

---Accomplishments this past year

In a previous progress report the concept of "facets" of diseases was introduced and the need of CADUCEUS to proceed from broad pathophysiological and pathobiochemical concepts to specific disease processes was emphasized. The need for better representation of anatomical information and for better time representation were pointed out.

Drs. Miller and Myers have continued in the development of a new format for the CADUCEUS knowledge base. A major goal in making the transition from the INTERNIST-I knowledge base to that of CADUCEUS has been to insure that there is continuity between the two: The CADUCEUS knowledge base will be derived from the information in the INTERNIST-I knowledge base, with significant additions made as necessary.

A screen-oriented editor program for entering and manipulating the knowledge base was written in Franz Lisp. Using the editor, a total of 52 diagnosis nodes have been created and a total of 282 findings have been defined. Due to the more complex nature of a finding, the 282 findings represent over 600 old INTERNIST-I style manifestations.

In the CADUCEUS knowledge base, the basic unit of observational information is called a finding. Unlike an INTERNIST-I manifestation, a finding can be assigned a status within a given patient, either "normal" or any number of forms of abnormal. For example, the status of the finding "heart murmur" can be either absent (normal) or present. Various qualifiers are allowed to modify a finding. For the finding "heart murmur", in a specific patient, a user might specify that it is heard at the second left interspace, that it is systolic, that it is heard in early systole only, that it is blowing, that it is grade 2 of 6, and its shape is crescendo-decrescendo. For findings whose values vary numerically, (e.g. SGOT-blood) the units of measurement and the normal range are specified so that a user may simply enter a number as the result of the test.

The concept of a disease profile has been carried over from the INTERNIST-I knowledge base. However, there are three separate diagnostic node types represented in the CADUCEUS knowledge base: the disease, the facet, and the subdivision. A disease is an entity whose presence should be reported if detected in a patient, and conceptually corresponds to the diseases mentioned in the separate chapters of standard medical textbooks. A subdivision is either a specific subtype of a disease (e.g. hepatitis B is a subtype of acute viral hepatitis, a disease) or a major specific organ system involvement by a multisystem disease (e.g. lupus nephritis and lupus cerebritis are subdivisions of the disease system lupus erythematosus).

The internal organization of disease, facet and subdivision profiles is identical. Apart from links to other nodes, there are nine essential components to each profile: disease parameters (e.g. prevalence of disease, specific sites it effects); demographic information about patients with the disease; general predisposing factors (which are interdependent, only one of which is likely to be present); independent risk factors (which often co-exist synergistically); general findings caused by the illness; specific findings which are relatively unique to the disease process; characteristic findings (e.g. a positive throat culture for beta hemolytic streptococcus in streptococcal pharyngitis); academically known but clinically contraindicated findings (e.g. one should not do a renal biopsy in patients with renal leptospirosis, but we know what the biopsy will show if it is done anyway); and manifestations whose presence make the diagnosis untenable (e.g. male sex makes pregnancy an invalid consideration). In addition to the aforementioned work in internal medicine, Drs. Gordon Banks and John Vries have been working on the development of a neurological diagnostic component for CADUCEUS. Dr. Banks has developed a neuroanatomic database which contains spatial descriptors for nearly 1,000 neuroanatomic structures and contains information as to their blood supply, and function. This database will allow anatomic localization of neurologic lesions. Some of this work for the peripheral nervous system has been done previously by students in our laboratory. The approach to the central nervous system has been to design a set of "symbolic coordinates". In constructing the neuroanatomic database, the human body, including the nervous system, is conceptually partitioned into a set of cubes (boxes). The largest cube, containing the entire body, is 2.187m on a side. This cube is divided into 27 smaller cubes, each 729mm on a side. Each of the smaller cubes is likewise subdivided until finally cubes that are each 1mm on a side are reached. Thus any cube has neighbors (of equal size) rostral, caudal, ventral, dorsal, left, and right of it, as well as a "parent" cube which contains it, and "daughter" cubes which it contains. Each of these cubes has the potential for being represented inside the computer program with a unique name (known as an atom in LISP, the language in which the database is programmed). Attached to each cube LISP atom

are lists of all of the anatomic structures that are completely and partially contained within the cube, as well as the blood supply to the region. This structure facilitates rapid retrieval of the location of a given anatomic structure as well as rapid localization of possible areas of involvement when there is evidence of dysfunction of one or more neural systems.

The hierarchical arrangement of the nested cubes ensures rapid convergence during searches, because if the sought object is not found in a parent cube, there is no need to search for it in any of the patient's children cubes. The addition of anatomic reasoning may allow parsimonious explanation of multiple manifestations arising from a single lesion, or allow the program to query the user regarding the presence of manifestations of involvement of areas that might be expected to be affected by whatever clinical state the program has under current consideration.

Dr. Vries has developed an imaging system using "octree encoding" to reconstruct n-dimensional images of the database as well as images of patients acquired by CT, NMR, and other neuroimaging techniques. Combining the database with the imaging system may open new areas of research, including clinical-pathological correlation of imaged lesions with symptoms, signs, and affected structures, automated reading of images, etc.

Dr. Miller in the last year completed work on a sub-project of CADUCEUS, called CPCS. He received support for this work from the National Library of Medicine New Investigator Program. The original objective of the project was to create a program, CPCS (for Computer-based Patient Case Simulator), to aid in the teaching of diagnosis to medical students. The INTERNIST-I/CADUCEUS knowledge base was to be used as the source of the program's medical expertise. This overall goal has been accomplished, and the program CPCS exists and runs on our VAX-11/780 using Franz Lisp. The CPCS project was a feasibility study to demonstrate that it is possible to construct a general case simulator. The project has been successful in that the CPCS program has been written, and runs quite well in its small test domain. But there is room for the future development of CPCS. Further construction of the CADUCEUS knowledge base, in areas beyond the current set of liver diseases, will significantly improve the utility of the CPCS program. As additional capabilities are added to CADUCEUS, the corresponding changes will be made in CPCS.

The medical knowledge base has continued to grow both in the incorporation of new diseases and the modification of diseases already profiled so as to include recent advances in medical knowledge. The knowledge base of 3/1/84 includes 591 individual disease profiles, 4,040 manifestations of disease, and about 3,500 "links" or interrelationships among diseases as well as a myriad of miscellaneous pieces of information which are essential for the correct operation of the system. Twenty new diseases have been profiled during the past year and the pediatrics knowledge base has continued to grow.

Recently the medical knowledge base (but not yet the diagnostic program) has been made available on line for use of the medical house staff at Presbyterian-University Hospital, our main teaching hospital in internal medicine, and at an affiliated community hospital in Pittsburgh, the Shadyside Hospital which operates a residency program in internal medicine. Preliminary reports indicate that the residents find the knowledge base useful.

----Research in progress

There are five major components to the continuation of this research project:

1. The enlargement, continued updating, refinement and testing of the extensive medical knowledge base required for the operation of INTERNIST-I.
2. The completion and implementation of the improved diagnostic consulting program, CADUCEUS, which has been designed to overcome certain performance problems identified during the past years of experience with the original INTERNIST-I program.
3. Institution of field trials of CADUCEUS on the clinical services in internal medicine at the Health Center of the University of Pittsburgh.
4. Expansion of the clinical field trials to other university health centers which have expressed interest in working with the system.
5. Adaptation of the diagnostic program and data base of CADUCEUS to subservise educational purposes and the evaluation of clinical performance and competence.

Current activity is devoted mainly to the first two of these, namely, the continued development of the medical knowledge base, and the implementation of the improved diagnostic consulting program.

D. List of relevant publications

1. Pople, Harry E.: *Knowledge-based Expert Systems: The Buy or Build Decision* IN Walter Reitman (Ed.), ARTIFICIAL INTELLIGENCE APPLICATIONS FOR BUSINESS. Proceeding of the NYU Symposium. Ablex Pub. Corp., May 1983, pp. 23-40.
2. Myers, J.D.: *Artificial Intelligence and Medical Education* The Medical Journal, St. Joseph Hospital, Houston. Vol. 18, December 1983, pp. 193-202.

E. Funding support

1. Clinical Decision Systems Research Resource
 Harry E. Pople, Jr., Ph.D.
 Associate Professor of Business
 Jack D. Myers, M.D.
 University Professor (Medicine)
 University of Pittsburgh
 Division of Research Resources
 National Institutes of Health

 5 R24 RR01101-07
 07/01/80 - 06/30/85 - \$1,607,717
 07/01/83 - 06/30/84 - \$369,484
2. CADUCEUS: A Computer-Based Diagnostic Consultant
 Harry E. Pople, Jr., Ph.D.
 Associate Professor of Business
 Jack D. Myers, M.D.
 University Professor (Medicine)
 University of Pittsburgh
 National Library of Medicine
 National Institutes of Health

5 R01 LM03710-04
 07/01/80 - 06/30/85 - \$817,884
 07/01/83 - 06/30/84 - \$196,710

3. Neurologic Consultation Computer Program
 Gordon E. Banks, M.D.
 Assistant Professor of Medicine
 National Library of Medicine - New Investigator
 National Institutes of Health

5 R23 LM03889-02
 04/01/82 - 03/31/85 - \$107,675
 04/01/83 - 03/31/84 - \$35,975
 04/01/84 - 03/31/85 - \$35,975

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

A,B. Medical Collaborations and Program Dissemination Via SUMEX

CADUCEUS remains in a stage of research and development. As noted above, we are continuing to develop better computer programs to operate the diagnostic system, and the knowledge base cannot be used very effectively for collaborative purposes until it has reached a critical stage of completion. These factors have stifled collaboration via SUMEX up to this point and will continue to do so for the next year or two. In the meanwhile, through the SUMEX community there continues to be an exchange of information and states of progress. Such interactions particularly take place at the annual AIM Workshop.

C. Critique of Resource Management

SUMEX has been an excellent resource for the development of CADUCEUS. Our large program is handled efficiently, effectively and accurately. The staff at SUMEX have been uniformly supportive, cooperative, and innovative in connection with our project's needs.

III. RESEARCH PLANS

A. Project Goals and Plans

Continued effort to complete the medical knowledge bases in internal medicine and pediatrics will be pursued including the incorporation of newly described diseases and new or altered medical information on "old" diseases. The latter two activities have proven to be more formidable than originally conceived. Profiles of added diseases plus other information is first incorporated into the medical knowledge base at SUMEX before being transferred into our newer information structures for CADUCEUS on the VAX. This sequence retains the operative capability of INTERNIST-I as a computerized "textbook of medicine" for educational purposes.

B. Justification and Requirements for Continued SUMEX Use

Our use of SUMEX will obviously decline with the installation of our VAX. Nevertheless, the excellent facilities of SUMEX are expected to be used for certain developmental work. It is intended for the present to keep INTERNIST-I at SUMEX for comparative use as CADUCEUS is developed here. Our team hopes to remain as a component of the SUMEX community and to share experiences and developments.

C. Needs and Plans for Other Computing Resources Beyond SUMEX-AIM

Our predictable needs in this area will be met by the dedicated VAX computer recently installed.

D. Recommendations for Future Community and Resource Development

Whether a program like CADUCEUS, when mature, will be better operated from centralized, larger computers or from the developing self contained personal computers is difficult to predict. For the foreseeable future it would seem that centralized, advanced facilities like SUMEX will be important in further program development and refinement.

II.A.2.2. CLIPR - Hierarchical Models of Human Cognition

Hierarchical Models of Human Cognition (CLIPR Project)

**Walter Kintsch and Peter G. Polson
University of Colorado
Boulder, Colorado**

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The two CLIPR projects have made progress during the last year. The prose comprehension project has completed one major project, and is designing a prose comprehension model that reflects state-of-the-art knowledge from psychology (van Dijk & Kintsch, 1983) and artificial intelligence. During the last year, Polson, in collaboration with Dr. David Kieras of the University of Arizona have have continued work on a project studying the psychological factors underlying device complexity and the difficulties that nontechnically trained individuals have in learning to use devices like word processors. They have develop formal representations of a user's knowledge of how of operate a device and of the user-device interface (Kieras & Polson, in Press) and have completed several experiments evaluating their theory.

B. Technical Goals

The CLIPR project consists of two subprojects. The first, the text comprehension project, is headed by Walter Kintsch and is a continuation of work on understanding of connected discourse that has been underway in Kintsch's laboratory for several years. The second, the device complexity project is headed by Peter Polson in collaboration with David Kieras of the University of Arizona, Tucson. They are studying the learning and problem solving processes involved in the utilization of devices like word processors or complex computer controlled medical instruments (Kieras & Polson, in Press)

The goal of the prose comprehension project is to develop a computer system capable of the meaningful processing of prose. This work has been generally guided by the prose comprehension model discussed by Kintsch and van Dijk (1978), although our programming efforts have identified necessary clarifications and modifications in that model (Miller & Kintsch, 1980, 1981; Kintsch & Miller, 1981; Miller, 1982). In general, this research has emphasized the importance of knowledge and knowledge-based processes in comprehension, and we are accordingly working with the AGE and UNITS groups at SUMEX toward the development of a knowledge-based, blackboard model of prose comprehension. We hope to be able to merge the substantial artificial intelligence research on these systems with psychological interpretations of prose comprehension, resulting in a computational model that is also psychologically respectable.

The goal of the device complexity project is to develop explicit models of the user-device interaction. They model the device as a nested automata and the user as a production system. These models make explicit kinds of knowledge that are required to operate different kinds of devices and the processing loads imposed by different implementations of a device. We feel that tools being developed at SUMEX--in particular AGE and the UNIT package--will dramatically facilitate our abilities to generate such models of the user-device interface.

C. Medical Relevance and Collaboration

The text comprehension project impacts indirectly on medicine, as the medical profession is no stranger to the problems of the information glut. By adding to the research on how computer systems might understand and summarize texts, and determining ways by which the readability of texts can be improved, medicine can only be helped by research on how people understand prose. Development of a more thorough understanding of the various processes responsible for different types of learning problems in children and the corresponding development of a successful remediation strategy would also be facilitated by an explicit theory of the normal comprehension process.

Note that our goal of a blackboard model is particularly relevant to the understanding of learning difficulties. One important aspect of a blackboard model is the separation of cognitive processes into a set of interacting subprocesses. Once such subprocesses have been identified and constructed, it would be instructive to observe the model's performance when certain of these processes are facilitated or inhibited. Many researchers have shown that there are a variety of cognitive deficits (insufficient short-term memory capacity, poor long-term memory retrieval, and such) that can lead to reading problems. Having a blackboard model in which the power of individual components could be manipulated would be a significant step in determining the nature of such reading problems.

The device complexity project has two primary goals: the development of a cognitive theory of user-device interaction in including learning and performance models, and the development of a theoretically driven design process that will optimize the relationships between device functionality and ease of learning and other performance factors (Polson & Kieras, 1983). The results of this project should be directly relevant to the design of complex, computer controlled medical equipment. We are currently using word processors to study user-device interactions, but principles underlying use of such devices should generalize to medical equipment.

Both the text comprehension project and the device complexity project involve the development of explicit models of complex cognitive processes; cognitive modelling is a stated goal of both SUMEX and research supported by NIMH.

The on-going development of the prose comprehension model would not be possible without our collaboration with the AGE and UNITS research groups. We look forward to a continued collaboration, with, we hope, mutually beneficial results. Several other psychologists have either used or shown an interest in using an early version of the prose comprehension model, including Alan Lesgold of SUMEX's SCP project, who is exporting the system to the LRDC Vax. We have also worked with James Greeno -- another member of the SCP project -- on a project that will integrate this model with models of problem solving developed by Greeno and others at the University of Pittsburgh. Needless to say, all of this interaction has been greatly facilitated by the local and network-wide communication systems supported by SUMEX. There has been considerable communication between members of the prose comprehension and AGE/UNITS groups as program bugs have been discovered and corrected; the presence of a mail system has made this process infinitely easier than if telephone or surface mail messages were required. The mail system, of course, has also enabled us to maintain professional contacts established at conferences and other meetings, and to share and discuss ideas with these contacts.

D. Progress Summary

The prose comprehension project has completed an initial version of a model of prose comprehension (Miller & Kintsch, 1980). This model has been applied to a large number of texts, and has yielded quite reasonable predictions of recall and readability. Psychologists from other universities have used this system to derive reading time and recall predictions for their own experimental materials. We are currently using the AGE and UNITS packages to extend this model toward one that can make use of world knowledge in its analyses; this model is discussed in Miller and Kintsch (1981) and Miller (1982). It is further developed in van Dijk and Kintsch (1983) has been applied to the domain of word arithmetic problems in our most recent work (Kintsch and Greeno, in Press).

The device complexity project is in it's third year. We have developed an explicit model for the knowledge structures involved in the user-device interaction, and we are developing simulation programs. Our preliminary theoretical results are described in Kieras & Polson (in Press). We have also completed several experiments evaluating the theory.

E. List of Relevant Publications

1. Kieras, D.E. and Polson, P.G.: *An outline of a theory of the user complexity of devices and systems*. Working Paper No. 1, Device Complexity Project, Universities of Arizona and Colorado, May, 1982.
2. Kieras, D.E. and Polson, P.G.: *The formal analysis of user complexity*. Int. J. Man-Machine Studies, In Press.
3. Kintsch, W. and van Dijk, T.A.: *Toward a model of text comprehension and production*. Psychological Rev. 85:363-394, 1978.
4. Kintsch, W. and Greeno, J.G.: *Understanding and solving word arithmetic problems*. Psychological Review, In Press.
5. Miller, J.R. and Kintsch, W.: *Readability and recall of short prose passages: A theoretical analysis*. J. Experimental Psychology: Human Learning and Memory 6:335-354, 1980.
6. Miller, J.R. and Kintsch, W.: *Readability and recall of short prose passages*. Text 1:215-232, 1981.
7. Miller, J.R.: *A Knowledge-based Model of Prose Comprehension: Applications to Expository Text*. IN B.K. Britton and J.B. Black (Eds.), UNDERSTANDING EXPOSITORY TEXT. Erlbaum, Hillsdale, NJ, 1982.
8. Polson, P.G. and Kieras, D.E.: *Theoretical foundations of a design process guide for the minimization of user complexity*. Working Paper No. 3, Project on User Complexity, Universities of Arizona and Colorado, June, 1983.
9. Polson, P.G. and Kieras, D.E.: *A formal description of users' knowledge of how to operate a device and user complexity*. Behavior Research Methods and Instrumentation.
10. van Dijk, T.A. and Kintsch, W.: *STRATEGIES OF DISCOURSE COMPREHENSION*. Academic Press, New York, 1983.

F. Funding Support Status

1. Text Comprehension and Memory
Walter Kintsch, Professor, University of Colorado
National Institute of Mental Health - 5 Rol MH15872-14-16
7/1/81 - 6/30/84: \$281,085
7/1/83 - 6/30/84: \$69,878
2. Understand and solving word arithmetic problems
Walter Kintsch, Professor, University of Colorado
National Science Foundation
8/1/83 - 7/31/86: \$200,000
3. User Complexity of Devices and Systems
David Kieras, Associate Professor, University of Arizona
Peter G. Polson, Professor, University of Colorado
International Business Machines Corporation
1/1/82 - 12/31/84: \$364,000
1/1/84 - 12/31/84: \$145,000

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

A. Sharing and Interactions with Other SUMEX-AIM Projects

Our primary interaction with the SUMEX community has been the work of the prose comprehension group with the AGE and UNITS projects at SUMEX. Feigenbaum and Nii have visited Colorado, and one of us (Miller) attended the AGE workshop at SUMEX. Both of these meetings have been very valuable in increasing our understanding of how our problems might best be solved by the various systems available at SUMEX. We also hope that our experiments with the AGE and UNITS packages have been helpful to the development of those projects.

We should also mention theoretical and experimental insights that we have received from Alan Lesgold and other members of the SUMEX SCP project. The initial comprehension model (Miller & Kintsch, 1980) has been used by Dr. Lesgold and other researchers at the University of Pittsburgh, as well as researchers at Carnegie-Mellon University, the University of Manitoba, Rockefeller University, and the University of Victoria.

B. Critique of Resource Management

The SUMEX-AIM resource is clearly suitable for the current and future needs of our project. We have found the staff of SUMEX to be cooperative and effective in dealing with special requirements and in responding to our questions. The facilities for communication on the ARPANET have also facilitated collaborative work with investigators throughout the country.

III. RESEARCH PLANS

A. Long Range Projects Goals and Plans

The use of SUMEX by the prose comprehension group was greatly reduced in the

two years, because the focus of the work during that period was on experimental work and book writing, rather than computer simulation. This will change in the fall of 1984, when a new research associate will join the project whose primary responsibility will be in continuing the modelling work started in previous years with J. Miller (who is no longer associated with us). Thus, we expect a level of activity comparable to previous years next fall.

The primary goal of the device complexity project is the development of a theory of the processes and knowledge structures that are involved in the performance of routine cognitive skills making use of devices like word processors. We plan to model the user-device interaction by representing the users processes and knowledge as a production system and the device as a nested automata. We are also studying the role of mental models in learning how to use them.

B. Justification and Requirements for Continued SUMEX Use

The research of the prose comprehension project is clearly tied to continued access to the AGE and UNITS packages, which are simply not available elsewhere. We hope that our continued use of these systems will be offset by the input we have been and will continue to provide to those projects: our relationship has been symbiotic, and we look forward to its continuation.

C. Needs and Plans for Other Computational Resources

We currently use two other computing systems located at the University of Colorado. One is the Department of Psychology's VAX 11/780, which is used primarily to run real-time experiments to be modeled on SUMEX. The second is the University of Colorado's CDC 6400, which is used for various types of statistical analysis.

When the ARPA-sponsored Vax/Interlisp project is completed, we would be most interested in experimenting with becoming a remote AGE/UNITS site. It would seem that this sort of development is the ultimate goal of the package projects, and this type of interaction, once it becomes feasible, would be a logical extension of our association with the SUMEX facility.

D. Recommendations for Future Community and Resource Development

Our primary recommendation for future development within SUMEX involves (a) the continued support of INTERLISP, which is needed for AGE and for other work we have underway on SUMEX and (b) the continued development of the AGE and UNITS projects. In particular, we would like to see an extension of AGE to include a wider variety of control structures so that our psychological models would not be confined to one particular view of knowledge-based processing. The limited physical capacity of SUMEX, both in terms of address space and overloading, is, as before, a major problem. The prose comprehension group can no longer use the publicly released AGE/UNITS system due to its severely limited address space, and has had to build a personal AGE system from a stripped-down version of Interlisp and a selected subset of AGE and UNITS. We heartily endorse the plans underway to obtain more computing capacity for the SUMEX project.

Given our acquisition of a VAX, we particularly support the ongoing and continued development of INTERLISP for the VAX, so that local use of AGE and UNITS would be possible. Since we, as well as other psychologists, need the real-time capability of VAX/VMS to run on-line experiments, we hope that the INTERLISP system to be developed will be compatible with VMS. Note that this need for real-time work coincides with real-world applications of SUMEX programs, in which a VAX might be devoted to both real-time patient monitoring and diagnostic systems such as PUFF or MYCIN.

II.A.2.3. Rutgers Research Resource

Rutgers Research Resource--Computers in Biomedicine

**Principal Investigators: Saul Amarel [1982-83],
Casimir Kulikowski, Sholom Weiss [1983-84].
Rutgers University, New Brunswick, New Jersey**

I. SUMMARY OF RESEARCH PROGRAM

A. Goals and Approach

The fundamental objective of the Rutgers Resource is to develop a computer based framework for significant research in the biomedical sciences and for the application of research results to the solution of important problems in health care. The central concept is to introduce advanced methods of computer science - particularly in artificial intelligence into specific areas of biomedical inquiry. The computer is used as an integral part of the inquiry process, both for the development and organization of knowledge in a domain and for its utilization in problem solving and in processes of experimentation and theory formation.

At present, the total number of investigators who participate in scientific activities of the Resource is 83, of these, 20 have Rutgers appointments, 21 are outside investigators who participate in collaborative research projects that are mainly located at Rutgers, and 42 are investigators from collaborative national AIM projects that are located in different parts of the country. In addition, the Resource has 12 other members in Administrative, Computer Systems/Operations and general programming and secretarial functions. Thus, the Rutgers Resource community numbers at present a total of 95 participants.

Resource activities include research projects (collaborative research and core research) training/dissemination projects, and computing services in support of user projects.

B. Medical Relevance and Collaborations

In 1983-84 we continued the development of several versatile systems for building and testing consultation models in biomedicine. The EXPERT system has had many of its capabilities enhanced in the course of collaborative research in the areas of rheumatology, ophthalmology, and clinical pathology.

In ophthalmology we have developed a knowledge representation scheme for treatment planning which is both natural and efficient for encoding the strategies for choosing among competing and cooperating treatment plans. This involves a ranking of treatments according to their characteristics and desired effects as well as contraindications. Kastner has generalized the scheme so that it is now being used for a number of reasoning models: infectious eye disease, primary eye care, and rheumatology management. Our main collaboration continues to be with Dr. Chandler Dawson of the Proctor Foundation, UCSF.

In rheumatology, our collaboration with Drs. Donald Lindberg and Gordon Sharp at the University of Missouri-Columbia has continued at a very active level. The model for rheumatological diseases which now includes detailed diagnostic criteria for 26 major

diseases, had the management advice and treatment planning developed further. Dr. Sharp's group continues to develop the knowledge base in this area, with formalization of the knowledge carried out in conjunction with Dr. Lindberg's group and the Medical Expert Systems Group at Rutgers. The Resource researchers have developed new representational elements for EXPERT in response to the needs of the rheumatology research, and Politakis has developed a coordinated system called SEEK (System for Empirical Experimentation with Expert Knowledge) which provides interactive assistance to the human expert in testing, refining and updating a knowledge base against a data base of trial cases. SEEK has been tested and extended during the past year.

In clinical pathology our main collaboration has been with Dr. Robert Galen (Cleveland Clinic Foundation), with whom we have developed the serum protein electrophoresis model which is incorporated into an instrument - the scanning densitometer manufactured by Helena Laboratories. This instrument with interpretive reporting capabilities has now been on the market for over a year, is located at over 100 clinical sites, and represents the first known spin-off of AI expert systems research in the field of laboratory instrumentation. We continue to refine the representational mechanisms used for this kind of model.

In biomedical modeling applications we are experimenting with several prototype models for giving advice on the interpretation of experimental results in the field of enzyme kinetics, in conjunction with Dr. David Garfinkel. His PENNZYME program has been linked to a model in EXPERT, which allows the user to interpret the progress of the model analysis.

C. Highlights of Research Progress

Expert Medical Systems (C. Kulikowski, S. Weiss)

Research has continued on problems of representation, inference and control in expert systems. Emphasis has been placed this year on problems of knowledge base acquisition, empirical testing and refinement of reasoning (the SEEK system), and treatment planning strategies over time. From a technological point of view the market availability of the interpretive reporting version of a scanning densitometer, and the development of models for eye care consultation that run on microprocessor systems (Apple IIe, IBM-PC) represents an important achievement for AIM research in showing its practical impact in medical applications. This was recognized by the award of a scientific exhibit prize at the Academy of Ophthalmology Annual Meeting in November 1983.

1.1) SEEK: A System for Empirical Experimentation with Expert Knowledge

SEEK is a system which has been developed to give interactive advice about rule refinement during the design of an expert system. The advice takes the form of suggestions for possible experiments in generalizing and specializing rules in an expert model that has been specified based on reasoning rules cited by a human expert. Case experience, in the form of stored cases with known conclusions, is used to interactively guide the expert in refining the rules of a model. The design framework of SEEK consists of a tabular model for expressing expert-modeled rules and a general consultation system for applying a model to specific cases. This approach has proven particularly valuable in assisting the expert in domains where the logic for discriminating two diagnoses is difficult to specify; and we have benefited primarily from experience in building the consultation system in rheumatology.

1.2) Treatment Planning

The ranking and selection strategies developed as a stand-alone system last year

have been incorporated into the EXPERT framework. Capabilities for expressing reasoning over time have been added, so stored chart reviews can be carried out automatically, summarizing various patterns of findings over time, and abstracting the major features of interest for prognostic advice or treatment recommendations. Applications have been in infectious eye disease modeling, rheumatology treatment, and sequential advice in interpretation and sequencing of cardiac enzyme tests (e.g. CPK/LDH isoenzymes).

1.3) Technology Transfer

Important technology transfer milestones have also been achieved this year: the instrument interpretation EXPERT program for serum protein has been widely disseminated after being made available by Helena Laboratories, based on the prototype program developed by us; and we have succeeded in transferring a large knowledge base in rheumatology (about 1000 findings, 400 hypotheses and 1000 rules) onto a microprocessor (Motorola 68000) based system - the WICAT - which is well within the means of clinical researchers and practitioners. This system has been on site at the University of Missouri during the last year for testing and refining of the knowledge base.

1.4) Learning with Prior Structural Knowledge

This approach to knowledge acquisition and representation has as its goal to allow the expert to specify just the elements that are to enter into the reasoning model, with a few causal and taxonomic relations. These should then be sufficient to guide a learning program which operates on a data base of cases with known end-points. Such an approach would be useful in situations where the expert either has little time to explicitly formulate decision rules, or finds it difficult to do so. Our program [Drastal and Kulikowski, 1982] uses a blackboard representation, with multiple knowledge sources to handle the different conclusions, and the formation of rules from the data that pertain to them. We have tested this scheme in the areas of glaucoma and rheumatology, and shown that there are some interesting tradeoffs between the degree of a-priori structure provided by the expert, and the complexity of rule generation.

In relation to a system like SEEK, this approach represents a preprocessing or alternative means of developing the prototype model. We are now investigating the role of additional medical semantic constraints on the strategies of rule generation.

2) Artificial Intelligence: Expertise Acquisition and Problem Reformulation (S. Amarel)

The main research activity in this area is concerned with improvements in problem solving expertise via shifts in problem representation, i.e., via reformulation.

In this research, we have concentrated on the developmental processes that lead to the formation of specialized high performance procedures in sub-domains of a problem class. Theory formation is a key task in these processes; and we are now studying several approaches to this task - both top-down, model guided, approaches and 'bottom-up' methods that are based on detailed analysis of individual cases.

D. Up-to-Date List of Publications

The following is an update of publications in the Rutgers Resource for the period 1983 and 1984 (only publications not listed in previous SUMEX annual reports are presented here).

1. Weiss, S.M. and Kulikowski, C.A. *A Practical Guide to Designing Expert Systems*, Rowman and Allanheld, 1984.

2. Kulikowski, C.A. contributor to the Knowledge Acquisition chapter edited by B. Buchanan in the book *Building Expert Systems* (F. Hayes- Roth, et al., eds) Addison-Wesley, 1983 (in press).
3. Yao, Y. and Kulikowski, C.A., " *Multiple Strategies of Reasoning for Expert Systems*", Proc. Sixteenth Hawaii International Conference on Systems Sciences, pp. 510-514 , 1983.*
4. Kulikowski, C.A. " *Progress in Expert AI Medical Consultation Systems: 1980 - 1989* ", Proc. MEDINFO '83 , pp. 499-502, Amsterdam, August 1983.*
5. Kastner, J.K., Weiss, S.M., and Kulikowski, C.A., " *An Efficient Scheme for Time-Dependent Consultation Systems*", Proc. MEDINFO '83, pp.619-622, 1983.*
6. Kulikowski, C.A. " *Expert Medical Consultation Systems*", Journal of Medical Systems, v.7, pp. 229-234, 1983.*
7. Weiss, S.M., Kulikowski, C.A., and Galen, R.S., " *Representing Expertise in a Computer Program: The Serum Protein Diagnostic Program*", Journal of Clinical Laboratory Automation, v.3, pp. 383-387, 1983.*
8. Kastner, J.K., Weiss, S.M., and Kulikowski, C.A., " *An Expert System for Front-line Health Workers in Primary Eye Care*", Proc. Seventeenth Hawaii International Conference on Systems Sciences, pp. 162-166, 1984.*
9. Kulikowski, C.A. " *Knowledge Acquisition and Learning in EXPERT*", Proc. 1983 Workshop on Machine Learning, Univ. of Illinois,Champaign-Urbana 1983.

Indicate by an asterisk (*) that the resource was given credit.

E. Funding Support

Since December 1983, the Rutgers Research Resource on Artificial Intelligence in Medicine is funded under grant RR 02230-01 from the Division of Research Resources, Biotechnology Resources Program. Principal Investigators are Casimir A. Kulikowski, Professor of Computer Science and Chairman of the Department of Computer Science [1984-87], and Dr. Sholom M. Weiss, Associate Research Professor of Computer Science.

The total direct costs for the period 1983-87 is \$3,198,075, with the total for the current period (December 1, 1983 - November 30, 1984) being \$ 989,276.

The Rutgers Resource was funded until December 1983 through an NIH grant entitled "Rutgers Research Resource on Computers in Biomedicine" - number P41RR643. The Co-Principal Investigators were Dr. Saul Amarel, Professor, Chairman of the Department of Computer Science, and Director of the Laboratory for Computer Science Research, and Dr. Casimir Kulikowski, Professor of Computer Science at Rutgers.

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

A. Medical Collaborations and Dissemination

The SUMEX-AIM facility provides a backup node where some of our medical collaborators can access programs developed at Rutgers. The bulk of the medical collaborative work outlined in I.B. above is centered at the Rutgers facility (the Rutgers-AIM node).

Dissemination activities continue to be an important responsibility of the Rutgers Resource within the AIM community. The following activities took place in the last year:

1. Ninth AIM Workshop (1983):

Organized by Dr. Casimir Kulikowski, it was held in Baltimore, in conjunction with the SCAMC 83 meeting. It consisted of a series of working group discussions followed by summary presentations by members of the AIM community on their conclusions.

2. Hawaii International Conference On Systems Sciences:

Dr. Weiss presented a paper on the expert system for front-line health workers, and Dr. Kulikowski chaired a session on knowledge based medical systems.

3. VII-Pan-American Congress on Rheumatology:

Dr. Sharp presented the rheumatology knowledge base and consultation program at this meeting.

4. At the AAI-82 meeting, S. Amarel was elected member of the Executive Council of AAI. He is also General Chairman of IJCAI-83 which was held in Karlsruhe, W. Germany in August 1983. Dr. Kulikowski was the organizer for an expert medical systems session at MEDINFO 83.

B. National AIM Projects at Rutgers

The national AIM projects, approved by the AIM Executive Committee, that are associated with the Rutgers-AIM node are the following:

1. INTERNIST/CADUCEUS project, headed by Dr. Myers and Dr. Pople from the University of Pittsburgh, has been using the Rutgers Resource as a backup system for development and experimentation.
2. Medical Knowledge Representation project, headed by Dr. Chandrasekaran from Ohio State University, is doing most of its research on the Rutgers system.
3. PURSUIT project, directed by Dr. Greenes from Harvard University, is doing most of its research on a Goal-Directed Model of Clinical Decision-Making at Rutgers.
4. Biomedical Modeling, by Dr. Garfinkel from the University of Pennsylvania.
5. Attending Project, directed by Dr. Perry Miller of the Yale Medical Center, is doing much of the research on critiquing a physician's plan of management at Rutgers.
6. MEDSIM project: This is a pilot project designed to provide resource-sharing and community building facilities for about 25 researchers in bio-mathematical modeling and simulation.

C. Critique of SUMEX-AIM Resource Management

Rutgers is currently using the SUMEX DEC-20 system primarily for communication with other researchers in the AIM community and with SUMEX staff, and

also for backup computing in demonstrations, conferences and site visits. Our usage is currently running at less than 50 connect hours per year at SUMEX, with an overall connect/CPU ratio of about 30.

Rutgers is beginning to place more emphasis on the use of personal computers, and on network support needed to make these effective. Sumex has been help in the following ways:

- The AIM Executive Committee allocated to the Rutgers-AIM node one of the Xerox Dolphins acquired by SUMEX, to help us develop experience in supporting personal machines. This machine was used almost entirely to help us develop and test network support(We are using Ethernet with the Xerox PUP networking protocols), and subsequently returned to SUMEX.
- Most of network software that we use was originally developed at SUMEX. Having this software available has saved us an enormous amount of time.
- Initially SUMEX was very helpful in giving us advice about setting up our Ethernet and the Dolphins.

III. RESEARCH PLANS

A. Project Goals and Plans

We are planning to continue along the main lines of research that we have established in the Resource to date. Our medical collaborations will continue with emphasis on development of expert consultation systems in rheumatology, ophthalmology and clinical pathology. The basic AI issues of representation, inference and planning will continue to receive attention. Our core work will continue with emphasis on further development of the EXPERT framework and also on AI studies in representations and problems of knowledge and expertise acquisition. We propose to work on a number of technology transfer experiments to micro processing that will be affordable by our biomedical research and clinical collaborators. We also plan to continue our participation in AIM dissemination and training activities as well as our contribution -- via the RUTGERS/LCSR computer -- to the shared computing facilities of the national AIM network.

B. Justification and Requirements for Continued SUMEX Use

Continued access to SUMEX is needed for:

1. Backup for demos, etc.
2. Programs developed to serve the National AIM Community should be runnable on both facilities.
3. There should be joint development activities between the staffs at Rutgers and SUMEX in order to ensure portability, share the load, and provide a wider variety of inputs for developments.

C. Needs and Plans for Other Computing Resources Beyond SUMEX-AIM

Our computing is going to move in the direction of personal computers. We will continue to use Sumex for backup purposes, however.

D. Recommendations for Future Community and Resource Development

Use of personal computers and minicomputers is continuing to grow in the AIM community. We find that the biggest challenge is supporting these systems. Although some central computing will continue to be needed for communication and coordination, we believe that over the next few years all AIM research projects and even individual collaborators will come to have their own hardware. However many of these community members (particularly the collaborators) will not be in a position to support hardware or software on their own. We would certainly expect SUMEX to continue to provide expert advice in this area. However we believe it would be helpful for SUMEX to have a formal program to support smaller computers in the field. We envision this as including at least the following items:

- A central source of information on hardware and software that is likely to be of interest to the AIM community. SUMEX might want to become a distribution point for certain of this software, and even help coordinate quantity purchase of hardware if this proves useful.
- Assistance in support of hardware and software in the field. Depending upon the hardware involved, this might involve advice over the telephone or actual board-swapping by mail. With our Dolphins we have found that there are a number of problems that can be resolved over the telephone if we can find someone with appropriate expertise.

II.A.2.4. SECS: Simulation & Evaluation of Chemical Synthesis

SECS - Simulation and Evaluation of Chemical Synthesis Project

Principal Investigator: W. Todd Wipke
Board of Studies in Chemistry
University of California
Santa Cruz, CA. 95084

Coworkers:

I. Kim	(Grad student)
D. Rogers	(Grad Student)
J. Chou	(Postdoctoral)
M. Hahn	(Grad Student)
M. Yanaka	(Postdoctoral)
I. Iwataki	(Postdoctoral)

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

With the SECS project our long range goal is to develop the logical principles of molecular construction and to use these in developing practical computer programs to assist investigators in designing stereospecific syntheses of complex bio-organic molecules. Our second area of research, the XENO project, is aimed at improving methods for predicting potential biological activity of metabolites and plausibility of incorporation and excretion of metabolites.

B. Medical Relevance and Collaboration

The development of new drugs and the study of drug structure biological activity relationships depends upon the chemist's ability to synthesize new molecules as well as his ability to modify existing structures, e.g., incorporating isotopic labels or other substituents into bio-molecular substrates. The Simulation and Evaluation of Chemical Synthesis (SECS) project aims at assisting the synthetic chemist in designing stereospecific syntheses of biologically important molecules. The advantages of this computer approach over normal manual approaches are many: 1) greater speed in designing a synthesis; 2) freedom from bias of past experience and past solutions; 3) thorough consideration of all possible syntheses using a more extensive library of chemical reactions than any individual person can remember; 4) greater capability of the computer to deal with the many structures which result; and 5) capability of computer to see molecules in a graph theoretical sense, free from the bias of 2-D projection.

The objective of using XENO in metabolism studies is to predict the plausible metabolites of a given xenobiotic in order that they may be analyzed for possible carcinogenicity. Metabolism research may also find this useful in the identification of metabolites in that it suggests what to look for. Finally, one may envision applications of this technology in problem domains where one wishes to alter molecules in order to inhibit certain types of metabolism.

C. Highlights of Research Progress

C.1 SECS Project Developments

The majority of our research has been aimed at strategic planning in chemical synthesis. Specific work has included the SST project for recognizing potential starting materials from a target, the MCS project for maximal common subgraph searching, and a project for rapid substructure search using parallelism.

C.1.a SST -- Starting Material Strategies. The importance of selecting good starting materials for a synthesis has been known for a long time, but only recently has work started on applying computer techniques to the selection process. The selection of starting material for a synthesis is frequently the major discovery in a synthesis and the process of converting the starting material to the target is minor by comparison. Last year we reported development of the SST program for selecting starting materials that are appropriate for a given synthetic target using a library of available chemicals, but without reference to reactions. SST handles problems of classes I-III given below:

I) Target = SM	Identical match
II) Target > SM	Superstructure match
III) Target < SM	Substructure match
IV) None of these	Similarity match

For a search over our abstracted file, the identical match means that the target and starting materials are identical except for functionalization. The superstructure match is the case where we must *make* carbon-carbon bonds during a synthesis. The substructure match is the case where the starting material is larger than the target, so carbon-carbon bonds have to be *broken*. Finally, the similarity match is where carbon-carbon bonds have to be both *made and broken* during the synthesis.

Our research in efficient starting material strategies has continued this past year in two different areas. In the first, we have explored the prospect of using a parallel computer in the graph matching process described in the following section and in the second we have developed a solution to the class IV problem (see above) which is described in a subsequent section.

C.1.c Subgraph Search Using Parallelism.

Subgraph matching is an important method used in many different computer applications in organic chemistry, including the recognition of functional groups, synthesis planning, constraint testing in structure generation, selection of starting materials for synthesis, and structure oriented retrieval. The fundamental problem is, given a *query substructure* (QS) and a *candidate superstructure* (CS), determine if there exists a mapping of the atoms (nodes) of the substructure onto the candidate superstructure such that the connected atom pairs in the query substructure are also connected in the superstructure, and that the atom and bond types also correspond.

Although substructure search is a non-numerical problem, it is computationally demanding because ultimately it involves establishing an atom by atom correspondence between the QS and the CS, and this problem is a member of the class of NP-complete problems. In a worst case for N atoms in the QS and M atoms in the CS ($M > N$), one may have to consider $N!/(M-N)!$ mappings for each CS. The objective of our research was to explore the feasibility of applying parallel processing to this problem.

Although the node matching process is an NP-complete problem, if we eliminate all backtracking, the order of the algorithm reduces to $O(N)$, where N is the number of