Genes in Pieces

Molecular biologists speculate on the meaning of a most unexpected discovery

About 2 years ago, molecular biologists discovered that the genes of animal viruses and of animals, including humans, occur in pieces, spread out along DNA. In between the gene fragments are long stretches of DNA whose functions are unknown. This discovery changed the face of molecular biology research. Old mysteries are being solved, and researchers finally are able to formulate some basic questions about how gene expression is controlled.

When people began studying the genes of higher organisms, their working models were bacterial genes. But bacterial genes are not spread out in pieces. So when the first fragmented genes were discovered in animal viruses and then in animals, researchers spent much of their time deciding whether this sort of gene organization is the exception or the rule in higher organisms. It now seems fairly certain that it is the rule. So far, few genes seem not to be in pieces. One exception is the histone genes, which are unusual for several other reasons as well.

Molecular biologists also have learned, in general terms, how these genes in pieces are expressed. First, an RNA copy is made of the entire length of DNA containing the fragments of a gene. There is so much of the extra DNA—called intervening sequences—that the RNA copy of the gene and its intervening sequences is often five to ten times longer than the sum of the lengths of the gene fragments.

Next, the copies of the extra DNA are snipped out of the long piece of RNA, and the remaining RNA copies of the gene fragments are sealed together. Finally, this shortened piece of RNA, which is now a copy of just the gene, moves from the cell nucleus to the cytoplasm, where the cell reads its genetic message and makes the protein that the gene codes for.

Even this outline of what goes on explains some previously puzzling aspects of DNA and gene expression. Researchers have known for years that the cells of higher organisms have huge quantities of DNA—far too much if the DNA consisted of just a string of genes plus some short sequences used to control gene expression, as it does in bacteria. No one had any idea where this extra DNA was in relation to the genes or what it was used for. With the discovery that genes are in pieces with the extra DNA between them, at least part of the mystery is solved.

Another outstanding problem was to decide why there are such enormous pieces of RNA in cell nuclei. Some of the pieces are far longer than genes, and no one was certain whether they contained copies of genes or were made for some other reason. It was known that only shorter, gene-sized pieces of RNA ever make their way from the cell nucleus to the cytoplasm. Researchers spent years trying to understand what was going on with the RNA in the cell nucleus. They now realize that the large nuclear RNA's are the copies of fragmented genes together with copies of the DNA that separates the fragments.

Satisfying as it is to have answers to these problems, the discovery of genes in pieces has given rise to a new question, whose answer, molecular biologists believe, is the key to understanding the whole problem of gene expression in higher organisms. The question is, Why are genes in pieces? So central is this question that someone in nearly every major molecular biology laboratory in this country and in Europe is trying to answer it.

One of the first to try to answer this question was Walter Gilbert of Harvard University. He argues on evolutionary grounds that the fragmented genes are far more easily shuffled to form new combinations than are genes that are all in one piece. The reasons for this are twofold, Gilbert explains. First, the farther apart two pieces of DNA are, the more likely it is that they will change places, or recombine. So when gene fragments are spread out, it is far more likely that they will be shuffled.

The second reason is that a certain form of recombination is easier if the DNA pieces that change places do not have to make genetic sense. This is because the genetic code is read in groups of three nucleotides. If two pieces of DNA break and rejoin during recombination, it would ordinarily be required that the code could still be read as before. In other words, the adjoining point would have to preserve the integrity of the triplet code.

But the DNA in the intervening sequences is not translated into proteins. In fact, its nucleotide sequence varies greatly from species to species and even from animal strain to animal strain. If two pieces of DNA break and rejoin within intervening sequences, they can break and rejoin anywhere. As a result, recombination is easier within intervening sequences than within genes.

According to Gilbert, what occurs is a trade-off. It is an extra burden for a cell to replicate and make RNA copies of all its extra DNA. But it has the advantage that more rapid evolution is possible. In bacteria, that advantage is not as important; bacteria divide about every 15 minutes and so they evolve rapidly anyway.

The same argument explains why histone genes are not in fragments, Gilbert says. Cells apparently need to make large amounts of histones very quickly at certain times. That is why they have so many copies of the histone genes. Since it takes time and energy to copy extra DNA sequences, it would not be efficient for the histone genes to be in pieces.

Gilbert suggests that gene fragments are not just random pieces of genes, but are themselves minigenes. Each gene fragment, he suspects, codes for a functional part of a protein. When the fragments are shuffled by recombination, these minigenes can be rearranged to make new, possibly better, gene combinations.

Stimulated by Gilbert's hypothesis, Sherman Beychock of Columbia University decided to look at one of the fragments of the globin gene to see if it codes for a functional part of the globin molecule. The globin gene is divided into three parts, which are separated by two sections of extra DNA. Beychock looked at the middle fragment and discovered that it codes for the part of the globin molecule that binds heme. Heme contains iron and is the oxygen-binding piece of the hemoglobin molecule.

Beychock explains that this discovery was not expected on the basis of the structure of the globin molecule. "The globin molecule does not look like it's made up of parts. The boundaries be-

0036-8075/80/0125-0392$00.50/0

tween [the sections of protein coded by the gene fragments] are within the helical sections of the protein. You needed the hint that a break point could occur there before you would think to look for it."

Another confirmation of Gilbert's hypothesis is the structure of the immunoglobulin or antibody genes. These genes are in pieces, and it seems that each piece codes for a distinct part of the antibody molecule. For example, these molecules have two regions, called variable and constant. The DNA coding for the variable region is separated from that coding for the constant region by an intervening sequence. Through recombination, different variable regions are brought nearer to a particular constant region, and this shuffling helps cells to make many different kinds of antibodies.

In addition, Lee Hood and his associates at the California Institute of Technology and, independently, Susumu Tonegawa and his associates at the Basel Institute for Immunology find that the constant region of one type of immunoglobulin gene, the heavy chain gene, is divided into four regions by intervening sequences. Each of these regions codes for a separate functional part of the molecule. It also seems likely that these gene fragments evolved as separate units. "As of now, the immunoglobulins are by far the nicest example of Gilbert's hypothesis," says Hood.

Gilbert notes that his hypothesis has allowed him to make a correct hypothesis about the origin of the insulin genes in the rat. The rat has two insulin genes. Each gene has an intervening sequence at its beginning, but one gene has a second intervening sequence as well. Gilbert believes that the present insulin genes arose when various functional minigenes were joined. Therefore, he theorized, the rat insulin genes duplicated during evolution and the second intervening sequence in one of the rat insulin genes was later lost. To test this theory, Gilbert and his colleagues recently looked at the chicken, which he says is the rat's ancestor and which has only one insulin gene. As he predicted, the chicken gene has two intervening sequences.

> "A number of molecular biologists believe there is more to the extra DNA than the evolutionary theories imply."

Not everyone thinks that all of the gene fragments can themselves be minigenes. Robert Schimke of Stanford University, for example, says some genes are so fragmented that it is hard to believe each piece codes for a functional part of the protein. He cites the recent work of Pierre Chambon at the Institut de Chimie Biologique in Strasbourg, France. Chambon found that the conalbumin gene of the chicken, which codes for a major protein in egg white, is divided into at least 17 fragments.

Philip Leder of the National Institute of Child Health and Human Development has proposed another evolutionary advantage of having genes in pieces. Certain genes, such as the globin genes, appear in several copies. If there are long stretches of repeated gene sequences on DNA, it is highly likely that recombination will take place within them. If it does, one of the gene copies will probably be lost. But if each of the identical genes is broken into fragments separated by intervening sequences of DNA, recombination between these identical genes is far less likely. So, says Leder, one advantage of having genes in pieces is that multiple copies of genes will be preserved.

A number of molecular biologists believe there is more to the extra DNA than the evolutionary theories imply. They reason that cells are not wasteful, that they would not keep something like the extra DNA around and use it only for evolution. The extra DNA could well have some other function, these investigators speculate, and the one that first comes to mind is that of controlling gene expression.

As a first step toward seeing how the extra DNA functions in the moment-to-moment working of the cell, several groups of investigators are asking what would happen if there were no extra DNA in certain genes. Would the RNA copies of unfragmented genes still function?

Among the first to try to answer this question were Ching-Juh Lai and George Khoury of the National Cancer Institute. They studied SV40, a monkey virus whose genes are fragmented, and isolated a mutant in which some of the extra DNA was deleted. They found that there seemed to be no RNA copies of the genes whose intervening DNA was deleted. This indicates that the RNA copies of the extra DNA might be used by the cell to prevent the breakdown of RNA copies of genes.

Dean Hamer and Philip Leder of the National Institute of Child Health and Human Development came to a similar conclusion when they studied the expression of α and β globin genes. They added these genes to SV40, added the SV40 to cells, and saw that the globin genes were expressed. Then they showed that if they flipped the globin gene sequences over so that the cell transcribed the opposite strands of the DNA for these genes, a stable RNA copy never got through the nucleus to the cytoplasm. Finally, they found that if they not only flipped the globin genes but also removed part of an extra DNA sequence of the SV40, the RNA copy of the globin genes was degraded.

Carrying these sorts of experiments even further, Peter Gruss of the National Cancer Institute and Khoury made an exact deletion mutant of SV40—one in which only one of the intervening DNA sequences was removed. They found that an RNA copy of the gene was made but that it never got from the nucleus to the cytoplasm. Still, says Khoury, "no one knows whether [RNA] transcripts of these intervening sequences have any role in the cell."

What lies ahead, then, is tackling a huge set of questions. For example, says Khoury, researchers would very much like to know just how general these fragmented genes really are. With few exceptions, all the fragmented genes looked at so far code for specialized cell products, like globin, insulin, or antibodies, or are viral genes. Schimke is one of the only researchers who has looked at anything else—he examined the enzyme dihydrofolate reductase and found it extremely fragmented. It has five intervening sequences that are so long, says Schimke, that this is the largest stretch of fragmented gene known relative to the size of the RNA copy that eventually reaches the cytoplasm. Still, this is only the beginning of an investigation of genes used by all cells, not just by specialized cells.

Other important questions, says Khoury, are, What are the nucleotide signals that so precisely define splice sites, and do they involve interactions between RNA molecules or between RNA molecules and proteins? Is there a limit to the size of the intervening sequences? Are intervening sequences removed all at once or piecemeal? And can the efficiency of splicing at a particular site regulate gene expression?

The existence of all these questions, says Phillip Sharp of the Massachusetts Institute of Technology, indicates the vitality of this area of research. "It's clear now that we can formulate the questions of what steps occur in gene regulation. Previously, we were in a fog."

—GINA BARI KOLATA