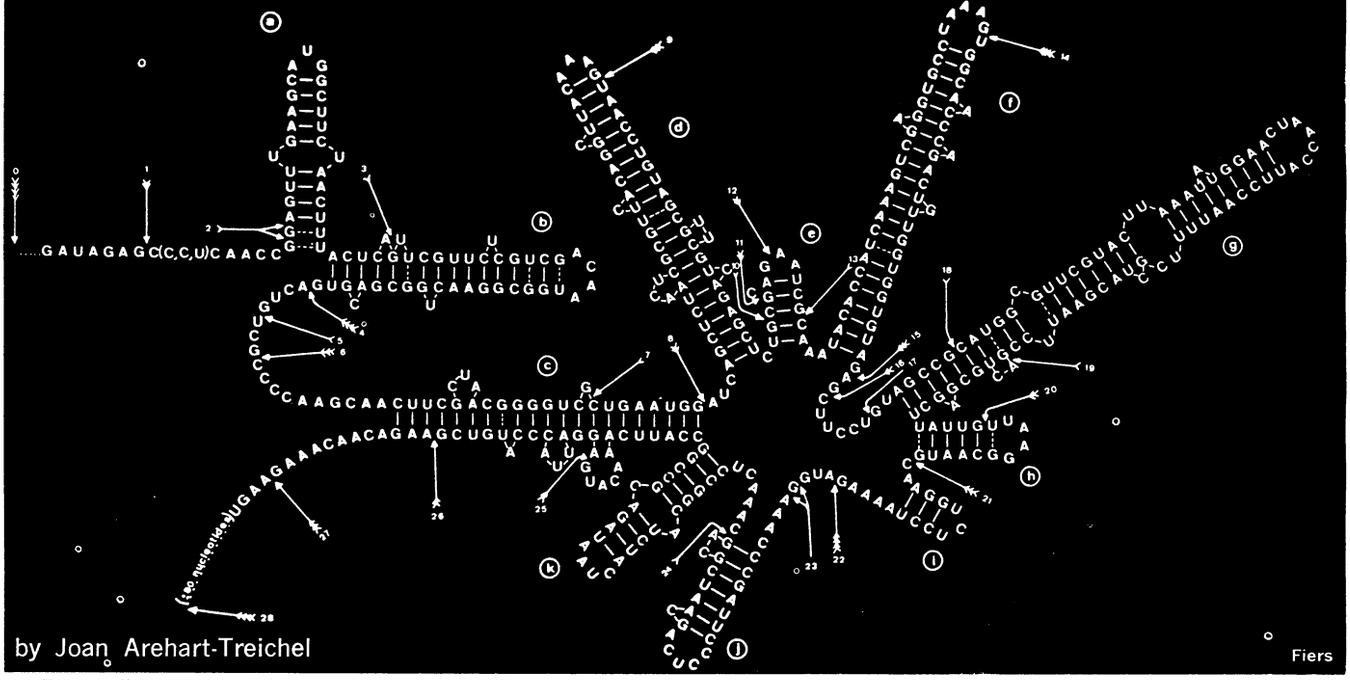
Molecular biology's flower child

A gene that codes for a protein has been sequenced. It folds up in a flower configuration when it is not coding.



Deoxyribonucleic acid (DNA) was recognized in chromosomes 70 years ago. But it wasn't until the 1950's that scientists identified its genetic function. Today, investigators are confident that DNA carries the genetic information in all living organisms, and that DNA, or a related nucleic acid, ribonucleic acid (RNA), carries the genetic information in viruses. They are also confident that the major role of genes—segments of DNA or RNA—is to code for the structure of proteins.

During the past several years, some artificial genes have been synthesized. Two years ago, for example, Nobel laureate Har Gobind Khorana, then of the University of Wisconsin and now of the Massachusetts Institute of Technology, used DNA material to construct a yeast gene. But no one had unraveled a real gene that dictates the production of a protein. Now researchers at the State University of Ghent in Belgium have done just that.

The journal NATURE has hailed the Belgians' work as one of 1972's outstanding achievements in molecular biology. The Belgians used a technique previously worked out by Nobel laureate Fred Sanger of Cambridge, England. Sanger told SCIENCE NEWS, "Theirs is good work indeed." Norton Zinder of Rockefeller University and a leader in RNA virus research, calls the Belgians' achievement "tremendous." He commends them for the energy and effort they directed to sequencing an entire gene. Khorana concurs: "It is great work."

Adds David Baltimore, a leading cancer virus researcher at MIT: "What they did was quite remarkable. It required great organization and technical expertise. You cannot say that now they have done that one can sequence anything. It is still hard and still requires a large amount of resources."

The Belgian research team, headed by Walter Fiers, also includes Willy Min Jou, Guy Haegerman and Marc Ysebaert. In a recent interview with SCIENCE NEWS in his laboratory in Ghent, Fiers explained the background for their work and described how they arrived at their achievement.

The genetic material Fiers' group chose to deal with was from the MS2 virus. The MS stands for "male specific." The virus invades male, not female bacteria. This virus contains all of its genetic information in one RNA molecule. Hence the advantage of dealing with it rather than with genetic information from animals or man. It is estimated that in each of the trillion or so cells in the human body, the total length of the DNA molecules amounts to more than 500,000 times the length of the MS2 RNA molecule.

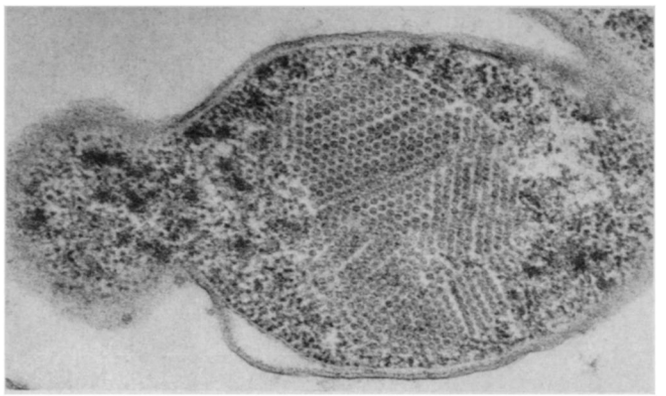
Because the MS2 virus makes three proteins, the Ghent scientists knew that the MS2 RNA molecule had to contain three genes—one to code for each protein. They also had an idea of how the three genes line up on the RNA molecule. The first gene is for the A protein. It makes a protein that helps the virus get into a bacterium host. The second gene is the coat gene. This gene codes for a protein that makes a coat around the virus. The third gene is known to be the polymerase gene. It makes an enzyme that helps the virus replicate itself inside a host bacterium. With these three genes, the MS2 virus is able to make 10,000 copies of itself in a host bacterium within 10 to 25 minutes.

The Belgians knew that the genes on the MS2 molecule had to be made up of the chemical molecules that all genes contain: nucleotides. There are four kinds of nucleotides: adenine (A), cystosine (C), guanine (G) and uracil (U). Three nucleotides comprise each word (codon or triplet) of the genetic code. Since there are four nucleotides, there are 64 word possibilities. Most of the triplets serve as code words for one or another of the 20 amino acids that make protein molecules. GCU, GCC, GCA and GCG specify the amino acid alanine. UCU, UCC, UCA and UCG specify the amino acid serine. And so forth.

The gene from the MS2 RNA molecule that the Belgians set out to sequence was the second gene on the molecule, the coat gene. They took a solution of MS2 RNA molecules and digested it with a small amount of enzyme, so that large fragments were obtained. These fragments were then separated, mainly by electrophoresis on gels. The fragments were again digested with

Joan Arehart-Treichel

Fiers: Now after cancer virus genes.



MS2 viruses in bacterium.

Chris Meijvisch

various enzymes so that smaller nucleotides were obtained. These smaller nucleotides were separated mainly by electrophoresis on chemically modified sheets of paper (Sanger's technique). This way the Belgian's were able to determine the chemical composition of the various fragments—what sequence of nucleotides made up the various fragments.

How did they figure out which of the identified nucleotides made up the coat gene? They knew the amino-acid sequence for the coat protein. They knew that since there are 130 amino acids in the coat protein, the coat gene would have to have 130 nucleotide triplets, or 390 nucleotides. They knew there were several codon possibilities for each amino acid in the protein. The challenge, then, was to fit hundreds of identified nucleotide fragments into a sequence of 390 nucleotides that would code for the coat protein.

"In a sense," Zinder says, "it was a test case for doing nucleotide sequencing where the protein sequence is known." The challenge, in other words, was none less than a molecular biology jigsaw puzzle—fitting the right nucleotide fragments together. "The fit was so good," Fiers recalls, "that we thought we had to be right."

Fiers and his colleagues knew that the nucleotides making up the coat gene had to be stretched out in a linear sequence in order to code for the coat protein. Physical-chemical observations, however, showed that this linear sequence folds up when the gene is not coding. In fact, the nucleotides appear to arrange themselves into many hairpin turns and loops, somewhat like the petals of a flower. So they dubbed the nonworking gene "the flower model."

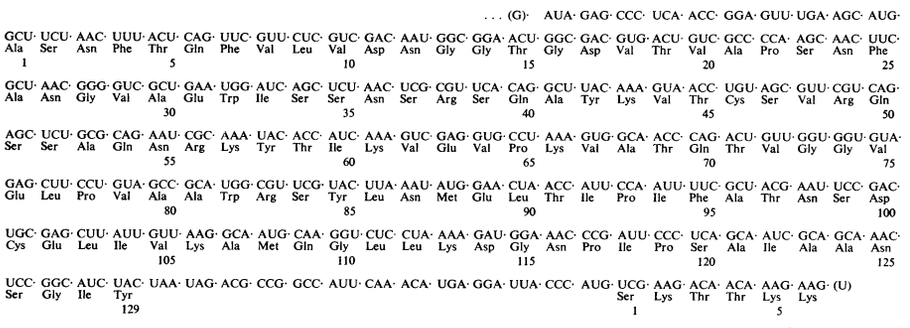
Fiers says figuring out how the linear sequence of coat gene nucleotides pair up, when not coding, to make a flower configuration was even knottier than sequencing the gene. Some clues helped them make a model. For example, past molecular biology experience told them that only the nucleotide bases alanine and uracil would pair up, and that only the nucleotides bases guanine and cytosine would pair up. They managed to put together a flower model of the nonworking gene using almost all of the 390 nucleotides. A few nucleotides, however, did not fit, and stuck out of the model like sore thumbs. Fiers concludes that the model of the nonworking gene is essentially correct, yet that further experiments, computer evaluations and improved theoretical approaches will be required to make the model completely accurate.

Although the Belgian team has been working seven years to sequence the coat gene, they have made their greatest strides over the past year or two. Meanwhile, they have also had some success in sequencing other sections of the MS2 RNA molecule. Thanks to previous work by Sanger and by Joan Argetsinger-Steiz, now at Yale University, they have been able to determine that 130 nucleotides precede the A protein gene on the molecule, and that 60 or more nucleotides follow the polymerase gene on the molecule. Since these nucleotides are not part of the genes that code for the structure of proteins, might they serve as specific recognition sites in RNA replication? Fiers thinks this is feasible. He and his colleagues have also sequenced the last 100 nucleotides in the A protein gene, and the first 100 nucleotides in the polymerase gene. This means, that in addition to the 390 nucleotides they have sequenced for the coat gene, that they have unraveled about 40 percent of the entire MS2 RNA molecule. Fiers estimates they will have completed the entire sequence within the next two years. He and his co-workers are also undertaking the sequencing of nucleotides in a cancer-causing virus.

There is little doubt that sequencing of genes holds powerful ramifications for the advance of medical science. As Baltimore points out, "Ultimately, I think, we are going to see the sequencing of almost all viral nucleic acids, and it would help one understand how they interact with machinery in the host cell. It would also tell what kinds of proteins they make and enable one to direct the manipulation of their genes."

Says Zinder: "It was nice Fiers and his group did this work. It had to be done. There are many people who will be doing this kind of thing." Rather than undertake the sequencing of whole genes, however, a number of researchers are now concentrating on sequencing only those nucleotides that act as starters and stoppers (punctuation marks) in gene coding for proteins.

It is also possible that gene sequencing will one day lead to clinical correction of genetic diseases. Fiers is hopeful that even within his lifetime the gene that codes for the protein molecule insulin might be isolated and sequenced. The amino-acid sequence for insulin has already been worked out and would provide a guideline. Once the sequence of the insulin gene is worked out, a synthetic copy of the gene might be made. The synthetic gene might be introduced into the cells of diabetic patients. Whether the gene would express itself (start making insulin) is not known. But bacterial genes have already been introduced into human cell cultures and have expressed themselves in those cells (SN: 10/23/71, p. 276). □



Fiers/Nature

Nucleotide sequence of coat protein gene with amino acids it specifies.