Disambiguation through Deduction). A related task of importance is parsing and transformation of sentences to "data base language." This is a point where we have been able to utilize existing programs. Thus we have written a simple Swedish grammar for Bill Wood's network parser and in order to increase efficiency developed a compiler for the parser. This compiler generates a LISP program which is equivalent to a given grammar. In parallel we have our own work on parsers both for natural and programming languages; the latter, e.g., in connection with the work of Mats Nordstrom, who uses LISP to give a formal definition of the semantics of SIMULA. A project on the applied level, where small data base methods are employed, aims at facilitating the communication between an ecologist and a large data base containing environmental data by means of a flexible query language.

Most important for the possibility of concentration on essential problems is the support by appropriate programming systems and general program modules. Thus we have devoted a considerable part of our resources on work in these areas.

The main effort on the base level is the development of BBN-Uppsala-LISP for IBM 360/370 performed by Jaak Urmi in co-operation with the Uppsala Computer Center. This is an interactive paged LISP system including compiler, break and editing facilities as in the original BBN-LISP, though some features of INTERLISP might not be incorporated. BBN-Uppsala-LISP is presently active and used within our group and will be released during the spring.

An improved version of Mats Norstrom's FORTRAN implemented LISP interpreter, formerly reported on in AISB Newsletter May 72 and up to now distributed to more than 75 places, is now completed and will be documented in the near future. Further projects on the base level are a language PLABS, with BASIC-resembling syntax and containing the symbol and list handling facilities of LISP, and an experimental recursive language REC.

On the intermediate level we are developing support for small data base applications with emphasis on program generators and manipulators rather than general programs. A major package here is PCDB which is a program generator for management of and deduction in a data base built from predicate calculus formulas. (Haraldson: PCDB - A Program which Generates Procedures for Maintaining a Data Base of Formulas in Predicate Calculus.) Other programs make reductions of automatically generated functions using substitution and partial evaluation, remove recursion (Risch: REMREC - A Program for Automatic Recursion Removal in LISP) or support flexible input/output of property lists.

Another line of research represented at Datalogilaboratoriet in Uppsala is programming support for conventional languages such as FORTRAN, etc. There is also a special group for socio-cybernetic research which is developing a simulation system for social processes. This system is characterized by being modular, not completely numerical, self-organizing, and using several levels of description.

A Nordic Summer School will be held at Datalogilaboratoriet, Uppsala University, June 3 - 14 on the topic of methods for small data bases. Lectures are held in Scandinavian languages and are open to Nordic graduate students and other persons with corresponding background.

E. THE NIH HEURISTICS LABORATORY

by
C. L. Chang
Division of Computer Research and Technology

The Heuristics Laboratory had a very fruitful year. More than twenty papers were either published or accepted, about three papers per man-year. Most problems undertaken by the staff of the Laboratory are concerned with the application of computers to biomedical fields. Since these problems are large and difficult, efficient and heuristic methods are needed to solve them. Therefore, the activities of the Laboratory are creative and at the frontier of computer application, and are different from those routinely implementing existing methods and systems. The projects worked upon by the Heuristics Laboratory during the fiscal year are summarized as follows:
(The names of members of the Laboratory involved in each project are listed in parenthesis.)

A NEW DESIGN FOR CHEMICAL STRUCTURE SEARCH (Hodes):

We are collaborating with the Walter Reed Army Institute of Research (WRAIR) on the design of a chemical substructure system in conjunction with the conversion of the obsolete WRAIR system from a second generation to a third generation computer. The new system will allow searching on-line for both structure and substructure, and will accomodate a file whose compounds can number in the millions. There are two basic new elements in our design. First is the establishment of a system of structure fragment screens. We have developed a uniform, iterative method for determining a complete set of screens from single atoms to quite large fragments. Second, is a new use of hash coding which for the first time brings to bear the full speed of random access on substructure searching. These two elements have been integrated into a rational framework based on new theoretical work on the discrimination power of screens. We have obtained an elegant and practical design whose performance is versatile, predictable and efficient, and an improvement over other chemical search systems. Preliminary results on establishing screening fragments have been gratifying.

MASS SPECTRAL SEARCH (Heller):

As of September 1, 1973 the NIH Mass Spec Search System (MSSS) was officially put on the international GE computer network by the United Kingdom Spec Data Centre. NIH's role in the future will be that of a consultant for the system. A copy of the system remains at NIH for internal use and further research and development.

INTERPRETATION OF MASS SPECTROMETRY DATA (Slagle, Chang and Heller):

A program was developed to reorganize a large file of mass spectral data. From the display of the reorganized data, we were able to define new groups of compounds by the presences and the absences of certain mass spectral peaks. So far, we have good results for four groups: thiolesters, sulfoxides, aromatic carbamates, and amino esters. The definition rules for these groups have been tested intensively, and found to be able to discriminate a compound in the groups from other compounds not in the groups.

MLAB - AN ON-LINE MODELING LABORATORY (Knott):

MLAB has been further developed. We have spent considerable time promoting MLAB and collaborating with users in setting up models and applying them. The third edition of the MLAB

manual has been prepared. Using MLAB jointly with Jean-Marie Ketelslager, we have obtained some results on hormone binding.

MICROBIOLOGY DATA BANK (Norton):
Collaborating with Dr. Krichevsky of NIDR, we revised the program for inputting data to the microbiology data bank to give considerably more extensive diagnostic messages in cases of deviations from the specified input format which it will not forgive.

TISSUE TYPING COMPUTER AIDS (Dixon and Norton):
We are collaborating with Dr. Kayhoe of NIAID on the evaluation of the 1330 trays using the tissue-typing program. The results demonstrated new patterns of serum-antigen specificity on the NIH tray. That is, the program enabled many "false pointers" previously interpreted as "noise" or laboratory error to be interpreted systematically in terms of (sometimes weak) serum reactions to antigens determinants.

COMPUTER MODEL OF THE HUMAN MENSTRUAL CYCLE (Dixon):
We are collaborating with Dr. Cargille of NICHD on the modeling of the human menstrual cycle. A computer model written in the DYNAMO language simulates the process of ovulation and the variation of hormones during the human menstrual cycle. Last year the model did a good job of simulating the normal cycle and a dynamic mechanism for atresia was postulated and put into the model. This year the model was completely rewritten. Many awkward equations were streamlined and the whole model was made simpler and easier to understand. This new version of the model was called MARK4. A simulation of pregnancy was also included in MARK4. A statistical version of the model called MARK5 was also written. MARK5 can be run for many cycles and statistics of hormone levels at various times during the cycle can be collected automatically. MARK6 was also written. This is a version of MARK5 which uses a larger time interval (DT) so long runs can be made at less expense. A set of about 10 validity tests was then prepared. The model passed half the validity tests without change. Modifications were then made to the model to get it to pass the rest of the tests. This work is still in progress.

COMPUTER INFERENCE AND LEARNING (Dixon):
The goal of this project is to develop a program to discover relationships in numerical data. It is hoped that this program will be of practical use to research workers in all fields for automatic analysis of experimental data. The basic idea is to improve the standard technique of linear multiple regression analysis by using products of predictors as if they were new predictors. One might call this technique non-linear regression analysis. Heuristics are necessary to reduce the very large number of possible combinations to a reasonable number. A basic program called ACID2 was written. ACID2 does a multiple regression analysis on a given data set by the stepwise method. All products of predictors are used. The models thus obtained are tested on a portion of the data not used in the regression analysis. This is called the acid test - hence the name of the program. A number of auxiliary programs have been written to collect, fabricate and manipulate data sets for this program. There are now 18 data sets on which to test this program. They include blood pressure data, tissue typing data, stock market averages, a computer simulated vehicle, a computer simulated radio, tables of random numbers, as well as others. The ACID2 program is in a primitive condition. No heuristics are used at present. It is hoped that the model will be able to learn its own heuristics. But preliminary results indicate that non-linear regression works better than linear regression.

AUTOMATIC PROGRAM-VERIFICATION (Slagle, Chang, and Lee):
The application of theorem proving techniques to program verification was considered. A new and direct technique for proving programs correct was obtained. Progress in this field may eventually enable us to test whether software is reliable or not.

AUTOMATIC PATTERN RECOGNITION (Slagle, Chang and Lee):
An algorithm for finding prototypes for a nearest neighbor classifier was developed. The algorithm was applied to 514 cases of liver disease, and only 34 prototypes were found necessary to achieve 100% of correct diagnosis.
In addition, many clustering algorithms were tested. The short spanning path algorithm was found to be simpler and comparable with the minimum spanning tree method.

PROFESSIONAL ACTIVITIES:
J. R. Slagle participated in the National Institute of Education Planning Conference on Productivity and Efficiency in Education in the United States. He gave a computer science seminar at IBM Research Center at Yorktown Heights, the State University of New York at Albany, and the University of Texas at Austin.
C. L. Chang gave seminars at the Rutgers University, IBM Research Center at Yorktown Heights, and IBM Research Laboratory at San Jose. He taught with Slagle a DCRT course on pattern recognition.
J. K. Dixon taught two DCRT courses on heuristic programming.
S. R. Heller gave a lecture at the NATO ASI on "Computer Representation and Manipulation of Chemical Information" in Holland, and a talk in the CODATA meeting in Freiburg, Germany.
L. Hodes gave a seminar at the University of Maryland.
G. D. Knott taught MLAB and graphics courses in DCRT. He has written 6 months worth of INTERFACE articles on various topics. Also, he taught the computer science component of the Computers in Clinical Medicine program sponsored by DCRT.
L. Norton participated in a government-wide briefing on DMB systems, giving a 1/2 hour talk on his indexing system.

PUBLICATIONS:
1. Chang, C. L., "Pattern Recognition by Piecewise Linear Discriminant Functions," IEEE Trans. Computer, Vol. C-22, No. 9, pp. 859-862, Sept. 1973.
2. Chang, C. L. "Finding Prototypes for Nearest Neighbor, Classifiers." To appear in IEEE Trans. on Computers.
3. Heller, S. R., "Computer Techniques for Interpreting Mass Spectra." To appear in NATO ASI Proceeding.

4. Heller, S. R., C. L. Chang, and K. Chu, "The Interpretation of Mass Spectrometry Data Using Cluster Analysis," Analytical Chemistry, May, 1974.
5. Heller, S. R., R. J. Feldmann, H. M. Fales, and G. W. A. Milne, "A Conversational MS Search System. IV. the Evolution of a System," J. Chem. Doc. 13, 130, 1973.
6. Heller, S. R., R. Katz, and A. E. Jacobson, "A MO Study of Norephrine and Dopamine," Mol. Pharm., 9, 486-496, (1973).
7. Heller, S. R., D. Koniver, H. M. Fales, and G. W. A. Milne, "A Conversational MS Search System. III. Display and Plotting of Spectra and Dissimilarity Comparison." In press.
8. Hodes, L. "Semiautomatic Optimization of External Beam Radiation Treatment Planning," Radiology, Vol. 110, No. 1, pp. 191-196, Jan. 1974.
9. Knott, G. D., "A numbering System for Combinations," CACM, Vol. 17, No. 1, pp. 45-46, Jan. 1974.
10. Knott, G. D., "Hashing Functions." To appear in the British Computer Journal.
11. Krichevsky and L. Norton, "Storage and Manipulation of Data by Computers for Determinative Bacteriology." To appear in Int. J. of Systematic Bacteriology.
12. Krichevsky, and L. Norton, "An On-Line Query Program for Interacting with Bacteriological Data," Proc. of First Int. Congress for Bacteriology.
13. Krichevsky, and L. Norton, "The World's Culture Collections as an Information System," Proc. of 2nd Int. Conference on Culture Collections.
14. Lee, R. C. T., and C. L. Chang, "Application of Minimum Spanning Trees to Information Storage," Proc. of Int. Symposium on Computers and Chinese Input/Output Systems, Aug. 1973.
15. Lee, R. C. T., C. L. Chang, and R. J. Waldinger, "An Improved Program-Synthesizing Algorithm and its Correctness," JCACM, Vol. 17, No. 4, pp. 211-217, April 1974.
16. Slagle, J. R., "Automatic Theorem Proving for Theories with Simplifiers, Commutativity, and Associativity." To appear in JACM.
17. Slagle, J. R., "Theorem Proving, Automated." To appear in Encyclopedia of Computer Science.
18. Slagle, J. R., C. L. Chang, and R. C. T. Lee, "Experiments with Some Cluster Analysis Algorithms." To appear in Pattern Recognition.
19. Slagle, J. R., J. Dixon, and T. Jones, "List Processing." To appear in Encyclopedia of Computer Science.
20. Slagle, J. R., and L. Norton, "Experiments with an Automatic Theorem Prover having Partial Ordering Inference Rules," CACM, 16, pp. 682-688, Nov. 1973.
21. Slagle, J. R., and L. Norton, "Automated Theorem-Proving for the Theories of Partial and Total Ordering," To appear in Computer Journal.

AI PROJECTS SPONSORED BY THE NATIONAL SCIENCE FOUNDATION

6

[Ed. Note: The following project summaries are reprinted from the "Summary of Awards" booklet published for fiscal year 1973 by the National Science Foundation's Division of Computer Research, formerly called the Office of Computing Activities.]

6a

PATTERN ANALYSIS AND RECOGNITION
Walter Freiberger
Div. of Applied Math
Brown University

6b

This project continues research in the area of pattern analysis and pattern recognition. Topics included are:

1) line pattern analysis,
2) statistical inference as pattern recognition,
3) statistical geometry,
4) stochastic feature selection,
5) stochastic elastic deformation models,
6) pattern analysis in modelling cerebral activities, and
7) analysis of biological growth patterns.

PROOF PROCEDURES IN PREDICATE CALCULUS AND TYPE THEORY
Peter B. Andrews, Donald W. Loveland
Dept. of Mathematics
Carnegie-Mellon U.

6c

Investigations in mathematical logic centering on proof procedures in predicate calculus and type theory are being continued. The primary goals of these investigations are the development of efficient proof procedures for computer implementation and the better understanding of existing procedures.

MECHANICAL PROBLEM SOLVING
R. B. Banerji, G. W. Ernst
Division of Engineering
Case Western Reserve U.

6d

This project deals with the automatic discovery of heuristics in problems and the automatic proof for straight-forward theorems involving many definitions. Specific activities include:

a) the development of an algorithm which will find invariant properties of states in a problem from the descriptions of problems,

b) the determination of additional general properties of solution techniques through the analysis of solutions to discrete problems occurring in operations research and artificial intelligence,

c) the development of a computer program which incorporates the results of the above,

d) the development of methods for breaking theorems into independent parts, and

e) the extension of the theorem-proving methods developed to higher-order logic.