© EYEWIRE

Communication Theory and Molecular Biology at the Crossroads

The Guest Editor Explains the State of Research

BY ELEBEOBA E. MAY

In 1948, Claude E. Shannon's idea that images, text, and various types of data can be transmitted using a series of binary digits transformed the communication industry and society as a whole. Several years later in 1953, James Watson and Francis Crick announced the discovery of the DNA double helix. (In addition to Watson and Crick's work, Rosalind Franklin's research contributed significantly to the discovery of the structure of DNA.) Their discovery led to the eventual realization that proteins and regulatory signals can be represented using a series of quaternary symbols, A, T, G, and C, corresponding to the nucleic acid bases adenine, thymine, guanine, and cytosine. Whereas the quantification of Shannon's treatise on information theory led to the birth of coding theory and has promoted advances in digital communication, satellite communication, storage technology, and biomedical imaging, the parallels between and potential impact of the intersection of Shannon's 1948 ideas with Watson and Crick's 1953 discovery is still being realized.

As evidenced by the eight articles in this issue, researchers are increasingly curious about the communication protocols of molecular systems. In this special issue we endeavor to explore ideas at the crossroads of communication theory and molecular biology from various disciplinary backgrounds and vantage points, providing an overview of the state of research and making compelling observations regarding the nature of biological information transmission in light of the principles of digital communication.

Shannon and the Double Helix

We commence with an intriguing article, "Claude Shannon: Biologist," in which Schneider describes how insights gained from the application of information theory to molecular biology suggest that Shannon's channel capacity theorem only applies to living organisms and products of living organisms. He goes on to argue that information theory is actually a theory about biology, which makes Shannon—the father of information theory—a biologist.

In the article "Should Genetics Get an Information-Theoretic Education?," Battail describes how information theory can provide a theoretical framework for understanding the role of information in living systems. This compelling discourse shows how the hypothesis that faithful communication of genetic information over geological time depends on error-correcting codes can be used to explain the evolutionary emergence of discrete species and taxonomical hierarchy as well as evolution's trend towards increased complexity.

Gupta provides a comprehensive overview of research at the intersection of Shannon's information theory and the Watson-Crick and Franklin double helix discovery. "The Quest for Error Correction in Biology" shows the breadth of information and coding theory motivated inquiries into biological phenomena. Gupta gives special focus to work that deals with the existence of error correction in biology.

Coding Theoretic Properties of Nucleic Acids

The degree to which principles from communication theory can be used to understand, describe, or explain molecular biology is an ongoing debate. A central question is whether evidence of coding theoretic properties exists in biology and if such evidence supports information or coding theory view of genetics. The next three articles explore this question from varying perspectives.

We begin with Mac Dónaill's critical assessment of the emergence of the DNA alphabet in "Digital Parity and the Composition of the Nucleotide Alphabet." Expressing nucleotide bases as four-digit binary numbers, this fascinating article examines nucleic acid replication from a coding theory perspective and shows how this framework explains the selection of A, C, G, and T as the optimal alphabet for encoding genetic information.

Beginning with a review of the structure, signal content, and mutation mechanisms that affect DNA, Rosen investigates how DNA protects itself in "Examining Coding Structure and Redundancy in DNA." In an effort to uncover potential coding properties in genomic sequences, Rosen develops a method for detecting linear dependencies and repetitive structures in DNA. Application of these methods to the analysis of coding theoretic properties of protein-coding and nonprotein-coding regions are discussed.

Gonzalez, Giannerini, and Rosa take an interesting approach in investigating the existence of error control mechanisms in genetic processes. In "Detecting Structure in Parity Binary Sequences," they encode the exons of a gene using a mathematical coding strategy that transforms the exons into binary parity strings. The encoded sequence is analyzed for

 0739-5175/06/$20.00©2006IEEE

Researchers are increasingly curious about the communication protocols of molecular systems.

dependency structures, which, if discovered, would help support the hypothesis of the existence of deterministic error control within genetic sequences.

Coding Theory and Gene Expression

Successful development of biological information and coding theory can provide a theoretical basis for understanding, quantifying, and engineering error control in natural and synthesized biosystems. In the future, we can envision diseases, including various forms of cancer, AIDS, and geriatric maladies quantified in terms of failures in the genetic error-control system. Thus the intersection of communication theory and molecular biology could potentially yield a quantitative framework for engineering fault-tolerant genes, proteins, and genomes that approach an organism's communication capacity. Although such ideas remain ahead of us, the final articles in this special issue illustrate immediate applications of coding theory and coding theoretic principles to genomics.

In "Finding Large Domains of Similarly Expressed Genes," Nicorici, Yli-Harja, and Astola present a minimum description length (MDL, useful for source encoding) method for finding large domains of similarly expressed genes. They discuss results of using their method to discover coexpressed genes in *Drosophila* and human genomes.

May, Vouk, and Bitzer present an error-control coding model for translation initiation in "Classification of *Escherichia coli* K-12 Ribosome Binding Sites." Modeling the messenger RNA as noisy encoded sequence and the ribosome as an error-control decoder, they construct a Bayesian classifier to distinguish between valid and invalid ribosome binding sites using an eleven-base classification window.

Acknowledgments

I would like to acknowledge the assistance of Mladen Vouk and Donald Bitzer who served as coeditors for this special issue and Robert Istepanian for assistance in the review process. I would also like to thank John Enderle, editor-in-chief of *IEEE Engineering in Medicine and Biology Magazine*, for giving us the opportunity to serve as guest editors for this special issue on communication theory, coding theory, and molecular biology, and I wish to acknowledge Raouf Naguib, Ron Summers, Robert Istepanian, and the 2003 IEEE Engineering in Medicine and Biology Conference Committee for allowing us to organize a special session on this topic, from which this issue results. I am indebted to all who participated in that workshop and to Sandia National Laboratories' Computer Science Research Institute (CSRI) for financial support of the workshop.

I sincerely thank all of the authors for their contributions and diligent efforts, which made this special issue possible. Your willingness to contribute to this work has hopefully shed greater light on the exciting and growing field of biological information and coding theory. We greatly appreciate all of the reviewers for their critical reading of the articles submitted. Finally the author would like to thank all involved for their tremendous patience and cooperation in completing this special issue.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.