The old axiom "one gene, one protein" no longer holds true. The more complex an organism, the more likely it became that way by extracting multiple protein meanings from individual genes

The Alternative Genome

BY GIL AST

GENES of mice and men are 88 percent alike. Many of the ways that humans differ from rodents arise from how we edit our genetic information.

Spring of 2000 found molecular biologists placing dollar bets,

trying to predict the number of genes that would be found in the human genome when the sequence of its DNA nucleotides was completed. Estimates at the time ranged as high as 153,000. After all, many said, humans make some 90,000 different types of protein, so we should have at least as many genes to encode them. And given our complexity, we ought to have a bigger genetic assortment than the 1,000-cell roundworm, *Caenorhabditis elegans,* which has a 19,500-gene complement, or corn, with its 40,000 genes.

When a first draft of the human sequence was published the following summer, some observers were therefore shocked by the sequencing team's calculation of 30,000 to 35,000 protein-coding genes. The low number seemed almost embarrassing. In the years since, the human genome map has been finished and the gene estimate has been revised downward still further, to fewer than 25,000. During the same period, however, geneticists have come to understand that our low count might actually be viewed as a mark of our sophistication because humans make such incredibly versatile use of so few genes.

Through a mechanism called alternative splicing, the information stored in the genes of complex organisms can be edited in a variety of ways, making it possible for a single gene to specify two or more distinct proteins. As scientists compare the human genome to those of other organisms, they are realizing the extent to which alternative splicing accounts for much of the diversity among organisms with relatively similar gene sets. In addition, within a single organism, alternative splicing allows different tissue types to perform diverse functions working from the same small gene assortment.

Indeed, the prevalence of alternative splicing appears to increase with an organism's complexity—as many as three quarters of all human genes are subject to alternative editing. The mechanism itself probably contributed to the evolution of that complexity and could drive our further evolution. In the shorter term, scientists are also beginning to understand how faulty gene splicing leads to several cancers and congenital diseases, as well as how the splicing mechanism can be used therapeutically.

Overview/*Cut-and-Paste Complexity*

- A gene's instructions can be edited by cellular machinery to convey multiple meanings, allowing a small pool of protein-coding genes to give rise to a much larger variety of proteins.
- That such alternative splicing of genetic messages is possible was long understood. But only when the genome sequences of humans and other organisms became available for side-by-side comparison did geneticists see how widespread alternative splicing is in complex organisms and how much the mechanism contributes to differentiating creatures with similar gene sets.
- Alternative splicing enables a minimal number of genes to produce and maintain highly complex organisms by orchestrating when, where and what types of proteins they manufacture. Humans, in turn, may soon be able to regulate our own gene splicing to combat disease.

Pivotal Choices

THE IMPORTANCE of alternative editing to the functioning of many organisms cannot be overestimated. For example, life and death depend on it—at least when a damaged cell must determine whether to go on living. Each cell constantly senses the conditions inside and outside itself, so that it can decide whether to maintain growth or to self-destruct in a preprogrammed process known as apoptosis. Cells that cannot repair DNA will activate their apoptotic program. Craig B. Thompson of the University of Pennsylvania and his colleagues have recently shown that a gene called *Bcl-x,* which is a regulator of apoptosis, is alternatively spliced to produce either of two distinct proteins, Bcl-x(L) and Bcl-x(S). The former suppresses apoptosis, whereas the latter promotes it.

The initial discovery that cells can give rise to such different forms of protein from a single gene was made some 25 years ago, but the phenomenon was considered rare. Recent genome comparisons have revealed it to be both common and crucial, adding a dramatic new twist to the classical view of how information

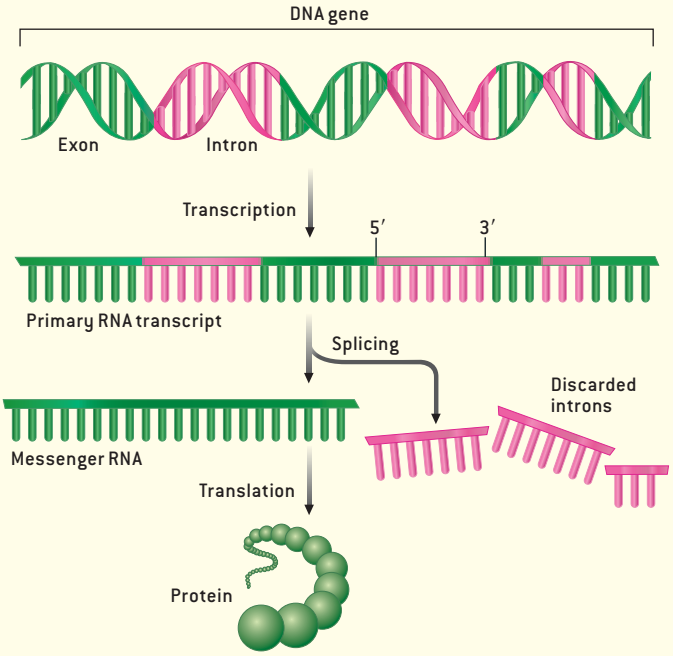
BILL STEELE (*preceding pages*)

ONE GENE, MANY PROTEINS

The classical view of gene expression was simple: a DNA gene is first transcribed into RNA form, then cellular splicing machinery edits out "junk" stretches called introns and joins meaningful portions called exons into a final messenger RNA (mRNA) version, which is then translated into a protein. As it turns out, these rules do not always apply. In complex organisms, the initial RNA transcript can be alternatively spliced—exons may be discarded and introns, or portions of them, retained—to produce a variety of mRNAs, and thus different proteins, from a single gene.

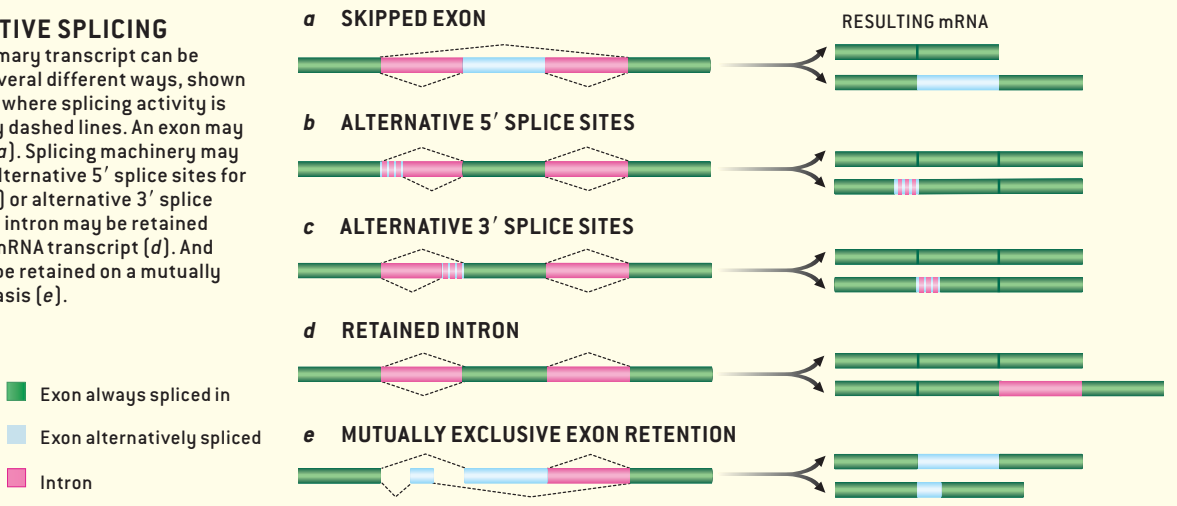
CLASSIC GENE EXPRESSION

A DNA sequence is transcribed into a single-stranded copy made of RNA. Cellular machinery then "splices" this primary transcript: introns—each of which is defined by distinctive nucleotide sequences at its beginning and end, known, respectively, as the 5′ (five-prime) and 3′ (three-prime) splice sites—are removed and discarded while exons are joined into an mRNA version of the gene that will be translated into a protein by the cell.



ALTERNATIVE SPLICING

A gene's primary transcript can be edited in several different ways, shown at the right, where splicing activity is indicated by dashed lines. An exon may be left out (*a*). Splicing machinery may recognize alternative 5′ splice sites for an intron (*b*) or alternative 3′ splice sites (*c*). An intron may be retained in the final mRNA transcript (*d*). And exons may be retained on a mutually exclusive basis (*e*).



stored in a gene is translated into a protein. Most of the familiar facts still hold true: whole genomes contain all the instructions necessary for making and maintaining an organism, encoded in a four-letter language of DNA nucleotides (abbreviated A, G, C and T). In human chromosomes, roughly three billion nucleotides are strung together on each of two complementary strands that form a double helix. When the time comes for a gene's instructions to be "expressed," the double-stranded zipper of DNA opens just long enough for a single-stranded copy of the gene's sequence to be manufactured from a chemical cousin, RNA. Each sequence of DNA nucleotides that gets transcribed into an RNA version in this manner is called a gene. Some of the resulting RNA molecules are never translated into proteins but rather go on to perform housekeeping and regulatory functions within the cell [see "The Unseen Genome: Gems among the Junk,"

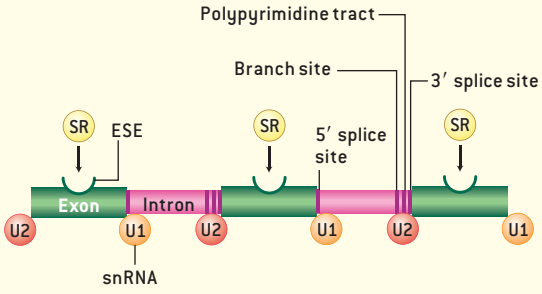
LUCY READING-IKKANDA

THE SPLICING MACHINE

Once a primary RNA transcript of a gene has been created, a structure called the spliceosome carries out RNA editing. In complex organisms, this process is controlled by splicing regulatory (SR) proteins that define exons and direct the spliceosome to specific splice sites. These regulatory molecules therefore determine when and how alternative splices of a gene will be generated. SR proteins are themselves produced in varying forms in different tissues and cell types or during different stages of development within the same tissue.

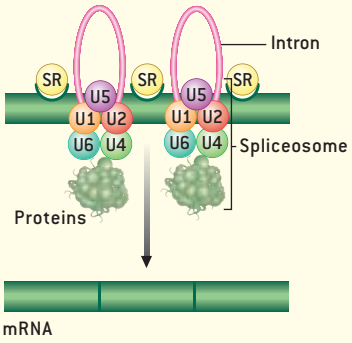
EXON DEFINITION

An SR protein binds to each exon in the transcript at a distinctive nucleotide sequence called an exonic splicing enhancer (ESE). The SR protein's binding defines the exon for the splicing machinery by recruiting small nuclear RNA (snRNA) molecules called U1 and U2 to splice sites on adjacent introns.



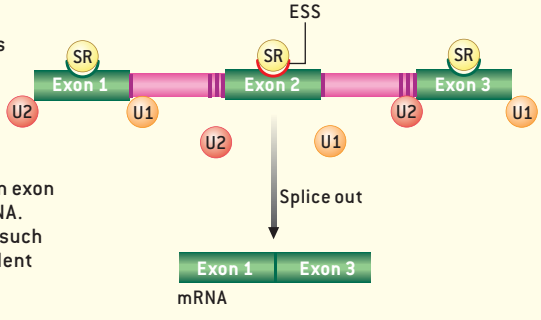
SPLICEOSOME FORMATION

When the original snRNAs have recognized the intron's splice sites, they form a complex with additional snRNAs and more than 100 proteins. This spliceosome complex snips out the introns and joins the exons to produce the mature mRNA.



SPLICING SUPPRESSION

An SR protein may also suppress rather than enhance the binding of snRNAs, in which case the sequence to which it binds is called an exonic splicing suppressor (ESS). The SR protein can thus cause an exon to be spliced out of the final mRNA. In humans and other mammals, such exon skipping is the most prevalent form of alternative splicing.



by W. Wayt Gibbs; Scientific American, November 2003]. The RNA transcripts of genes that do encode a protein will ultimately be read by cellular machinery and translated into a corresponding sequence of amino acids. But first the preliminary transcript must undergo an editing process.

In 1977 Phillip A. Sharp of the Massachusetts Institute of Technology and Richard J. Roberts of New England Biolabs discovered that these initial, or primary, RNA transcripts are like books containing many nonmeaningful chapters inserted at intervals within the text. The nonsense chapters, called introns, must be excised and the meaningful chapters connected together for the RNA to tell a coherent story. In the cutting-and-ligation process, known as splicing, the introns are snipped out of the primary transcript and discarded. Segments of the transcript containing meaningful protein-coding sequences, called exons, are joined together to form a final version of the transcript, known as messenger RNA (mRNA) [*see box on preceding page*].

But by 1980 Randolph Wall of the University of California at Los Angeles had shown that this basic view of pre-mRNA splicing, in which all introns are always discarded and all exons are always included in the mRNA, does not invariably hold true. In fact, the cellular machinery can "decide" to splice *out* an exon or to leave an intron, or pieces of it, in the final mRNA transcript. This ability to alternatively edit pre-mRNA transcripts can significantly increase any gene's versatility and gives the splicing mechanism tremendous power to determine how much of one type of protein a cell will produce over the other possible types encoded by the same gene.

In 1984 Tom Maniatis, Michael Green and their colleagues at Harvard University developed a test-tube procedure to reveal the molecular machinery that performs the cutting of introns and pasting together of exons. Details of its workings, and of the regulatory system controlling it, are still being filled in, but this research is unveiling an exquisitely intricate system with fascinating origins.

The Splicing Machine

In complex organisms, two distinct levels of molecular equipment are involved in splicing pre-mRNA transcripts. The so-called basal machinery, which is found in all organisms whose genomes contain introns, has been highly conserved through evolutionary time, from yeast to humans. It consists of five small nuclear RNA (snRNA) molecules, identified as U1, U2, U4, U5 and U6. These molecules come together with as many as 150 proteins to form a complex called the spliceosome that is responsible for recognizing the sites where introns begin and end, cutting the introns out of the pre-mRNA transcript and joining the exons to form the mRNA.

Four short nucleotide sequences

LUCY READING-IKKANDA

within introns serve as signals that indicate to the spliceosome where to cut [*see box on opposite page*]. One of these splicing signals sits at the beginning of the intron and is called the 5′ (five-prime) splice site; the others, located at the end of the intron, are known as the branch site, the polypyrimidine tract and, finally, the 3′ (three-prime) splice site.

A separate regulatory system controls the splicing process by directing the basal machinery to these splice sites. More than 10 different splicing regulatory (SR) proteins have been identified. Their forms may vary in different tissues or stages of development in the same tissue. SR proteins can bind to short nucleotide sequences within the exons of the pre-mRNA transcript. These binding sites are known as exonic splicing enhancers (ESE) because when the appropriate SR protein binds to an ESE, that action recruits the basal machinery's snRNAs to the splice sites adjacent to either end of the exon. Yet an SR protein can also bind to an exonic splicing suppressor (ESS) sequence within the exon, which will suppress the basal machinery's ability to bind to the ends of that exon and result in its being spliced out of the final mRNA.

The effect of skipping just one exon can be dramatic for an organism. In fruit flies, for example, alternative splicing regulates the sex-determination pathway. When a gene called *Sex-lethal* is expressed, a male-specific exon may be skipped during splicing, leading to the synthesis of a female-specific Sex-lethal protein. This protein can then bind to any subsequent pre-mRNA transcripts from the same gene, ensuring that all further splicing events will continue to cut out the male-specific exon and guaranteeing that only the female-specific protein will be synthesized. If the male-specific exon is spliced in during the first round of editing, however, a nonfunctional mRNA results, which commits the fly's cells to the male-specific pathway.

The effect of skipping just one exon can be dramatic for an organism.

Exon skipping is the most common type of alternative splicing found in mammals. But several other kinds have also been identified, including one that causes introns to be retained in mature mRNA, which is most prevalent in plants and lower multicellular lifeforms. Intron retention is probably the earliest version of alternative splicing to have evolved. Even today the splicing machinery of single-celled organisms, such as yeast, operates by recognizing introns, in contrast with the SR protein system of higher organisms, which defines exons for the basal machinery.

In the unicellular system the splicing machinery can recognize only intronic sequences of fewer than 500 nucleotides, which works fine for yeast because it has very few introns, averaging just 270 nucleotides long. But as genomes expanded during evolution, their intronic stretches multiplied and grew, and cellular splicing machinery was most likely forced to switch from a system that recognizes short intronic sequences within exons to one that recognizes short exons amid a sea of introns. The average human protein-coding gene, for example, is 28,000 nucleotides long, with 8.8 exons separated by 7.8 introns. The exons are relatively short, usually about 120 nucleotides, whereas the introns can range from 100 to 100,000 nucleotides long.

The size and quantity of human introns—we have the highest number of introns per gene of any organism—raises an interesting issue. Introns are an expensive habit for us to maintain. A large fraction of the energy we consume every day is devoted to the maintenance and repair of introns in their DNA form, transcribing the pre-mRNA and removing the introns, and even to the breakdown of introns at the end of the splicing reaction. Furthermore, this system can cause costly mistakes. Each miscut and ligation of pre-mRNA leads to a change in the gene transcript's protein-coding sequence and possibly to the synthesis of a defective protein.

For instance, an inherited disease that I am investigating, familial dysautonomia, results from a single-nucleotide mutation in a gene called *IKBKAP* that causes it to be alternatively spliced in nervous system tissues. The resulting decreased availability of the standard IKBKAP protein leads to abnormal development of the nervous system, and about half of all patients with this disease die before the age of 30. At least 15 percent of the gene mutations that produce genetic diseases (and probably certain cancers as well) do so by affecting pre-mRNA splicing. So why has evolution preserved such a complicated system that is capable of causing disease? Perhaps because the benefits outweigh the risks.

Advantages in Alternatives

BY GENERATING more than one type of mRNA molecule and, therefore, more than one protein per gene, alternative splicing certainly allows humans to manufacture more than 90,000 proteins without having to maintain 90,000 genes. On average, each of our genes generates about three alternatively spliced mRNAs. Still, that number does not explain our need for so many introns

THE AUTHOR

GIL AST is a senior lecturer in the department of human genetics and molecular medicine at the Tel Aviv University Medical School in Israel. His research focuses on the molecular mechanics of pre-mRNA splicing, the evolution and regulation of alternative splicing, and splicing defects associated with cancers and inherited diseases. He recently collaborated with scientists at Compugen to develop a bioinformatics system for predicting alternative-splicing events to detect novel protein variants.

and why they occupy the vast majority of real estate within genes, leaving exonic sequences to make up only 1 to 2 percent of the human genome.

After the sequencing teams had revealed this seemingly empty genomic landscape in 2001, yet another enigma arose when the mouse genome was published in 2002. It turned out that a mouse possesses almost the same number of genes as a human. Although approximately 100 million years have passed since we had a common ancestor, the vast majority of human and mouse genes derive from that ancestor. Most of these share the same intron and exon arrangement, and the nucleotide sequences within their exons are also conserved to a high degree. So the question becomes, if so little differs between the genomes of humans and mice, what makes us so vastly different from the rodents?

Christopher J. Lee and Barmak Modrek of U.C.L.A. recently revealed that one quarter of the alternatively spliced exons in both genomes are specific either to human or mouse. Thus, these exons have the potential to create species-specific proteins that could be responsible for diversification between species. Indeed, one group of alternatively spliced exons is unique to primates (humans, apes and monkeys) and might have contributed to primates' divergence from other mammals. By studying the process whereby such an exon is born, we can begin to see the advantages of introns in general, and the energy we expend to sustain them seems justified.

These primate-specific exons derive from mobile genetic elements called Alus, which belong to a larger class of elements known as retrotransposons—short sequences of DNA whose function seems to be generating copies of themselves and then reinserting those copies back into the genome at random positions, rather like little genomic parasites. Retrotransposons are found in almost all genomes, and they have had a profound influence by contributing to the genomic expansion that accompanied the evolution of multicellular organisms. Almost half the human genome is made up of transposable elements, Alus being the most abundant.

CHIMPANZEES AND HUMANS share 99 percent of their genomes, including tiny mobile genetic elements, called Alus, found only in primates. Alus may have given rise, through alternative splicing, to new proteins that drove primates' divergence from other mammals. Humans' divergence from other primates may also be thanks in part to alternative splicing: recent studies have shown that the nearly identical genes of humans and chimps produce essentially the same proteins in most tissues, except in parts of the brain, where certain human genes are more active and others generate significantly different proteins through alternative splicing of gene transcripts.

Alu elements are only 300 nucleotides long with a distinctive sequence that ends in a "poly-A tail." Our genome already contains some 1.4 million Alu copies, and many of these Alu elements are continuing to multiply and insert themselves in new locations in the genome at a rate of about one new insertion per every 100 to 200 human births.

The Alus were long considered nothing more than genomic garbage, but they began to get a little respect as geneticists realized how Alu insertion can expand a gene's protein-generating capacity. About 5 percent of alternatively spliced exons in the human genome contain an Alu sequence. These exons most likely originated when an Alu element "jumped" into an intron of a gene, where the insertion normally would not have any negative consequence for the primate because most introns are spliced out and discarded. Through subsequent mutation, however, the Alu could turn the intron in which it resides into a meaningful sequence of genetic information—an exon. This can happen if changes in the Alu sequence create a new 5′ or 3′ splice site within the intron, causing part of the intron to be recognized as "exon" by the spliceosome. (Such mutations usually arise during cell division, when the genome is copied and a "typo" is introduced.)

If the new Alu exon is only alternatively spliced in, the organism can enjoy the best of two worlds. By including the Alu exon, its cells can produce a novel protein. But the new capability does not interfere with the gene's original function, because the old types of mRNA are also still synthesized when the Alu exon is spliced out. Only when a mutated Alu becomes spliced constitutively—that is, the Alu exon is always spliced into all the mRNAs produced from the gene—does it become problematic, because it can trigger genetic diseases caused by the absence of the old protein. To date, three such genetic illnesses caused by misplaced Alu sequences have been identified: Alport and Sly syndromes and OAT deficiency.

My colleagues and I have shown

MIKEL ROBERTS *Corbis Sygma*

that all it takes to convert some silent intronic Alu elements into real exons is a single-letter change in their DNA sequence. At present, the human genome contains approximately 500,000 Alu elements located within introns, and 25,000 of those could become new exons by undergoing this single-point mutation. Thus, Alu sequences have the potential to continue to greatly enrich the stock of meaningful genetic information available for producing new human proteins.

RNA Therapy

MORE THAN 400 research laboratories and some 3,000 scientists worldwide are trying to understand the very complex reactions involved in alternative splicing. Although this research is still at a very early stage, these investigators agree that recent findings point toward future therapeutic applications, such as new gene therapy strategies that exploit the splicing mechanism to treat both inherited and acquired disorders, such as cancer.

One approach might be to direct a short stretch of synthetic RNA or DNA nucleotides, called antisense oligonucleotides, to bind to a specific target on the patient's DNA or RNA. Antisense oligonucleotides could be delivered into cells to mask either a specific splice site or some other regulatory sequence, thereby shifting the splicing activity to another site. Ryszard Kole of the University of North Carolina at Chapel Hill first demonstrated this technique on human blood progenitor cells from patients with an inherited disorder called beta-thalassemia, in which an aberrant 5′ splice site causes oxygen-carrying hemoglobin molecules to be deformed. By masking the mutation, Kole was able to shift splicing back to the normal splice site and restore production of functional hemoglobin.

Later, Kole showed that the same technique could be used on human cancer cells grown in culture. By masking a 5′ splice site of the *Bcl-x* apoptosis-regulating gene transcript, he was able to shift splicing activity to generate the Bcl-x(S) form of mRNA rather than the Bcl-x(L) form, decreasing the cancer cells' synthesis of the antiapoptotic protein and enhancing synthesis of the proapoptotic protein. In some cancer cells, this change activates the apoptotic program; in others, it enhances the apoptotic effects of chemotherapeutic drugs administered along with the oligonucleotides.

Why has evolution preserved a complicated system that can cause disease?

Another way to use the alternative splicing mechanism for therapy was demonstrated in 2003 by Adrian Krainer and Luca Cartegni of Cold Spring Harbor Laboratory in Long Island, N.Y., who found a way to induce cells to splice in an exon that would otherwise be skipped. They created a synthetic molecule that can be programmed to bind to any piece of RNA according to its sequence, then attached the RNA-binding part of an SR protein. This chimeric molecule can therefore both bind to a specified sequence on the pre-mRNA and recruit the basal machinery to the appropriate splice signal. Krainer and Cartegni used this method on human cells grown in culture to correct splicing defects in mutated versions of the *BRCA1* gene, which has been implicated in breast cancer, and of the *SMN2* gene, which causes spinal muscular atrophy.

Yet a third approach capitalizes on the ability of the spliceosome to join two different pre-mRNA molecules from the same gene to form a composite mRNA. Termed trans-splicing, this event is common in worms but occurs only rarely in human cells. Forcing the spliceosome to trans-splice could allow a mutated region of pre-mRNA responsible for disease to be precisely excised and replaced with a normal protein-coding sequence. Recently John Englehardt of the University of Iowa used this technique in cell culture to partially correct the pre-mRNA of a gene that produces a defective protein in the airway cells of cystic fibrosis sufferers.

Before the human genome was decoded, few scientists believed that organisms as complex as humans could survive with a mere 25,000 genes. Since the sequence was completed, alternative splicing has emerged as the pivotal process that permits a small number of genes to generate the much larger assortment of proteins needed to produce the human body and mind while precisely orchestrating their manufacture in different tissues at different times. Moreover, splicing explains how the tremendous diversity among humans, mice and presumably all mammals could originate in such similar genomes.

Evolution works by presenting organisms with new options, then selecting to keep those that confer an advantage. Thus, novel proteins created by the splicing in of new Alu-derived exons probably helped to make humans the species we are today. And further investigation of the alternative splicing process promises still greater improvements in our quality of life. SA

MORE TO EXPLORE

Alternative Splicing: Increasing Diversity in the Proteomic World. B. R. Graveley in *Trends in Genetics,* Vol. 17, No. 2, pages 100–107; February 2001.

Splicing Regulation as a Potential Genetic Modifier. M. Nissim-Rafinia and B. Kerem in *Trends in Genetics,* Vol. 18, No. 3, pages 123–127; March 2002.

The Hidden Genetic Program of Complex Organisms. John S. Mattick in *Scientific American,* Vol. 291, No. 4, pages 60–67; October 2004.

How Did Alternative Splicing Evolve? Gil Ast in *Nature Reviews Genetics,* Vol. 5, pages 773–782; October 2004.