Report on April 1991 Visit to ICOT

Ann M. Barber, M.D.

Laboratory of Mathematical Biology, NCI
National Institutes of Health, Bg 10, Rm 4B56
Bethesda, MD 20892 USA

I. Introduction

We first met last year when Dr. Shunichi Uchida, Dr. Katsumi Nitta and Dr. Takashi Chikayama demonstrated the PSI machine at the US National Institutes of Health. Afterwards you invited me to visit ICOT for two weeks starting April 15, 1991, because as you said, you wanted to learn more about genetic information processing. In turn, I welcomed this opportunity to learn more about logic programming.

My particular area of interest is multiple sequence alignment. As genetic data accumulate, there is more and more need for aligning multiple sequences. Multiple sequence alignment is used to analyze macromolecular structures and functions. In addition, it is important for aligning DNA fragments from a DNA sequencer. The latter is a major task required by the Human Genome Project. So far, no algorithm is able to produce a complete answer for the multiple sequence alignment problem. My hope in coming to ICOT was to address this important problem on your fast Fifth Generation Computer Systems.

II. Summary of the visit

A. Two week Schedule.

Masato Ishikawa arranged the schedule of my visit to ICOT. This included multiple discussions with the Seventh, First and Third Research Laboratories. In addition, it included the following:

MApr 15 Multi-PSI hardware - Kentaro Onizuka
T Apr 16 VTR English translation prepared with Masato Ishikawa
to describe the PIM, parallel inference machine
WApr 17 Bunkyo-ku, Tokyo Metropolitan Institute of Medical Science,
RINSHO-KEN, Yoko Nadaoka
T Apr 18 Kappa/QUIXOTE - Hidetoshi Tanaka
F Apr 19 Kyoto University, Institute for Chemical Research -
Prof. Minoru Kanehisa and Kenta Nakai
MApr 22 Kanagawa, NEC Corporation - Akihiko Konagaya
T Apr 23 GIP Working Group - Dr. Katsumi Nitta
WApr 24 Tsukuba, RIKEN Gene Bank Life Science Center -

Dr. Akinori Sarai, Dr. Miyajima and Dr. Murakami
T Apr 25 Logic programming - Nobuyuki Ichiyoshi
F Apr 26 KL1, PIMOS - Dr. Kazuo Taki and Dr. Koichi Furukawa

B. Presentations.

Over the course of two weeks, I presented three talks about my current research work:

1. Interactive Graphics Software for Analyzing Multiple Macromolecular Sequences. I demonstrated various aspects of the SequenceEditingAligner software that I wrote in the C programming language for Silicon Graphics workstations. This software performs interactive display, editing, alignment, and analysis for over 300 sequences simultaneously. Presentations were given last week for GIP and Yoko Nadaoka at RINSHO-KEN, Tokyo Institute of Medical Science in Bunkyo-ku. An additional presentation was given this week for Dr. Akinori Sarai's laboratory at RIKEN Life Science Center in Tsukuba.

2. Biologists Query Sequence Analysis. To convey the needs of biologists for computer analysis, I detailed the biological questions that I have addressed while working at NIH. These issues have included DNA analysis for promoters and enhancers, protein homology model building of HIV-1 reverse transcriptase, and multiple sequence alignment of CAP DNA binding sites. I gave this talk at Kyoto University for Prof. Minoru Kanehisa's laboratory.

3. CAP Binding Sites and Results of Multiple Sequence Alignments I discussed some results from my multiple sequence alignments. For instance, (1) Sequence analysis of CAP binding sites revealed two structural classes; this result was supported by my subsequent gel electrophoresis mobility shift assays. (2) Multiple sequence alignment of retroviral reverse transcriptases produced a model of this important enzyme in HIV-1. And (3) Analysis of p53 proteins revealed functional domains. Furthermore, for this talk I also analyzed the steps needed to develop multiple sequence alignment algorithms. Alignments use a metric distance, a scoring function, an alignment algorithm, and an evaluation of statistical significance. My discussion also included ways to evaluate the quality of the alignment programs. I presented this seminar for ICOT's Genetic Information Processing Working Group.

C. Discussions with ICOT members

My many meetings and discussions with the members of ICOT were enjoyable and fruitful. Some of these are detailed here by Laboratory.

C1. Seventh Research Laboratory - Algorithms for genomic analysis

With the Seventh Research Laboratory I have had extensive discussions about algorithms for genetic and structural analysis. The GIP group has made good progress since last year; the sequence alignment algorithms are state-of-the-art. Masato Ishikawa, Masaki Hoshida, and Makoto Hirosawa adapted the Needleman-Wunsch dynamic programming algorithm for three-dimensions. Since this algorithm is one of the best for two sequence alignment, it makes sense to adapt this to three-dimensional analysis. Unfortunately, time and space requirements make it difficult to use this algorithm for four- or more- dimensional analysis. Therefore, they cluster and merge multiple three-dimensional alignments by aligning similar sequences of different alignments. Finally, recognizing that the alignments are not optimal they test different gaps to maximize alignment scores. Together we explored additional algorithms, their advantages, disadvantages, and time requirements. More about this later.

In addition to algorithms for multiple sequence alignment, members of the Seventh Laboratory are developing algorithms for macromolecular structural analysis. Makoto Hirowawa and Masato Ishikawa described Feldman's protein folding simulation. Then, we discussed how research efforts in this area will reveal parameters important for protein structural predictions. We also considered ways to evaluate the results and how to modify the algorithm if results fall short of expectations.

Additional algorithms are being explored by the Seventh Laboratory. For instance, I was pleased to talk with Kentaro Onizuka about his work applying fractals to genome analysis. By using protein crystal coordinates he calculated the number of spheres needed to cover a structure. Then, he relates this number to the resulting fractal dimension and coorelates this dimension with the structural conformation. For instance, he found that a helix-turn-helix structure had fractal dimension of about 3.0, its amino terminus had fractal dimension about 1.0, and a beta-turn-beta structure had dimension about 2.0. In addition, Kentaro Onizuka is applying wavelet transformations to genome sequences and in this way aims to better understand hydrophobic sequence alignments. As wavelet transformations have been studied in the past and have been used for other applications, it makes sense to apply them now to genome analysis.

Furthermore, Dr. Katsumi Nitta and I discussed knowledge processing for genetic applications. I believe the quality of knowledge processing depends greatly on the quality of its rules. Although I feel that the first priority should go to multiple sequence alignment, I realize that when firm analytical rules start accummulating, then knowledge processing will be an important tool.

C2. Third Research Laboratory - Object-oriented relational database

Besides algorithms for genomic analysis with the Seventh Lab, I also enjoyed talking with Hidetoshi Tanaka of the Third Research Laboratory about QUIXOTE, Kappa, genetic databases, and user interfaces for these. I was pleased by the detail of the genetic data items. This will make genetic analysis considerably easier than seen in currently available genetic databases. Furtheremore, the user interface seems detailed but also intuitive. This rare combination of features will allow biologists to use your genetic databases with pleasure.

C3. First Research Laboratory - logic programming, parallel reasoning

Finally, I appreciated the chance to discuss KL1, PIMOS, logic programming, and parallel reasoning with various members of the First Research Laboratory. Dr. Kazuo Taki and Inamura of the First Lab and earlier Nobuyuki Ichiyoshi of the Seventh Lab explored with me the strengths and weaknesses of KL1 and logic programming. In addition, they helped me understand how best to approach the Multiple Sequence Alignment problem using the Parallel Inference Machine. In addition, I enjoyed my interactions with Dr. Koichi Furukawa. After I presented an Exhaustive Search Algorithm to solve the Multiple Sequence Alignment Problem, he cleverly and quickly implemented this algorithm with a 14 statement Prolog program. More about this later.

III. Research

A. Goals

My purpose in coming to ICOT was to develop better methods for Multiple Sequence Alignment by using ICOT's Parallel Inference Machine. As discussed above, this problem has major significance for both the Human Genome Project and basic scientific research. The first goal is to produce alignments which cannot be improved by either person or machine. The second goal is to do this efficiently.

For now, it is important to separate the tasks of scoring and aligning. The scoring function that assesses similarily can have different forms. A general form might be:

$$score(loc) = (match\ weight)(\#matches) - (mismatch\ weight)(\#mismatches) - (gap\ penalty)(\#gaps).$$

In this way, similarity can be defined, for instance, as identity, mutation conservation, hydrophobicity, or a certain number of mismatches with any of these matrices. Nevertheless, no matter how similarity is defined, the aligning process should proceed by a single complete and efficient algorithm.

B. Results

As for aligning algorithms (independent of similarity functions) reasonable approaches include the following:

1. Needleman-Wunsch Dynamic Programming Algorithm
   Certainly, a multi-dimensional Needleman-Wunsch algorithm would solve the problem. Unfortunately, time increases to the order of $O(L \wedge n)$, where $L = sequence\ lengths$ and $n = \#sequences$. Although we discounted this approach because of this exponential time constraint, we should remember that what a 370KLIPS PSI-II can do in 3 min, a 20MLIPS PIM/p does in three seconds. In addition, a dynamic programming algorithm that yields a complete alignment of four sequences is a significant contribution to current biological research.

2. Pinning with an Exhaustive Search Algorithm
   Again, to address the first goal of a finished product I suggested Pinning an alignment by motifs shared by all sequences. This Exhaustive Search Algorithm lends itself easily to parallel programming. In addition, due to the motif philosophy of protein development such an approach yields meaningful biological results. So, I was pleased when Dr. Koichi Furukawa implemented this plan by writing a 14 Prolog statement program.

3. Pinning with a Layered Stream bottom up Search Algorithm
   Finally, we would also like to address the second goal, that is, not only an algorithm that works but also an efficient one. In order to decrease the number of calculations,

Dr. Kazuo Taki suggested a Layered Stream Search Algorithm. This approach starts with individual regions of similarity and builds these up into larger and larger regions of similarity until the whole length of the sequences are aligned. This is a sound approach and I look forward to seeing this implemented.

C. Future

1. Aligning algorithms
   Algorithm development will continue as discussed above in III.B The Seventh Research Laboratory is working on the Exhaustive Search Pinning Algorithm. In addition, the First Research Laboratory is implementing the Layered Stream Search Pinning Algorithm. Dr. Kazuo Taki and I hope to publish the results of the Layered Stream approach.

2. Efficiency
   Since increasing the number of sequences rapidly becomes intolerable in multiple sequence alignments, we might also consider ways to limit search time. In particular, faster run times can result from (1) limiting the number of sequences, (2) pruning search space by selecting regions for alignment, and (3) approximating alignments by restricting the similarity functions.

   With KL1 and parallel processing the Parallel Inference Machine is able to do multiple calculations at the same time. This could reduce the constant factor for $O(L \wedge n)$. Thus, if limits are made on the number of sequences, then we may be able to find a complete alignment in a reasonable amount of time.

   A different way to handle time constraints is to select short subsequences and then to run an exhaustive search. One way to implement this approach is to ask the investigator interactively to select a region of interest for aligning. Thus, by pruning the search space, run time is decreased. Of course, additional speedup could be found with lower level language implementations or with hardware modifications.

   Finally, another way to decrease run time is to seek approximate solutions without exponential growth. Here too, restricting the similarity function by interactive techniques may be effective.

3. Knowledge Processing
   Makoto Hirosawa suggested using a motif knowledge search space. I agree with him that a motif dictionary is valuable and encouraged him to consider the Prosite database.

IV. Conclusions

1. Much progress has been made since last year. I applaud the interaction between ICOT and the biologists in the GIP Working Group.

2. Multiple Sequence Alignments have important implications in biological research because they help us understand unknown functions and structures of macromolecules. In addition, aligning multiple sequences is a well-defined problem whose answer is essentailly limited by speed. Combined with knowledge processing, this becomes a powerful tool.

3. Unfortunately, our work this month in Multiple Sequence Alignment reveals an important issue in the development of Fifth Generation Computer Systems. I wonder whether a basic software system been developed for FGCS. In addition, when do you think there will be a "programming environment that reduces the cognitive demands on human beings?" I notice that very few people, if any, feel comfortable writing programs in this Kernel Language 1 and most ICOT members describe KL1 as difficult. I wish there were a programming environment in which I could easily write alignment programs.

4. I enjoyed working with ICOT and the Genetic Information Processing staff. I appreciate our conversations about Multiple Sequence Alignment, I value our discussions about Logic Programming, and I treasure the way you shared Japan with me.

V. Acknowledgments

Special appreciation goes to the Genetic Information Processing members, including Katsumi Nitta, Masato Ishikawa, Masaki Hoshida, Makato Hirosawa, Kentaro Onizuka, Tomoyuki Toya, Hidetoshi Tanaka, and Kazumasa Yokota, for good discussions and fun times. I send thanks to the management and directors of ICOT, including Executive Director Hiroichi Hiroshige, Chief of the Seventh Laboratory Dr. Katsumi Nitta, Director of the Research Center Dr. Kazuhiro Fuchi, Deputy Director Dr. Koichi Furukawa and Manager of Research Dr. Shunichi Uchida for supporting my stay at ICOT. Also, I am grateful to Kazuhide Iwata, Emiko Higuchi, and Kumiko Karakawa for assistance and kindnesses during my stay in Japan. Thank you all for making my visit to Japan both fun and productive.

Ann M. Barber, M.D.

Laboratory of Mathematical Biology
NCI, NIH, Bg 10, Rm 4B56
Bethesda, MD 20892
301 496-4781 or -2495

Laboratory of Mathematical Biology
NCI, Bg 469, Rm 151, P.O. Box B
Frederick, MD 21702-1201
301 846-5576 or -5532

9101 Clewerwall Drive
Bethesda, MD 20817
301 469-8340

EDUCATION

Stanford Univ, Stanford, CA 94305	B.S. 1974 Mathematics
Stanford Univ, Stanford, CA 94305	M.S. 1974 Mathematics
Northwestern Univ, Chicago, IL 60611	M.D. 1981 Medicine

EMPLOYMENT

National Institutes of Health, DCRT, Bethesda, MD
Mathematician, 1974-79

Massachusetts General Hospital, Lab of Computer Science, Boston, MA
Program Analyst Engineer II, 1976-77

Northwestern University Medical Center, Dept of Medicine, Chicago, IL
Internal Medicine Resident Physician, 1981-84

Columbus Hospital, Chicago, IL
Emergency Room Physician, 1983-84

National Institutes of Health, NCI, Medical Branch, Bethesda, MD
Medical Staff Fellow, 1984-1987

National Institutes of Health, NCI, Laboratory of Mathematical Biology
Senior Staff Fellow, 1987-present