Artificial Intelligence and Molecular Biology

Lawrence Hunter

National Library of Medicine
Lister Hill Center, Mail Stop 54
Bethesda, MD 20894
HUNTER@NLM.NIH.GOV

The twin goals of artificial intelligence research are to understand the computational mechanisms underlying human-like thought and behavior, and to design and synthesize comparable machines. Despite their inherent generality, these goals are best served by decomposing this overwhelmingly complex task into more tractable subproblems, often addressing quite specific behaviors or cognitive abilities. These decompositions generate the collection of tasks and domains that defines AI research at a particular point in time.

The particular collection of tasks and domains that AI researchers pursue has a strong relationship to the progress made towards the ultimate goals of our field. Tasks and domains define not only what problems need solution, but also the evaluation metrics by which proposed approaches are compared. There is no direct method for addressing the question "What role does computational mechanism X (say, deduction) play in intelligence?" We must be satisfied with being able to evaluate claims about the role of a particular mechanism in addressing a particular set of tasks.

Balancing the AI truism that "In order to think about thinking, you have to think about thinking about something," is the fact that what you think is true about thinking about something might, alas, not be true about thinking about something else. The everpresent risk of a divide and conquer strategy is an infelicitous division, precluding the assembly of solutions to subproblems into a coherent whole. Nevertheless, it is still apparently unreasonable to expect solutions to the general problem of computational intelligence to spring fully formed from the head of Zeus (or some other equally powerful intellect). The tension between solving a problem that is ungeneralizably narrow and taking on one that is insolubly broad pervades the recent arguments about both methodology (e.g. [Cohen, 1991]) and scaling up (e.g. [Schank, 1991]), as well as other controversies in AI.

The selection of tasks and domains is, therefore, a central problem for AI researchers. How can one evaluate the merit of a particular choice? I suggest several criteria, some more tangible than others. First, the problems in the area must be at the appropriate level for the current state of the art in AI. Ideally, the problems posed would be just a little bit beyond the ability of current methods, providing challenges but not conundrums. Second, the area ought to suggest mechanisms for comparing alternative approaches. The methodological problems in AI are real, and domains with clear performance metrics offer an advantage over those without. Third, the domain ought to encompass a wide variety of tasks related to intelligence. Many of the current problems in AI involve questions about integration of multiple abilities and about scaling or generality of methods. A domain or task area appropriate to the current state of the field should facilitate the application of a broad variety of techniques and suggest opportunities for integration of multiple methods. A final criterion is practicality: The problem area ought to provide as much of the necessary infrastructure for doing the AI research as possible. Widely available raw data or models, existing computational methods, eager collaborators, and institutional support all make a difference in how productive a task or domain is likely to be. Related to the question of institutional support is the apparent significance of the problems addressed by the field. Although good work has clearly been done in domains with no practical significance, demonstrating accomplishments in real-world tasks is a good idea for a field that, as Drew McDermott put it, is not getting any younger.

There are, no doubt, many domains that meet the desiderata enumerated above, and no single domain will suffice for all AI research. However, I believe that the current state of molecular biology provides such a good match to the current state of AI that it is worthy of consideration by many researchers.

To be honest, my initial attraction to the domain had little to do with any of the above rationales. There is a great deal of intellectual excitement in molecular biology right now. There has been an explosion of new knowledge about the living world, due to the advent of the Human Genome Program, new insights into elegantly informative model organisms such as the nematode worm, *c. Elegans*, startling technical advances such as inexpensive gene sequencing, PCR methods for exponential amplification of tiny amounts of DNA, exquisitely sensitive immunological labelling and dozens of other abilities. The answers to age old mysteries about such compelling topics as the mechanisms of aging, the origin of life, and causes of cancer seem suddenly much closer. It is reasonable to expect a technological (and social) revolution based on biology to occur in the coming decades and for it to be at least as widespread and significant as the introduction of computers. *That* is what got me going.

A key advantage of AI research is that one can study anything one thinks is interesting and still be doing AI – all one has to do is call the current interest a "domain" and, magically, developing some expertise in the area is part of doing AI research. Now, clearly, not everyone is as turned on to this domain as I am (those who are are already working in the area). Nevertheless, molecular biology has a lot to offer AI researchers, even those less enamored of the study of life.

Perhaps the most attractive thing about the domain is the breadth of interesting problems appropriate for AI methods. Trying to enumerate them all is an impossible task, but the possibilities touch on nearly every active area of AI research. A great deal of biological knowledge is symbolic and relational; qualitative simulation and reasoning about biological phenomena such as biochemical pathways, physiology and organismal development are natural applications. Many of these systems pose scaling challenges: for example, can any qualitative modelling approach handle intermediary metabolism, a system with over 1,000 objects and 10,000 relations, many of which involve multistep, multipathway feedback loops? Already work on this problem has led to innovative and general new methods for reasoning about pathways through complex feedback systems [Mavrovouniotis, 1988].

Perhaps the best known computational biology problem involves the prediction of the shape of a folded protein from its (linear) amino acid sequence. Physical simulation, although theoretically solvable from first principles, is computationally intractable. However, clever representations and Monte Carlo methods have led to some progress in simulation. Machine learning methods for learning the mapping directly have also shown promise; the reigning champion prediction approach is a neural network [Qian & Sejnowski, 1988], which is now used widely by biologists (perhaps the most prominent use of a neural network to date is the prediction of the structure of the principle neutralizing determinant of HIV, which appeared on the cover of *Science* on 24 August 1990). However, even the best current performance on this task is surprisingly poor overall. The potential for developing a machine learning approach that can take cope with the tremendous amount of available data and take advantage of existing approximate knowledge of folding mechanisms has become a challenge problem to a significant portion of the ML community. Other molecular biology problems, such as promoter recognition (a subset of the gene recognition problem described below), have already become widely used testbeds for comparison of machine learning techniques. The promoter dataset is interesting because it involves both data and an incomplete and partially incorrect domain theory; experiments with this dataset led to the creation of an influential method that combines knowledge-based and connectionist approaches [Towell, Shavlik, & Noordewier, 1990].

Laboratory robotics and automated experiment planning are turning out to be extremely important problems for the Human Genome effort. In order to sequence the huge amount of DNA involved, there are many hierarchical steps that involve breaking the original DNA down into more manageable pieces. Coordinating this process involves tracking thousands of pieces at many grain sizes, ensuring complete coverage at every level, checking for contamination by the molecules used to shuttle the target DNA around, and more. All of this execution monitoring, reaction and replanning is currently done by hand, while commercial robots manipulate the materials. Significant opportunities exist for the application (and evaluation) of AI planning ideas here, and freeing humans from these mind-numbing and difficult tasks would significantly advance the genome sequencing effort. For those who care about such matters, it may be interesting to note that although American robotics hardware is widely used in biological applications, the Japanese appear to have made the greatest strides in integrated laboratory management software for genetic sequencing projects.

The results of these genetic sequencing projects are large amounts of raw data which must then be annotated in order to be useful. The problems of annotating raw sequence include finding gene boundaries, splice sites, and a variety of other signals that are incompletely understood and genuinely noisy. There are likely to be novel signals in these sequences that we do not yet recognize. This general problem is termed "gene recognition" and there are many approaches to it, including rule- and pattern-based ones, statistical methods and neural networks. There are several widely used programs which have strong AI roots. One of these systems, GRAIL [Uberbacher, 1991], uses a neural network to combine outputs from a variety of approximate methods into a prediction that is more reliable than any of the individual methods alone. Methods for combining evidence from many partial theories have also found applications in the protein structure prediction problem and elsewhere. GRAIL illustrates another advantage of the domain: shortly after internet announcement of the system there were hundreds of users, all citing it in their publications.

There are some quite interesting problems in large scale knowledge representation and inference that arise as the result of exponential growth in the molecular biology databases. Although the official organizations that manage these databases have adopted datastructure standards and made the move to modern database management systems, there are still a wide variety of challenges to handling a diverse collection of unusual datatypes that is already over 100 megabytes and doubling yearly. Although not always reflected in the databases, there are multiple and often complex relationships between items within a database and across databases that must be inferred. Object-oriented and deductive database technologies appear to offer some promise in extending the functionality of the backbone relational databases, and additional promise appears to lie in the application of frames or other more advanced knowledge representation technology. However, the sheer size of these databases overwhelms all of the AI knowledge representation systems I have tried to use. Just to convey an idea of the order of magnitude of the task, it may be worth noting the approximate size of a few biological systems of interest: there are O(3000) genes in the

bacterium *e. Coli*, and O(100,000) in people. There are O(5000) protein building blocks and O(1500) protein families currently known. We have about 60,000 protein sequences in the databases, and about 600 protein structures—each structure has the positions in 3-space of O(1000) atoms. There are 933 cells in the best characterized multicellular organism, *c. Elegans*, about 300 of which are neurons; the complete developmental pathway, from egg to adult (called the cell fate map) is known for this animal, and there is a large library of well-characterized mutants as well. The size of these problems place them a bit beyond the AI state of the art, but not completely out of reach.

Although, of course, I can't mention them all, nearly every AI approach seems to have a commensurate molecular biology task. For natural language researchers, it turns out that many genetic phenomena are well characterized by formal grammars; see e.g. [Searls, 1988]. Anyone who has tried to assimilate even a portion of the vast amount of existing biological knowledge recognizes the need for sophisticated tutoring systems. There are a variety of interesting vision problems associated both with the automatic parsing and recognition of the patterns of lines and spots on polyacrylimide gels used in sequencing and other molecular biology experiments, and in the management of laboratory robotics generally. Case-based reasoning's ability to handle generalizations and significant exceptions appears to be a good fit to the character of biological knowledge. Information theoretic analyses have recently made predictions about DNA binding processes that were borne out experimentally. And so on and so on.

In addition to all of the challenge problems that molecular biology poses, there are important institutional and social advantages to the domain. In this era of shrinking defense budgets, there is now potential funding for AI from the National Center for Human Genome Research, the Department of Energy, the National Library of Medicine (as well as elsewhere in the National Institutes of Health), and the Biological and Behavioral Sciences division of the National Science Foundation. None of these organizations is interested primarily in AI, but neither is the DoD.

The institutional synergies are not merely financial. There is a thriving molecular biology community on the internet, where databases, software and expert consultation are freely available. The scientific usefulness of the BIONET bulletin board system and its high signal to noise ratio could be a model for the rest of usenet. Biologists were the first and most active academic community to take advantage of Thinking Machine's Wide Area Information Service (WAIS) technology, and they are eager consumers of other state of the art computer techniques.

I would be remiss if I did not also point out some of the challenges of working in the domain. First of all, biology is a large and complex field, and in order to be productive in it, a computer scientist has to learn a significant amount of vocabulary and basic background. Fortunately, this is mitigated somewhat by training programs and generous fellowships targeted at computer scientists (and others, predoctoral to mid-career) interested in getting involved in the area, available from the National Center for Human Genome Research. Despite the plethora of good problems, finding a specific match between an AI research issue and a doable molecular biology problem is not always easy. And there is the frustration of reduced performance of AI methods on the difficult real-world problems in the domain. Finding a good place to publish can also be a challenge, as computer scientists often see this kind of work as too application-oriented, and biologists understandably don't care as much about the computational method as the biological result. Fortunately this problem seems to be abating. Another challenge is the clash of cultures between biology and AI. For example, biology conferences rarely produce proceedings, and conference papers are not taken as real publications; in AI, conference papers are stringently refereed, and publication in conference proceedings is often significant. A more difficult cultural clash comes from the differing styles of collaboration. AI's history of expert systems research has promoted a rather distant form of interaction between domain expert and knowledge engineer [Forsythe, 1989]. Successful collaboration with increasingly computer-sophisticated biologists seems to involve a more balanced and collegial collaboration, and a greater frequency of interaction than has been traditional in AI.

In conclusion, I would like to suggest a further relationship between AI and molecular biology. One of the key functions of every living system is to process information so as to facilitate its goals (ultimately, survival and reproductive success); this is a deep and productive metaphor linking AI systems and living systems. AI has taken advantage of this connection by pursuing genetic algorithms and other approaches to "artificial life" [Langton, 1989]. The renown biophysicist Harold Morowitz suggests the metaphor will go even further: He claims that computer science will be to biology as mathematics is to physics. Let us hope.

References

Cohen, P. 1991. A Survey of the Eighth National Conference on Artificial Intelligence: Pulling Together or Pulling Apart? *AI Magazine*, 12(1): 16-41.

Forsythe, D. & Buchanan, B. 1989. Knowledge Acquisition for Expert Systems: Some Pitfalls and Suggestions. *IEEE Trans. on Systems, Man and Cybernetics*, 19(3): 435-442.

Langton, C., ed. 1989. *Artificial Life*. Redwood City, CA: Addison Wesley.

Mavrovouniotis, M. 1988. Computer-Aided Design of Metabolic Pathways. PhD diss., MIT.

Qian, N. & Sejnowski, T. 1988. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *Journal of Mol. Bio.*, 202: 865-884.

Schank, R. 1991. Where's the AI? *AI Magazine*, 12(4): 38-49.

Searls, D. 1988. Representing genetic information with formal grammars. In Proceedings of the 7th National Conference on AI pp. 386-391. AAAI Press.

Towell, G.; Shavlik, J. & Noordiwier, M. 1990. Refinement of approximately correct domain theories by knowledge-based neural networks. In Proc. of 8th National Conference on AI pp. 861-866. Boston, MA: AAAI Press.

Uberbacher, E. (1991). *GRAIL* Oak Ridge, National Laboratory. For info, send mail to grailmail@ornl.gov