AI Magazine Volume 11 Number 5 (1990) (© AAAI)

Artificial Intelligence and Molecular Biology

Lawrence Hunter

Molecular biology is emerging as an important domain for artificial intelligence research. The advantages of biology for design and testing of AI systems include large amounts of available online data, significant (but incomplete) background knowledge, a wide variety of problems commensurate with AI technologies, clear standards of success, cooperative domain experts, non-military basic research support and percieved potential for practical (and profitable) applications. These considerations have motivated a growing group of researchers to pursue both basic and applied AI work in the domain.

More than seventy-five researchers working on these problems gathered at Stanford for a AAAI sponsored symposium on the topic. This article provides a description of much of the work presented at the meeting, and fills in the basic biology background necessary to place it in context.

Although molecular biology was the domain of several early AI systems, most notably MOLGEN and DENDRAL, the area had been relatively quiet until recently. The last few years have seen an explosion of data, knowledge and analytical techniques in molecular biology, which have triggered a renaissance of AI research in the domain. The growing community of computer scientists and biologists doing research in the area gathered together for the first time at the 1990 AAAI Spring Symposia at Stanford in March. The AI/MB symposium received applications from more than 150 researchers from more than half a dozen countries, representing ongoing research efforts at more than 30 institutions.

The work presented at the symposium spanned a wide range, from very basic research in machine learning and the automatic generation of representations, to the application challenges of squeezing out that last 1% of error in the automated interpretation of sequencing gels. The AI techniques employed by the presenters spanned even a greater range: They included A*, inductive category formation, Bayesian inference, computational linguistics (both applied to texts and to DNA sequences), neural networks, qualitative modelling, hierarchical pattern recognition, case-based reasoning, deductive inference in prolog, Connection Machine-ism, model-directed visual processing, constraint propagation, expert systems, knowledge-base maintenance, minimal length encoding, object-oriented databases, simulation, induction of context free grammars, and various knowledge acquisition technologies. I believe that more different AI technologies have been applied to molecular biology problems in the last few years than have been brought to bear in any other specific application domain in AI.

The symposium covered not only the science currently being done, but also spent half a day discussing what might be termed the sociology of the new community. Representatives of NIH, NSF and the Human Genome Project spoke about the kinds of grant money, fellowships and training support available for AI and MB work. And four of the most senior members of this young field, David Searls, Doug Brutlag, Peter Friedland and Joshua Lederberg, concluded the meeting with a panel discussion about where the field has been and where it is headed.

Although there have been other meetings

0738-4602/91/$4.00 ©1991 AAAI

of computer scientists and molecular biologists in recent years, this symposium was among the first to focus on results in computer science, rather than on biological findings. Although this was no doubt frustrating to several biologists in the audience, who repeatedly pointed out examples of glaring biological naïveté, it was reassuring to researchers worried that subfields defined by domain risk downplaying basic research issues in AI.

In fact, this domain provides some impressive advantages for basic AI research. Side by side comparisons of diverse AI techniques on real biological problems proved to be quite illuminating. Those problems involved lots of data and solutions that, for the most part, can be empirically verified. These problems also seem to be at about the right level of difficulty for AI approaches: they put functional pressure on theories, but don't seem to require a complete model of cognition to be solved. And biologists have shown themselves to be both eager collaborators and tough critics, a good mix for working with AI researchers.

The symposium provided clear, head-on comparisons of various AI techniques applied to a single problem. The DNA sequence analysis presentations included presentations from groups using expert systems, A*, model-driven vision, formal inference, grammar induction, minimal length encoding, case-based reasoning and a combined neural network/explanation-based reasoning model, all trying to recognize and characterize signals in DNA sequences. The ability to compare such radically different techniques against a single set of problems clearly identified strengths and weaknesses of the approaches.

The protein structure prediction session featured the head-on comparison of several machine learning techniques with neural networks. In addition to being able to compare the performance of different approaches, the domain also makes possible objective performance measurements. For example, none of the programs presented could predict protein secondary structure more than 70% accurately. On a more positive note, one of the DNA sequence analysis programs (the one built in using techniques from model-driven vision) found a gene that had been overlooked in a sequence that had been previously analyzed by expert molecular biologists.

The symposium highlighted the many factors that make molecular biology a good domain for AI. Unlike many other areas in science, molecular biology has a significant non-quantitative element. Systems of biochemical reactions are often described in symbolic or semi-quantitative terms, rather than by differential equations. Objects in biological systems can often be represented in compact inheritance hierarchies (evolution is, of course, the source of the inheritance hierarchy metaphor). Other kinds of biological inference seem to map well onto neural network topologies. Many of the problems molecular biologists are facing seem to require programs that can manage shifts in representation, cleverly search through huge conceptual spaces, or organize and use large, complex systems of knowledge. Successful approaches to these domain problems are likely generalize into new and useful AI technologies.

Molecular biology also provides a great deal of real-world data for use in machine learning, intelligent information retrieval, case-based reasoning, and other information-hungry AI research areas. A single NMR protein structure experiment can produce more than 500 megabytes of data that takes human experts weeks or even months to decode. Macromolecular sequence databases have exceeded the 100 megabyte mark, and are growing rapidly. And the scientific literature of molecular biology encompasses more than a thousand journals and on the order of thirty thousand articles per year, far too large for unaided humans to stay current with. These large datasets are often in machine readable form, and supplemented by extensive, systematic (but incomplete and perhaps partially incorrect) background knowledge. This easy accessibility is a powerful incentive for AI researchers in search of domains to explore their theories.

The Domain: Problems in Molecular Biology

So what are the problems in molecular biology, and what are the AI efforts being made to address them? There are thousands of open research topics in the field, but here I will describe a few where there is a significant amount of online data available and a potential match to existing AI techniques. All of these problems were addressed at the symposium by one or more presenters.

Protein Structure Prediction

Proteins are the main functional units in living systems. They *catalyze* (make possible) nearly all reactions in living organisms, trans-

port energy, oxygen and nutrients, and provide the mechanisms for communication within and between cells, movement, structural support, storage, and defense against outside invaders. Proteins are chains of amino acid residues. There are twenty different amino acid building blocks, and active entities range from very short polypeptides of five or six residues to complex, multi-domain proteins containing more than a thousand.

The sequence of amino acids that makes up a protein is called its *primary structure*. The functioning of a protein, however, is determined only indirectly by its primary structure; it is the three dimensional shape of a protein that confers its function. This shape, called the *tertiary structure* (I'll describe secondary structure in a moment), is created when the chain of amino acids folds up, exposing some residues to the protein's environment, hiding others, and facilitating the creation of bonds between residues that are not adjacent in sequence. Recent technological breakthroughs have made the determination of a protein's primary sequence relatively inexpensive and quick; however, the determination of a protein's tertiary structure is still quite difficult.

In most cases, a given amino acid sequence will fold up into a single, most energetically advantageous shape. It should therefore be possible to predict a protein's shape from its sequence. Unfortunately, the space of possible shapes is so large that molecular dynamics calculations, which use physical laws to minimize the free energy of evolving positions of the thousands of atoms in a protein, are currently millions to billions of times too slow (even on supercomputers) to be practical for general protein structure prediction. An important class of AI applications to molecular biology, therefore, are those attempting to heuristically predict protein structure from sequence.

Predicting the tertiary structure (shape) of a protein involves assigning a location to every one of the thousands or tens of thousands of atoms in the protein, an immense task. Linus Pauling observed in the 1950's that proteins can be divided up into locally coherent structures can be grouped into classes: there are corkscrew-like structures called *alpha helices,* extended planar structures called *beta sheets,* and everything else, termed *random coils*. Description of a protein in terms of these classifications form its *secondary structure,* and can be almost as useful as tertiary structure in understanding function. For example, two alpha helices with a 90 degree turn between them are commonly found in proteins that bind to DNA.

Secondary and tertiary structures are known from X-ray crystallography for more than 400 proteins. Many of these structures are either very similar to each other or have poor resolution, leaving about 100 distinct, detailed protein structures in publicly accessible databases. The primary structure (sequence) is known for all of these proteins and thousands of others. Using the crystallographic database as solved cases, several different machine learning and pattern recognition approaches have been tried in an attempt to learn a mapping between sequence and secondary structure. Neural networks, hierarchical pattern induction systems, inductive category formation and expert systems have all been applied to this problem. The effectiveness of various systems reported on at the symposium ranged from 60% to almost 70% correct, which is somewhat better than the original Chou and Fastman algorithm used by biochemists in the mid 1980s, but still not all that good.

One of the reasons for the limited effectiveness of these systems is that they all use strictly local information to make predictions; they slide a "window," looking at about a dozen adjacent amino acids at a time, down the protein and make predictions based on the residues in the window. It is possible that residues distant in the primary sequence play a significant role in secondary structure formation, hence limiting the effectiveness of sliding window prediction. Another possible source of the accuracy limitation may be that helices, turns and sheets are not the right level of description for prediction. Zhang and Waltz, taking the latter stance, presented an alternative to traditional secondary structure classes, based on novel scheme for automatically generating representations. They built an autoassociative neural network trained by backpropagation on a detailed characterization of some of the structure data. The input and output layers consisted of 120 binary nodes for representing the position of a residue, and the dataset could be accurately stored using 20 hidden nodes. Once trained, the values of the hidden nodes could be used as a compact representation for the structures presented at the input. They used k-means classification to group the generated representations, which formed a more fine grained classification than traditional helices, sheets and coils. Although they did not present methods for mapping sequence into their structure representations for prediction, they

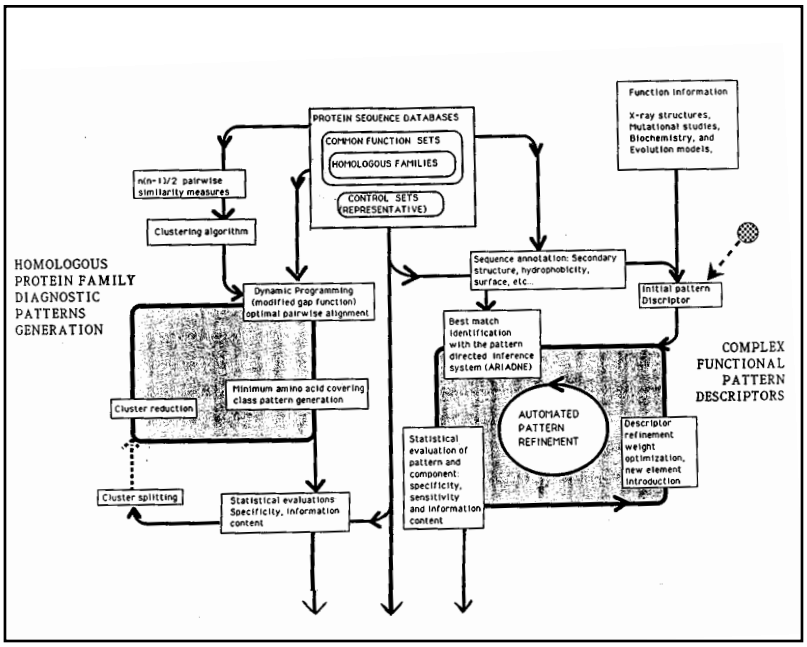

Figure 1: Overall flowchart of system for the automatic generation of primary sequence patterns from sets of related protein sequences, from the abstract submitted by Smith, Lathrop, et al.

did provide some indirect evidence that their induced representation was biologically significant.

Secondary structure is not all there is to protein shape. One amino acid (cysteine) can form disulphide bonds with other cysteines, pulling together large pieces of a protein which are not adjacent in sequence. Also, whether a given residue or secondary structure group is on the inside or on the surface of a protein is relevant in assessing the protein's function. A group at UC San Francisco School of Pharmacy's Computer Graphics Laboratory presented the results of experiments using neural networks to identify which cysteines are likely to form disulphide bonds, and which residues will be internal and which external. These networks performed at greater than 80% accuracy, and they hope that analyses of the functioning of the hidden nodes may lead to new understanding of the chemical processes involved.

Protein shape is not the only way to find relationships to function, either. Temple Smith, Rick Lathrop and a large group of collaborators at Harvard and MIT have been searching for protein sequence motifs (patterns) that can be used to predict certain kinds of functionality directly from primary sequence. They have had some significant successes, and have built a large system that attempts to automatically infer sequence motifs (see figure 1) from related sets of protein sequences. Lathrop's PhD thesis work at the MIT AI lab focuses on parallel pattern induction for protein sequence motifs using a Connection Machine.

In addition to AI work dedicated to making predictions about protein structure, significant effort has also gone into knowledge based systems for selecting and applying traditional tools of protein structure analysis. Dominic Clark, Chris Rawlings, and others at the British Imperial Cancer Research Foundation presented work describing a prolog based system for knowledge based orchestration of protein sequence analysis. Their system is capable of retrieving data from several different databases and applying a variety of algorithms to that data for such tasks as calculating hydrophobicity plots, making estimates of secondary and tertiary structure, searching for sequence motifs, and so on (see figure 2). Kuhara and colleagues from the Graduate School of Genetic Resources Technology in Kyushu, Japan described a similar system called GENAS.

The problem of predicting protein structure (and function) from sequence has vast scientific, medical and economic significance. The size of the space of possible protein shapes is immense, yet every time a living thing synthesizes a new protein, the problem is solved in the order of seconds. Researchers from many different areas of biology, medicine, pharmacology, physics and computer science are attacking this problem; many people at the symposium hoped that AI systems or techniques would help find the answer.

NMR

One of the reasons that predicting structure from sequence is so important is that it is very hard to find out a protein's structure directly. The only currently available method is X-ray crystallography, which requires growing a relatively large, very pure crystal of the protein whose structure is to be analyzed. The process is difficult and time consuming at best, and for some proteins it is simply impossible. In principle, nuclear magnetic resonance (NMR) experiments can be used to determine protein structure; if protein NMR were to become practical, the problems of having to crystalize proteins would go away. One of the main difficulties with protein NMR is data analysis. A typical two dimensional NMR experiment on a protein produces more than 500 megabytes of real numbered data. These data are printed out as

large maps, and human experts pour over them, trying to extract relationships among the points on the map that indicate two and three atom "spin systems." These systems often overlap, and the researchers have to, among other tasks, distinguish between noise and overlapping data points and then assemble the small interactions into larger and larger systems. All of this work is now done by hand, and it can take weeks to months to analyze a single experiment. In the hotly competitive world of protein NMR, an small advantage in data analysis can lead to significant differences in the size or number of the proteins that can be analyzed. In addition, the overlapping data point problem limits the size of the proteins that NMR researchers can analyze.

AI groups led by vision researcher Terry Weymouth at University of Michigan and cognitive modeler Derek Sleeman at the University of Aberdeen in Scottland are working on rather different systems to aid in or automate analysis of protein NMR experiments. Members of these groups met for the first time at the symposium, and publicly discussed the distinctions between their approaches. Although mainstream computer science techniques have produced systems that are of some utility to NMR researchers, more sophisticated methods, like the ones described at the meeting, will clearly be needed.

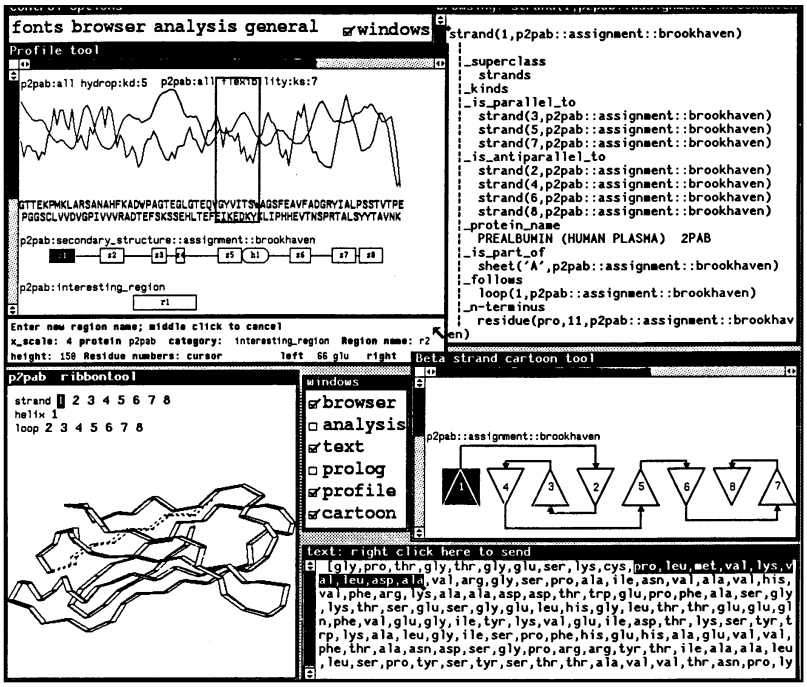

Figure 2: Screen dump showing graphical tools from the protein structure analysis tool PAPAIN, from the abstract submitted by Clark, Rawlings, et al.

Sequence Gels

Unlike the finding protein structure, recent discoveries have made it possible to read the "blueprints" for proteins relatively cheaply and quickly. The blueprints are DNA molecules, long sequences of four symbols (nucleotides) that code for the amino acids in every protein (they also contain other information). DNA sequencing experiments produce white acrylamide gels with dark spots on them, each spot indicating a particular nucleotide at a particular position. There are more than 3 billion nucleotides in the human genome, and between overlapping sequencing of the human genome for accuracy, and sequencing the genomes of other organisms, there will be tens of billions of spots to be read. This is clearly a task for machines, not people. And although not as difficult as many vision problems, there are quite a few technical challenges for gel reading systems. Ross Overbeek of Argonne National Laboratory, one of the program committee members for the symposium, described an imposing variety of hard problems he found lurking in those gels. Conventional technology can now read gels at the 95-99% accuracy level and above (depending on whose figures you believe), but each percentage point shy of perfect translates into more than 30 million added errors in the human genome sequence. Automated gel reading must be perfected soon for the human genome project to succeed, and there is a great deal of emphasis on technology development in this area. AI techniques may prove valuable in attacking the few remaining sources of error in this task.

DNA Sequence Analysis

Sequencing an entire genome creates an online record of *all* of the genetically transmitted information about an organism in a long strings taken from a four letter alphabet. That information includes not only the blueprints for all of the proteins that will be created by the organism, but also all of the control information that will be used to regulate the production of those proteins throughout its lifespan. Other information that is evolutionarily relevant but not expressed in a particular organism can also be found in these sequences. The complete DNA sequence of a bacterium is nearly at hand. The genomes of increasingly complex organisms will be available soon. Even achieving the goal of the human genome project, to acquire the

sequence of the human genome, is just the beginning of a larger task. Knowing the sequence of an organism's genome is not the same as understanding it. All of that data will have to be analyzed and understood. Many of the systems described at the symposium were designed to assist in that understanding.

The 3 billion nucleotides of human DNA contain about 100,000 genes. Each gene contains DNA that codes for both a particular protein and for a variety of other information about that protein. Some of the signals in gene sequences are used to control when the protein will be expressed, and in what quantities. Others indicate regions of the gene (introns) that should be spliced out before the sequence is translated into a protein. These control signals are crucial: they are the markers used by the systems that determine (among other things) whether a cell will be a neuron or a liver cell, and they play an important role in various diseases, including cancer. There may be still more signals in the genome whose role we do not yet understand. There is also a question of parity: Amino acids are coded for by triples of DNA nucleotides–each sequence therefore has three possible "reading frames." And, since DNA molecules are asymmetric, genes can be read in either of two directions. One stretch of DNA can therefore code for several genes, possibly using overlapping reading frames, going in opposite directions, and interrupted by different introns (which can cause shifts of reading frames within a single protein).

The complexity of huge DNA sequences has motivated many AI researchers to pursue automated methods of finding interesting signals in all of that data. The AI techniques presented at the symposium for this task included A*, an algorithm inspired by work in model-driven vision, an expert system, a formal mathematical algorithm, a system for inducing context free grammars, a case-based reasoning system, and a novel integration of explanation-based learning with neural nets.

Chris Fields and collaborators from New Mexico State University presented a program that looks for genes based on a model-driven vision algorithm. The program, GM, works by looking for a variety of low level signals that indicate potential regions of interest, and then using a one dimensional geometric model of gene contents to interpret the low level patterns and find complete genes. The model makes it possible for the program to use highly sensitive signal recognizers yet avoid being plagued by false alarms. Fields' program incorporates much of what is currently known about genetic signals and found a previously overlooked gene of the worm *C. elegans* in a sequence that had been analyzed by human experts. The program has been publicly released, and is now being used by working biologists.

Jude Shavlik and collaborators reported that the first test of their new machine learning algorithm was conducted on a DNA sequence analysis task. The algorithm is a combination of explanation-based learning and neural networks called kbann (for Knowledge-Based Artificial Neural Networks). The algorithm maps an explanation of an set of examples generated by an explanation-based learning algorithm into a topology and an initial set of approximately correct weights for a feedforward neural network. The network is then trained using backpropagation on the set of examples to fine tune the weights. They claim that this algorithm is more robust than explanation-based algorithms, and converges faster and avoids local minima better than traditional neural networks. The domain of application was the identification of promoters in bacterial DNA sequences. Promoters are control regions upstream of protein coding regions which bind to transcription factors that control the amount of protein produced. Many different algorithms have been tried on this problem. Mick Noordewier, who presented the paper, reported that kbann's error rate in a "leave-one-out" experiment was 5.3%, lower than standard backpropagation (9.2% errors), an ID3 based promoter predictor (19% errors) and a nearest neighbor (simple case-based reasoning) prediction algorithm (12.3% errors).

Another important use to which DNA sequence information has been put is reconstructing evolutionary relationships. Genes and control signals did not suddenly spring into being; they evolved from earlier genes and control mechanisms. Gene sequences undergo a variety of transformations as they evolve: One nucleotide can be substituted for another, forming a point mutation. Subsequences of nucleotides can be inserted or deleted from a gene. Pieces of sequence can be inverted, transposed or even move from one chromosome to another. And, of course, evolutionary change involves the composition of many instances of each of these changes. An important and difficult sequence analysis task requires figuring out how several different sequences are related to each other. It is this ability to find related gene sequences that enabled the discovery that certain onco-

genes (cancer-causing genes) are point mutations of normal growth factors. The ability to find more complex relationships among genetic sequences is likely to have even greater repercussions in biological understanding.

Several systems were described for doing evolutionary analyses of multiple genetic sequences. Kundu and Mukherjee from Louisiana State University described a system that used A* to find optimal covers of a set of sequences. They define a cover of a sequence S as a sequence S' that contains S as a not-necessarily-consecutive subsequence. A minimal cover of a set of sequences is the shortest sequence that covers all of them. Although it takes a restricted view of evolutionary change, covering can be seen as a kind of ancestral relationship. Kundu and Mukherjee showed that A* can be used to find an optimal cover for a set of sequences more efficiently than previously known algorithms, which used variations on dynamic programming.

David Sankoff described a system called DERANGE that represented all known biological transformations of sequences (inversion, duplication, transposition, translocation, etc.) and embodied an algorithm, called alignment reduction, that takes an initial set of linkages between two sequences, and reduces them to the minimal set of linkages that captures the transformation. The relative costs of various rearrangement mechanisms can be set, and the system works on both sequences and on the less specific characterizations found in genetic maps. The algorithm itself is a highly efficient branch and bound search which completely solves the general case. His code is available as a Macintosh application.

In a joint session with the symposium on minimal length encoding, Peter Cheeseman of NASA Ames Research Center presented an MLE method for reconstructing evolutionary trees. Cheeseman uses an MLE method to balance the trade off between the complexity of possible models of the evolution of a set of related sequences and the power of each model to account for similarities among the sequences. This work has several formal advantages over work which tries to assess the similarity among sets of sequences without an explicit evolutionary model.

Other DNA sequence analysis programs described included: An expert system, developed by MacInnes and collaborators at Los Alamos National Laboratory, that helped users select among and correctly apply different analytical tools in the widely distributed GCG sequence analysis suite. Like the work of Clark, *et al,* and Kuhara, *et al*, for protein sequences, this system uses AI technologies to help biologists use non-AI algorithms appropriately. Chris Overton at Unisys Co. presented a case-based reasoning system that retrieves cases of genes similar to an input (indexed by sequence, taxonomic and developmental factors) for identifying control and protein coding regions in the input gene. Finally, Park and Huntsberger of the University of South Carolina, presented a system that treated DNA sequences as sentences composed of parts like introns, exons and promoters, and induced grammars intended to be used to recognize genes in unanalyzed sequence data.

The potential of DNA sequence analysis as a domain for AI research is impressive. We can be sure there are important patterns in the data. We have substantial background knowledge how DNA molecules work, and about the physics and chemistry that constrain their activities. We have a lot of questions that we hope these sequences will shed light on. AI algorithms from A* to case-based reasoning to KBANN have already been applied to some of these questions, with some notable successes. And there is interest and support from biologists, who clearly recognize the need for automated assistance in sequence analysis. I expect a good deal more exciting AI research in understanding DNA sequences will emerge in the 1990s.

Modelling Biological Processes

The sequence and structure of individual molecules are of great biological importance, but living systems are complex, interacting systems of large numbers of different molecules. Work in qualitative reasoning, object oriented models, simulations and other AI technologies for representing and manipulating models of biological systems and processes was the basis for another session in the symposium.

A great deal of work in molecular biology involves investigation of the biochemical reaction pathways that make up the internal workings of all living things. These systems of reactions accomplish the tasks such as extracting energy and raw materials from foods (called *catabolic reactions),* synthesizing the building blocks of cells and doing mechanical or chemical work (called *anabolic reactions)* and also the macromolecular assembly and maintenance of proteins, nucleic acids and cellular organelles. Taken as a whole, all of these reactions are the *metabolism* of the

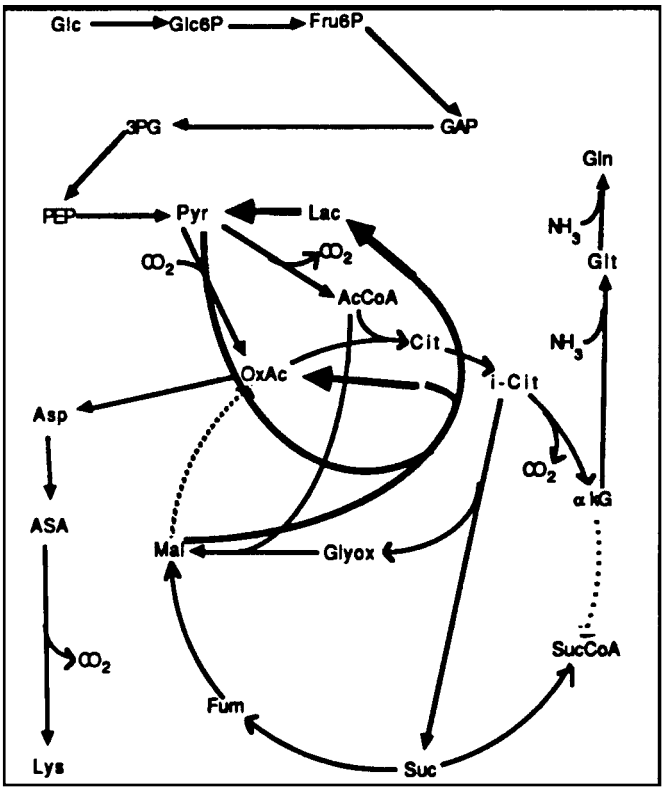

Figure 3: A program discovered this pathway for converting glucose to lysine (without malate dehydrogenase) that uses lactate dehydrogenase in the direction oppose to the one traditionally believed to be operating in bacteria. From the abstract submitted by Michael Mavrovouniotis.

organism. Understanding these pathways is crucial to understanding the basis of living systems, and manipulating them is at the core of pharmacology and the emerging field of molecular medicine.

Taken together, these reactions form a large, densely interconnected system of objects (called *substrates* by biochemists) and primitive actions *(reactions)* called the *metabolic map*. The reactions generally require catalysts, which are substances that affect a reaction without being changed by it. Many of the reactions in the map have complex preconditions involving available energy, substrates and environmental conditions such as temperature or pH. Although the complexity of the system is daunting, it is not overwhelming. There is a core map, shared among all life forms, from bacteria to people, that contains on the order of a thousand reactions.

Representing and manipulating the metabolic map, although very difficult for traditional database technologies, is a natural task for a variety of AI technologies. The metabolic map is a kind of semi-qualitative process representation, similar in some ways to the tradition of AI work in naive physics and device modelling. Researchers applying those methods in the domain gave several presentations at the symposium.

Michael Mavrovouniotis, a recent MIT graduate in both AI and Chemical Engineering, now at the University of Maryland Systems Research Center, presented his work on designing metabolic pathways. Using a qualitative representation of a large portion of the metabolic map as its knowledge base, his program is able to transform constraints on the inputs, products and reactions into a complete set of all of the possible pathways through the map that meet those constraints. The program embodies a novel algorithm for efficient generation of constrained pathways through a graph with certain topological features. It is useful both for detecting previously overlooked alternative pathways in living systems, and for the design of reaction pathways for chemical synthesis in industrial settings (see figure 3).

Dale Moberg, of The Ohio State University's Laboratory of AI Research (LAIR), presented work done in collaboration with University of Maryland philosopher of science Lindley Darden. They applied the functional representation and generic task methodologies developed by Chandrasakaren to problems in biological hypothesis formation. Darden had previously developed an account of theory change in science, using the example of the origins of modern genetics. Moberg used the LAIR work to represent the classical theory of genetics, and then use Darden's account to model how the theory changed to include an account of anomalous experimental results (see figure 4). This work is interesting not only for its claims about representation and reasoning regarding biological systems, but as an example of a productive collaboration between a philosopher of science and an AI researcher in implementing a philosophical theory.

Peter Karp, a recent Stanford PhD now at the National Library of Medicine, described his thesis work, which also embodied a model of theory change in the face of anomalous data. His work recapitulated the discovery of a form of control of gene expression called attenuation. He built a qualitative model of gene expression in bacteria representing the state of knowledge before the discovery of attenuation. His hypothesis generation program was able to use that model and a description of an unexpected experimental result to backtrack and find all

of the changes to the model which would have accounted for the result. He could then run the inferred variants in a forward direction to see if any implications of the different variant theories could be used to chose among them by experiment. One interesting result was Karp's observation that experimental conditions themselves are only hypotheses. That is, one of the ways that a theory can be wrong is in mischaracterizing the "ingredients" in an experiment.

Managing the Scientific Literature

It is important to recognize that knowledge of molecular biology far exceeds what can be found in sequence databases and the metabolic map. In fact, there are more than a thousand academic journals dedicated to issues related to molecular biology, and more than fourteen thousand journals related to topics in biomedicine generally. Even scanning just the table of contents of each of these journals would take more time than active scientists can spare, yet information relevant to any particular problem might be found almost anywhere in the literature. AI techniques in natural language understanding, image understanding and information retrieval are all being brought to bear in addressing this important problem.

The National Science Foundation recently awarded Robert Futrelle at Northeastern University a large grant to begin work on an automated mechanism for managing the biological literature. The preliminary work he presented at the symposium focused on two areas. First, he is working on a representations of the information typically found in figures, charts and graphs. In many scientific articles, the diagrams contain information that does not appear elsewhere in the text, and that can be very important to readers. Second, he is assembling a complete on-line, full text and diagrams library of articles on bacterial chemotaxis to use as a testbed for future literature management systems. His hope is that it will be possible to acquire, represent, annotate, index and cross-link the biological literature in a semi-automated way, and create sophisticated browsing tools for retrieving desired information and exploring the scientific literature generally.

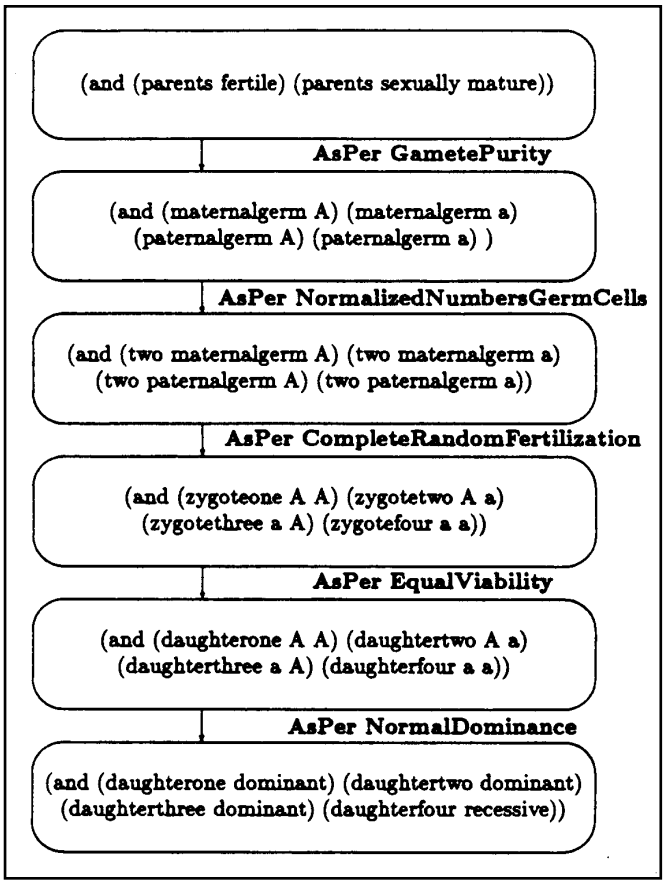

Figure 4: Representation of Medelian inheritance in gene theory in the Functional Representation system, from the abstract submitted by Dale Moberg and Lindley Darden.

Institutional Structure: Money, Careers & Resources

The AI and molecular biology symposium provided more than a forum for the exchange of ideas: it was also intended to facilitate the establishment of a new community. As with any new, interdisciplinary field, there is a need to open lines of communication and explore the institutional structures that shape the research. As is often the case at the spring symposia, there were opportunities for researchers and funding agents to meet informally. In addition to that traditional mechanism, a formal panel at the symposium specifically addressed the institutional issues of money, careers and resources.

This panel was motivated by the tremendous diversity in the AI and molecular biology research community. Symposium attendees included academics from biology, computer science and engineering departments; commercial interests from giants like UNISYS, Amoco and Bellcore to small AI/biotech startups like ARRIS Pharmaceuticals; and government scientists and funding agents from NIH, DOE and NSF. The 70 or so researchers came from seven countries spanning North America, Europe and the Pacific.

Research funding in biotechnology has a

somewhat different culture than that found in traditional AI support. There are a plethora of new players and new programs for AI researchers in search of support to consider: the National Center for Biotechnology Information, the National Center for Human Genome Research, the National Library of Medicine, and NSF's Biological Computing and Scientific Database initiatives. Speakers representing NIH and NSF discussed some of the ins and outs of acquiring funding through these agencies. In addition to traditional grant support, the National Center for Human Genome Research has announced a program for providing training in molecular biology for researchers from allied fields such as computer science. These training programs are available at all levels, from graduate and undergraduate fellowships to mid-career or sabbatical programs. A great deal of practical information was exchanged in a sometimes heated question and answer session.

In addition to financial resources, there are a wide variety of complex and rapidly evolving databases, analysis software and other research tools available to the community at large. Many of the database have the potential to be used as training sets, case-bases, or knowledge sources for AI projects. David Landsman from the National Center for Biotechnology Information described these resources in some detail, including information on how to acquire them. He also described an National Library of Medicine project, still in progress, that will integrate large segments of these databases, and track new biological results as they are published in the literature.

Also mentioned during this panel was the set of loosely coordinated researchers working on the "matrix of biological knowledge." Originating in a National Academy of Sciences report from 1985 and a six week summer workshop in 1987, the biomatrix group has brought together researchers from the AI and biology communities working on building knowledge bases in all aspects of biology. The Biomatrix group will hold an international conference July 9-11 at George Mason University in the Washington, DC area.

Conclusion

The meeting ended with a panel including Joshua Lederberg, Peter Friedland, David Searls and Doug Brutlag, the pioneers of AI and molecular biology. Each of these men are representative of the field, not only in their work, but also in their careers. David Searls, the originator of linguistic analysis of DNA sequences, has been working quietly and consistently on these problems in a corporate research environment for many years, and recently achieved as series of important milestones, including publication of a key proof in the 1988 AAAI conference and received a large DOE genome grant. Doug Brutlag is also reaping rewards for long term effort in the domain. Peter Friedland, who started in AI on the MOLGEN project at Stanford, and went on to be one of the founders of Intelligenetics, proved that research in AI and molecular biology does not preclude an eventual shift to other areas in AI: he now is the director of artificial intelligence research at NASA. Despite the career change, Friedland's talk was entitled "Why molecular biology is a good domain for AI research." Moving in the opposite direction, Joshua Lederberg, a Nobel laureate who collaborated with Ed Feigenbaum on the original DENDRAL project, is stepping down as president of Rockefeller University and is returning to AI and biology work after a decade long hiatus. His talk outlined a proposal for research into theory formation, and, more specifically, into automatically locating potential weaknesses in theories.

In addition to providing a forum discussion and meetings, the symposium has led to a variety of collaborations and other concrete projects. A database of researchers and funding agents in the area has been established, and is now publicly available. The AAAI Press will soon be publishing an edited volume based on work presented at the meeting. Several AI researchers will be presenting new work at the Hawaii International Conference on System Sciences's Biotechnology Computing minitrack in January, 1991. There is also the possibility of positive interaction between AI and molecular biology researchers and the growing community doing research in artificial life.

Overall, the symposium demonstrated more agreement on interesting problems than on interesting solutions, and common goals rather than common background knowledge or approaches. To my mind, the symposium displayed all the signs of the emergence of an exciting young field. The natural question is: How will it all evolve?

About the Author

Lawrence Hunter is director of the Machine Learning group at the National Library of Medicine, Bethesda, Maryland where he does research on systems which plan to acquire knowledge. He received his PhD at Yale University under the direction of Roger Schank.