Machine Learning, 21, 5–9 (1995)

Introduction

JUDE SHAVLIK shavlik@cs.wisc.edu
Department of Computer Sciences, University of Wisconsin, Madison, WI, 53706

LAWRENCE HUNTER hunter@nlm.nih.gov
National Library of Medicine, Bethesda, MD, 20894

DAVID SEARLS dsearls@cbil.humgen.upenn.edu
Department of Human Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA, 19104

The field of molecular biology is rich in problems that seem almost tailor-made for machine learning approaches. Indeed, molecular biology was one of machine learning's earliest application areas (Stormo, Schneider, Gold, & Ehrenfeucht, 1982). While biological systems are notoriously complex, there exists a level of interpretation at which the informational molecules—DNA, RNA, and proteins—that underlie such systems are simply strings of symbols. Determining how such strings fold up in three dimensions, how they are recognized as signals by other macromolecules, or how they otherwise participate in the physiology of the cell, are challenging problems both for experimental biologists and for computational biologists who attempt to make inferences or derive clarifying generalizations from examples.

Relating molecular sequence to structure, and structure to function, are the ultimate goals, and all such analyses take place against the backdrop of evolution; the fact that many of these macromolecules are more or less related to each other by various types of "family trees" adds another whole set of clues, tasks, and complications to the field. Even finding the relevant information can be challenging: for example, genes (regions of DNA that code for proteins) are scattered sparsely throughout the DNA of higher organisms, and are themselves interrupted by non-coding sequences, so that detecting the coding regions is a thorny task that has only in recent years been addressed with increasing success by machine learning techniques (Craven & Shavlik, 1994). In fact, arguably the machine learning application with the largest impact on molecular biology is Uberbacher and Mural's (1991) GRAIL system, which uses neural networks to find genes in DNA sequences submitted via the Internet.

DNA sequences are drawn from a four-letter alphabet of molecules called bases. The primary role of DNA is to code for protein sequences, which have a 20-letter alphabet of molecules called amino acids. It is the manner in which these proteins fold that determines their behavior in the cell. The folded shape of a protein and the chemical properties of the amino acids exposed on the protein's surface determine which biochemical reactions the protein can participate in. The translation between the alphabets of DNA and protein takes place via intermediary macromolecules called RNA, which also fold up and which have their own set of computational challenges (although these

challenges are not addressed in this issue). Further background on molecular biology can be found in Hunter (1993), as well as this issue's articles.

This collection of articles is an outgrowth of the *First International Conference on Intelligent Systems for Molecular Biology* (Hunter, Searls, & Shavlik, 1993), which initiated what has become a successful annual conference. The eight papers herein address a range of important problems involving DNA and proteins. The papers employ a variety of machine learning methods, and they not only improve the state-of-the-art in molecular biology, but they also contribute new, general-purpose machine learning techniques. Each paper underwent rigorous review by at least one molecular biologist and one machine learning researcher. We have ordered the papers in a biologically meaningful way—from collecting DNA sequences in the biological laboratory, to finding interesting portions of these sequences, to predicting the structure of DNA and proteins, and finally to predicting the function of proteins.

In the first paper, Parsons, Forrest, and Burks address a problem of great pragmatic interest to the genome community. While DNA consists of strings that typically extend for millions of bases, biologists receive the data from their automated sequencing machines in short fragments that are several hundred bases long, representing randomly selected substrings of a much longer region. Reassembling the longer string from the many overlapping fragments is not as straightforward as it might seem; it is in fact a permutation-ordering problem related to the traveling-salesman task. Parsons et al. apply genetic algorithms to this task, using specially-tailored operators and fitness functions, and compare their results with the greedy algorithms that are the current state-of-the-art. Analogous problems recur at many different levels and scales in genome analysis, and techniques such as GA's and simulated annealing have found wide use in these so-called "contig construction" tasks.

Milosavljevic's article describes an approach to machine discovery in DNA analysis. He presents an imaginative treatment of a fundamental problem in computational biology: detecting similarities between strings, indicative of a possible common ancestry. The problem with traditional common-string search methods is that they can be confounded by regions of low complexity, that is, substrings that have many short repeats and are thus impoverished in "information content." The common practice is to mask out such regions to avoid false positives, essentially giving up on finding repetitive but subtly related strings. Milosavljevic proposes the use of algorithmic mutual significance in comparing strings. In essence his technique applies Occam's Razor to decide whether it is more or less parsimonious to describe strings together, in terms of their mutual similarity, or separately.

Bailey and Elkan address motif discovery from macromolecular sequences, a key problem in genome analysis, and also suggest a general-purpose method for expanding the range of problems that can be solved with the Expectation Maximization (EM) approach. In molecular biology, a *motif* is a pattern (often described as a regular expression or a probability matrix) that appears in many molecular sequences because it plays a particular functional role. The innovation in Bailey and Elkan's approach is to relax the constraint that each training sequence contain exactly one instance of the pattern to be induced. In real molecular data, it is likely that some of the examples used in training

will be included in error (i.e., not contain the motif at all), and that other examples will contain multiple copies of a motif. Further, since there are many different motifs present in the data, Bailey and Elkan propose a method for finding multiple motifs from the same sequences. They convincingly demonstrate the effectiveness of their methods using two DNA testbeds, and their results suggest that the method is likely to be useful in other pattern-finding tasks as well.

Cohen, Kulikowski, and Berman's article addresses the structure of DNA molecules. Their task is to predict where water molecules are likely to be found in relation to a DNA molecule (so-called "hydration patterns"). Starting with information derived from X-ray crystallography—the classic, rather time-consuming technique for determining three-dimensional molecular structures—Cohen et al.'s algorithm learns to predict hydration patterns for new structures. An intriguing aspect of their approach is that the system itself designs inductive learning experiments. Because of uncertainties in the domain and limited numbers of exemplars, their algorithm must choose appropriate sets of training cases, attributes, and classes. It combines the results of such experiments with domain knowledge and domain-independent heuristics, then iterates. The resulting classifiers prove to be superior to those derived by other techniques.

The next three papers address the prediction of protein structure. The ultimate goal of protein structure prediction is to predict the *tertiary structure* of the protein—namely, the spatial locations of every atom in the protein after the protein has folded up in a solvent. However, most research has studied less-daunting tasks as stepping stones to this goal. The most common task has been to predict what is called *secondary structure*, a much simpler, local description of a protein's spatial structure (Qian & Sejnowski, 1988; Rost & Sander, 1993).

Lapedes, Steeg, and Farber describe an interesting approach to the automatic discovery of novel protein secondary-structure classes that are more predictable from amino-acid sequences than are the three standard classes (*alpha*-helix, *beta*-sheet, and "random" coil). They use an unsupervised learning method to optimize a measure of correlation between the outputs of two feed-forward neural networks; one network receives a partial amino-acid sequence as its input, while the other's input is a local representation of the protein's tertiary structure. Following training, the first network can classify the amino-acid sequences of new proteins. Their results indicate that such methods can discover non-trivial classifications that share some features of the standard secondary-structure motifs, yet are easier to predict from amino-acid sequences. Their novel classes may also prove useful as an intermediate step in algorithms that estimate the full tertiary structure of proteins.

Conklin investigates a task closely related to that of Lapedes et al. He addresses the recognition of protein structural motifs: sequences of amino acids that are predictive of local three-dimensional structure. Conklin uses a spatial description logic and techniques for structured concept formation to investigate the discovery and organization of motifs in the major protein-structure database (Protein Data Bank). An important aspect of his motif representation is that it includes both spatial coordinates and amino-acid properties. His method discovered several interesting motifs predictive of local tertiary structure,

which is significant since few such predictive motifs have been previously identified. His article also provides a substantial review of work on recognizing protein motifs.

Ioerger, Rendell, and Subramaniam also address the identification of a protein's tertiary structure from its amino-acid sequence. They improve an important step in a method commonly used by biologists called *homology modeling*, which is a type of nearest-neighbor classification. Often the best way to predict a protein's structure is to find a similar protein sequence whose tertiary structure is already known and then adapt that structure. By considering possible representations for protein sequences, Ioerger et al. essentially search for improved distance metrics—for comparing two protein sequences—that lead to better recognition of proteins with similar structure. Their approach increases predictive accuracy on their testbed by several percentage points. They also discuss the general problem of using domain knowledge to improve the input representation used in a machine learning task.

An extremely important task is to predict the function of a protein given its amino-acid sequence. While inferring the protein's tertiary structure can aid this task, it is also possible to try to learn to predict function directly from sequence. In the final paper, Wu, Berry, Shivakumar, and McLarty train neural networks to classify protein sequences into functional classes known as "superfamilies." Members of a given protein superfamily bear not only a functional resemblance to each other, but share moderately similar amino-acid sequences as well. The challenge in *superfamily recognition* is to find a good input representation for the variable length (and often quite long) protein sequences. Starting with an n-gram frequency table, Wu et al. apply a novel singular value decomposition method to generate a highly compressed representation of a given protein sequence. This compressed input representation efficiently captures the similarity among the strings of a superfamily, and using this representation improves accuracy of the classifying neural network over more commonly-used input representations.

As can be seen, computational biology is an important application area for machine learning. It is one where a wide variety of machine learning approaches are applicable, interesting unanswered research questions exist, and large testbeds are publicly available via the Internet. We hope you enjoy reading the papers in this special issue and will perhaps be tempted to participate in this exciting field.

References

Craven, M. & Shavlik, J. (1994). Machine learning approaches to gene recognition. *IEEE Expert, 9*, 2–10.

Hunter, L. (1993). Molecular biology for computer scientists, In L. Hunter (Ed.), *Artificial Intelligence and Molecular Biology*, Cambridge, MA: AAAI/MIT Press.

Hunter, L., Searls, D., & Shavlik, J., editors. (1993). *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, Bethesda, MD: AAAI Press.

Qian, N. & Sejnowski, T. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology, 202*, 865–884.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology, 232*, 584–599.

Stormo, G., Schneider, T., Gold, L., & Ehrenfeucht, A. (1982). Use of the perceptron algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research, 10*, 2997–3011.

Uberbacher, E. & Mural, R. (1991). Locating protein coding regions in Human DNA sequences using a neural network – multiple sensor approach. *Proceedings of the National Academy of Sciences (USA), 88,* 11,261–11,265.