# REPORT ON VISIT TO ICOT

David Haussler
Department of Computer and Information Sciences
University of California
Santa Cruz, California 95064 USA
haussler@cis.ucsc.edu

October 1990

I visited ICOT from Thursday October 11 until Friday October 19, 1990. On October 16, I gave a prepared talk entitled "Probably Approximately Correct Learning" at ICOT. During my stay in Japan I also gave the prepared talk "Learnability and the Metric Dimension: Decision Theoretic Generalizations of the PAC Learning Model" at the ALT'90 conference (October 9), and at IIAS-Fujitsu (October 18). I also a few short impromptu lectures at ICOT, including one describing my work in the analysis and classification of genetic sequences to Mr. Ishizaka and Dr. Nitta (October 11), and one about PAC learning on October 19.

In addition to giving lectures, I also read papers and saw demonstrations of some of the work that has been done at ICOT. In particular, Mr. Fujita demonstrated the automatic synthesis of a sorting program from the axiomatic specification of the sorting problem, Mr. Iwayama demonstrated a

parallel implementation of a truth maintenance system, and I also heard some of the ICOT presentations for the Second Joint ICOT/DTI-SERC Workshop on decomposition of parallel applications, and benchmarking and evaluations of parallel systems.

Most important, however, were the discussions I had with the researchers at ICOT during my stay here, and at IIAS-Fujitsu on October 18 and ALT'90 on October 8, 9 and 10. I will describe the discussions at ICOT in more detail. These discussions were on three main topics:

1. Application of machine learning techniques in genome analysis
2. Application of machine learning techniques in non-monotonic reasoning
3. Induction of propositional PROLOG programs from examples and membership queries

I had discussions on the application of machine learning techniques in genome analysis on October 11 with Mr. Ishizaka and Dr. Nitta, on October 15 with Dr. Abe and Dr. Yamanishi of NEC, and on October 16 with Drs. Abe, Yamanishi and Konagaya of NEC. On October 12, I heard a presentation on the programs for multiple sequence alignment that have been developed by Dr. Nitta, Mr. Ishikawa, Mr. Hoshida, Mr. Hirosawa, and Mr. Toya at ICOT. The first program uses a dynamic programming technique, and can only practically be used to align two or three sequences at a time. The second program can do multiple alignment of many sequences at once, using a simulated annealing approach. In this approach, an "energy" function is defined that gives a numerical value for each possible alignment, with better alignments having lower energies. The simulated annealing algorithm then searches for the alignment with minimum energy. The most critical aspect of this method is the definition of the energy function. If the energy function is not defined correctly, then the resulting alignment will not be the alignment preferred by biologists.

In my discussions with Drs. Abe, Yamanishi and Konagaya, we developed an energy function based on the MDL (Minimum Description Length) method of machine learning that I feel will perform better than the energy function currently being used in the simulated annealing-based sequence alignment program at ICOT. This new energy function incorporates the background knowledge of protein mutation frequencies given by Dayhoff's odds matrix, as does the energy function currently being used, but it uses this knowledge in a more consistent manner, so that the overall energy of the alignment has a simple interpretation as the negative logarithm of the probability of the alignment, given a certain stochastic model, plus a term representing the complexity of that model. This stochastic model is in fact a hidden Markov model, of the type that has been used successfully in other problems in genome analysis, and in related problems in speech recognition. I hope that experiments can be conducted using this new energy function, and that some of the related ideas we discussed can also be explored.

I discussed the application of machine learning techniques in non-monotoni reasoning with Mr. Satoh on October 17. Here the problem is to determine preferences among logical formulae based on empirical data. These preferences are then used in a system for non-monotonic reasoning developed by Mr. Satoh. I suggested that a preference ordering on a set $F$ of propositional formulae could be defined by saying that one formula $f$ is preferred over another formula $g$ if a "typical" randomly drawn truth assignment is more likely to satisfy $f$ than $g$. Such "typical" truth assignments could be derived directly from observations of randomly chosen objects or situations in the domain of interest. Given a large number of "typical" truth assignments drawn at random in this fashion, with high probability we can obtain a good approximation to this preference ordering on $F$, so long as $F$ is not too large. This is done as follows: From the random truth assignments, an empirical "score" is calculated for each formula in the set $F$. This score is simply the

fraction of truth assignments that satisfy the formula. Formula $f$ is preferred to formula $g$ if $f$ has a higher score than $g$. A theorem in probability theory known as Hoeffding's lemma gives bounds on the number of random truth assignments needed so that, with high probability, the preference ordering obtained in this manner is a good approximation to the desired preference ordering.

Finally, I discussed the problem of learning propositional PROLOG programs from random examples and membership queries with Mr. Ishizaka. This problem has been solved recently by Angluin, Frazier and Pitt. We discussed the technique that they used, and compared it to the technique that Mr. Ishizaka used in his algorithm to learning simple deterministic languages. We decided that it may be fruitful to look at generalizations of these techniques in the future.

I was impressed by the depth and scope of the work that has been done at ICOT. In particular, Mr. Fujita's work on automatic program synthesis is outstanding. My general impression (based on my somewhat limited exposure to the work at ICOT) is that the most significant progress has been made on relatively clean and well-defined problems of this type, which are most natural for the logic programming approach. However, to achieve comparable results for other types of problems such as speech and image recognition, natural language processing, and decision support, more attention should be focused on numerical and statistical approaches. It would be ideal if a paradigm could be found that incorporates the best features of logic-based knowledge representation and the numeric or "sub-symbolic" methods of representation found in neural networks and other statistical paradigms. While I can't predict what such a paradigm might look like at this time, I will predict that the winning approach to knowledge representation will be one in which knowledge can be obtained and refined by the process of learning from examples. A fully usable knowledge base will simply be too complex and

to dynamic to ever be built and maintained entirely by direct programming. Hence I would encourage further investigation of the potential of machine learning, and of the proper theoretical foundation for machine learning. In my conversations with researchers at ICOT, I felt a growing interest in these topics, and so I look forward to many future collaborations in these areas.

Let me close by thanking Dr. Fuchi, Dr. Furukawa, Dr. Hasegawa, Dr. Iwata, Mr. Ishizaka, and all the other members of the group at ICOT for their outstanding hospitality. My visit to ICOT has been a wonderful experience for me and for my wife.

## Curriculum Vitae for David Haussler

### Personal Data

**Home Address: 402 Nobel Dr., Santa Cruz, CA 95060**
**Tel: (408) 429-9472**
**Work: Associate Professor of Computer Science**
**Department of Computer and Information Science**
**University of California, Santa Cruz, CA 95064**
**Tel: (408) 459-2105**
**Email: haussler@saturn.ucsc.edu**

### Education

Connecticut College, New London, CN,
B.A. in Mathematics, 1975, Magna Cum Laude, Phi Beta Kappa.
Julia Bower Mathematics Award, 6/75.
California Polytechnic State University at San Luis Obispo,
M.S. in Applied Mathematics, 1979.
Mathematics Award, 6/79. University of Colorado at Boulder,
Ph.D. in Computer Science, 1982.
University of Colorado Doctoral Fellowship, '79, '80, '81.
University of Colorado Graduate Student Research Award, 6/82.

### Professional Appointments

Assistant Professor, Department of Mathematics and Computer Science, University of Denver, Denver, CO, 9/82 to 6/86. Assistant Professor, Department of Computer and Information Sciences, University of California, Santa Cruz, CA, 7/86 to 6/89. Associate Professor, Department of Computer and Information Sciences, University of California, Santa Cruz, CA, 7/89 to present.

### Consultancies and Visiting Positions

INTERACTIVE Systems Corp., Santa Monica, CA, 6/81-10/81. Seville Technology, Boulder, CO, 4/84-10/84. Universite de Haute Normandie, Rouen, France, 4/86-5/86. Xerox Inc., Xerox Park, Palo Alto, CA, 1/87-1/89. Mathematical Sciences Research Institute, Berkeley, CA 9/91-10/91.