MIT Libraries | DSpace@MIT

MIT Open Access Articles

# *Introduction to the special issue on information theory in molecular biology and neuroscience*

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

**Citation:** Milenkovic, O. et al. "Introduction to the Special Issue on Information Theory in Molecular Biology and Neuroscience." Information Theory, IEEE Transactions On 56.2 (2010) : 649-652. ©2010 IEEE.

**As Published:** http://dx.doi.org/10.1109/TIT.2009.2036971

**Publisher:** Institute of Electrical and Electronics Engineers

**Persistent URL:** http://hdl.handle.net/1721.1/62169

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

**Terms of Use:** Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Massachusetts Institute of Technology

# Introduction to the Special Issue on Information Theory in Molecular Biology and Neuroscience

INFORMATION theory—a field at the intersection of applied mathematics and electrical engineering—was primarily developed for the purpose of addressing problems arising in data storage and data transmission over (noisy) communication media. Consequently, information theory provides the formal basis for much of today's storage and communication infrastructure.

Many attempts were made to branch the field of information theory out so as to include topics from more diverse research areas, such as computer science, physics, economics, sociology, neuroscience, and more recently, bioinformatics and systems biology. One such effort was started by Henry Quastler, who launched the area of "information theory in biology" in 1949 (just a year after the 1948 landmark paper of Shannon and four years before the inception of molecular biology shaped by the work of Crick and Watson), in a paper written with Dancoff, "The information content and error rate of living things." Continuing this effort, Quastler organized two symposiums on "Information Theory in Biology." These attempts were rather unsuccessful as argued by Henry Linschitz, who pointed out that there are difficulties in defining information "of a system composed of functionally interdependent units and channel information (entropy) to "produce a functioning cell" (cf. L. E. Kay, *Who Wrote the Book of Life*, 2000).

Looking back at Shannon's pioneering work, it is apparent that successful application of information-theoretic ideas outside communication theory (in particular, in life sciences) requires new directions and analytical paradigm shifts. Shannon's theory does not have adequate extensions for handling complicated nonstationary, nonergodic, and time-varying systems such as living cells. In addition, classical information theory does not take into account the *context and semantics of information*, *its spatial structure*, *timeliness of information delivery*, or *multidimensional resource constraints on information processing systems*. Delay management and control, structure-based information delivery, and resource limitations are ubiquitous phenomena in biology—they appear in the context of gene regulatory networks and self-organizing networks, cascade-signalling in cells, and spatio–temporal code reconstruction in neuroscience.

The importance of addressing new challenges in information theory motivated by applications in life sciences was emphasized many times in the life science research community. Manfred Eigen, Nobel laureate in chemistry, opined: "The differentiable characteristic of living systems is information. Information assures the controlled reproduction of all constituents, thereby ensuring conservation of viability." He went on further to describe "life is an interplay of energy, entropy, and information," to be studied jointly by physicists, biologists, chemists, and information theorists.

While most fundamental problems at the interface of information theory and biology remain unsolved, there have been many success stories. Information-theoretic methods have been employed in several biological applications, such as predicting the correlation between DNA mutations and disease, identifying protein binding sequences in nucleic acids, and analyzing neural spike trains and higher functionalities of cognitive systems. As the natural sciences become increasingly diverse, a paradigm of union and cooperation between these fields and information theory will lead to even greater breakthroughs. In particular, information theory has the potential to galvanize the fields of molecular biology and the neurosciences, and these disciplines can bolster each other towards new insight and discoveries in the natural sciences as well as information theory itself

This Special Issue of the IEEE TRANSACTIONS ON INFORMATION THEORY explores these areas and presents a new platform for engaging both the life science and information theory communities toward collaborative further work in these disciplines. That such collaborations are indeed possible is exemplified by the guest editors, whose joint expertise includes analysis of algorithms, bioinformatics, coding, control, and information theory, neuroscience, and systems biology. The goal of the guest editor team was to select a set of papers for the special issue that illustrate how novel ideas in information theory may lead to novel and efficient solutions of problems in life sciences. Paper selection was performed based on two-sided reviews performed by experts in biology and information theory. The 20 papers in this issue are the result of this yearlong effort.

The papers in this Special Issue are divided in two classes. The first class of papers addresses problems in genomics, proteomics, and systems biology (molecular biology) using information- and coding-theoretic techniques. The second class of papers addresses questions from the field of computational and model-based neuroscience (neuroscience).

Two papers are invited contributions, illustrating the fact that classical information theory may not be adequate for addressing certain questions in molecular biology and neuroscience. The paper by Don H. Johnson, entitled "Information theory and neural information processing," deals with the inadequacy of classical information theory for characterizing the capacity of non-Poisson processes that represent good models for neural signals. By introducing the notion of an information sink, and via a key premise that in neuroscience, the result of information processing is an action, the author proposes a novel approach to non- Poisson characterization of neuronal populations. The paper by David J. Galas, Matti Nykter, Gregory W. Carter, Nathan D. Price, and Ilya Shmulevich, entitled "Biological information as set-based complexity," describes a class of mea-

Digital Object Identifier 10.1109/TIT.2009.2036971

0018-9448/$26.00 © 2010 IEEE

sures for quantifying the contextual nature of the information in sets of objects, based on Kolmogorovs intrinsic complexity. Such measures discount both random and redundant information and do not require a defined state space to quantify the information.

For the first time in the IEEE TRANSACTIONS ON INFORMATION THEORY, we introduce a special paper category termed "Opinions." "Opinions" are paper formats sometimes used by the biology community, and their goal is to raise challenging and potentially controversial issues in the research community. The paper by Gerard Battail, "Heredity as an encoded communication process," certainly raises important and interesting questions regarding the evolution of the genetic code. The premise of the paper is that the conservation of genomes over a geological timescale and the existence of mutations at shorter intervals can be conciliated assuming that genomes possess intrinsic error-correction codes.

In the area of molecular biology, papers explore three general areas: sequencing technology methods, interaction analysis methods, and theoretical sequence analysis.

Sequencing technology methods examined include: quantitative polymerase chain reaction (the paper by Haris Vikalo, Babak Hassibi, and Arjang Hassibi, entitled "Limits of performance of quantitative polymerase chain reaction (qPCR) systems"), sequence-related data compression (the paper by Pavol Hanus, Janis Dingel, Georg Chalkidis, and Joachim Hagenauer "Compression of whole genome alignments" and the paper by Yaniv Erlich, Assaf Gordon, Michael Brand, Gregory J. Hannon, and Partha P. Mitra, entitled "Compressed genotyping"), and DNA sequence base-calling (Xiaomeng Shi, Desmond S. Lun, Muriel Médard, Ralf Kötter, James C. Meldrim, and Andrew J. Barry, "Joint base-calling of two DNA sequences with factor graphs"). Interaction analysis method papers include the work by Joan-Josep Maynou-Chacón, Joan-Josep Gallardo-Chacón, Montserrat Vallverdú, Pere Caminal, and Alexandre Perera, "Computational detection of transcription factor binding sites through differential Rényi entropy," on detecting transcription factor binding sites as well as estimation of protein and domain interactions by Faruck Morcos, Marcin Sikora, Mark S. Alber, Dale Kaiser, and Jesús A. Izaguirre, "Belief propagation estimation of protein and domain interactions using the sum-product algorithm." In the theoretical analysis section, Andrzej K. Brodzik describes an approach for rapid sequence homology assessment in "Rapid sequence homology assessment by subsampling the genome space using difference sets." Lorenzo Galleani and Roberto Garello's work is focused on determining the minimum entropy mapping spectrum of a DNA sequence in the paper entitled "The minimum entropy mapping spectrum of a DNA sequence." A rigorous analysis of classification errors with applications to genomic data analysis is provided in the paper "Joint sampling distribution between actual and estimated classification errors for linear discriminant analysis" by Amin Zollanvari, Ulisses M. Braga-Neto, and Edward R. Dougherty. An interesting example of how statistical physics, coding theory, and biological network analysis can be studied jointly is described in an overview paper by Igor Zinovik, Yury Chebiryak, and Daniel Kroening, entitled "Periodic orbits and equilibria in Glass models for gene regulatory networks."

In the area of neuroscience, the focus of the seven papers is on modeling and analyzing the properties of coding mechanisms employed by ensemble of neurons, capacity analysis of neural channels, as well as cochlear implant analysis. In the first category, the paper by Aurel A. Lazar, "Population encoding with Hodgkin–Huxley neurons," considers ensemble coding advantages in terms of system capacity. A similar approach to studying neural systems is pursued in the paper by Prapun Suksompong and Toby Berger, "Capacity analysis for integrate-and-fire neurons with descending action potential thresholds," with an emphasis on integrate-and-fire neurons. In "A mathematical theory of energy efficient neural computation and communication," the authors Toby Berger and William B. Levy extend the analysis on integrate-and-fire neurons by determining the long-term probability distribution of the action potential of this ensemble of neurons. The cooperative coding capability of neurons is modeled and analyzed in the overview paper by Mehdi Aghagolzadeh, Seif Eldawlatly, and Karim Oweiss, entitled "Synergistic coding by cortical neural ensembles." The paper by Michael C. Gastpar, Patrick R. Gill, Alexander G. Huth, and Frédéric E. Theunissen, "Anthropic correction of information estimates and its application to neural coding," describes new bias correction methods for estimating asymmetric mutual information, while the paper "Symmetry breaking in soft clustering decoding of neural codes" by Albert E. Parker, Alexander G. Dimitrov, and Tomáš Gedeon describes how bifurcation analysis can be used to study optimal quantization schemes for minimizing neural signal information distortion. Finally, the paper "A channel model for inferring the optimal number of electrodes for future cochlear implants" by Mark D. McDonnell, Anthony N. Burkitt, David B. Grayden, Hamish Meffin, and Alex. J. Grant, considers the problems of determining the number of electrodes that achieves optimal hearing performance in patients with cochlear implants.

The guest editors would like to thank all authors that submitted their manuscripts to this unique special issue and they are grateful to all diligent and thorough reviewers that evaluated the papers. The editors also hope that this special issue will set directions for research in information theory for many years to come.

OLGICA MILENKOVIC, *Guest Editor-in-Chief*

GIL ALTEROVITZ, *Guest Editor*

GERARD BATTAIL, *Guest Editor*

TODD P. COLEMAN, *Guest Editor*

JOACHIM HAGENAUER, *Guest Editor*

SEAN P. MEYN, *Guest Editor*

NATHAN PRICE, *Guest Editor*

MARCO F. RAMONI, *Guest Editor*

ILYA SHMULEVICH, *Guest Editor*

WOJCIECH SZPANKOWSKI, *Guest Editor*

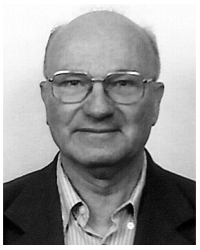
**Olgica Milenkovic** (S'01–M'08) received the M.S. degree in mathematics and Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, in 2001 and 2002, respectively.

She is an Assistant Professor of Electrical and Computer Engineering at University of Illinois, Urbana-Champaign (UIUC). From 2002 to 2006, she was with the Department of Electrical and Computer Engineering at the University of Colorado, and in 2006, she also worked as Visiting Professor at the University of California, San Diego. Her teaching and research interests lie in the areas of algorithm theory, bioinformatics, coding theory, discrete mathematics, and signal processing.

Dr. Milenkovic is a recipient of the NSF Career Award and the DARPA Young Investigator Award.

**Gil Alterovitz** (S'02) received the B.S. degree is in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, the S.M. degree from the Massachusetts Institute of Technology (MIT), Cambridge, in electrical engineering and computer science, where he was a NDSEG Fellow, and the Ph.D. degree in electrical and biomedical engineering from MIT through the Harvard/MIT Division of Health Sciences and Technology.

He is now a Faculty Member at Harvard Medical School with the Children's Hospital Informatics Program at the Harvard/MIT Division of Health Sciences and Technology (HST). He was a U.S. Fulbright to Canada (University of Toronto) in 1998–1999. He has worked at Motorola (where he won the Motorola Intellectual Property Award), at IBM, and as a consultant for several national clients. He wrote the "Proteomics" section for the Wiley Encyclopedia of Biomedical Engineering and has two textbooks published in the field of bioinformatics. He has appeared or has been cited for achievements in several national media outlets, including three separate editions of *USA Today*. In 2001, he was selected as one of approximately 20 international delegates to the Canada25 forum (to discuss healthcare/technology) covered by CBC radio, a national TV special, and Canada's Maclean's.



**Gerard Battail** (M'98) was born in Paris, France, on June 5, 1932. He graduated from the Faculté des Sciences (1954) and the Ecole nationale supérieure des Télécommunications (ENST) in 1956, both in Paris, France.

After his military duty, he joined the Centre national d'Etudes des Télécommunications (CNET) in 1959. He worked there on modulation systems and especially on frequency modulation, using fundamental concepts of information theory to understand its behavior in the presence of noise, namely, the threshold effect. In 1966, he joined the Compagnie Française Thomson-Houston (later become Thomson-CSF) as a scientific advisor to technical teams designing radioelectric devices. There he interpreted channel coding as a diversity system for designing decoders, especially soft-input ones. He also worked on source coding, frequency synthesizers, mobile communication, and other problems related to the design of industrial radio communication devices. In 1973, he joined ENST as a Professor, where he taught modulation, information theory, and coding. He has been involved in research activities in the same fields with special emphasis on adaptive algorithms as regards source coding and, for channel coding, on soft-in soft-output decoding of product and concatenated codes. He was led to criticize the conventional criterion of maximizing the minimum distance of a code, and proposed instead a criterion of closeness of the distance distribution with respect to that of random coding. These rather unorthodox views are now recognized as having paved the way to the invention of turbocodes by Berrou and Glavieux in the early 1990s. After his retirement in 1997, he started working on applications of information theory to the sciences of nature. He especially investigated the role of information theory and error-correcting codes in genetics and biological evolution, showing that the conservation of genomes needs error-correcting means. He believes that engineering and biology have much more in common than generally believed and pleads for a tight collaboration of geneticists and communication engineers. He applied for many patents, wrote many papers, and participated in many symposia and workshops. He authored a textbook on information theory published by Masson in 1997.

Dr. Battail is a Member of the Société de lElectricité, de l'Electronique et des Technologies de lInformation et de la Communication (SEE). Before his retirement, he was a Member of the Editorial Board of the Annales des Télécommunications. From 1990 to 1997, he was the French official member of Commission C of URSI (International Radio-Scientific Union). From June 2001 to May 2004, he served as Associate Editor at Large of the IEEE TRANSACTIONS ON INFORMATION THEORY.

**Todd P. Coleman** (S'01–M'05) received the B.S. degrees in electrical engineering (*summa cum laude*), as well as computer engineering (*summa cum laude*) from the University of Michigan, Ann Arbor, in 2000, along with the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 2002, and 2005.

During the 2005–2006 academic year, he was a Postdoctoral Scholar at MIT's Department of Brain and Cognitive Sciences and Massachusetts General Hospital's Neuroscience Statistics Research Laboratory in computational neuroscience. Since fall 2006, he has been on the faculty in the Electrical and Computer Engineering Department and Neuroscience Program at the University of Illinois-Urbana Champaign. His research interests include information theory of timing channels in computer and neural systems; the intersection between statistics and information theory; and theoretical and practical advances in decentralized control and feedback information theory to design high-performance brain-machine interfaces.

Dr. Coleman is a National Science Foundation Graduate Research Fellowship recipient, was awarded the University of Michigan College of Engineering Hugh Rumler Senior Class Prize in 1999 and was awarded the MIT EECS Departments Morris J. Levin Award for Best Masterworks Oral Thesis Presentation in 2002. Beginning Fall 2009, Coleman has served as a co-Principle Investigator on an NSF IGERT interdisciplinary training grant for graduate students, titled "Neuro-engineering: A Unified Educational Program for Systems Engineering and Neuroscience". He also has been selected to serve on the DARPA ISAT study group for a three-year term from 2009 to 2012.

**Joachim Hagenauer** (M'79–SM'87–F'92) received the degrees from the Technical University of Darmstadt, Germany. He held a postdoctoral fellowship at the IBM T. J. Watson Research Center, Yorktown Heights, NY, working on error—correction coding for magnetic recording. Later he became a Director of the Institute for Communications Technology at the German Aerospace Research Center DLR, since 1993, he held a chaired professorship at the TU Munich from which he retired in 2006. During 1986–1987, he spent a sabbatical year as an "Otto Lilienthal Fellow" at Bell Laboratories, Crawford Hill, NJ, working on joint source/channel coding and on trellis-coded modulation for wireless systems. He served as an editor and guest editor for the IEEE and for the "European Transactions on Telecommunications (ETT)". His research interests concentrate on the turbo principle in communications and on the application of communication principles to genetics.

Dr. Hagenauer is a Fellow and a "Distinguished Lecturer" of the IEEE. He served as President of the IEEE Information Theory Society. Among other awards he received in 1996, he received the E.H. Armstrong-Award of IEEE COMSOC, in 2003 the IEEE "Alexander Graham Bell Medal" and an Honorary Doctorate from the University Erlangen-Nuremberg in 2006.

**Sean P. Meyn** (S'85–M'87–SM'95–F'02) received the B.A. degree in mathematics (*summa cum laude*) from the University of California, Los Angeles (UCLA), in 1982 and the Ph.D. degree in electrical engineering from McGill University, Canada, in 1987 (with Prof. P. Caines, McGill University).

After a two-year Postdoctoral Fellowship at the Australian National University in Canberra, Australia, he moved to the Midwest, where he is now a Professor in the Department of Electrical and Computer Engineering, and a Research Professor in the Coordinated Science Laboratory at the University of Illinois-Urbana Champaign. He is coauthor with Richard Tweedie of the monograph *Markov Chains and Stochastic Stability* (Springer-Verlag, 1993). His new book, *Control Techniques for Complex Networks* (Cambridge University Press). He has held visiting positions at universities all over the world, including the Indian Institute of Science, Bangalore during 1997–1998 where he was a Fulbright Research Scholar. During his latest sabbatical during the 2006–2007 academic year, he was a visiting professor at MIT and United Technologies Research Center (UTRC). His research interests include stochastic processes, optimization, complex networks, and information theory. His is currently funded by NSF, Motorola, DARPA, and UTRC.

Dr. Meyn received jointly with Tweedie the 1994 ORSA/TIMS Best Publication In Applied Probability Award. The 2009 edition is published in the Cambridge Mathematical Library.

**Nathan Price** (M'09) received the Ph.D. degree in bioengineering from University of California, San Diego (UCSD), where he coauthored 20 journal articles on the analysis of genome-scale biomolecular networks under the mentorship of Bernhard Palsson.

He is an Assistant Professor in the Department of Chemical and Biomolecular Engineering, Institute for Genomic Biology, and Center for Biophysics and Computational Biology at the University of Illinois, Urbana-Champaign. He then worked on systems approaches to medicine under the guidance of Lee Hood at the Institute for Systems Biology, funded by a fellowship from the American Cancer Society.

Dr. Price was featured in Genome Technology's initial naming of "Tomorrow's PIs," received the Howard Temin Pathway to Independence Award in Cancer Research from the NIH, and an NSF CAREER award focused on genome-scale systems and synthetic biology for bioenergy applications. He is an Associate Editor for *PLoS Computational Biology and BMC Systems Biology*, and a member of the Scientific Advisory Board of TetraVitae Bioscience.

**Marco F. Ramoni** (M'07) received the Ph.D. degree in biomedical engineering and the B.A. degree in philosophy (epistemology) from the University of Pavia, Italy, and his postdoctoral training from McGill University, Montreal, Canada.

Since 1999, he has been on the faculty of Harvard Medical School, where he is currently Associate Professor of Pediatrics and Medicine (Bioinformatics) and an affiliated faculty of the Harvard-MIT Division of Health Sciences and Technology. He also is the director of the Biomedical Cybernetics Laboratory at the Harvard-Partners Center for Genetics and Genomics, where he serves as Associate Director of Bioinformatics, and the Director of the Training Fellowship in Biomedical Informatics at Children's Hospital Boston.

**Ilya Shmulevich** (SM'04) received the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 1997.

From 1997 to 1998, he was a Postdoctoral Researcher at the Nijmegen Institute for Cognition and Information, University of Nijmegen and the National Research Institute for Mathematics and Computer Science, University of Amsterdam, The Netherlands, where he studied computational models of music perception and recognition. In 1998–2000, he worked as a Senior Researcher at the Tampere International Center for Signal Processing, Signal Processing Laboratory, Tampere University of Technology, Tampere, Finland. From 2001 to 2005, he was an Assistant Professor at the Cancer Genomics Laboratory, Department of Pathology, The University of Texas M. D. Anderson Cancer Center and an Adjunct Professor in the Department of Statistics, Rice University. Currently, he is a Professor at The Institute for Systems Biology and an Affiliate Professor in the Departments of Bioengineering and Electrical Engineering, University of Washington, the Department of Signal Processing, Tampere University of Technology, Finland, and the Department of Electrical and Electronic Engineering, Strathclyde University, Glasgow, U.K. His research interests include systems biology, nonlinear signal and image processing, and computational learning theory.

Dr. Shmulevich is an Associate Editor of the *EURASIP Journal on Bioinformatics and Systems Biology*.

**Wojciech Szpankowski** (F'04) is a Professor of Computer Science and (by courtesy) Electrical and Computer Engineering at Purdue University, West Lafayette, IN, where he teaches and conducts research in analysis of algorithms, information theory, bioinformatics, analytic combinatorics, random structures, and stability problems of distributed systems.

In 2001, he published the book *Average Case Analysis of Algorithms on Sequences* (Wiley, 2001). He has been a guest editor and an editor of technical journals, including *Theoretical Computer Science*, the *ACM Transaction on Algorithms*, the IEEE TRANSACTIONS ON INFORMATION THEORY, *Foundation and Trends in Communications and Information Theory*, *Algorithmica, Combinatorics, Probability, and Computing*. He chaired a number of conferences, including Analysis of Algorithms, Gdansk and Berkeley, the NSF Workshop on Information Theory and Computer Science Interface, Chicago, and the workshop Information Beyond Shannon, Orlando and Venice.

Dr. Szpankowski held several Visiting Professor/Scholar positions, including McGill University, INRIA, France, Stanford, Hewlett-Packard Labs, Universite de Versailles, University of Canterbury, New Zealand, and Ecole Polytechnique, France. He is the Erskine Fellow and a recipient of the Humboldt Research Award.