Editorial Introduction

AI and Bioinformatics

Janice Glasgow, Igor Jurisica, and Burkhard Rost

■ This article is an editorial introduction to the research discipline of bioinformatics and to the articles in this special issue. In particular, we address the issue of how techniques from AI can be applied to many of the open and complex problems of modern-day molecular biology.

This special issue of *AI Magazine* focuses on some areas of research in bioinformatics that have benefited from applying AI techniques. Undoubtedly, bioinformatics is a truly interdisciplinary field: Although some researchers continuously affect wet labs in life science through collaborations or provision of tools, others are rooted in the theory departments of exact sciences (physics, chemistry, or engineering) or computer sciences. This wide variety creates many different perspectives and terminologies. One result of this Babel of languages is that there is no single definition for what the subject of this young field really is. Even the name of the field varies: *Bioinformatics, theoretical biology, biocomputing,* or *computational biology* are just a few of the terms used. In fact, this lack of a precise definition is not of the type, "I recognize it when I see it"; rather, different representatives of the field have fairly different ideas about what it actually is.

Here, we do not attempt to impose any specific definition of the field. The particular collection of reviews presented constitutes a sparse sampling from the broad activities in the area. Larry Hunter ("Life and Its Molecules: A Brief Introduction") describes some of the concepts and terms prevalent in today's molecular biology. If you find the plethora of technical terms overwhelming, be assured that modern-day biology is far more complex than suggested by the simplified sketch presented here. In fact, researchers in life sciences live off the introduction of new concepts; the discovery of exceptions; and the addition of details that usually complicate, rather than simplify, the overall understanding of the field.

Possibly the most rapidly growing area of recent activity in bioinformatics is the analysis of microarray data. The article by Michael Molla, Michael Waddell, David Page, and Jude Shavlik ("Using Machine Learning to Design and Interpret Gene-Expression Microarrays") introduces some background information and provides a comprehensive description of how techniques from machine learning can be used to help understand this high-dimensional and prolific gene-expression data. The authors point out that it is natural to apply machine learning to such data, but it is also challenging because of its complexity.

The term *protein function* is not well defined; it encompasses a wide spectrum of biological contexts in which proteins contribute to making an organism live. (Note that the term *gene function* is somehow a misnomer in the sense that it means "the function of the protein encoded by a particular gene.") This intrinsic complexity of terminology makes it extremely difficult to build databases with controlled vocabularies for function. Furthermore, the vast majority of experimental data is buried in free-text publications. Mining free text, such as MEDLINE abstracts and machine learning interpretations of controlled vocabularies, constitutes another area of increasing activity. Rajesh

… AI techniques from knowledge representation, machine learning, knowledge discovery, and reasoning are at the forefront in addressing the important questions that are arising in molecular biology.

Nair and Burkhard Rost ("Annotating Protein Function through Lexical Analysis") review a few of the recent methods that have begun influencing experimental research. They observe that to date the technically simplest tools appear to be the most successful ones and that the seemingly most simple problem—identifying the gene-protein name from a publication—constitutes one of the major bottlenecks in incorporating free-text mining systems into everyday MEDLINE searches. Ross King ("Applying Inductive Logic Programming to Functional Genomics") reviews applications of inductive logic programming that address the problem of predicting some aspects of protein function. In particular, he reviews a method that combines the mining of controlled vocabulary with machine learning to render genomewide annotations of function.

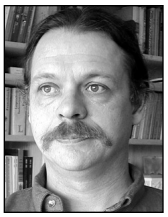
High-throughput experiments targeting the genome have become almost a standard tool for experimental biology over the last decade (for example, large-scale sequencing, micro-arrays, RNAi, two-hybrid methods, mass spectrometry). In contrast, the first comprehensive attempt at realizing high-throughput experiments for proteins—structural genomics—is still in the phase of pilot projects. One goal of structural genomics is to experimentally determine a structure for each representative protein. This seemingly simple objective hides an avalanche of bottlenecks and problems. Many of these will benefit from AI-driven solutions. One such bottleneck—protein crystallization—is addressed in the final two papers. Bruce Buchanan and Gary Livingston ("Toward Automated Discovery in the Biological Sciences") focus on the use of a novel data-mining technique to extract relationships from the data on crystal-growing experiments. Igor Jurisica and Janice Glasgow ("Applications of Case-Based Reasoning in Molecular Biology") demonstrate how case-based reasoning can be applied to assist in the planning of such experiments. They also provide an overview of several other applications in molecular biology that have benefited from case-based reasoning.

An alternative to experimental methods for determining protein structure is the application of automated techniques for predicting structure from sequence. The paper by Claus Andersen and Soren Brunak ("Amino Acid Subalphabets Can Improve Protein Structure Prediction") presents some novel research that illustrates an interesting application of AI geared toward learning about the relation between amino acid alphabets and protein. In particular, this work demonstrates the importance of knowledge representation in extracting and integrating information in biological databases.

A common theme among the articles is that biological data are complex, and the quantity of such data is growing at an unprecedented rate (and arguably outgrowing the central processing unit and storage capacity of computers). The problems that are faced in understanding molecular sequence, structure, and function rely on the ability to manage and understand these data. Thus, it is not surprising that AI techniques from knowledge representation, machine learning, knowledge discovery, and reasoning are at the forefront in addressing the important questions that are arising in molecular biology.

Janice Glasgow is a professor in the School of Computing at Queen's University, Canada, where she holds a research chair in biomedical computing. Currently, she is on sabbatical and is a senior visiting research fellow at the Institute of Advanced Studies, University of Bologna. She sits on the editorial board for several top journals in AI, cognitive science, and bioinformatics; is a past president of the Canadian Society for Computational Intelligence; and until recently was the vice-chair for the AI technical committee for the International Federation for Information Processing. Her e-mail address is janice@cs.queensu.ca.

Igor Jurisica is an assistant professor in the departments of computer science and medical biophysics at the University of Toronto and the department of the school of computing at Queen's University. In addition to his position as a scientist at the Ontario Cancer Institute/Princess Margaret Hospital, Division of Cancer Informatics, Jurisica holds a visiting scientist position at the IBM Canada Toronto laboratory. He is recognized for his work in computational biology, including representation, analysis, and visualization of high-dimensional data generated by high-throughput biology experiments. His e-mail address is juris@cs.toronto.edu.



Burkhard Rost has been an associate professor at Columbia University since 1999. After graduating from the Institute of Theoretical Physics, Heidelberg, he was part of EMBL Heidelberg (1990–1994, 1996–1998), EBI Cambridge (1995), and LION Biosciences (1998). His research focuses on methods predicting protein structure and function from sequence. The major goals of his research are to develop tools that can be applied in the context of entirely analyzing sequence organisms.