

Report on the collaborative project between ICOT and the NIH

Richard J. Feldmann
Division of Computer Research and Technology
National Institutes of Health
Bethesda, Maryland 20892

The collaborative project between ICOT and the NIH must be considered at several levels. At the highest level the project is meant to provide a vehicle for developing friendship and understanding between Japanese and American scientific workers. By means of visits to each others laboratories, by almost daily fax and e-mail messages we have begun to understand each other's ways of thinking.

Two specific scientific projects were used as the scientific substrate for the project: Genetic Information Processing and Protein Folding.

The Genetic Information Processing work is very much influenced by the ICOT style of logic programming. This work is being done also in collaboration with workers at the Argonne National Laboratory (ANL) and the Lawrence Berkeley National Laboratory (LBNL). Over the last year and a half, four workshops have been held. The emergence of the InterNet means that workers can come together in one physical place to meet and talk but still use the computers and databases in their own laboratories. A very graphical interface and database program, called GenoGraphics, has developed from these workshops. GenoGraphics which is the work of Ross Overbeek and Ray Hagstrom from ANL, started with the data representation of George Michaels (NIH) on the *E. coli* genome and the work of Kaoru Yoshida (ICOT and LBNL) and Cassandra Smith (LBNL) on human chromosome 21. The *E. coli* data was collected by Ken Rudd who is now a member of the National Center for Biomedical Communication and Information (NCBI) in the National Library of Medicine (NLM) of the NIH. Dr. Michaels has just recently held such a workshop at the NIH. Workers from all over the USA and from England came together to increase the range of genomes which can be handled by GenoGraphics. During this workshop the genome for yeast pombe collected by the workers at the Imperial Cancer Research Fund (ICRF) laboratory in London, England was introduced into the GenoGraphics logic programming data format. The ICRF workers had spent almost half a year developing programs and organizing their data. During the week-long workshop they were successful in transferring their data to logic programming format.

Dr. Michaels expects that in the next year there will be fragments of about 50 genomes for various small organisms entered into the GenoGraphics format. GenoGraphics as a result of Dr. Hagstrom's work is now written in C and runs on any PC compatible machine. We expect that GenoGraphics will become a world standard tool for the representation, manipulation and investigation of genomes.

The collaboratory model which George Michaels has developed is a powerful outcome of our interaction with ICOT and the other US national laboratories. Scientific workers can now come together from all over the world and using the InterNet can work together effectively for a short period of time.

In the Protein Folding portion of our collaboration with ICOT we have built and analyzed several models for the representation of protein structure. The collection and analysis of x-ray crystallographic data sets was begun in our laboratory almost 20 years ago. The relationships between protein structure, function and folding pathway have been very difficult to elucidate. The protein folding problem is the key technology which will enable biological system design. During the collaborative project workers in both countries engaged in the design and construction of both physical and computer models. Physical models provide simple, visual, trans-cultural vehicles for communication. Computer models can be constructed to represent salient features of the physical models. Using both logic programming and conventional machines we have investigated the statistics and dynamics of these models. The resolution of a protein model and its water environment is a critical determinant of the computer power required to simulate folding. Parallel computational techniques for simulating protein folding using logic programming machines have been developed by our ICOT collaborators, Makoto Hirose, Masato Ishikawa and Masaki Hoshida. Hirose-san spent his whole winter vacation programming and running the folding algorithm. At the NIH, David Rawn (Towson State University) and I have made progress towards finding a topological principle which unites the water seeking (hydrophilic) and water avoiding (hydrophobic) aspects of protein structure. A complete and simple topological model would reduce the N^2 portion calculations to the number of amino acids in a given protein. With such a model we would hope to be able to fold proteins on many different types of computers.

Discussion with the ICOT workers has also focused on computer languages, style of operating environment and network connectivity. Using the PSI II and III machines loaned to the NIH under the auspices of this collaboration, it has been possible to evaluate the state of development of the hardware and software produced by the Fifth Generation Project. Any user who decided to accept a research machine must know that it will be a lot of work. The FGCS conference shows that at the end of the project, much more of the potential of the hardware and software is now usable. During the continuation year we at the NIH would expect to make much greater use of the capabilities of the PIM machines at ICOT. Discussions during the conference brought out

the problems which other collaborators have been experiencing in the early utilization of the ICOT hardware and software. The decision by MITI announced at the conference is a clear indication that the Japanese viewpoints of the utility of international collaboration are rapidly changing. The PIMOS operating system should be ported to world-standard machines so that scientists all over the world can begin to do program development in KL1.

During this collaborative project we have come to the opinion that even more than any single or parallel computer, the network is the most powerful artifact created by man. In this trans-global project we have experience the transition from paper letters to fax letters to fully electronic messages to interactive use of remote computers. In the beginning of the project the InterNet between Bethesda and Tokyo was rather slow and unreliable. Machines at either end had trouble talking to each other. As the project proceeded, the ability to communicate both electronically and intellectually rose to very high levels. The network gives us the ability to reason with each other about problems of mutual concern. The reasoning which we can do by sending messages is, however, rather limited. The InterNet is coming to the point where scientists can use databases and run processes on computers all over the world. New classes of tool for utilizing the network are being developed in many places in the world. We would hope to use these network tools to more strongly couple our collaborative research efforts.

We thank the administrators of ICOT for making such a strong and exciting collaboration possible. We expect to continue our collaborative work long after the formal end of the project.

The collaborative project between ICOT and the NIH

Richard J. Feldmann
Division of Computer Research
and Technology
National Institutes of Health
Bethesda, Maryland 20892

June 3, 1992

Our project started in year 9

October 1988	First contact
October 1989	First visit to NIH
July 1990	First visit to ICOT
October 1990	PSI II delivered
October 1991	PSI III delivered

Considered at several levels

Two scientific projects

Developing friendship and understanding
between Japanese and American
scientific workers.

Genetic Information Processing

Scientific projects

Protein Folding

General conclusions

Genetic Information Processing

Included other national laboratories

Argonne National Laboratory (ANL)

Lawrence Berkeley National Laboratory (LBNL)

National Library of Medicine (NLM)

Imperial Cancer Research Fund (ICRF)

Produced a new class of genetic
information manipulation tool

GenoGraphics

Prototyped in Prolog on Sun

Distributed in C on PC

Whole genomes of Bacterium, Yeast
and Human

Protein Folding

Built and analyzed several models
for the representation
of protein structure.

Used both logic programming and
conventional machines

Made progress toward understanding
protein folding

General Comments

Hardware is irrelevant

Software is everything

Mass action of programmers

Graphical User Interface

International Interface Languages

English and C

Balance between C and Logic Programming

Data structures Independent
of processing language

Still need highest level tools
to help thinking

Communications defines the 1990s

The network is the most powerful artifact
built by man

Collaboration will continue