The First International Conference on Intelligent Systems for Molecular Biology

David Searls, Jude Shavlik, and Lawrence Hunter

Immediately preceding the 1993 National Conference on Artificial Intelligence (NCAI) in Washington D.C., a new conference series on the application of AI to molecular biology was inaugurated in neighboring Bethesda, Maryland. The First International Conference on Intelligent Systems for Molecular Biology (ISMB-93), held 6–9 July 1993 at the Lister Hill Center of the National Library of Medicine (NLM), attracted over 200 computer scientists and biologists from 13 countries.

As organizers of the conference, we saw it as the culmination of a series of successful meetings and colloquia, including workshops by the American Association for Artificial Intelligence, that, taken as a whole, indicate that molecular biology is one of the most rapidly growing application areas of AI and warrants a dedicated conference. AAAI was a co-sponsor of the meeting and published the proceedings (AAAI Press, Menlo Park CA, ISBN 0-929280-47-4, $45). Extensive additional support in the form of grants was provided by the National Institutes of Health (NIH), primarily through NLM but also through the Division of Computer Research and Technology, and by the Department of Energy Office of Health and Environmental Research (which, like NIH, is heavily involved in the Human Genome Project). Further support was provided by the Biomatrix Society, a group that has a predilection for AI approaches to biological data.

The conference was preceded by a series of well-attended tutorials, including introductions to AI for biologists and to molecular biology for computer scientists. Other tutorials covered genetic algorithms, neural networks, and linguistic methods for sequence analysis. Invited speakers, who opened each day of the conference, were Temple Smith (Boston University), speaking on the classification of protein structure cores; Leroy Hood (University of Washington), speaking on sequencing technology and recent results in immune system gene regions; and Harold Morowitz (George Mason University), speaking on a new theory of the evolutionary development of intermediary metabolism. A lively poster session and reception concluded the first day of the technical program. A total of 27 talks and 25 posters were accepted after rigorous review and are represented by 9-page papers in the proceedings (designed more on the computer science model than the biological, where conference proceedings play a much smaller role). Plans are under way to adapt the best of these papers for publication in an upcoming edition of *Machine Learning*.

Day One

The emphasis of the first day was on an aspect of the protein-folding problem, called *secondary-structure prediction,* and on classification or clustering into families of other tertiary (that is, three-dimensional) substructures. It was evident that the repertoire of machine-learning and probabilistic techniques being applied to these problems is widening dramatically. Represented at the conference were papers based on constructive induction, probabilistic networks, unsupervised learning, case-based reasoning, hidden Markov models, and megaclassification techniques, as well as the now-traditional approach of neural networks.

Protein secondary-structure prediction has attained the status of a classic problem in computational biology and machine learning. The primary (linear) sequence of proteins folds into tertiary (three-dimensional) structures whose shape is essentially determined by the primary sequence. Although the general protein-folding problem is extremely difficult, shorter secondary-structure elements called alpha helices and beta sheets can be predicted with some success, using algorithms that examine a window of surrounding primary sequence. Beginning with the NETTALK-like approach of Ning Qian and Terrence Sejnowski, then of Johns Hopkins University (1988, "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models," *Journal of Molecular Biology* 202, pp. 865–884), a succession of machine learning and statistical techniques have been applied to this problem, with asymptotically improving results now reaching the range of 70-percent accuracy. A special-interest group at the ISMB conference met to formalize what has amounted to an unofficial contest to attack aspects of the protein-folding problem, using standard data sets, performance metrics, and cross-validation criteria.

Day Two

The second day of the conference concentrated on nucleic-acid sequence analysis at a wide variety of levels, ranging from neural networks trained to interpret the output of DNA-sequencing machines all the way to combinatorial algorithms for

genomic map integration. Novel approaches were presented to the problems of sequence assembly and restriction-site mapping, gene identification, interpretation of DNA crystallography data, knowledge discovery in biological-sequence databases, and RNA-structure prediction. AI techniques from the fields of machine learning, genetic algorithms, and knowledge base design played central roles in these approaches.

The final day of the conference covered a variety of AI techniques such as constraint reasoning and qualitative modeling and biochemical applications, such as simulation of metabolic pathways, study of gene regulation, and automated analysis of biological literature. The poster papers, presented the first evening of the meeting, addressed many other aspects of the topics covered during the three days of talks.

Throughout the conference, enthusiasm for this emerging field ran high (despite outside temperatures exceeding 100° F), and planning is already under way for the second conference in the series, to be held at Stanford University on 14–17 August 1994. Those with an interest in attending should send e-mail to ismb@nlm.nih.gov to be placed on a mailing list.

David Searls is co-director of the Computational Biology and Informatics Library at the University of Pennsylvania School of Medicine, where he is research associate professor of genetics, with a secondary appointment in the Department of Computer and Information Science. He received a Ph.D. in biology from Johns Hopkins University in 1980 and an M.S. in computer science from the University of Pennsylvania in 1985.

Jude Shavlik is assistant professor of computer sciences at the University of Wisconsin at Madison. In 1980, he received an M.S. in biophysics from Yale University and, in 1988, a Ph.D. in computer sciences from the University of Illinois. He is author of *Extending Explanation-Based Learning by Generalizing the Structure of Explanations* (Pitman, 1990) and co-editor of *Readings in Machine Learning* (Morgan Kaufmann, 1990).

Lawrence Hunter is the director of the Machine Learning Project at the National Library of Medicine. He received his Ph.D. in computer science from Yale University in 1989 and is the editor of *Artificial Intelligence and Molecular Biology* (AAAI Press, 1993).