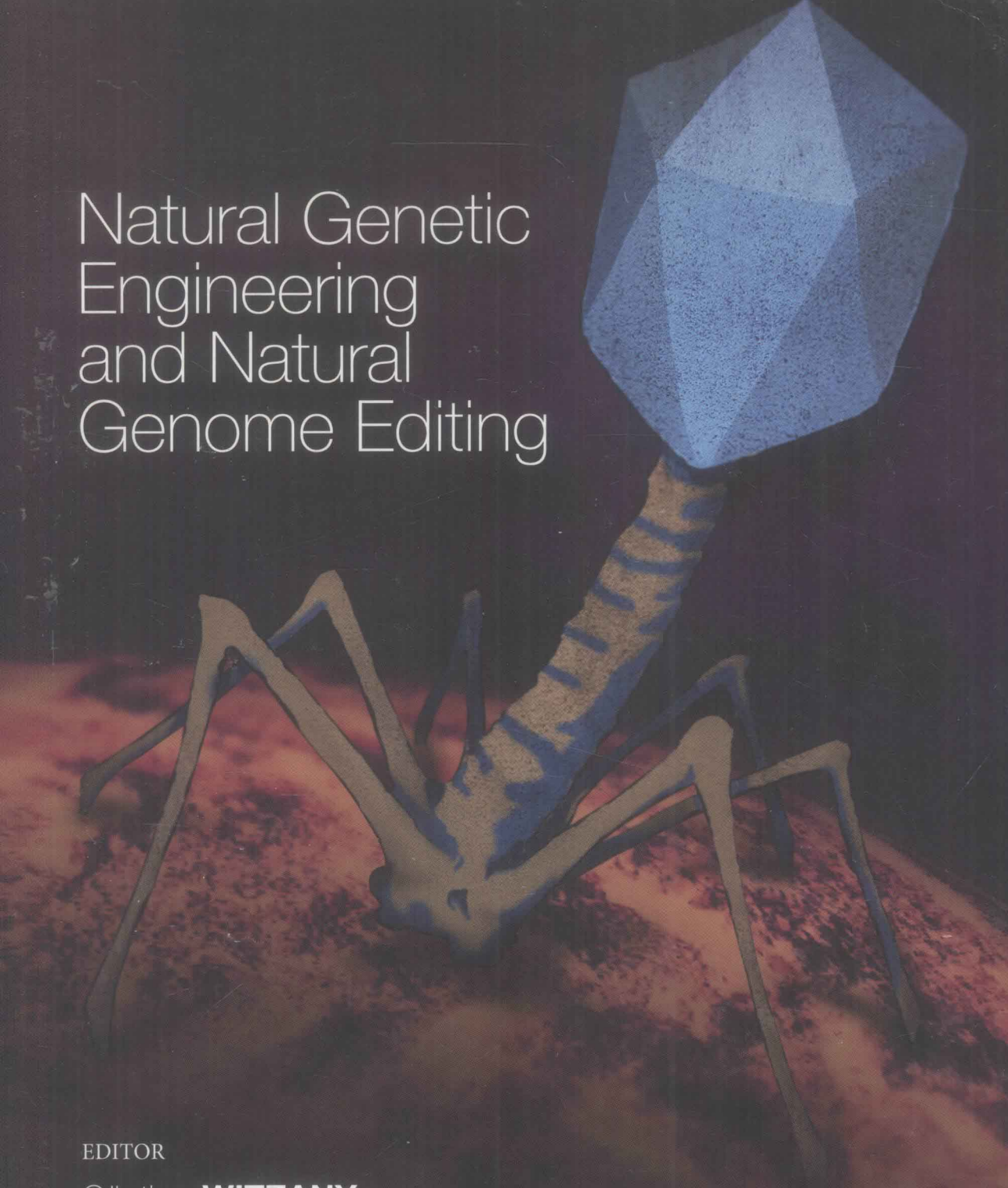
Natural Genetic Engineering and Natural Genome Editing



EDITOR

Günther **WITZANY**

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES VOLUME 1178

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Volume 1178

Natural Genetic Engineering and Natural Genome Editing

Edited by
GÜNTHER WITZANY

Published by Blackwell Publishing on behalf of the New York Academy of Sciences
Boston, Massachusetts
2009

The *Annals of the New York Academy of Sciences* (ISSN: 0077-8923 [print]; ISSN: 1749-6632 [online]) is published 32 times a year on behalf of the New York Academy of Sciences by Wiley Subscription Services, Inc., a Wiley Company, 111 River Street, Hoboken, NJ 07030-5774.

MAILING: The *Annals* is mailed standard rate. POSTMASTER: Send all address changes to *ANNALS OF THE NEW YORK ACADEMY OF SCIENCES*, Journal Customer Services, John Wiley & Sons Inc., 350 Main Street, Malden, MA 02148-5020.

Journal Customer Services: For ordering information, claims, and any inquiry concerning your subscription, please go to interscience.wiley.com/support or contact your nearest office:

Americas: Email: cs-journals@wiley.com; Tel: +1 781 388 8598 or 1 800 835 6770 (Toll free in the USA & Canada).
Europe, Middle East and Asia: Email: cs-journals@wiley.com; Tel: +44 (0) 1865 778315
Asia Pacific: Email: cs-journals@wiley.com; Tel: +65 6511 8000
Information for Subscribers: The *Annals* is published in 32 issues per year. Subscription prices for 2009 are:
Print & Online: US$4862 (US), US$5296 (Rest of World), €3432 (Europe), £2702 (UK). Prices are exclusive of tax. Australian GST, Canadian GST and European VAT will be applied at the appropriate rates. For more information on current tax rates, please go to www3.interscience.wiley.com/aboutus/journal_ordering_and_payment.html#Tax. The price includes online access to the current and all online back files to January 1, 1997, where available. For other pricing options, including access information and terms and conditions, please visit www.interscience.wiley.com/journal-info.

Delivery Terms and Legal Title: Prices include delivery of print publications to the recipient's address. Delivery terms are Delivered Duty Unpaid (DDU); the recipient is responsible for paying any import duty or taxes. Legal title passes to the customer on despatch by our distributors.

Membership information: Members may order copies of *Annals* volumes directly from the Academy by visiting www.nyas.org/annals, emailing membership@nyas.org, faxing +1 212 298 3650, or calling 1 800 843 6927 (toll free in the USA), or +1 212 298 8640. For more information on becoming a member of the New York Academy of Sciences, please visit www.nyas.org/membership. Claims and inquiries on member orders should be directed to the Academy at email: membership@nyas.org or Tel: 1 800 843 6927 (toll free in the USA) or +1 212 298 8640.

Printed in the USA.

The *Annals* is available to subscribers online at Wiley InterScience and the New York Academy of Sciences' Web site. Visit www.interscience.wiley.com to search the articles and register for table of contents e-mail alerts.

ISSN: 0077-8923 (print); 1749-6632 (online)
ISBN-10: 1-57331-765-9; ISBN-13: 978-1-57331-765-8

A catalogue record for this title is available from the British Library.

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Volume 1178

The New York Academy of Sciences
7 World Trade Center
250 Greenwich Street, 40th Floor
New York, New York 10007-2157

Natural Genetic Engineering and Natural Genome Editing

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Volume 1178

Natural Genetic Engineering and Natural Genome Editing

Editor
GÜNTHER WITZANY

This volume presents manuscripts stemming from the conference "Natural Genetic Engineering and Natural Genome Editing" held on July 3–6, 2008 at the St. Virgil Conference Center, Salzburg, Austria.

CONTENTS

Part III. Communal Evolution

Part IV. Modular Interacting Agents

Part V. Epigenetic Control

Introduction

A Perspective on Natural Genetic Engineering and Natural Genome Editing

In 1983, when I finished my studies of philosophies of language and science—in particular pragmatic action theory—I did not know the direction my research interests would take. In the ensuing four years, I studied a number of articles concerning different subjects in biology and was struck by the key vocabulary that was used for the description of the essential activities of cellular life, such as "genetic code," "genetic information," "cell–cell communication," "nucleotide sequences," "protein coding sequences," "self/nonself recognition"—all of which connote themes of communication and exchange of information, similar to the themes I had encountered in my studies of philosophy and action theory. Of particular influence on my thinking were the articles and books of Karl von Frisch, who received the Nobel Prize for his work on the language of bees; and a book by Manfred Eigen, in which he developed a profound argument for the idea that the genetic code functions not only as an analogue of natural human language but that both the evolution of life and the evolution of the mind crucially depend on the characteristic features of languages.

According to these and many other results of the discourse in the philosophy of science in the 20th century (especially between 1920 and 1980), it was clear that if the genetic code functions like a natural language then a variety of consequences follow because several preconditions must have been met. First, considering the genetic code as a natural language requires there to be a repertoire of signs (indices, icons, symbols)[a] that can be combinatorially arranged according to syntactic rules, similar to words composed of the characters of the alphabet, to generate information. Without syntactic rules to determine correct sequence order, the combination of signs could not carry informational content, that is, meaning (similarly, if our natural language had no syntactic rules, meaning could not be ascribed to randomly generated collections of words). In other words, a coherent syntax excludes randomly derived mixtures of characters of an alphabet; and for humans, coherent syntax is generated by humans who are competent with the syntactic rules. This book, for example, could contain the same characters but in a random order; such a book would be meaningless. Without competent authors who combine the characters according to a set of coherent syntactic and semantic rules, meaning does not exist.

Signs cannot exist or function without sign-using agents, and agents generate signs to communicate. In communicative action, agents can both exchange messages about

[a] According to the founder of semiotics, Charles Sanders Peirce, we are able to differentiate three different kinds of signs: indices, icons, and symbols. Indices are, in most cases, abiotic stimuli from the environment that are interpreted in the realm of memory, for example, a plant root identifies nutrients as being relevant, just as the plant shoot does with the angle of sunlight. Icons are biotic one-to-one signals (analogue) that need no further explanation, for example, plant cells identify auxin in a hormonal coordination process. There are also symbols, that is, signs or sequences of signs, like characters of an alphabet used to generate words, sentences, codes, which do not indicate by themselves what they mean (what their function could be) but are signs through natural or cultural conventions. Such sequences may also be sequences of behaviors, like the dance of the honey bees in colder hemispheres.

Natural Genetic Engineering and Natural Genome Editing: Ann. N.Y. Acad. Sci. 1178: 1–5 (2009).
doi: 10.1111/j.1749-6632.2009.05021.x

something and coordinate (organize) common behavior, that is, group behavior. A necessary condition for coordinated behavior among more than one agent is reciprocal communication that relies on pragmatic rules of information exchange that enables successful interaction. An additional element is that these pragmatic rules of information exchange must be embedded in the real-life contexts of the sign-using agents. As we know through modern communicative action theory, one agent alone (*solus ipse*) would not be able to generate signs with which it could establish sign-mediated interactions. In other words, sign use within communicative actions is inherently interconnected with group behavior and the cultural history of group identity.

According to the needs of communicating agents, syntactically correct combined signs can transport messages with meaning (semantics). But the meaning of signs depends on the context wherein agents are interwoven. As the context varies, agents will use the same signs to transport different messages. In plant communication, for example, the hormone auxin may transport various messages either in the context of neuronal-like cell–cell communication, as a hormone to transport messages between root and shoot, or as a morphogenic sign. Real-life languages or codes that are used in communication processes to coordinate and organize appropriate group behavior may differ slightly within the same species according to different habitats. The specific circumstances of a habitat of living populations for growing up and socialization may lead to special dialects, that is, the signs that are used to communicate are identical, but the meaning of the transported messages differs depending on regional customs. This phenomenon of dialects is definitively investigated in the communication of ants, honeybees, and bacteria, but transferred to the level of genetic content arrangements (i.e., the distribution of information in the genome) may enable us to investigate species-specific characteristics of genome organization.

To explore this possibility further, I developed the theory of "communicative nature": Living nature is organized and coordinated by communication processes, that is, sign-mediated interactions within and among cells, tissues, organs, and organisms, using a variety of single and group behaviors. If the genetic code has language-like features, it must embody syntactic (combination), pragmatic (context), and semantic (content) semiotic rules, rules that are conserved but under certain circumstances can be changed, rearranged, or even developed *de novo*. The linguistic feature of the genetic code excludes randomly derived sequences. No language-like text emerges as a random mixture of the alphabet or a randomly derived mixture of characters. If the genetic code has language-like features, there must be competent "agents" that generate syntactically correct nucleotide sequences that arrange and rearrange them, repair them if they are damaged, and integrate and delete them. Because no language is spoken by itself and no code can code itself, I tried to identify linguistically competent "agents" in natural nucleotide sequence editing that are responsible for generating meaningful nucleotide sequences that code for the constituent proteins required by all life forms. Editing of meaningful sequences needs "editors" and "agents": Where and what are such agents?

Fortunately, in the last few decades, it has become increasingly evident that cell functions consist of an abundant diversity of accompanying activities that are orchestrated by a variety of regulations. Together, these interconnected and fine-tuned networks represent a "high-definition information-processing organelle" (J. Shapiro). The systematics of interconnections among these cellular activities was deciphered by excellent researchers who followed the lead of Barbara McClintock. I found an excellent description of such

information processing capabilities in the articles of James Shapiro. One of his articles was published in the proceedings of a congress held in 2001 called "Contextualizing the Genome: The Role of Epigenetics in Genetics, Development and Evolution." The other articles in the proceedings outlined the fact that through different reading patterns ("alternative splicing pathways generate different mRNAs from a single gene") of the genetic text induced by epigenetic regulation, such as methylation, histone modification, etc., the DNA information storage medium contains not just one meaning/function but many meanings/functions for different purposes, such as those needed in different developmental stages. The philosophical consequences of this fact were outlined by the editors, Getrudis Van de Vijver and colleagues in the first chapter of the 2001 congress proceedings. The role of repetitive elements in particular attracted my attention because they seemed to be one class of the natural genetic engineering elements proposed by James Shapiro.

After reading articles by Eva Jablonka and, later on, Frank Ryan, my attention was drawn to a book by the virologist Luis P. Villarreal, *Viruses and the Evolution of Life*. After reading the book it immediately became clear that viruses could be responsible for natural genome editing, that is, viruses would function as competent agents of integrating genetic content and genomic architecture. That is to say, viruses especially the great abundance and variety of persistent nonlytic settlers of host genomes or cytoplasm, are "linguistically competent" to combine, recombine, arrange, rearrange, repair, insert, delete, or even generate nucleotide sequences *de novo*. In addition they are evolutionarily older than cellular life forms and are coevolutionary obligate settlers of all cellular life.

I mentioned this to Alfred Winter, who works for the government of Salzburg County in Austria as one of the most successful managers in cultural affairs, and as he did with the "Gathering in Biosemiotics 6," held in 2006 in Salzburg, he immediately encouraged me to organize a new symposium and invite experts. Together with Erich Hamberger (Communication Science, University of Salzburg) and Hiltrud Oman (head administrator of the Gathering in Biosemiotics 6), I began to organize the symposium "Natural Genetic Engineering and Natural Genome Editing," which met in 2008 in Salzburg. Several articles I produced between 2006 and 2007 on biocommunication in plants, corals, fungi, bacteria, and viruses helped identify participants and many of the keynote speakers. In addition, James Shapiro and Luis Villarreal suggested other experts.

The official goals for this symposium made it clear that the presentations could serve as an outlook to the 21st century of life sciences. Over the past several years, the concept of natural genetic engineering has been advanced to encompass biochemical functions that make up the cellular toolbox for changing genome sequence composition and organization. Natural genetic engineering activities range from the introduction of point mutations by mutator polymerases, to large-scale chromosome rearrangements mediated by transposable elements and nonhomologous end joining, to incorporation of viral and microbial DNA into the genomes of host organisms. Recent literature on whole-genome sequences provides abundant evidence for the action of natural genetic engineering in evolution. Discoveries about natural genetic engineering have coincided with rapid progress in our understanding of epigenetic control and RNA-directed chromatin formation. Both natural genetic engineering and chromatin formatting exemplify the "read–write" potential of the genome as an information storage organelle. Special attention needs to be paid to the role of viruses and other so-called parasitic elements in the origin of genome formatting and natural genetic engineering capabilities. A critical

question concerns the role of nonrandom genetic change operators in the production of complex evolutionary inventions. The purpose of the symposium was to bring together scientists working on genome organization, genome restructuring, genome formatting, and virus research to discuss how we could integrate these discoveries into our basic understanding of evolution, development, and disease.

The constructed sections that shaped the symposium are nearly identical to those presented in this volume. In the first part, "Information Processing Replaces Mechanics," James Shapiro pictures the radical discoveries between the elaboration of the central dogma of molecular biology and the current understanding of cell function that contradicts atomistic pre-DNA ideas of genome organization and violates the central dogma at multiple points. John Mattick proposes a new view on the evolution and genetic programming of complex organisms, suggesting that they have largely evolved by constructing more elaborate regulatory networks transacted by regulatory RNAs. Eugene Koonin describes a model of a precellular stage of biological evolution of inorganic compartments that harbored a diverse mix of virus-like genetic elements in which he not only recapitulates early ideas of J.B.S. Haldane but argues that key components of cellular life originated as components of virus-like entities. Patrick Forterre and David Prangishvili confirm the major roles of viruses in biological evolution: If capsid-encoding organisms (viruses) and ribosome-encoding organisms (cells) are the major types of living entities on our planet, it seems logical to conclude that their conflict has been the major engine of biological evolution. Eshel Ben-Jacob demonstrates that bacteria together have developed strategies of cooperation through intricate communication capabilities, such as quorum sensing, chemotactic signaling, and the exchange of genetic information.

The second part, "Viral Infection-driven Eukaryotic Evolution" opens with Philip Bell and his viral "eukaryogenesis" hypothesis in which he proposes a key role for viruses in the emergence of eukaryotes from a prokaryotic world environment. František Baluška reflects on cell–cell channels, viruses, and evolution, with the result that infection, parasitism, and symbiosis played major roles in the evolution of higher levels of biological complexity. This contribution is confirmed by two more detailed investigations: Manfred Heinlein reports about methods of viral infection through plant cell–cell channels via virus-encoded movement proteins with a variety of signal-mediated interactions, such as protection of viral RNA against host plant defense, and genetic as well as epigenetic changes in the progeny of infected plants. Amin Rustom investigates tunneling nanotube-related membrane connections among mammalian cells, which function in a similar way as plant cell–cell channels of plants to serve cellular transport needs and viral spread.

The third part, "Communal Evolution," begins with Lorraine Olendzenski and Peter Gogarten. In their contribution on the tree/web of life in light of horizontal gene transfer, they show that gene transfer between divergent organisms may provide an adaptive advantage that is even more pronounced within closely related organisms. The latter transfers may be neutral or nearly neutral for the recipient. Bruce Webb, Tonja Fisher, and Tyasning Nusawardani report on polydnaviruses as obligate symbionts of some parasitic wasps, with dramatic effects on both the viral genome and the delivery of viral genes into the wasp genome. Mariana Varela, Thomas Spencer, Massimo Palmarini, and Frederick Arnaud investigate friendly viruses that are observed in the special relationship between sheep and retroviruses, which clearly demonstrates the co-evolution between endogenous retroviruses and their mammalian hosts. Mahmoud El Hefnawi, Wessam

H. El Behaidy, Lamya A. El Housseiny, and Suher Zada show that via natural genetic engineering of hepatitis C virus NS5a for immune system counterattack, this protein plays an important role in destabilizing the cell environment and facilitating cancer.

In the fourth section, "Modular Interacting Agents," Jürgen Brosius favors Ed Trifonov's definition of a code as a more complex view of the gene as an entity composed of many subgenic modules. Luis Villarreal suggests that a consortium of persistent, nonlytic viruses constituted the adaptive immune system in jawed vertebrates for the first time. The origin of the adaptive immune system most likely occurred by massive colonization events with endogenous retroviruses. According to this unexpected and novel view on the evolution of the adaptive immune system the editor encouraged Luis Villarreal to write a more detailed contribution. Jean-Nicolas Volff demonstrates that cellular genes in animal genomes are derived from retrotransposons and retroviruses. Günther Witzany proposes some agents of natural genome editing. Noncoding RNAs could be adapted versions of persistent viral agents that now act as modular tools for cellular needs. Nika Lovšin and Matija Peterlin demonstrate that the APOBEC3 protein family protects against infections of some retroviruses, which indicates that this protein family is a remnant of a persistent retroviral infection event that now wards off competing genetic parasites.

In the fifth part, "Epigenetic Control," I. King Jordan and Ahsan Huda report on some of the many ways that transposable elements have contributed to the epigenetic regulation of human genes and are distributed nonrandomly along chromosomes. Marcella Faria and Maria Carolina Elias investigate epigenetic controls in *Trypanosoma cruzi* and show how RNA interference is lacking in their genomes, which could be seen as a symptom of alternative epigenetic controls orchestrated by parasite–host interactions. Ehud Lamm shows that the network perspective dissolves the distinction between regulatory architecture and regulatory state, consistent with the theoretical impossibility of distinguishing *a priori* between "program" and "data" and its consequences for understanding the evolution of biological categories such as epigentic–genetic. Gertrudis Van de Vijver reflects on the meaning of current epigenetic developments in biology and the consequences for the idea of a contextual, stratified determination of living systems.

In addition to the authors that have contributed to this volume, Gil Ast, Nigel Goldenfeld, Shiv Grewal, Erich Hamberger, Kalin Vetsigian, Jerica Sabotic, and Reinhard Vlasak gave presentations. The symposium was organized in cooperation with Schatzkammer Land Salzburg, Kulturelle Sonderprojekte (Alfred Winter), and in partnership with the Leopold Kohr Academy. I would also like to thank the members of the organizing committee, Peter Eckl and Nikolaus Bresgen (Department of Cell Biology, University of Salzburg); Hiltrud Oman and Alexandra Parigiani (head of administration, organization, and reception); and Pierre Madl (Department of Biophysics, University of Salzburg) for technical support. The symposium was sponsored by the Schatzkammer Land Salzburg, Kulturelle Sonderprojekte, Tecan Sales Austria, the Austrian Federal Ministry of Science and Research, the University of Salzburg, and the Leopold Kohr Academy. We gratefully acknowledge this financial support. Finally, I would like to thank the *Annals* staff of the New York Academy of Sciences for guiding this book to press with patience and professional fine-tuning, especially Douglas Braaten and Ralph Brown.

GÜNTHER WITZANY
Telos-Philosophische Praxis
Salzburg, Austria

Revisiting the Central Dogma in the 21st Century

James A. Shapiro

Department of Biochemistry and Molecular Biology, University of Chicago, Gordon Center for Integrative Science, Chicago, IL, USA

Since the elaboration of the central dogma of molecular biology, our understanding of cell function and genome action has benefited from many radical discoveries. The discoveries relate to interactive multimolecular execution of cell processes, the modular organization of macromolecules and genomes, the hierarchical operation of cellular control regimes, and the realization that genetic change fundamentally results from DNA biochemistry. These discoveries contradict atomistic pre-DNA ideas of genome organization and violate the central dogma at multiple points. In place of the earlier mechanistic understanding of genomics, molecular biology has led us to an informatic perspective on the role of the genome. The informatic viewpoint points towards the development of novel concepts about cellular cognition, molecular representations of physiological states, genome system architecture, and the algorithmic nature of genome expression and genome restructuring in evolution.

Key words: **biological theory; evolutionary theory; genome system architecture; cognition; informatics**

The Irony of Molecular Biology

When the structure of DNA was figured out in 1953, there was a strong belief among the pioneers of the new science of molecular biology that they had uncovered the physicochemical basis of heredity and fundamental life processes.[1] Following discoveries about the process of protein synthesis, the consensus view was most cogently summarized a half-century ago in 1958[2] (and then again in 1970[3]) by Crick's declaration of "the central dogma of molecular biology." The concept was that information basically flows from DNA to RNA to protein, which determines the cellular and organismal phenotype. While it was considered a theoretical possibility that RNA could transfer information to DNA, information transfer from proteins to DNA, RNA, or other proteins was considered outside the dogma and "would shake the whole intellectual basis of molecular biology."[3] This DNA/nucleic acid-centered view is still dominant in virtually all public discussions of biological questions, ranging from the role of heredity in disease to arguments about the process of evolutionary change. Even in the technical literature, there is a widespread assumption that DNA, as the genetic material, determines cell action and that observed deviations from strict genetic determinism must be the result of stochastic processes.

The idea of a "dogma" in science has always struck me as inherently self-contradictory. The scientific method is based upon continual challenges to accepted ideas and the recognition that new information inevitably leads to new conceptual formulations. So it seems appropriate to revisit Crick's dictum and ask how it stands up in the light of ongoing discoveries in molecular biology and genomics. The answer is "not well." The last four decades of biomolecular investigation have brought a wealth of discoveries about the informatics of living systems

Address for correspondence: James A. Shapiro, Department of Biochemistry and Molecular Biology, University of Chicago, Gordon Center for Integrative Science, 929 E. 57th Street, Chicago, IL 60637, USA. Voice: 773-702-1625; fax: 773-947-9345. jsha@uchicago.edu

Natural Genetic Engineering and Natural Genome Editing: Ann. N.Y. Acad. Sci. 1178: 6–28 (2009).
doi: 10.1111/j.1749-6632.2009.04990.x

and made the elegant simplifications of the central dogma untenable. Let us review what some of these discoveries have been and see how they revolutionize our concepts of information processing in living cells. The great irony of molecular biology is that it has led us inexorably from the mechanistic view of life it was believed to confirm to an informatic view that was completely unanticipated by Crick and his fellow scientific pioneers.[1]

Basic Molecular Functions

The molecular analysis of fundamental biochemical processes in living cells has repeatedly produced surprises about unexpected (or even "forbidden") activities. A short (and partial) list of these activities provides many illustrative complications or contradictions of the central dogma.

- **Reverse transcription.** The copying of RNA into DNA was predicted by Temin from his studies of RNA tumor viruses that pass through a latent DNA stage.[4] Crick published his 1970 formulation of the central dogma in response to the announcement by Temin and Mitzutani of the discovery of an RNA-dependent DNA polymerase, now called reverse transcriptase.[5] Thus, information can flow from RNA to DNA. We now know that reverse transcriptase activity is present in both prokaryotic and eukaryotic organisms and fulfills a number of different functions related to the modification or addition of genomic DNA sequences. Genome sequencing has revealed abundant evidence of the importance of reverse transcription in genome evolution.[6–8] Indeed, over one-third of our own genomes comes from DNA copies of RNA.[9]
- **Posttranscriptional RNA processing.** Early in the studies of RNA biogenesis, it became apparent that RNA was modified after it was copied from DNA. In some cases, such as tRNA, the modifications altered the individual nucleotides and also involved its cleavage from precursor transcripts.[10,11] With the advent of recombinant DNA technology, it was discovered that many messenger RNAs encoding proteins are processed from initial transcripts by internal cleavage and splicing of intervening sequences.[12,13] We now recognize that differential splicing is an important aspect of biological regulation and differential expression of genomic information.[14,15] In addition, processes of trans-splicing were found to join pieces of two different transcripts[16,17] and RNA editing could alter the base sequence of transcripts.[18,19] Thus, the information content of RNA molecules has many potential inputs besides the sequence of the DNA template for transcription.
- **Catalytic RNA.** Studies of RNA processing by Altman and Cech revealed that some RNA molecules could undergo structural changes in the absence of proteins.[10,20] These discoveries opened the floodgates on the recognition that RNA molecules can have catalytic processes in many ways analogous to those of proteins. This means that RNA plays a more direct role in determining cellular characteristics than the limited protein-coding role assigned by Crick.
- **Genome-wide (pervasive) transcription.** In a widely cited 1980 article published with Leslie Orgel, Crick applied the central dogma view to discriminate genomic DNA into classes that do and do not encode proteins, labeling the latter as "junk DNA" unable to make a meaningful contribution to cell function.[21] One criterion propounded to distinguish informational DNA is whether it is transcribed into RNA. Employing this criterion, the evidence for functionality of all regions of the genome has recently been extended by a detailed investigation of 1% of the human genome.[22] This

study has indicated that virtually all DNA in the genome, most of which does not encode protein, is transcribed from one or both strands.[23] So the central dogma-based notion that the genome can be functionally discriminated into transcribed (informational, coding) and nontranscribed (junk) regions appears to be invalid. There are other reasons for discounting the notion that only protein-coding DNA contains biologically meaningful information.[24]

- **Posttranslation protein modification.** In the early days of molecular biology, it was expected that the rich structural information in protein sequences was sufficient to determine their functional properties. However, biochemical analysis quickly revealed that proteins were subject to functional modulation via an enormous range of covalent alterations after translation on the ribosomes. These modifications included proteolytic cleavage,[25–27] adenylylation,[28] phosphorylation,[29–32] methylation,[33] acetylation,[34,35] attachment of peptides,[36] addition of sugars and polysaccharides,[37–40] decoration with lipids,[41,42] and cis- and trans-splicing.[43] Thus, like RNA, the information content of protein has many potential inputs other than the sequence code maintained in the DNA. It is significant to note that these protein-catalyzed modifications are critical to cellular signal transduction and regulatory circuits. They clearly fall into one of Crick's excluded catgories.[3]
- **DNA proofreading and repair.** In the early days of molecular biology and the central dogma, the stability of genomic information was assumed to be an inherent property of the DNA molecule and the replication machinery. Studies of mutagenesis have revealed that cells possess several levels of protein-based proofreading and error correction systems that maintain the stability of the genome, which is subject to chemical and physical damage, replication errors, and collapse of the replication complex leading to broken DNA molecules.[44–46] In some cases, these protein systems are also responsible for making specific localized changes in the DNA sequence.[47] Thus, the maintenance of genomic information during the replication loop in the central dogma has protein inputs as well.

Cellular Sensing and Intercellular Communication

A major achievement of molecular biology has been the identification of molecules that cells use to acquire information about their chemical, physical, and biological environment and to keep track of internal processes. Many of the biological indicators include molecules produced by the cells themselves. Recognizing the chemical basis for sensing and communication constitutes a major advance in understanding how cells are able to carry out the appropriate actions needed for survival, reproduction, and multicellular development.

- **Allosteric binding proteins.** One of the key triumphs of early molecular biologists was deciphering how small molecules regulate protein synthesis through interactions with DNA-binding transcription factors.[48] This accomplishment was expanded by the more general theory of allosteric transitions in proteins that bind two or more ligands.[49] Binding of one ligand alters the protein shape and alters the interaction with the second ligand. Through these structural and functional alterations, allosteric proteins serve as microprocessors that can transmit information from one cellular component to another.
- **Riboswitches and ribosensors.** The discovery of catalytic RNA led to a dynamic view of RNA structure and function.[50] Information is contained in three-dimensional structure as well as

one-dimensional nucleotide sequence. One aspect of this dynamic view is the realization that RNA can also bind ligands and behave allosterically. Riboswitches, the RNA molecules that bind small molecule ligands and then interact with nucleic acids or proteins, can intervene at all steps in information transfer between the genome and the rest of the cell.[51]

- **Surface and transmembrane receptors.** The first allosteric proteins and RNAs to be studied operated as soluble molecules in the cytoplasm or (in eukaryotic cells) nucleoplasm. Embedded in cell membranes and attached to the cell surface, molecular biologists have identified a wide variety of receptor proteins for detecting extracellular signals, including those indicating the presence of other cells.[52,53] Either the receptors themselves or associated proteins span the cell membrane(s) and transmit external information to the cytoplasm and other cell compartments, including the genome.[54,55]
- **Surface signals.** Complementary to receptors are molecular signals attached to the cell surface that indicate the presence and status of the cell.[56,57] These signals include proteins, polysaccharides, and lipids, and their presence or precise structure can change depending upon cellular physiology, stress, or differentiation. They interact with cognate receptors on other cells.[58] Thus, a great deal of metabolic, developmental, and historical information can be conveyed from one cell to another.[59] Without this kind of information transfer between cell surfaces, successful multicellular development would not be possible.[60]
- **Intercellular protein transfer.** In some cases, multiprotein surface structures serve as conduits for the transmission of proteins from the cytoplasm of one cell to another[61] (see also papers by Baluska, Heinlein, and Rustom from this symposium). Such molecular injections are basic to interkingdom communication in microbial pathogenesis and symbiosis with multicellular hosts.[62–64]
- **Exported signals.** In addition to cell-attached signaling, there is intercellular communication that occurs by molecular diffusion through the atmosphere or aqueous environments. Molecular classes as diverse as gases,[65,66] amino acids or their derivatives,[67] vitamins,[68] oligopeptides,[69] and larger proteins (often decorated with polysaccharide or lipid attachments) serve as alarm signals, hormones, pheromones, and cytokines to carry information between cells that are not in direct contact. Both prokaryotes and eukaryotes use these signals to regulate genetic exchange, homeostasis, metabolism, differentiation, multicellular defense, and morphogenesis.
- **Internal monitors.** The sensory capabilities of cells are not exclusively dedicated to the external chemical or biological environments. Monitoring internal processes and detecting actual or potential malfunctions are critical for reliable cellular reproduction. Molecular studies have revealed a wide range of functions that provide information about the accuracy of DNA replication,[44–46] protein synthesis,[70] membrane composition,[71] and progress through the cell cycle.[72] Current ideas about aberrations in the control of cellular proliferation in cancer attribute a major role to breakdowns in these internal monitoring processes, which often lead to uncontrolled proliferation and genomic instability.

Cellular Control Regimes

As genetic and molecular analysis of cell and organismal phenotypes progressed in the 1970s and 1980s, it quickly became evident that each character depends as much on the cellular functions that regulate expression of genomic information as on the functions that execute the underlying biochemical processes. It is now

taken for granted that every cell process is subject to a control regime that operates algorithmically to adjust to the changing contingencies of both the external and internal environments. Many features of these control regimes have been identified over the past few decades, but it is important to note that we still lack a comprehensive theory of cellular regulation.

- **Feedback regulation circuits.** The molecular analysis of metabolism and protein synthesis at the cellular and multicellular levels has revealed repeated patterns of positive and negative feedback circuitry that is used to achieve and maintain distinct states necessary for reproduction and development.[73] These patterns occur in the control of all cell processes (e.g., replication, transcription, posttranscriptional processing, translation, posttranslational processing, enzyme activity, RNA and protein turnover, etc.), but it is remarkable that the diversity of the molecular components is compatible with a relatively limited set of formal logical descriptions.
- **Signal transduction networks.** Molecular studies of cell growth and differentiation have shown that information about the response to external or internal signals can be transmitted along multimolecular pathways by processes such as sequential protein modifications.[30] These informational transmission chains are often interconnected, so it is more appropriate to describe and analyze them as signal transduction networks than as separate pathways.
- **Second messengers.** In many signal transduction networks, information is transmitted in the form of a small, freely diffusible molecule in the cytoplasm, such as cAMP (used both in pro- and eukaryotes). These cytoplasmic molecules are called second messengers,[74,75] and they constitute chemical symbols of various conditions. In *Escherichia coli*, for example, elevated levels of cAMP represent an absence of glucose in the external environment.[76]
- **Checkpoints.** An important conceptual advance in understanding emergency responses and regulation of the cell cycle was the concept of a checkpoint, a monitoring system that halts progress through the cell cycle until essential preliminary steps have been completed.[77] Concerning the genome, checkpoints have been identified that monitor DNA integrity, completion of DNA replication, and alignment of chromosomes at metaphase.[72] The same concept can be applied to other complex biological processes, such as cellular differentiation and morphogenesis.
- **Epigenetic regulation.** A major focus of current studies on genomic regulation is the control of chromosome regions by alternative chromatin structures. Since chromatin states do not alter DNA sequence but are heritable over many cell generations, and also because chromatin restructuring plays a critical role in cellular differentiation, this control mode is now included under the rubric "epigenetic."[78,79] Epigenetic processes encompass many phenomena, including parental imprinting and erasure of expression states,[80] higher order regulation of multiple linked genetic loci,[81] restriction of genome expression in differentiation,[82] silencing of mobile genetic elements and nearby genetic loci,[83] chromosome position effects,[84] and X chromosome inactivation in mammals.[85] Biochemical analysis has revealed a large number of protein- and DNA-modifying activities that can reformat chromatin from one state to another, often in response to particular stimuli[86,87] or after nuclear transfer.[88]
- **Regulatory RNAs.** Although regulatory RNA molecules had been known for several decades in bacteria, the realization in the 1990s that certain animal "genes" had RNA rather than protein products stimulated extensive research into the role