Guest Editors’ Introduction

Computer Science and Biology: An Unlikely Pair

Lawrence Hunter and Richard H. Lathrop

The very existence of this second installment of Intelligent Systems in Biology is a tribute to the strength and vibrancy of that community. The original Call for Papers produced an outpouring of community support that ultimately involved some 200 scientists and 52 manuscripts. The volume of high-quality manuscripts spilled over from the first installment (IEEE Intelligent Systems, Nov./Dec. 2001, vol. 16, no. 6) to this second one, and several accepted manuscripts will spill over to future issues of IEEE Intelligent Systems. This outstanding community support reflects the interest and excitement that permeates the field. In a very real and immediate sense, our biomedical colleagues today are running the experiments and generating the data that will unlock the deepest secrets of biology and medicine. As indicated by the cover illustration and its image of harvesting the fruits of knowledge, intelligent systems will play a crucial role in transforming these into tangible biomedical benefits that will improve the lot of all mankind. To some, there is a deep irony in the fact that computational intelligence has such a crucial role in the understanding of life.

Computer science and biology are, at first blush, an unlikely pairing: abstract, symbolic–numeric computation and wet, evolved living things. But the depth of the relationship between computing and life has only begun to be plumbed, and already the marriage has changed both fields forever. One day, the relationship between biology and computer science might be seen to be as deep and abiding as the relationship between mathematics and physics.

Advanced computation

Biology itself is undergoing a revolutionary change that would be impossible without advanced computation. Although the study of living things has engaged scholars since the beginning of history, the last generation has been particularly fruitful. Until recently, gathering any information genuinely relevant to the function of living things took heroic effort. Increases in the breadth and scale of data generation now reach every corner of biology. As argued in the Guest Editor's Introduction in the first installment, new data generation technologies have brought a “high-throughput” era to biology, creating rapid and dramatic growth in the opportunities for intelligent systems.

For example, consider the radical changes now taking place in the pharmaceutical business, perhaps the most profitable and research-driven large industry in human history. The pharmaceutical industry has transformed itself recently from being driven by biology and chemistry to being driven by information about biology and chemistry. Pharmaceutical companies have invested dramatically in systems to acquire, manage, store, and process unprecedented amounts of heterogeneous complex data about incompletely understood biomedical processes. Intelligent systems from robotics to automated inference engines have become widespread.

Similarly, across a wide spectrum of computing, major hardware and software suppliers are making significant investments in the life sciences market. IBM has made perhaps the largest corporate commitment in the area, with Sun, Compaq, Oracle and many others also devoting significant new resources to life sciences. A plethora of computational biology startups working in the area, including Genomica, Molecular Mining, and DoubleTwist, are complemented by biotechnology companies with significant computational investment.
such as Celera and Gene Logic. Every one of these efforts has an intelligent systems component.

Conversely, biology has proven to be a rich source of computational metaphors for intelligent systems, besides being a treasure trove of problems and data. Biological systems are themselves highly evolved, extremely robust, and amazingly effective information processing systems. Their properties have inspired a number of new computing paradigms. Artificial neural networks, modeled on biological neurons, have become highly competitive machine-learning tools. Genetic algorithms mimic the Darwinian optimization program of natural selection, and are the tool of choice in many optimization problems too ill-defined or poorly understood for more classical approaches. Artificial immune systems have been devised to detect computer viruses. DNA computing has solved NP-hard problems in linear time, and under certain plausible conditions can be shown to be Turing complete. Living cells are being engineered to perform directed computations, using chemical concentrations as state variables. Doubtless, biology will continue to produce potent metaphors for intelligent systems. The two fields have so much to offer each other.

In this issue

This issue opens with a perspective on "Frontiers at the Interface Between Computing and Biology," the focus of a new project of the Computer Science and Telecommunications Board at the National Academies of Science. From the policy context to the technical opportunities, Marjory Blumenthal discusses the forces behind the current environment for change and examines the intersection of computing science and biology research.

The research articles that follow showcase high points from some of the more interesting and exciting research in the field today. As before, the potential role of intelligent systems is so broad, and the opportunities so many, that any small volume can present only the tip of the iceberg of today’s intelligent systems in biology. “Bayesian Methods for Elucidating Genetic Regulatory Networks” shows how Bayesian reasoning and graphical probabilistic models—familiar AI tools for reasoning under uncertainty—can be used to unravel the mysteries of genetic regulation and control. Gene expression systems generate raw data describing the context-specific activity of many thousands of individual genes across many different environments and conditions. Bayesian network methods are well suited to this domain for their expressive power and their robustness. The article extends the semantics of Bayesian networks, derives principled scoring methods in the presence of genomic data, and demonstrates how alternative explanatory models can be compared rigorously.

“The Frame-Based Module of the SUISEKI Information Extraction System” combines the statistical analysis of protein interactions, the analysis of the syntactical structure of the phrase, and a frame-based module dedicated to the detection of protein and gene names. The result is a system that mines the free-text biomedical literature to build interaction networks for a protein–protein interaction database.

“Multidimensional Data Integration and Relationship Inference” illustrates the pharmaceutical industry’s quest for intelligent systems to acquire, manage, store, and process heterogeneous complex data. The article presents a schema for multidimensional data analysis based on an industrial-strength data warehouse infrastructure. Relationships are inferred for genes extracted from functional pathways based on cluster analysis of diverse heterogeneous data.

“A Machine Learning Strategy for Protein Analysis” provides a brief overview of the application of machine-learning methods to proteomics problems. Proteomics treats the whole field of protein structure and function relationships. Machine-learning methods have proven extremely effective in teasing out important relationships. The research described here already has produced one of the most successful protein secondary structure prediction methods in the world, using machine-learning methods. The article outlines a novel strategy for the complete prediction of protein 3D coordinates.

“Information Retrieval Meets Gene Analysis” considers mining the free-text biomedical literature to help understand genetic regulatory networks, in an elegant synthesis of the themes treated in the preceding two articles. The article uses the literature to establish functional relationships among genes on a genome-wide basis, with an ultimate goal to understand the complex biological relationships among all discovered genes and proteins.

In addition, several articles have overflowed this volume and will appear in later issues of IEEE Intelligent Systems.

“Intelligent System for Vertebrate Promoter Recognition” addresses elucidation of the molecular control signals that underly genetic regulation and control in genetic networks, a topic treated by several of the articles listed earlier. Vertebrate genetic control signals are more subtle and complicated than those in lower organisms, leading to a more difficult computational task, but are more relevant to human biology and medicine. The article uses a collection of models based on multisignal integration and artificial neural networks, leading to increased accuracy on a large and diverse human sequence set.

“Multi-Dimensional Distribution Analysis using Linear Information Compression, Applied to Structural Biology” proposes a novel data compression technique that should be applicable to many problems faced by intelligent systems. It avoids the explosion in number of parameters (coefficients) representing a multidimensional distribution as the number of dimensions increases. It is applied to knowledge-based empirical (mean-force) potentials and compares favorably with conventional 1D distance-based potentials.

“Computational Challenges in Cell Simulation” seeks to harness the enormous diversity of the components involved and the complexity of the interactions between them to produce a multilevel multiscale simulation of an entire living cell grounded at the molecular level. The article surveys the challenges involved, reviews a number of related efforts, and describes the E-Cell project and simulation environment.

Enjoy.
ACKNOWLEDGMENTS

Two people behind the scenes deserve extra special thanks: Margaret Wyvill kept track of all the manuscripts and reviews, and Mario Espinoza was the Web master for the review Web site. Special thanks to Nigel Shadbolt for his vision in initiating this special issue and to the IEEE Intelligent Systems editorial staff for their help and encouragement. My deepest debt goes to the volunteer referees, who are responsible for the high scientific quality of these pages and have my shining gratitude and devoted thanks. The following honored colleagues are hereby all named “Associate Guest Editor” of this issue: Barb Bryant, Philipp Bucher, Tim Ting Chen, David M. Cooper, Rich Cooper, John Corradi, Terence Critchlow, Steve Culp, Dan Davison, Tom Defay, Francisco M. De La Vega, Valentina Di Francesco, Paolo Frasconi, Frederique Galisson, Richard Goldstein, Harvey Greenberg, Debraj GuhaThakurta, Reece Hart, Dennis Kibler, Mark Lacy, Franz Lang, Gerald Loeffler, Satoru Miyano, Uwe Ohler, Shoba Ranganathan, Isidore Rigoutsos, Paolo Romano, Burkhard Rost, Andrey Rzhetsky, Hershel Safer, Herbert Sauro, Steffen Schulze-Kremer, Vijaya Tirunagaru, Herbert Treutlein, Josif Vaisman, David Wild, and Tau-Mu Yi.

The Authors

Lawrence Hunter is the director of the Center for Computational Pharmacology at the University of Colorado School of Medicine, and an associate professor in the Pharmacology, Computer Science, and Preventative Medicine and Biometrics departments. He is also a founder and director of the Molecular Mining Corporation. His research interests span from cognitive science to rational drug design. His primary focus recently has been the application of machine-learning techniques to data generated by high throughput molecular biology. He received his PhD in computer science from Yale University. He spent over 10 years at the National Institutes of Health, ending as the Statistics and Bioinformatics Section at the National Cancer Institute. He inaugurated two academic bioinformatics conferences, ISMB and PSB, and was the founding President of the International Society for Computational Biology. Contact him at UCHSC, Campus Box C236, 2817b, 4200 E. Ninth Ave., Denver, CO 80262; Larry.Hunter@uchsc.edu, http://compbio.uchsc.edu/hunter.

Richard H. Lathrop is vice-chair of undergraduate education in the Information and Computer Science Department at the University of California, Irvine. His research interests include applying intelligent systems and advanced computation to problems in molecular biology, especially protein structure prediction, protein-DNA interactions and genetic regulation, rational drug design and discovery, bio-nanotechnology, and other molecular structure/function relationships. In addition to a PhD in artificial intelligence, he holds degrees in electrical engineering, computer science, and mathematics. Contact him at ICS Dept. #3425, UCI, Irvine, CA, 92697-3425; rickl@uci.edu; www.ics.uci.edu/~rickl.