The Knowledge Engineering Review, Vol. 8:4, 1993, 349–357

Current Research in Artificial Intelligence and Molecular Biology

DOMINIC. A. CLARK, CHRISTOPHER J. RAWLINGS AND SIMON PARSONS
Imperial Cancer Research Fund, P.O. Box 123, Lincoln's Inn Fields, London, WC2A 3PX, UK.

1. Introduction, background and motivation

This document is an overview of the "1st international conference on Intelligent Systems for Molecular Biology" that was held at the Lister Hill Center, NIH, Bethesda, Maryland, July 6-9th 1993 and the "AI and [the] Genome" workshop at IJCAI-13, Chambery, August 30th 1993.

Molecular biology has enjoyed phenomenal success since the 1950s both as an experimental and theoretical science. One manifestation of this is that the ability to generate low level data such as DNA sequences has accelerated beyond all expectations in the last decade or so. Part of this success has led to the instigation of the Human Genome Project (HGP) whose aim is to sequence the entire human genome by the year 2000, thereby creating the opportunity for major advances in our understanding of the molecular and genetic basis of disease. Overall, the complexity of biological systems and the rapid expansion of the amount of data makes molecular biology an area of unparalleled challenge and opportunity for AI and informatics in general.

The molecular sciences, together with medicine, have been two of the most important areas for developing innovative applications of AI. Initially, these activities were confined to a small number of centres. The most well-known founding projects were DENDRAL (Buchanan and Feigenbaum 1978) and MOLGEN (Stefik 1978, 1981a, 1981b). However, following the increased level of participation in symposia on AI and Molecular Biology at Biomatrix and AAAI events the idea of hosting an international conference was born in 1991 and realized at the first international conference for Intelligent Systems in Molecular Biology (ISMB-93). Of 70 papers submitted from researchers in the US, Europe and Japan, 27 were accepted for oral presentation and 25 as posters, with 3 invited speakers.

1.1. Generic problems

The generic problems of interpreting data from molecular biology experiments arise from a number of sources including: volume of data; complexity of the domain; paucity of general theories; paucity of well behaved experimental systems; sparseness of information; combinatorial complexity of the hypothesis spaces and the requirement for consistency across intersecting data and methodologies. The combination of these factors means that computational techniques to address problems in the biological sciences must remain robust in the face of many partial or uncertain theories and possibly conflicting observations. In the generic sense these difficulties are common to many areas of problem solving and thus molecular biology provides a challenging domain where new approaches to developing intelligent systems can be developed and evaluated.

1.2. The AI approaches to problem solving

Molecular biology is a key area for AI research because of both the scientific/technical challenges and the potential significance of scientific results. In consequence, many AI technologies have been applied:

Machine Learning: Applications of machine learning have generally been to aspects of pattern recognition and discovery in nucleotide or amino acid sequences - particularly the recognition (prediction) of structural features in protein sequence data. The particular machine learning approaches include probability based methods, neural nets and symbolic learning schemes (e.g. rule induction).

Uncertainty Management and Minimization: Uncertainty manifests itself in molecular biology data in a variety of different ways. Many techniques have been used to manage or minimize this uncertainty in applications. These include bayesian networks, hidden markov models, simulated annealing, genetic algorithms and hybrid technologies such as Constraint Logic Programming (henceforth CLP).

Knowledge Representation Schemes: Some public databases of molecular biology and genetics have recently moved to a relational model (e.g. the Genome Data Base) while many collections of data (e.g. the Brookhaven PIR) are still released as structured text files. More recent AI oriented approaches are looking at object oriented databases and deductive databases.

2. Conference Presentations

2.1. ISMB-93 Invited Speakers

Temple Smith (Boston University) surveyed the hard problems in the field of protein folding and reviewed the different approaches taken to address them. He identified pattern matching, neural nets, threading and other techniques and highlighted the need for large computational resources both in terms of processor speed and memory.

Leroy Hood (University of Washington) argued that sequencing technology was not going to be the worst bottleneck facing the HGP. He argued that since current machines can individually produce 36,000 base pairs per day we can expect future development and investment to yield the throughput required to sequence the genome. Hood argued that the real problems lie in building management software that will support the scale and complexity of the laboratory research enterprise required by genome sequencing and mapping laboratories.

Harold Morowitz (George Mason University) addressed the need to build models of cellular biochemistry. The biochemical pathways in the cell have evolved over the millennia into a highly complex and interactive system. He presented a theory for the evolutionary process and argued that such a model was a necessary underpinning to all intelligent systems in molecular biology and as such was a challenge to AI research.

2.2. Genome Analysis: The Problem and Conference papers

The genetic make-up of an organism can be analysed at many different levels of resolution, with the most detailed being the genetic code - the sequential order of DNA bases represented by their initial letters A,C,G and T. The total number of bases believed to comprise the human genome is of the order of 10^9, however, current DNA sequencing techniques can typically only deal directly with sequences of several hundred bases. Consequently, sequences of order 10^4 and above can only be constructed by knitting together shorter fragments. A variety of strategies for achieving this are being adopted.

In the top-down strategy, genetic and physical maps are constructed to provide a long range view of the genome by locating markers (short sequences of DNA of the order of 10-100 bases long which are treated as points) on particular chromosomes. Some of these are positioned as a result of genetic linkage analysis (e.g. studying the inheritance patterns of genes), others by physical analysis of the genome. The arrangements of these markers are called genetic maps.

In Genetic Linkage Analysis alternative forms of a marker (alleles) are studied within a pedigree (family) to determine their pattern of inheritance. Multiple markers can be examined and then statistical methods used to estimate the likelihood that they are physically linked and therefore must lie close together on the same chromosome. Matise et al. described a system called MultiMap intended to reduce the amount of user interaction typically involved in the construction of genetic linkage maps. They describe a heuristic approach implemented in Lisp and C which has produced maps at least as good as those requiring greater user intervention (and therefore the possibility of data entry error) in substantially shorter times.

The rationale of physical mapping is to recursively produce finer grain hierarchical maps by cross locating the relative positions of DNA clones of decreasing length. Assuming the high level maps are correct and clones at the lowest level are sufficiently short, it is then theoretically possible to sequence the entire genome by sequencing the minimal subset of the clones which in combination span the whole genome and then arrange their sequences according to the order provided in the higher level maps.

Conversely, sequence fragment assembly uses a bottom-up strategy. Here longer regions of DNA are broken into short fragments for direct sequencing. The process of sequence fragment assembly seeks to combine the sequence fragments into longer stretches of contiguous sequence. Parsons et al. describe a genetic algorithm which they apply to the problem of DNA sequence fragment assembly. The assembly

of overlapping sets of fragments into large segments of contiguous sequence requires the pairwise alignment of fragments to detect overlapping sections, followed by the synthesis of these pairwise alignments into an optimal global alignment, expressed as a consensus sequence. Parsons et al. treat this as a variant of the travelling salesman problem in which a cost (or goodness of fit) function is minimized. Results obtained using genetic optimization algorithm were similar to those obtained using other optimization techniques such as simulated annealing. However, it was argued that alternative cost functions might be more appropriate in this domain.

Restriction fragment mapping exploits the property of certain enzymes to recognise specific (short) base sequences and introduce a break in the DNA at or near the recognition site. The size of the fragments can potentially be used to construct an ordered map of the recognition sites. Having established a map of the right resolution, the shorter fragments may then be sequenced. The difficulties in constructing a restriction map, however, come from at least three sources: not all possible fragment lengths will be represented in the sample set (incompleteness); there is error in measurement of the fragment lengths and there is enormous variety in the lengths of the fragments such that the length of some fragments will be less than the error associated in the measured lengths of others. Skiena and Gundaram approach restriction map construction as an instance of the Turnpike Construction Problem. Their algorithm was able to deal with data sets with up to 7% error in the measured lengths and up to 8 missing fragments.

2.3. Finding Genes

Having obtained the sequence from a large piece of the genome, the most important question is to determine the meaning of the message conveyed in the genetic codes. In particular, to identify those parts of the genome which code regions that express proteins and those involved in other aspects of the gene expression and control systems. This is complex for at least three reasons: (1) Redundancy: The genetic code has a great deal of in-built redundancy and there is considerably variability in the use of alternative redundant codes among different species; (2) Signal detection: Most genomic DNA does not code for protein - it is estimated that perhaps as much as 95% of human DNA may be non-coding and (3) Complexity: As well as coding the messages that specify the amino acid sequence of proteins, DNA also carries the information that controls the expression of these proteins. Recognising the control elements can be particularly difficult because they can be sequentially remote from the coding region (60k base pairs being the largest documented).

Accurate recognition of coding regions (exons) and noncoding regions (introns) in large segments of uncharacterized genomic DNA is an unsolved but important problem. The key to this is the successful recognition of the sites where the DNA changes from being coding to being non-coding (exon-intron junctions). Solovyev and Lawrence describe a technique for exon-intron junction recognition based on an analysis of local oligonucleotide (8-mer or 9-mer) composition. Their approach takes advantage of the facts that (a) there is a recognisable difference in oligonucleotide composition in the leading and trailing sequences of the junction and (b) that junctions must alternate along the genome. Their method achieved 96-97% accuracy on a test set selected from the public DNA sequence database Genbank.

Mephu Nguifo and Sallantin describe the use of a symbolic learning algorithm that employs a number of different styles of argumentation for the prediction of primate splice junctions in gene sequences. The output forms part of an interactive dialog with the user which may be critiqued so that the final hypothesis is a synthesis of human knowledge and the empirical data.

2.4. RNA structure and function

RNA is an intermediary in the process of transcription from DNA to the protein sequence and consists of a single strand of nucleotides which folds into an energetically stable structure. Traditionally, molecular biologists have viewed RNA simplistically as a composition of helical elements. However, recent discoveries have resulted in a more complicated view of RNA structures and a realization that their biological importance has been underestimated. Furthermore, as RNA is a vital reagent and intermediary in genetic sequencing and engineering techniques, their effectiveness could be improved by a more accurate understanding of the three dimensional structures.

Klinger and Brutlag describe a comparative analysis of 1208 known tRNA sequences, which recognise the three-letter codes and transfer the particular protein building blocks (amino acids) during the synthesis of a protein. Their approach employs a flexible representation scheme using classifications of

the protein building blocks based on physical and chemical properties. Meanwhile, Altman describes and demonstrates the applicability of a probabilistic algorithm to the prediction of the topological structures of tRNA molecules (in which each base is approximated to a sphere) using constraints derived from physical chemistry (covalent bond lengths) of known tRNA crystal structures and other analyses such as that of Klinger and Brutlag. Their results support the view that these constraints provide enough information to define the topological structure of tRNAs. In contrast with other approaches this method gives measures of uncertainty associated with the final prediction.

2.5. Protein structure and function

Having ascertained the precise roles and locations of the various DNA components involved in coding for proteins, the next problem is to understand how proteins fold up into their final biologically active three-dimensional structure. Unfortunately, the prediction of the three dimensional structure of a protein from its amino acid sequence is a complex scientific problem that has not yet yielded to a completely satisfactory computational solution. There are many aspects to protein folding. The following are a subset currently addressed by AI techniques.

- The most accurate method of protein structure prediction is model building by homology. Given a sequence, if we can identify one or more homologous proteins of known structure, then we have the basis for building the structure of the new protein on the structure of the known one. The problems are (a) finding the best structure-sequence alignment, (b) dealing with local mismatches and (c) that the approach is very frequently not feasible because no homologues are identified.
- Protein Sequence Analysis aims to identify the functionally important regions or patterns within the protein sequence known to be associated with specific biological activities or structural motifs.
- Protein folding is thought to be mediated by the formation of intermediate locally stable secondary structures such as helices and sheets. Secondary Structure Prediction (SSP) aims to predict these from the sequence. Currently, the best methods based on sequence information alone attain only approximately 70% accuracy in the prediction of the secondary structure of individual amino acids. There is therefore a need to develop better methods. It is generally acknowledged that a major problem for SSP concerns longer range interactions.
- Proteins generally fold into one or more structurally distinct domains. A major problem in structure prediction is to identify the domain structure. For fairly short proteins (<120 amino acids) this is trivial since such proteins tend to fold into a single domain. However, for larger proteins, the domain structure may be more complex. Techniques which search for internal repeats can be useful in so far as they potentially identify domains produced by gene duplication, while techniques which search for trans-membrane regions can also provide constraints. However, the problem is complicated by the fact that structural domains may not be composed of contiguous sequence.
- Having located the domains within a protein sequence, a natural progression is to look at the general architecture of the protein fold within each domain. Protein Topological Class Prediction aims to identify the topological folding class of the protein. That is, whether the secondary structures fold into an α/β-sheet, an α-helical bundle, an α/β-barrel and so on.
- Given a topological folding class and a set of SSPs, topology prediction can be used to produce hypotheses about the secondary structure packing based on generic principles from known structures.
- The Inverse Folding Problem addresses the question, given a structure, which sequences could fold into that structure, with the intention that for each new structure determined by crystallography/NMR, one can potentially search for homologous sequences. Finally, even when NMR spectra or electron density maps are available, determining the protein's structure is by no means trivial.

One of the common observations in the field of protein structure analysis is that although similar sequences always have similar structures (the basis for model building), similar structures do not always have similar sequences. This makes it difficult to identify potential homologues on the basis of sequence analysis alone. Ioerger et al. attempt to deal with this situation by arguing that the amino acid sequence itself is poor for assessing such similarities and therefore that feature based learning approaches are not effective. Instead they propose a constructive inductive scheme for learning more detailed semantic relationships that can be used as the basis of sequence analysis.

Delcher et al. describe several experiments using probabilistic networks for SSP and compared their results with other techniques on common data sets. Such probabilistic networks are very different in their operation from the more usual window based techniques, though results showed prediction accuracy similar to other methods such as neural nets. Leng et al. describe a two-level case-based

reasoning architecture for SSP. Testing on the dataset of Qian and Sejnowski (1988) they produced 69.3% accuracy. However, with an alternative set representative proteins the accuracy rate was 67.3%.

Zhang et al. report a study intended to produce an alternative categorization of secondary structural elements using an auto-associative neural net system. In this task the aim is to get the neural net output to match its input though a small number of hidden units through which all information flows. Because this is a narrow channel (in relation to the number of inputs and outputs), the encoding must be an efficient one which captures the most salient structural features of the data. This encoding is then extracted using a clustering algorithm and different numbers of hidden units compared. The results suggested that the most efficient categorization was in terms of six secondary structural types. However, these corresponded very closely to existing secondary structural classifications, corresponding roughly to start, middle and end of each of β-strand and α-helix.

Brown et al. reported the use of Dirichlet mixture priors to derive Hidden Markov Models for protein families. This is a statistical technique related to Bayesian methods. The technique was illustrated for database scans for a sequence motif known to be indicative of a structural feature (called an EF-hand) that enables a protein to bind calcium. Both Ferran et al. and Wu et al. described neural net applications for clustering protein, DNA or RNA sequences into families. This provides a basis for the rapid classification of new protein sequences, while States et al. describe how to achieve similar goals using a graph theoretic approach. In their paper on Representation for Discovery of Protein Motifs, Conklin et al. proposed a spatial description logic for representing and combining sequential and structural motifs using a single uniform semantics. From a more empirical perspective Onizuka et al. presented a multi-level description scheme of protein conformation which provides a spatial analogue of fourier transforms, in which a given structure is represented as a set of overlapping topological vectors at differing levels of structural resolution.

Clark et al. address both the problem of identifying structural homologues for model building and the *de novo* prediction of protein topology. They described two programmes for predicting protein α/β-sheet and β–sheet topologies from secondary structure assignments and folding constraints using a parallel constraint logic programming language (ElipSys). The more recent programme addresses the problems of dealing with uncertain rules. This process is aided if the topological folding space is known. Clues about the this can come from both SSP and functional classification of the protein. Dubchak et al. reported a study in which they trained neural nets to recognise the topological folding class based upon protein sequence features only. Specifically, a test set of proteins was divided into 4-helical bundles, α/β barrels, α/β sheets, all-β proteins and others. Both the neural net techniques tested were generally good at predicting folding families.

2.6. Modelling Biochemical Pathways

The importance of understanding cellular structures and events at the molecular level in the context of the functional biochemical was emphasised by Harold Morovitz. The use of mathematical models of biochemical pathways to simulate normal and abnormal human biochemistry has a long history. Mathematical models have been successfully used in computer teaching aids and incorporated into medical instrumentation. However, intelligent systems for molecular biological problem-solving require a different approach. Biochemical pathways are often represented in symbolic terms as a succession of transformations between reactants and products collectively called metabolites (Mavrovouniotis, 1993a). In order to model these pathways it is necessary to explicitly represent many aspects of the biochemical environment and reason with changes over time as well as physical parameters such as concentration. It therefore provides a challenging application domain for qualitative modelling techniques and reasoning.

Previously, Karp (1993) has described the issues of representing biological knowledge for use in multiple problem solving tasks. Karp and Riley surveyed representations for several metabolic databases with detailed conclusions about the additional necessary knowledge.Mavrovouniotis (1993b) addressed the thermodynamic feasibility problem as it relates to complete biochemical pathways and presented an algorithm that analyses individual reactions and selective construction of larger sub-pathways to uncover localized and distributed thermodynamic bottlenecks.

2.7. Databases and Knowledge Bases for Molecular Biology

There are a number of problems that arise when constructing databases and knowledge bases for molecular biology, particularly when they are required to support an intelligent reasoning system. The problems come from a number of sources: the vast amount of data (e.g. number and length of biological sequences); noise in the data, errors, inconsistencies and missing data; the different levels of structural and functional organization needed for intelligent reasoning; the distributed nature of biological knowledge and data collections; and the huge body of published literature in the biological sciences

The biological literature is currently growing at a rate of 600,000 papers per year and contains a vast amount of information relating to all aspects of the science. However, much of the information in these papers (e.g. precise experimental procedures) is not contained in the title, keywords or abstract and therefore difficult to access through the usual reference services. Given, however, that research papers are highly structured documents, Baclawski et al. are developing an object oriented database for information from the materials and methods sections.

Genbank is one of the international repositories for nucleic acid sequence data and contains information on virtually every sequence that has been published. Given its size and breadth it is unlikely not to contain errors, omissions and inconsistencies. As well as the raw sequence data, database entries contain a complex tabulation of the positions of biological features observed in the sequence. A number of problems can arise in the feature tables that make automatic processing difficult. These include: missing features; mislabelled features; incompatible boundary specifications among multiple features and relegation of significant information to free text in the comments field.

Aaronson et al. described the application of a linguistic approach to recognise and rectify these problems using a DNA grammar to analyse the feature table. DNA grammars can be thought of as sets of re-write rules for parsing gene structures. However, whereas a natural language parse tree might consist of a noun phrase and verb phrase broken down into smaller language components, a DNA parse tree will consist of genetic sequence elements such as expression-promoter sites, the repeating structure of introns and exons etc. In general, such grammars provide a very powerful formalism with derivable properties for representing the high level structure of genes. Aaronson et al. applied their parser to the Genbank sequence features for the family of globin proteins and identified 30% more introns and 40% more exons than in the original author entries.

In reasoning about biological systems there are many different levels of structural, functional, genetic and other information that may be relevant. For most systems this knowledge may be distributed across many different databases. Veretnik and Schatz described the Worm Community System (WCS) which contains extensive information from many different sources and a set of tools for interacting with this system that provide an "ideal environment for dry biology".

2.8. IJCAI-93: AI and Genome Workshop

The IJCAI-13 'AI and Genome' workshop was very much in the spirit of ISMB-93, but on a smaller scale. The invited speaker, Ed Uberbacher, described the GRAIL system. This uses neural networks to identify protein-coding exons in DNA sequences with 90-95% accuracy.

In protein sequence analysis, Ishikawa et al. extended the work of Tanaka et al. (1993a) by parallelising the iterative alignment process and applying a genetic algorithm to tune the alignment strategy while Tanaka et al (1993b) presented an algorithm to build the hidden markov model described by Tanaka et al. (1993a). Onizuka et al. (1993b) described how to apply stochastic methods to the classification of proteins based on the multi-level description proposed by Onizuka et al. (1993a).

Steeg et al. were concerned with identifying high level features in protein sequences by discovering commonly occurring patterns of amino acids and degrees of correlation between such patterns. They have developed some fast probabilistic methods for doing so. Tchoumatchenko et al. on the other hand applied neural networks to the problem of SSP. However, unlike other neural network applications, such as that of Wu et al. and Parsons et al. (in this case) they emphasised extracting the decision rules discovered by the neural network. This approach is an important advance from the view of neural nets as a black box whose internal mysteries should not be examined. Understanding the decision rules discovered by neural networks is an important step to integrating with knowledge-based technologies.

Topics related to genetic mapping were well-represented. Doursenot et al. presented a novel approach to the construction of long range physical genetic maps from hybridization fingerprinting data and other constraints using the parallel CLP language ElipSys. They showed that both the constraint handling and the parallel logic language features of ElipSys could be used to address a massively combinatorial problem and provide an approach to the integration of additional mapping information during map construction. They concluded that parallelism and constraint handling complemented each other well in the ElipSys system and this enabled large scale and scientifically relevant applications of declarative programming methods to be developed. Graves presented tools for designing genetic knowledge bases based on graph logic and constructive type theory. The integration of mapping information obtained at all the possible levels of resolution was an important consideration. This approach, with its graphical representation reminiscent of semantic networks provides an interesting contrast with the other papers presented on data and knowledge representation, the majority of which were object-oriented.

Integration of different types of mapping information was also the concern of the papers by Medigue et al. and Dorkeld et al. The first paper discusses the need to integrate knowledge about different types of problem solving knowledge and software and describes a system that uses an object oriented knowledge representation scheme to provide a co-operative problem solving environment. This system, SCARP, is related to that described by Dorkeld et al. who have developed the SHIRKA system for the object oriented representation of biological entities and maps. Future work will integrate the two. The representation language described by Cui et al. is also object oriented, but extends the paradigm to allow active processing. That is, it provides a means for changes in the data to be detected and propagated so that, for instance, when a nucleic acid sequence is entered, reactive programs are invoked to seek homologous sequences. The database programming language they describe is also a specification language which allows database operations to be validated as well as executed.

The final paper on modelling was that of Glasgow et al., who addressed the representation of the structure of molecules, in particular amino acids, using a combination of semantic networks and frames in order to provide a model of shape and the spatial relations between atoms at all levels of complexity. They are also interested in integrating the various possible sources of molecular information, in much the same way as Graves is interested in representing different information about order and Medigue et al. are interested in integrating information about processes and software packages.

3. Assessments, Projections and Conclusions

Following the 1990 AAAI symposium on AI and Molecular Biology, Hunter (1991) observed that "more different AI technologies have been applied to molecular biology problems... than have been brought to bear in any other specific application domain in AI". The ISMB and IJCAI meetings suggest that this view is even truer in 1993 than it was in 1990 with a broad range of sometimes complementary, sometimes competing, computational techniques being tested. Among all this enthusiasm, however, we should remember that molecular biology is not a single computational problem domain, but a collection of biologically related problems which may in fact not yield to a single elegant solution, but to a variety of methods. Molecular biology has many of the awkward properties of the "real world" which is why it should be considered an important application domain for AI research. Furthermore, given the transient nature of hypotheses in molecular biology, it may be that some "problems" turn out not to require computational solutions, but need major shifts in the way in which the problem is conceived. The importance of computational techniques in "normal" science (where the problem and possible solution space are known) is clear. The extent to which AI techniques can contribute to higher level "revolutionary" paradigm shifts is far less clear. Whether these are solely the preserve of human biologists is a challenging but unanswered question.

Hunter (1991) also observed that "the [1990] symposium displayed all the signs of the emergence of an exciting young field. The natural question is: How will it evolve?" Viewing the field with 3 years hindsight, several areas of evolution can be identified. In terms of applications, given the international commitment to the HGP, it is not surprising that the principal new application area that has arisen is genome mapping. In terms of technological trends at least four areas deserve mention: benchmarking, parallelism, hybrid techniques and databases.

Benchmarking and comparison of results from computational molecular biology research relies upon data from large public domain databases. These are regularly updated and old data sets become difficult

to obtain. Consequently, it is often difficult to compare the relative efficacy of published techniques. By contrast, computer science has a much stronger history of benchmarking which is making its way into the interdisciplinary field. It was therefore encouraging to see a number of papers applying their methods to historical control sets (e.g. Qian & Sejnowski 1988, for SSP). Future developments of benchmark data sets for other topic (e.g. gene identification) would be of great benefit.

Molecular Biology is a science concerned with a number of interrelated entities (genes, DNA, proteins etc.). It is therefore encouraging to see research aimed at integrating this information either in databases (e.g. the Worm Community System) or in analysis (e.g. combining different genetic maps). Hybrid methods are also in the ascendent. Aaronson et al. use high level declarative schemes for generic knowledge rich reasoning and low level algorithmic procedures for data manipulation and many researchers now accept that a "pure AI" approach is not likely to be practical for intelligent systems that address hard biological science problems. On a similarly eclectic theme, it is interesting to note that many of the algorithms described have been parallelised or are amenable to parallelisation. The potential advantages in performance from parallelism in problems which involve search (in a hypothesis space or across a database) cannot be emphasized enough.

The one other area of technology which has developed apace over the last three years is databases - particularly the adoption of object oriented approaches. The continued rapid accumulation of molecular biology data and the need for detailed and complex cross-referencing systems means that database and knowledge-base technology development must remain a central part of all major initiatives.

The final point concerns the very nature of the relationship between AI and molecular biology. Hitherto, the field has been characterized by the close coordination between biologists and computer scientists with most of the benefits accruing to the computer scientists as their naivete is exposed and the delights and frustrations of molecular biology are made clear. From this hard core of collaborations are now coming some scientific results of clear importance and so the balance of benefit is changing. These conferences have shown that close interdisciplinary research can be mutually rewarding and there is a continuing need to bring new scientists into the AI and computational molecular arena. Continuing this theme, it is anticipated that ISMB-94 will also be held in the USA but that the venue for ISMB-95 will be a European one. Detailed reviews of many of the application areas above can be found in Hunter (1993).

Acknowledgements

Funding for this research and conference participation was provided by the CEC under ESPRIT III projects 6708 "APPLAUSE" and 6333 "IDEA" and by the Imperial Cancer Research Fund.

References

Aaronson, J, Haas, J and Overton, GC, 1993, "Knowledge discovery in GenBank", *ISMB-93*, 3-11.

Altman, R, 1993, "Probabilistic structure calculations: a three dimensional tRNA structure from sequence correlation data", *ISMB-93*, 12-20.

Baclawski, K, Futrelle, R, Fridman, N and Pescitelli, MJ, 1993, "Database techniques for biological materials and methods", *ISMB-93*, 21-28.

Brown, M, Hughey, R, Krogh, A, Mian, IS, Sjolander, K and Haussler, D, 1993, "Using Dirichlet mixture priors to derive hidden Markov models for protein families", *ISMB-93*, 47-55.

Buchanan, BG and Feigenbaum, EA, 1978 "Dendral and Meta-Dendral: Their Application Dimension", *Artificial Intelligence*, **11** (1), 5-24.

Clark, DA, Rawlings, CJ, Shirazi, J, Véron, A and Reeve, M, 1993, "Protein topology prediction through parallel constraint logic programming", *ISMB-93*, 83-91.

Conklin, D, Fortier, S and Glasgow, J, 1993, "Representation for discovery of protein motifs", *ISMB-93*, 101-108.

Cui, Z, Fox, J and Hearne, C, 1993, "Knowledge based systems for molecular biology: the role of advanced databases and formal specification", *IJCAI-93*, 87-97.

Delcher, AL, Kasif, S, Goldberg, HR and Hsu, W, 1993, "Protein structure modelling with probabilistic networks", *ISMB-93*, 109-117.

Dorkeld, F, Perrière, G and Gautier, C, 1993, "Object-oriented Modelling in Molecular Biology", *IJCAI-93*, 99-106.

Doursenot, S, Clark, DA, Rawlings, CJ and Véron, A, 1993, "Contig Mapping using ElipSys", *IJCAI-93*, 3-11.

Dubchak, I, Holbrook, S and Sung-Hou, K, 1993, "Comparison of Two Variations of Neural Network Approach to the Prediction of Protein Folding Pattern", *ISMB-93*, 118-126.

Ferran, EA, Ferrara, P and Pflugfelder, B, 1993, "Protein classification using neural networks", *ISMB-93*, 127-135.

Glasgow, JI, Fortier, S, Allen, FH, 1993, "Representation for Molecular Scene Analysis", *IJCAI-93*, 77-86.

Graves, M, 1993a, "Integrating order and distance relationships from heterogeneous maps", *ISMB-93*, 154-162.

Graves, M, 1993b, "Knowledge Base Design Techniques for Genome Mapping", *IJCAI-93*,59-66

Hunter, L, 1991, "Artificial Intelligence and Molecular Biology", *AI Magazine*, Winter 1991, 27-36.

Hunter, L, 1993, *Artificial Intelligence and Molecular Biology*, AAAI/MIT Press, Menlo Park CA.

Ioerger, TR, Rendell, L and Surbramaniam, S, 1993, "Constructive induction and protein structure prediction", *ISMB-93*, 198-206.

Ishikawa, M, Toya, T and Totoki, Y and Konagaya 1993, "A Parallel Iterative Aligner with Genetic Algorithm", *IJCAI-93*, 13-21

Karp, P, 1993, "A Qualitative Biochemistry and its Application to the Regulation of the Tryptophan Operon. In Hunter (1993).

Karp, P and Riley, M, 1993, "Representations of metabolic knowledge", *ISMB-93*, 207-215.

Klinger, TM and Brutlag, DL, 1993, "Detection of Correlations in tRNA Sequences with Structural Implications", *ISMB-93*, 225-233.

Leng, B, Buchanan, BG and Nicholas, HB, 1993, "Protein secondary structure prediction using two level case-based reasoning", *ISMB-93*, 251-259.

Matise, TC, Perlin, M and Chakravarti, A, 1993, "MultiMap: an expert system for automated genetic linkage mapping", *ISMB-93*, 260-265.

Mavrovouniotis, ML, 1993a, "Identification of Qualitatively Feasible Metabolic Pathways. In Hunter (1993).

Mavrovouniotis, ML, 1993b, "Identification of Localized and Distributed Bottlenecks in Metabolic Pathways", *ISMB-93*, 275-283.

Médigue, C, Willamowski, J, Schmeltzer, O, Uvietta, P, Rechenmann, F, Chevenet, F, Perrière, G and Gautier, C, 1993, "Modelling tasks for problem solving in molecular biology", *IJCAI-93*, 67-75.

Mephu Nguifo, E and Sallantin, J, 1993, "Prediction of primate junction gene sequences with a cooperative knowledge acquisition system", *ISMB-93*, 292-300.

Onizuka, K, Ishikawa, M, Wong, STC and Asai, K, 1993a, "A multi-level description scheme of protein conformation", *ISMB-93*, 301-309.

Onizuka, K, Asai, K and Ishikawa, M, 1993b, "A Scheme for Protein Tertiary Structure Prediction Based on Stochastic Reasoning", *IJCAI-93*, 49-58.

Parsons, R, Forrest, S and Burks, C, 1993, "Genetic algorithms for sequence assembly", *ISMB-93*, 310-318.

Qian, N and Sejnowski, TJ, 1988, "Predicting the Secondary Structure of Globular Proteins using Neural Network Models, *J. Mol. Biol*, **202**, 865-884.

Skiena, S and Sundaram, G, 1993, "A partial digest approach to restriction site mapping", *ISMB-93*, 362-370.

States, DJ, Harris, N and Hunter, L, 1993, "Computationally efficient cluster representation in molecular sequence mega classification", *ISMB-93*, 387-394.

Steeg, EW, Robinson, D, Deerfield, DW, 1993, "The Efficient Detection of Higher-Order Features in Protein Sequence Data", Working Paper (Abstract in *IJCAI-93*).

Stefik, M, 1978, "Inferring DNA Structures from Segmentation Data", *Artificial Intelligence*, **11**, 85-114.

Stefik, M, 1981a, "Planning with Constraints [MOLGEN: Pt 1]", *Artificial Intelligence*, **16**, 111-140.

Stefik, M, 1981b, "Planning and meta-planning [MOLGEN: Pt 2]", *Artificial Intelligence*, **16**, 141-169.

Tanaka, H, Asai, K, Ishikawa, M and Konagaya, A, 1993a, "Hidden Markov Models and Iterative Aligners: Study of Their Equivalence and Possibilities", *ISMB-93* 395-401.

Tanaka, H, Onizuka, K and Asai, K, 1993b, "Classification of Proteins via Successive State Splitting of Hidden Markov Network, *IJACI-93*, 25-30.

Tchoumatchenko, I, Vissotsky, F and Ganascia, J.-G, 1993, "How to Make Explicit A Neural Network Trained to Predict Proteins Secondary Structure", *IJCAI-93*, 31-47

Veretnik, S and Schatz, B, 1993, "Pattern discovery in gene regulation; designing an analysis environment", *ISMB-93*, 411-419.

Wu, C, Berry, M, Fung, Y-S and McLarty, J, 1993, "Neural networks for molecular sequence classification", *ISMB-93*, 429-437.

Zhang, X, Fetrow, JS, Rennie, WA, Waltz, DA and Berg, G, 1993, "Automatic derivation of sub-structures yields novel structural building blocks in globular proteins", *ISMB-93*, 438-446.

ISMB-93 = Proceedings of the First International Conference on Intelligent Systems for Molecular Biology (eds. L. Hunter, D. Searls and J. Shavlik) AAAI/MIT Press, Menlo Park CA, 1993.

IJCAI-93 = Proceedings of the "AI and the Genome" workshop, IJCAI, Chambery, France.