THE HIDDEN GENETIC

PROGRAM of COMPLEX ORGANISMS

By John S. Mattick

Biologists assumed that proteins alone regulate the genes of humans and other complex organisms. But an overlooked regulatory system based on RNA may hold the keys to development and evolution

BACTERIA AND HUMANS differ greatly in their structural and developmental complexity, but biologists have long assumed that all organisms used the same genetic mechanisms. Yet new work hints that complexity arises from an additional program hidden in "junk" DNA.

Assumptions can be dangerous, especially in science. They usually start as the most plausible or comfortable interpretation of the available facts. But when their truth cannot be immediately tested and their flaws are not obvious, assumptions often graduate to articles of faith, and new observations are forced to fit them. Eventually, if the volume of troublesome information becomes unsustainable, the orthodoxy must collapse.

We may be witnessing such a turning point in our understanding of genetic information. The central dogma of molecular biology for the past half a century and more has stated that genetic information encoded in DNA is transcribed as intermediary molecules of RNA, which are in turn translated into the amino acid sequences that make up proteins. The prevailing assumption, embodied in the credo "one gene, one protein," has been that genes are generally synonymous with proteins. A corollary has been that proteins, in addition to their structural and enzymatic roles in cells, must be the primary agents for regulating the expression, or activation, of genes.

This conclusion derived from studies primarily on bacteria such as *Escherichia coli* and other prokaryotes (simple one-celled organisms lacking a nucleus). And indeed, it is still essentially correct for prokaryotes. Their DNA consists almost entirely of genes encoding proteins, separated by flanking sequences that regulate the expression of the adjacent genes. (A few genes that encode RNAs with regulatory jobs are also present, but they make up only a tiny fraction of most prokaryotes' genetic ensembles, or genomes.)

Researchers have also long assumed that proteins similarly represent and control all the genetic information in animals, plants and fungi—the multicellular organisms classified as eukaryotes (having cells that contain nuclei). Pioneering biologist Jacques Monod summarized the universality of the central dogma as "What was true for *E. coli* would be true for the elephant."

Monod was only partly right. A growing library of results reveals

that the central dogma is woefully incomplete for describing the molecular biology of eukaryotes. Proteins do play a role in the regulation of eukaryotic gene expression, yet a hidden, parallel regulatory system consisting of RNA that acts directly on DNA, RNAs and proteins is also at work. This overlooked RNA-signaling network may be what allows humans, for example, to achieve structural complexity far beyond anything seen in the unicellular world.

Some molecular biologists are skeptical or even antagonistic toward these unorthodox ideas. But the theory may answer some long-standing riddles of development and evolution and holds great implications for gene-based medicine and pharmaceuticals. Moreover, the recent discovery of this system affords insights that could revolutionize designs for complex programmed systems of all kinds, cybernetic as well as biological.

RNAs AND PROTEINS may communicate regulatory information IN PARALLEL.

The Ubiquitous Junk

A DISCOVERY in 1977 presaged that something might be wrong with the established view of genomic programming. Phillip A. Sharp of the Massachusetts Institute of Technology and Richard J. Roberts of New England Biolabs, Inc., and their respective colleagues independently showed that the genes of eukaryotes are not contiguous blocks of protein-coding sequences. Rather they are mosaics of "exons" (DNA sequences that encode fragments of proteins) interspersed with often vast tracts of intervening sequences, or "introns," that do not code for protein. In the nucleus, a gene is first copied in its totality as a primary RNA transcript; then a process called splicing removes the intronic RNAs and reconstitutes a continuous coding sequence—messenger RNA, or mRNA—for translation as protein in the cytoplasm. The excised intronic RNA, serving no apparent purpose, has been presumed to be degraded and recycled.

But if introns do not code for protein, then why are they ubiquitous among eukaryotes yet absent in prokaryotes? Although introns constitute 95 percent or more of the average protein-coding gene in humans, most molecular biologists have considered them to be evolutionary leftovers, or junk. Introns were rationalized as ancient remnants of a time before cellular life evolved, when fragments of protein-coding information crudely assembled into the first genes. Perhaps introns had survived in complex organisms because they had an incidental usefulness—for example, making it easier to reshuffle segments of proteins into useful new combinations during evolution. Similarly, biologists have assumed that the absence of introns from prokaryotes was a consequence of intense competitive pressures in the microbial environment: evolution had pruned away the introns as deadweight.

One observation that made it easier to dismiss introns—and other seemingly useless "intergenic" DNA that sat between genes—as junk was that the amount of DNA in a genome does not correlate well with the organism's complexity. Some amphibians, for example, have more than five times as much DNA as mammals do, and astonishingly, some amoebae have 1,000 times more. For decades, researchers assumed that the underlying number of protein-coding genes in these organisms correlated much better with complexity but that the relationship was lost against the variable background clutter of introns and other junk sequences.

But investigators have since sequenced the genomes of diverse species, and it has become abundantly clear that the correlation between numbers of conventional genes and complexity truly is poor. The simple nematode worm *Caenorhabditis elegans* (made up of only about 1,000 cells) has about 19,000 protein-coding genes, almost 50 percent more than insects (13,500) and nearly as many as humans (around 25,000). Conversely, the relation between the amount of nonprotein-coding DNA sequences and organism complexity is more consistent.

Put simply, the conundrum is this: less than 1.5 percent of the human genome encodes proteins, but most of it is transcribed into RNA. Either the human genome (and that of other complex organisms) is replete with useless transcription, or these nonprotein-coding RNAs fulfill some unexpected function.

This line of argument and considerable other experimental evidence suggest that many genes in complex organisms—perhaps even the majority of genes in mammals—do not encode protein but instead give rise to RNAs with direct regulatory functions [see "The Hidden Genome," by W. Wayt Gibbs, SCIENTIFIC AMERICAN, November and December 2003]. These RNAs may be transmitting a level of information that is crucial, par-

Overview/*Revising Genetic Dogma*

- A perplexingly large portion of the DNA of complex organisms (eukaryotes) seems irrelevant to the production of proteins. For years, molecular biologists have assumed this extra material was evolutionary "junk."
- New evidence suggests, however, that this junk DNA may encode RNA molecules that perform a variety of regulatory functions. The genetic mechanisms of eukaryotes may therefore be radically different from those of simple cells (prokaryotes).
- This new theory could explain why the structural and developmental complexity of organisms does not parallel their numbers of protein-coding genes. It also carries important implications for future pharmaceutical and medical research.

JEFF JOHNSON (*DNA background*); MOREDUN ANIMAL HEALTH LTD/SCIENCE PHOTO LIBRARY (*bacterium*); LENNART NILSSON (*embryo*) (*preceding pages*)

ticularly to development, and that plays a pivotal role in evolution.

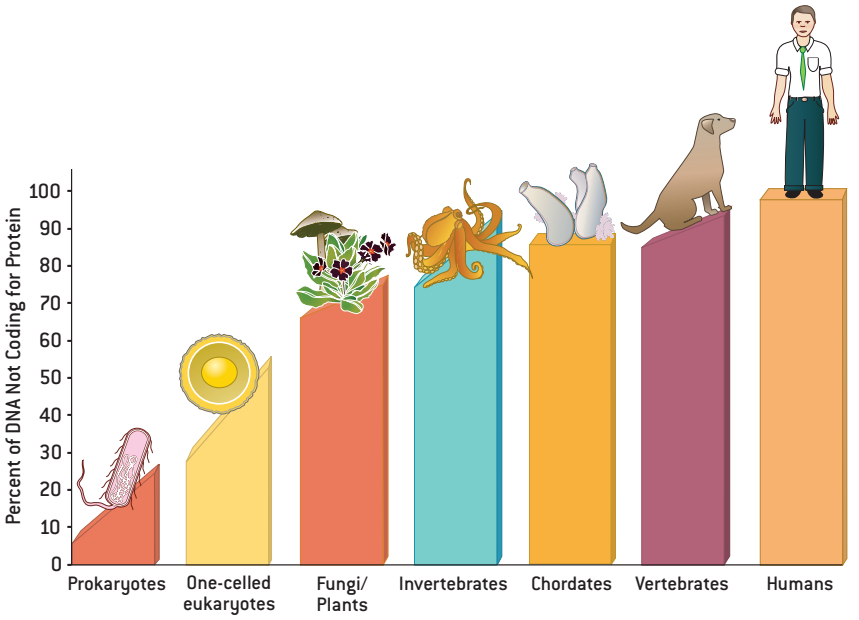
From Parasites to Parallel Controls

THE CLUE to understanding this point may lie in a new interpretation of introns. Contrary to early assumptions that introns generally date back to the dawn of life, evidence amassed more recently indicates that these sequences invaded the genes of higher organisms late in evolution. Most likely, they derived from a type of self-splicing mobile genetic element similar to what are now called group II introns. These elements are parasitic bits of DNA that have the peculiar ability to insert themselves into host genomes and to splice themselves out when expressed as RNA.

Group II introns are found only occasionally in bacteria, and it is easy to see why. Because bacteria lack a nucleus, transcription and translation occur together: RNA is translated into protein almost as fast as it is transcribed from DNA. There is no time for intronic RNA to splice itself out of the protein coding RNA in which it sits, so an intron would in most cases disable the gene it inhabits, with harmful consequences for the host bacterium. In eukaryotes, transcription occurs in the nucleus and translation in the cytoplasm, a separation that opens a window of opportunity for the intron RNA to excise itself. Introns can thus be more easily tolerated in eukaryotes.

Of course, as long as introns needed to splice themselves in and out of genomes, their sequences could not have deviated much from that of group II introns. But a further leap in intron evolution may have accompanied the evolution in eukaryotes of the structure called the spliceosome. This is a complex of small catalytic RNAs and many proteins; its job is to snip intron RNA out of messenger RNA precursors efficiently.

By freeing introns from the need to splice themselves, the spliceosome would in effect have encouraged introns to proliferate, mutate and evolve. Any random mutation in an intron that proved beneficial to the host organism would have been retained by natural selection. Intronic RNAs would therefore be evolving independently and in parallel with proteins. In short, the entry of introns into eukaryotes may have initiated an explosive new round of molecular evolution, based on RNA rather than protein. Instead of being junky molecular relics, introns could have progressively acquired genetic functions mediated by RNA.



NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.

If this hypothesis is true, its meaning may be profound. Eukaryotes (especially the more complex ones) may have developed a genetic operating system and regulatory networks that are far more sophisticated than those of prokaryotes: RNAs and proteins could communicate regulatory information in parallel. Such an arrangement would resemble the advanced information-processing systems supporting network controls in computers and the brain.

We may have totally misunderstood THE NATURE OF THE GENOMIC PROGRAMMING.

Functional jobs in cells routinely belong to proteins because they have great chemical and structural diversity. Yet RNA has an advantage over proteins for transmitting information and regulating activities involving the genome itself: RNAs can encode short, sequence-specific signals as a kind of bit string or zip code. These embedded codes can direct RNA molecules precisely to receptive targets in other RNAs and DNA. The RNA-RNA and RNA-DNA interactions could in turn create structures that recruit proteins to convert the signals to actions.

The bit string of addressing information in the RNA gives this system the power of tremendous precision, just as the binary bit strings used by digital computers do. It is not too much of a stretch to say that this RNA regulatory system would be largely digital in nature.

The evidence for a widespread RNA-based regulatory system is strong, albeit still patchy. If such a system exists, one would expect that many genes might have evolved solely to express RNA signals as higher-order regulators in the network. That appears to be the case: thousands of RNAs that never get translated into protein (noncoding RNAs) have been identified in recent analyses of transcription in mammals. At least half and possibly more than three quarters of all RNA transcripts fit this category.

One would also expect that many of

LUCY READING; SOURCE: R. J. TAFT AND J. S. MATTICK

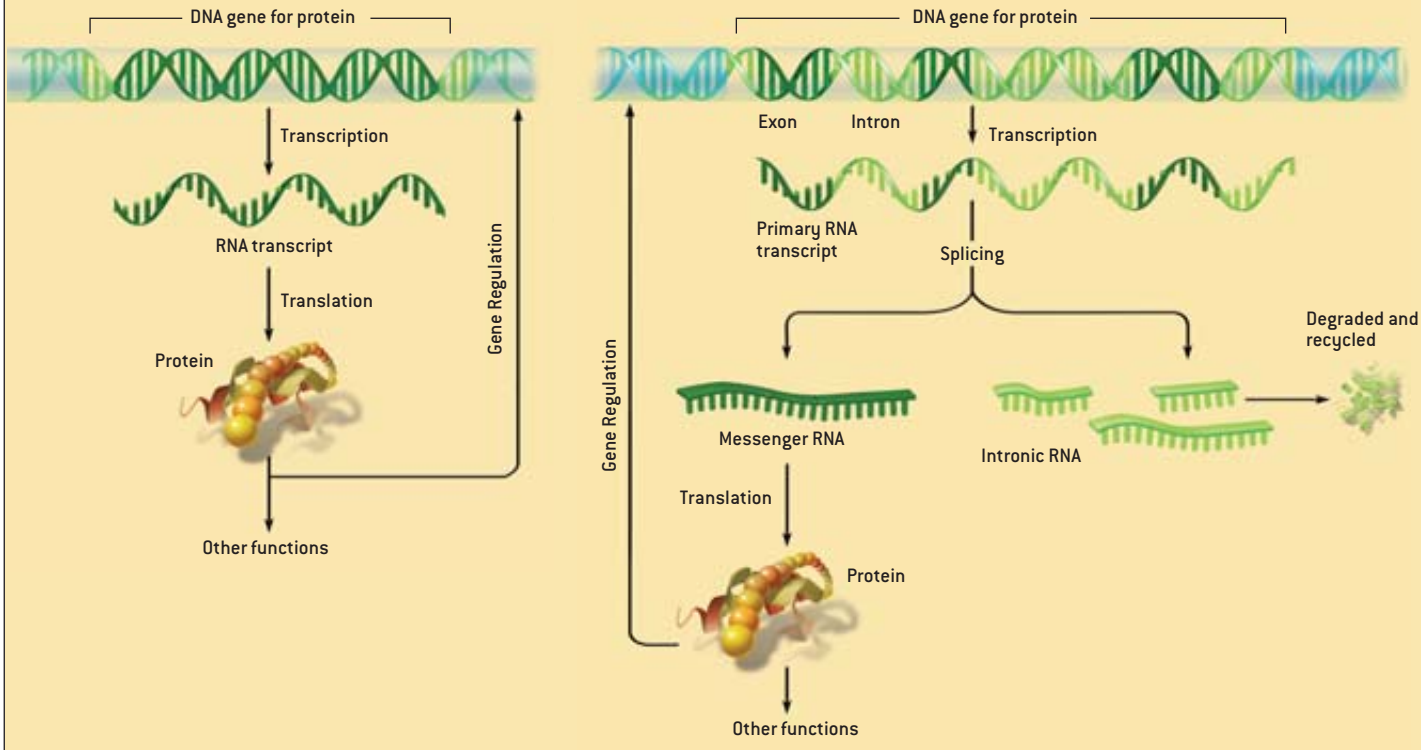
AN EVOLVING VIEW OF GENE ACTIVITY

GENE ACTIVITY IN PROKARYOTES

Prokaryotes (bacteria and other simple cells) have DNA that consists almost entirely of protein-coding genes. When those genes are active, they give rise to RNA transcripts that are immediately translated into proteins, which in turn regulate genetic activity and provide other functions.

TRADITIONAL VIEW OF GENE ACTIVITY IN EUKARYOTES

In the DNA of eukaryotes (complex organisms), individual genes comprise "exon" sequences that code for segments of protein separated by noncoding "intron" sequences. When a gene is active, it is entirely transcribed as RNA, but then the intronic RNA is spliced out and the exonic RNA is assembled as messenger RNA. The cell translates the messenger RNA into protein while breaking down and recycling the intronic RNA, which serves no purpose.



these RNAs might be processed into smaller signals capable of addressing targets in the network. Hundreds of "microRNAs" derived from introns and larger nonprotein-coding RNA transcripts have in fact already been identified in plants, animals and fungi. Many of them control the timing of processes that occur during development, such as stem cell maintenance, cell proliferation, and apoptosis (the so-called programmed cell death that remodels tissues). Many more such small RNAs surely await discovery.

These RNA signals, by finding targets on other RNAs, DNA and proteins, could influence a cell's genetic program in many ways. For example, they could inform various genes that a particular protein-coding sequence has been transcribed, and that feedback could trigger a host of parallel adjustments. More important, however, the RNA signals could serve as a powerful feed-forward program embedded in the genetic material that controls the trajectories of gene expression. If so, they could explain some of the deep mysteries surrounding cell differentiation and organism development.

Regulating Development

CONSIDER WHAT HAPPENS during human embryonic development: a single fertilized cell progresses to become a precisely structured, beautifully sculptured organism of an estimated 100 trillion cells with distinct positions and functions. The pattern of gene expression that makes this transformation possible relies heavily on two phenomena: modification of chromatin and alternative splicing.

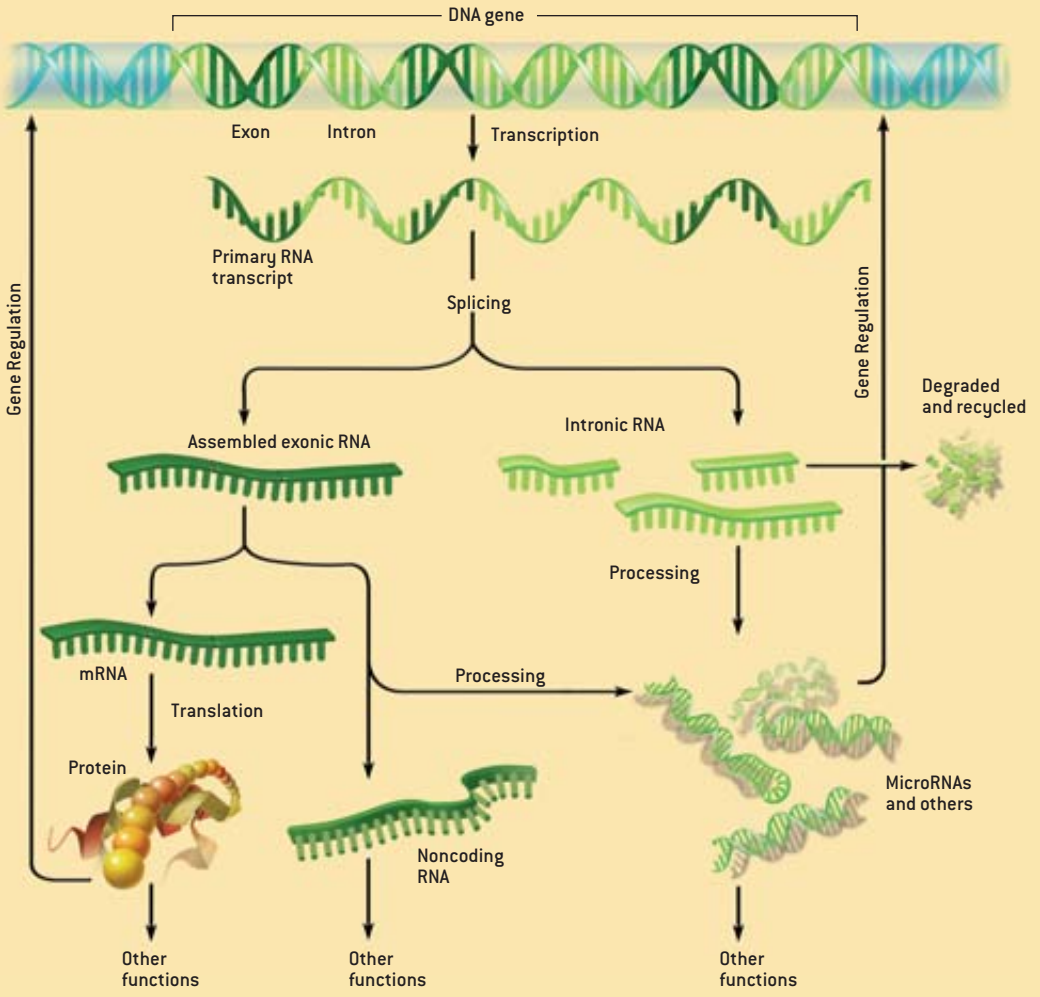
Chromatin is the material that makes up chromosomes; it consists of DNA complexed with proteins. Within cells, small chemical tags (such as methyl and acetyl groups) can attach to segments of the DNA and to the chromatin proteins and thereby determine whether the genes in the associated DNA will be accessible for transcription or will stay silent. Recent results indicate that RNA signaling directs the tagging of the chromatin and thus gene expression. Indeed, a number of complex chromosomal processes, such as mitosis (cell division) and meiosis (the formation of sperm and egg precursors), as well as a range of complex genetic phenomena appear to depend on biochemical pathways that affect RNA processing.

Alternative splicing generates divergent repertoires of RNAs and proteins in the cells of a body's different tissues, all of which share a common set of genes. Most protein-coding transcripts are alterna-

TERESE WINSLOW

NEW VIEW OF GENE ACTIVITY IN EUKARYOTES

Some of the intronic RNA and even some of the assembled exonic RNA may play a direct regulatory role by interacting with the DNA, other RNA molecules or proteins. By modifying protein production at various levels, these noncoding RNAs may superimpose additional genetic instructions on a cell.



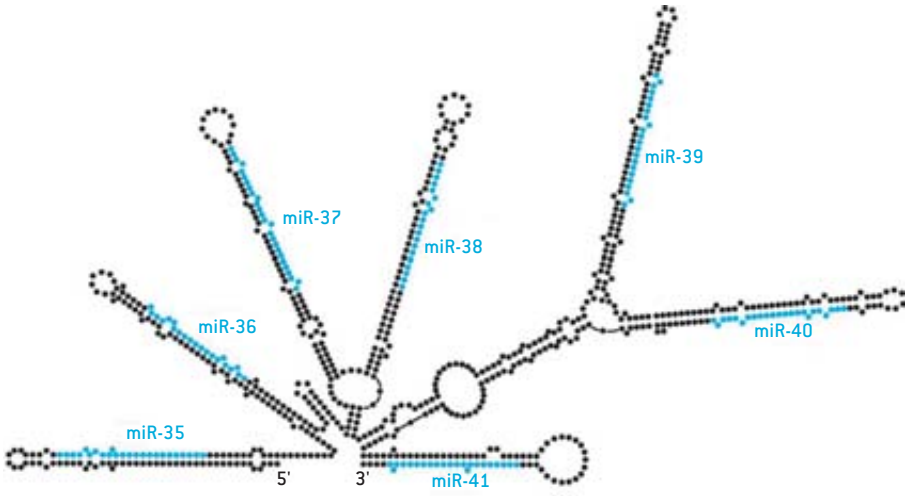
tively spliced in mammals. When intron RNA is spliced out of a gene's transcript, the protein-coding RNA regions may be assembled in more than one way to yield more than one type of protein. The phenomenon is of fundamental importance to animal and plant development, but no one yet understands how cells specify which form of a protein they will make. Few protein factors that control the alternative splicing of specific genes have been found. Consequently, researchers have usually supposed that subtle combinations of general factors activate or repress alternative splicing in different contexts. But no strong evidence has backed up that presumption.

A more likely and mechanistically appealing possibility, however, is that RNAs regulate the process directly. In principle, these molecules could exert exquisitely flexible control by tagging or grabbing particular sequences in primary gene transcripts and steering how the spliceosome joins the pieces. In keeping with that idea, DNA sequences at the intron-exon junctions where alternative splicing occurs are often resistant to change during evolution. Also, a number of laboratories have demonstrated that artificial antisense RNAs designed to bind to such sites can modify splicing patterns in cultured cells, as well as in whole animals. It is perfectly plausible that this phenomenon occurs naturally in vivo, too, but has just not yet been detected.

Controlling Complexity

SUCH CONSIDERATIONS lead naturally to a more general consideration of what type of information, and how much of it, might be required to program the development of complex organisms. The creation of complex objects, whether houses or horses, demands two kinds of specifications: one for the components and one for the system that guides their assembly. (To build a house, one must specify the needed bricks, boards and beams, but one must also have an architectural plan to show how they fit together.) In biology, unlike engineering, both types of information are encoded within one program, the DNA.

The component molecules that make up different organisms (both at the individual and the species levels) are fundamentally alike: around 99 percent of the proteins in humans have recognizable equivalents in mice, and vice versa; many of those proteins are also conserved in other animals, and those involved in basic cellular processes are conserved in all



PROPOSED PRECURSOR molecule for microRNAs is a primary RNA transcript that may produce multiple small RNAs (*blue*). The structure of the precursor might guide the excision of these small RNA signals.

eukaryotes. Thus, the differences in animals' forms surely arise more fundamentally from differences in the architectural information.

Protein-coding genes obviously specify the components of organisms, but where does the architectural information reside? Biologists have widely assumed that the instructions for assembling complex organisms are somehow embedded in the diverse combinations of regulatory factors within cells—that is, in the permutations of regulatory proteins interacting with one another and with the DNA and RNA. Yet, as Daniel C. Dennett of Tufts University has observed, although such combinatorics can generate almost endless possibilities, the vast majority will be chaotic and meaningless—which is problematic for biology. Throughout their evolution and development, organisms must navigate precise developmental pathways that are sensible and competitive, or else they die. Generating complexity is easy; controlling it is not. The latter requires an enormous amount of regulatory information.

Both intuitive and mathematical considerations suggest that the amount of regulation must increase as a nonlinear (usually quadratic) function of the number of genes. So, as the system becomes more complex, an increasing proportion of it must be devoted to regulation. This nonlinear relation between regulation and function appears to be a feature of all integrally organized systems. Therefore, all such systems have an intrinsic complexity limit imposed by the accelerating growth of their control architecture, until or unless the regulatory mechanism changes fundamentally.

Generating COMPLEXITY is easy; CONTROLLING IT IS NOT.

In agreement with this prediction, the number of protein regulators in prokaryotes has been found to increase quadratically with genome size. Moreover, extrapolation indicates that the point at which the number of new regulators is predicted to exceed the number of new functional genes is close to the observed upper limit of bacterial genome sizes.

Throughout evolution, therefore, the complexity of prokaryotes may have been limited by genetic regulatory overhead, rather than by environmental or biochemical factors as has been commonly assumed. This conclusion is also consistent with the fact that life on earth consisted solely of microorganisms for most of its history. Combinatorics of protein interactions could not, by themselves, lift that complexity ceiling.

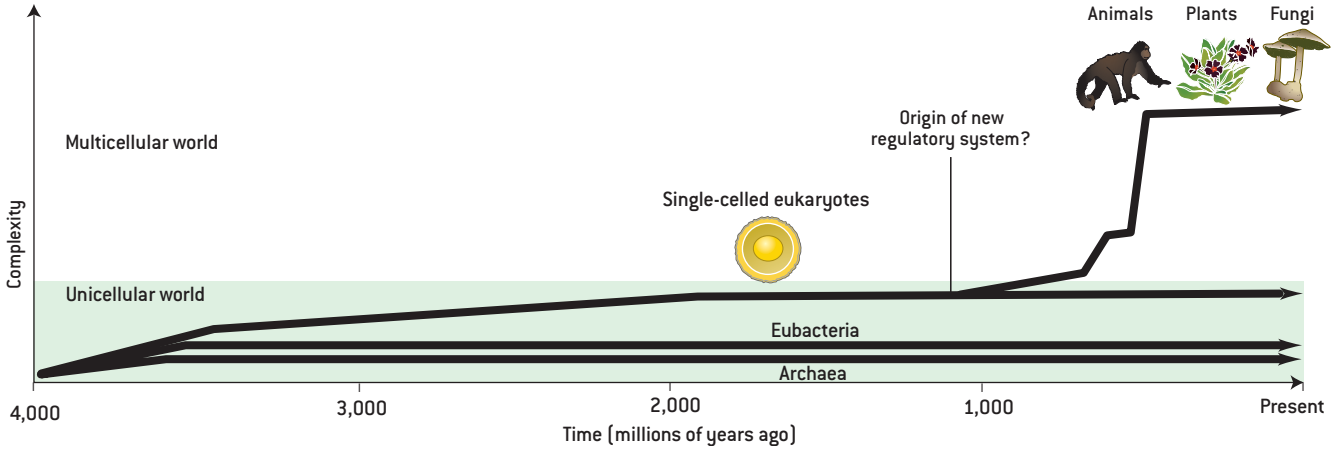
Eukaryotes must have found a solution to this problem. Logic and the available evidence suggest that the rise of multicellular organisms over the past billion years was a consequence of the transition to a new control architecture based largely on endogenous digital RNA signals. It would certainly help explain the phenomenon of the Cambrian explosion about 525 million years ago, when invertebrate animals of jaw-dropping diversity evolved, seemingly abruptly, from much simpler life. Indeed, these results suggest a general rule with relevance beyond biology: organized complexity is a function of regulatory information—and, in virtually all systems, as observed by Marie E. Csete, now at Emory University School of Medicine, and John C. Doyle of the California Institute of Technology, explosions in complexity occur as a result of advanced controls and embedded networking.

The implications of this rule are staggering. We may have totally misunderstood the nature of the genomic programming and the basis of variations in traits among individuals and species. The rule implies that the greater portion of the genomes in complex organisms is not junk at all—rather it is functional and subject to evolutionary selection.

The most recent surprise is that vertebrate genomes contain thousands of noncoding sequences that have persisted virtually unaltered for many millions of years. These sequences are much more highly conserved than those coding for proteins, which was totally unexpected. The mechanism that has frozen these sequences is unknown, but their extreme constancy suggests that they are involved

THE AUTHOR

JOHN S. MATTICK, born and raised in Sydney, today is a professor of molecular biology at the University of Queensland and director of the Institute for Molecular Bioscience. Formerly he was also foundation director of the Australian Genome Research Facility. His accomplished career includes the development of Australia's first genetically engineered vaccine. In 2001 Mattick was appointed an Officer in the Order of Australia, and in 2003 he was awarded the Australian government's Centenary Medal. Married with three sons, he enjoys walking and body surfing as time permits.



UNICELLULAR LIFE, primarily prokaryotes, ruled the earth for billions of years. When multicellular life appeared, however, its complexity rose with dizzying speed. The evolution of an additional genetic regulatory system might explain both the jump to multicellularity and the rapid diversification into complexity.

in complex networks essential to our biology. Thus, rather than the genomes of humans and other complex organisms being viewed as oases of protein-coding sequences in a desert of junk, they might better be seen as islands of protein-component information in a sea of regulatory information, most of which is conveyed by RNA.

The existence of an extensive RNA-based regulatory system also has ramifications for pharmacology, drug development and genetic screening. Traditional genetic diseases such as cystic fibrosis and thalassemia are caused by catastrophic component damage: one of the individual's proteins simply doesn't work. Yet many, if not most, of the genetic variations determining susceptibility to most diseases and underpinning our individual idiosyncrasies probably lie in the noncoding regulatory architecture of our genome that controls growth and development. (Noncoding RNAs have already been linked with several conditions, including B cell lymphoma, lung cancer, prostate cancer, autism and schizophrenia.)

Such defects will not be easy to identify by molecular genetic epidemiology, nor will they necessarily be easy to correct. But understanding this regulatory system may ultimately be critical to understanding our physical and psychological individuality, as well as trait variation in plants and animals. It may also be the prelude to sophisticated strategies for medical intervention to optimize health and for truly advanced genetic engineering in other species.

Aside from introns, the other great source of presumed genomic junk—accounting for about 40 percent of the human genome—comprises transposons and other repetitive elements. These sequences are widely regarded as molecular parasites that, like introns, colonized our genomes in waves at different times in evolutionary history. Like all immigrants, they may have been unwelcome at first, but once established in the community they and their descendants progressively became part of its dynamic—changing, contributing and evolving with it.

Good, albeit patchy, evidence suggests that transposons contribute to the evolution and genomic regulation of higher organisms and may play a key role in epigenetic inheritance (the modification of genetic traits). Moreover, this past July Erev Y. Levanon of Compugen and colleagues elsewhere announced an exciting discovery involving a process called A-to-I (adenosine-to-inosine) editing, in which an RNA sequence changes at a very specific site. They demonstrated that A-to-I editing of RNA transcripts is two orders of magnitude more widespread in humans than was previously thought and overwhelmingly occurs in repeat sequences called Alu elements that reside in noncoding RNA sequences. A-to-I editing is particularly active in the brain, and aberrant editing has been associated with a range of abnormal behaviors, including epilepsy and depression.

Although RNA editing occurs to some extent in all animals, Alu elements are unique to primates. An intriguing possibility is that the colonization of the primate lineage by Alu elements made it possible for a new level of complexity to arise in RNA processing and allowed the programming for neural circuitry to become more dynamic and flexible. That versatility may have in turn laid the foundation for the emergence of memory and higher-order cognition in the human species.

Finally, understanding the operation of the expanded and highly sophisticated regulatory architecture in the genomes of complex organisms may shed light on the challenges of designing systems capable of self-reproduction and self-programming—that is, true artificial life and artificial intelligence. What was dismissed as junk because it was not understood may well turn out to hold the secrets to human complexity and a guide to the programming of complex systems in general. SA

MORE TO EXPLORE

Darwin's Dangerous Idea. Daniel C. Dennett. Simon and Schuster, 1995.

Challenging the Dogma: The Hidden Layer of Non-Protein-Coding RNAs In Complex Organisms. John S. Mattick in *Bio£ssays,* Vol. 25, No. 10, pages 930–939; October 2003.

The Unseen Genome: Gems among the Junk. W. Wayt Gibbs in *Scientific American,* Vol. 289, No. 5, pages 46–53; November 2003.

Noncoding RNAs: Molecular Biology and Molecular Medicine. Edited by J. Barciszewski and V. A. Erdmann. Landes Bioscience/Eurekah.com, Georgetown, Tex., 2003.

More information and lists of publications will be found at the author's Web site (under construction) at **http://imb.uq.edu.au/groups/mattick**

LUCY READING