Report on a Visit to ICOT

Toni Kazic
Institute for Biomedical Computing, Washington University, St. Louis MO USA

This report summarizes my activities and impressions during my recent visit to Japan at the invitation of ICOT researchers, December 14-28, 1994. I was an invited researcher at ICOT from December 17 to the 28th. The two weeks were mainly occupied by attendance at two meetings and extensive discussions with ICOT researchers in the area of computational biology. This is my third visit, and comes as the activities of the Follow-On Project are concluding. I therefore think it is appropriate to briefly discuss ICOT's impact on Japanese science, computational biology, and the global scientific enterprise. Particularly now, as the role of science in many societies is being considered afresh, it is important to determine what lessons the ICOT experiment can teach.

1 Meetings

1.1 International Symposium on Fifth Generation Computer Systems 1994, Tokyo

Although the Symposium on Fifth Generation Computer Systems (FGCS '94) was held before my stay as an invited researcher, I would like to discuss it briefly, as it effectively provided background for my work at ICOT and ably illustrates the quality of the environment and work there. Also, one motivation of my trip was to present our recent results on modeling biological function at a workshop on computational biology held during FGCS '94. That workshop proved extremely lively, and first-rate discussions occurred. The work performed at ICOT on protein representation, protein folding, and multiple sequence alignment has moved beyond repetition of existing work for self-instruction, through adaptation of existing algorithms and ideas, and is now poised to make unique contributions to international science. The effort on parallel, nested relational database systems has demonstrated the utility of more sophisticated data models, and is likely to serve as an important reference point as database size and complexity continue to increase through genome sequencing efforts. It is important to realize that many biological problems will not be accurately or expressively represented by hierarchical models. In particular, any system which requires an at least context-dependent language (in the formal Chomskyian sense) cannot conform to an inherently context-free language such as those represented by hierarchical models. In this sense, I believe that the next generation of databases will draw from more complex languages.

The workshops were preceded by plenary sessions at which the major results from the Follow-On Project's work were summarized. I will confine myself to the artificial intelligence applications, as I am only slightly acquainted with the details of many issues in the design and methodology of parallel computing. However, the work on parallel methods of multiple sequence alignment deserves mention. The use of parallel methods in multiple sequence alignment has been obvious to many people, but surprisingly has not been well done in a generally available manner. ICOT's implementation of several parallel alignment algorithms and the alignment workbench, and their porting to clusters of UNIX workstations, is likely to have a significant impact on practical genome analysis. The most notable results in artificial intelligence have been the development and porting of Quixote and Helios, and their application to biological and legal problems. The case that the representational facilities of Quixote are *required* (as opposed to advantageous) for certain kinds of knowledge has never been convincingly made.

For example, the biological models treated so far have been at relatively low levels of resolution, so that the full capacities of Quixote are not required. A side-by-side comparison of the same problem, represented at equal resolution, in Quixote and another language such as Prolog, would go a long way towards illustrating Quixote's expressiveness and in making a solid case for its computational efficiency. That said, however, its inclusion of constraints and subsumption make Quixote an attractive system for certain kinds of problems, particularly those involving complex and varied inference mechanisms. Quixote is particularly significant because it represents an early and well-conceived attempt to build a consistent parallel hybrid system incorporating a superset of first-order logic and object-oriented inferencing relations. The lessons of combination were generalized to Helios, a more extensive system which employs modular agents to reconcile different inferencing modes. The general notion of agents cooperating to solve problems is not unique to ICOT, but needs careful exploration by the world community. As a testbed, Helios offers a good opportunity to explore the issues of expressiveness, efficiency, encapsulation, and problem-solving methodologies. The preliminary results on legal reasoning are encouraging; and it strikes me that Helios might be a good model to work out many issues in database federation. The transaction model, though naturally suggested by the agent metaphor and parallel processing, may not be the most appropriate model for complex reasoning tasks as it is too discrete.

1.2 Genome Informatics Workshop 1994, Yokohama

I also attended the Genome Informatics Workshop, which provided a good overview of the current state of Japanese computational biology. In general, Japanese computational biology is rapidly coming of age. Though many projects suffer from imitation and an insufficient attention to the current literature, first-rate work is being done. The work of Akutsu *et al.* (which includes several ICOT people) on alternative representations for protein secondary structures and hashing schemes is significant. Several provocative applications of machine learning were presented, including two papers on tree grammars (Mamitsuke and Abe, and Kobayashi and Yokomori), and a new learning algorithm for protein motifs which apparently correctly infers generalized motifs from only positive examples (Arimura *et al.*). ICOT projects showed well in this venue, and testify to the progress of ICOT researchers in this area.

2 Discussions

During the unprogrammed part of my visit I enjoyed several stimulating discussions with ICOT researchers. The topics covered ranged from our work on representation of biological function and structure (*Klotho* and *Atropos*), to the prediction of biochemical reactivity, searching databases of three-dimensional structures, and the prediction of protein tertiary structure. Hirosawa-san and I discussed recent work in the representation of biochemical reactions; I found the discussion clarified several points I have been thinking about somewhat muddle-headedly. I have found the work on vectorial descriptions of protein structural elements very stimulating, and discussed with Tanaka-san several applications of the underlying ideas. I believe continuing this dialogue will prove very important for both of us in the coming years. Asai, Ishikawa, Tanaka, Wise, and I also considered several parts of the protein structure prediction problem, both in the workshop and later at ICOT. We agree on the importance of attempting to predict tertiary structure *de novo*, the need for close contact with experimentalists, and the possible significance of long-range interactions (though how long-range, and whether these can be mediated by layers of ordered water molecules, none of us now has any clear idea). The dominant theme which emerges is the need to find appropriate representations which permit a more clear-headed attack of the central problems in these areas.

3 General Comments

In considering what lessons ICOT has taught us, I believe these fall into three general categories.

The first is the technical results. I have already discussed the technical results of the computational biology section in section 1.1 above, so I will not repeat them here. ICOT experimented with a variety of parallel architectures and achieved some notable successes in parallel processing. The idea of linking hardware and software development remains a novel idea, not widely adopted by industry. In time, however, I suspect we will see more examples of this type of coupling. To some extent, the availability of dedicated CPUs for specialized applications is one indication of this trend. ICOT's emphasis on the importance of efficient execution of complex inferences remains a watershed for the computational community. The major result here was a deepening understanding of what it means to "parallelize" inference, and how difficult this is. In retrospect this should not be surprising: we still understand very little about how humans simultaneously carry out multiple chains of reasoning, or otherwise sharply limit the search space of available proof trees, thereby achieving at least the impression of parallelism. Though some results appear negative now, I believe in time they will serve as an important springboard to a clearer understanding of parallel inference — and the success of MGTP, and the promise of Helios, show that the required new paradigms can be developed. I suspect future work in this area will need to take more cognitive science into account. In this regard, I find Nitta-san's focus on reasoning in criminal law to be courageous and ultimately highly pragmatic: this is an excellent model system in which to work out the answers to many of the most important and challenging problems.

The second is the development of human capital. ICOT has been a successful garden of human beings. Projects, software, and people are dispersing across Japan and beginning a significant fertilization process for academia and industry. For example, it is very apparent from the Symposium that ICOT has spawned a generation of Japanese computational biologists who can do first-class work on internationally significant problems. They bring a much-needed engineering perspective to questions in computational biology, and are strongly focused on fundamental research in contrast to infrastructure development. ICOT researchers are now asking biologically significant questions, and originating new approaches. Their development is perhaps best illustrated by the quality of the workshop they organized: there was a high-level discussion of the importance and pitfalls of hidden Markov models, open questions in protein folding, and the complexities of representing biological function. Their grasp of the biological literature continues to improve, and they are often much better read in the computational biology literature than others of their Japanese colleagues. This is not a trivial achievement, since there are multiple language barriers they must cross. These scientists (Asai, Hirosawa, Ishikawa, Onizuka, and H. Tanaka) represent an important pool of talent which deserves careful nurturing. Extended stays abroad, close contact with leading experimentalists, and a solid network of collaborations would accelerate their development. The Internet and the planned virtual laboratory will facilitate their progress, but direct, sustained contact will still be required. They have become full participants in the world computational biology community.

The third is the development of new scientific approaches. ICOT itself represented a significant risk on the part of MITI and the Japanese government. Clear enunciation of lofty goals invites criticism and doom-saying, and inevitably even the most successful projects fall short in some way. In response to this challenge, ICOT developed a unique research style — more open and relaxed than is common in many Japanese laboratories (and the outreach to foreign scientists has from the beginning been uniquely successful), and yet with an emphasis on teamwork and personnel development which is unusual in the West. Some of the results are easy to see: the professional and psychological mobility of ICOT researchers as they return to the wider world; the clear understanding of the many varieties of collaboration; and the ease of exchanging ideas. Most important, however, is how this environment is fostering the devel-

opment of indigenous innovation. Japanese people are often criticized in the West for being at best adaptors; and indeed, this is taught to Japanese children both as self-explanation and as a positive character trait. Perhaps in consequence of this, and the influence of mercantilism, a tradition of pure research is largely absent from Japanese science (with of course some notable exceptions like Tomanaga). But a mercantile outlook can also fuel indigenous innovation in science, as it previously did in the arts during the Tokugawa shogunate. By both adapting and innovating, ICOT researchers are inventing a scientific style unique to Japan and possibly the world: they may be the generation which goes beyond stereotypes to a style which combines the strengths of both approaches. It is far too early to say what impact they will have had in fifty years; but the potential to research difficult questions, such as climate modeling, urban planning, and protein folding, in ways which produce both intellectually substantive and practically useful results has clearly been developed. In the short term, this augurs well for science as the social contract between it and society is renegotiated. Over the long term, such indigenous innovation can help produce the artifacts of a world in which everyone can feel equally comfortable.

4 Thanks

I wish to express my thanks to everyone who helped make my stay possible and enjoyable, both intellectually and personally. These include H. Tanaka, M. Ishikawa, and K. Susaki; N. Namikoshi, K. Karakawa, and M. Kitakata; and K. Nitta and Chikiyama. ICOT alumnae were also significant facilitators as well as colleagues: K. Asai, M. Hirosawa, and K. Onizuka. My congratulations to all of these, and to S. Uchida, on ICOT.

Toni Kazic

Institute for Biomedical Computing
Washington University School of Medicine
Box 8036, 700 South Euclid Avenue
St. Louis, MO 63110

tel: 314-362-3121
fax: 314-362-0234
email: toni@athe.wustl.edu

Research Experience

1992–present *Instructor, Institute for Biomedical Computing, Washington University School of Medicine, St. Louis, Missouri.* Representation of knowledge on biological function for databases to support automated inference in organismal modification and design. Analysis of metabolic organization and function, including development of new approaches to pattern recognition for this area.

1989–1992 *Instructor, Department of Genetics, Washington University School of Medicine, St. Louis, Missouri.* Design of a prototypic computational model of *E. Coli physiology, and implementation of a prototype database of biochemical reactions for this organism. Research on representation of biological information to support automated inference and modeling in biochemistry and physiology. Studies of large-scale structure of the E. Coli chromosome.*

1991 Consultant, Division of Computer Research and Technology, National Institutes of Health, Bethesda, Maryland. *Applications of logic programming to the analysis of genomic data in E. Coli, human chromosome 21 and* Drosophila melanogaster.

1990 Visiting scientist, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois. *Began education in computer science, mathematical logic, and computational linguistics; developed preliminary design of a computational approach to bacterial physiology.*

1986–1989 Research associate, Department of Microbiology and Immunology (now Molecular Microbiology), Washington University School of Medicine, St. Louis, Missouri. *Analyzed deletion formation in E. Coli, and determined that deletion frequency varies with chromosomal location as well as palindrome length and sequence. Devised a method for determining the growth rate of hundreds of cultures simultaneously. Determined that translation of the tetracycline-resistance protein is essential for supercoiling in pBR322, and suggested that protein intercalation into the cellular membrane provides an essential anchor for transcription-driven supercoiling.*

Teaching Experience

1982 Instructor, Department of Biology, University of Pennsylvania. *Introductory microbiology for biology majors.* Approximately 30 students.

1980 Instructor, Department of Biology, University of Pennsylvania. *Introductory biology for nonmajors.* Approximately 40 students.

Organized and presented all lectures for both courses, and supervised three laboratory staff in the microbiology course.

Education

1984–1986 Postdoctoral fellow, Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania. Identified, characterized, and sequenced a hyper-recombinant allele of the *E. Coli uracil-DNA glycosylase* (ung) *gene. Constructed and characterized other* ung *alleles. Bruce K. Duncan, advisor.*

1975–1984 *Ph.D., genetics. University of Pennsylvania, Philadelphia, Pennsylvania. Graduate student. Cloned, mapped, and sequenced the promoter and regulatory regions of the purine biosynthetic locus* purHD *of E. Coli. Studied regulation of the locus using* purHD-lacZ *and* purHD-galK *fusions. Joseph S. Gots, thesis advisor.*

1971–1975 *B.S., microbiology. University of Illinois, Urbana, Illinois.*