Antibodies: Getting Their Genes Together

Assembly of antibody genes requires a well-orchestrated series of gene rearrangements that serve to increase the diversity of the gene products

The genomes of higher organisms gain a great deal of potential versatility from the fact that most of the genes occur in pieces, rather than as continuous stretches of DNA. And nowhere has that potential been fulfilled more dramatically than in the genes coding for the protein chains of antibody molecules. Within the past few years, research has shown that assembly of the individual genes by combining separate segments of DNA accounts for much—although definitely not all—of the almost limitless diversity of antibodies.

The discoveries have helped to solve one of the major problems of immunology, namely, how the mammalian genome could contain the information needed to produce a million or so different antibodies. "Three billion nucleotides are not very many when it comes to building a mouse or man," says Philip Leder of the National Institute of Child Health and Human Development. "If you use them only in a linear way, there is not enough information. The solution is to shuffle the genes and use the DNA a piece at a time."

Direct demonstration of the importance of antibody gene rearrangements emerged about 3 years ago from studies of the DNA segments coding for the lighter of the two types of protein chains of which antibody molecules are composed (*Science*, 20 October 1978, p. 298). The information needed to make a single light chain, which consists of about 220 amino acid residues, turned out to be encoded in three separate DNA segments in germ-line cells that had not yet begun to develop into antibody-producing cells.

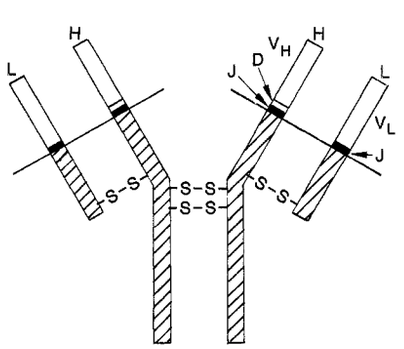

Fig. 1. Antibody molecule structure

The letters H and L designate the heavy and light chains, respectively; V, D, J, and C refer to the variable, diversity, joining, and constant regions of the appropriate chains.

Because of the structural organization of antibody chains, the possibility that each might be encoded in two separate DNA segments was not unexpected. Each light chain consists of two regions of roughly equal length: the variable region, which differs from one chain to the next, and the constant region, which is the same for all light chains of the same class (Fig. 1). Similarly, heavy chains, which contain about 440 amino acid residues, have a variable region consisting of about 110 amino acids; the remainder of the molecule comprises the constant region. Together, the variable regions of the heavy and light chains form the business end of the antibody, the site that binds a specific antigen.

In 1965, William Dreyer of the California Institute of Technology and J. C. Bennett of the University of Alabama School of Medicine proposed that encoding the variable and constant regions separately would be an efficient way to provide for a situation in which a single constant region might be combined with any of numerous different variable regions. Direct examination of antibody genes by Susumu Tonegawa of the Basel Institute of Immunology and Leder, among others, showed that the variable and constant regions of light chains are separately encoded but that two distinct gene segments, not just one, are needed to produce the variable region (Fig. 2). One segment (designated V) codes for the first 95 or so amino acids of the variable region; and a second (called J because its product joins the constant and variable regions) codes for the remaining 15. Information for the constant region is carried by a third segment (C).

When a germ-line cell begins to develop, the V and J gene segments are joined to one another but remain separate from the C region segment. The final connection is made later, after the DNA is copied into an RNA messenger.

Although the discovery of a separate J region segment came as a surprise, it was not an unpleasant one, because its existence increases the amount of diversity that can be programmed into the variable region. According to current estimates, the mouse genome carries perhaps 300 V gene segments and four J gene segments that can combine to produce variable regions for light chains of the κ class. (Actually, there are five J segments, but the fifth cannot produce a functional product.) The κ chains constitute about 95 percent of all the light chains of mice and about 60 percent of those of humans; the rest belong to the λ class. Simple V-J joining can thus generate about 1200 different κ variable regions.

But more than that, the ends of the V and J regions are not always accurately connected. Leder says, "There is some flexibility in V-J joining. It can change the protein sequence in a very critical region around amino acid 96." Studies of the three-dimensional structure of antibodies have shown that this amino acid is one of those in contact with bound antigen. An amino acid change in this region might alter the antigen specificity of the antibody.

Information about rearrangement of the genes coding for antibody heavy chains has been slower in coming, partly because these proteins are twice as large as light chains. However, within the past year or so, researchers have obtained evidence that heavy chain genes are also stitched together from separate DNA pieces, although there is a significant difference from the pattern set by light chains.

Three pieces, not two, must be joined to form the gene coding for a complete heavy-chain variable region (Fig. 3).

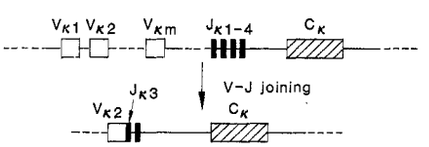

Fig. 2. Light chain gene assembly

The upper diagram shows the arrangement of κ chain gene segments in the DNA of germ-line cells. During V-J joining any of the variable gene segments ($V_{κ1}$ through $V_{κM}$) can be connected to any of the joining gene segments ($J_{κ1-4}$), as shown in the lower diagram for the $V_{κ2}$-$J_{κ3}$ combination. The variable region gene thus assembled remains separate from the constant region gene ($C_κ$) until this whole stretch of DNA is copied into RNA transcripts. Then the left end of the variable region gene is spliced to the right end of the constant region gene, and the resulting messenger RNA is translated to yield a κ light chain.

0036-8075/81/0529-1015$00.50/0

Two groups, one led by Leroy Hood of the California Institute of Technology (Caltech) and the other by Tonegawa, noted that the variable regions of the heavy chains they were studying contained a short stretch of amino acids not encoded by the appropriate V and J gene segments. They suggested that a third piece of DNA would have to be inserted between the V and J segments to form the gene for the entire variable region of a heavy chain. This third segment has been desigated D for diversity, because it codes for amino acids in one of the most variable sections of the heavy-chain variable region. Such D segments have now been identified both in germline and in differentiated immune cells. They appear to be located, as expected, between the V and J segments. Like the V-J joining of light chains, V-D-J joining occurs after an undifferentiated immune cell begins to develop into an antibody-producing cell.

"The D segments contribute a lot to diversity," says Tonegawa. "We don't know how many there are, but there are probably families of segments." His group has identified two families. The better studied of the two has eight members, which have 17 nucleotides each and have similar but different structures, in germ-line cells. The size of the second family and the total number of families are unknown.

Tonegawa has suggested that combination of two D gene segments to form D regions of different lengths may contribute to the diversity of heavy-chain variable regions. This suggestion and the model he has proposed for D-D joining are controversial.

Not controversial is the identity of the DNA sequences that serve as signals for V-J and V-D-J joining. In all cases, for both κ and λ light chains and for heavy chains, the signals are the same. Investigators, including Hood, Leder, and Tonegawa, observed that each V region segment is flanked on the right, and each J region segment on the left, by a seven-nucleotide sequence that is a palindrome (it reads the same in both directions). Moreover, Hood and Philip Early, also of Caltech, found that the J region palindrome is followed first by a spacer consisting of 11 or 22 nucleotides and then by a 10-nucleotide sequence that is rich in adenine- or thymine-containing nucleotides (Fig. 4). This same sequence, although inverted, appears on the left-hand flanks of the V, D, and J gene segments.

The size of the spacer regions has interesting implications for V-J and V-D-J joining. As a rule, segments with an 11-nucleotide spacer can pair only with segments having a 22-nucleotide spacer. Hood, noting that one turn of a DNA helix contains 11 nucleotides, has speculated that the gene segments to be spliced are brought together by proteins that first recognize and bind to one or two turns of the DNA helix and then bind to each other. If so, the cell might control the rearrangements of an antibody gene by making the proteins available at the appropriate time during its development.

The assembly of complete variable regions from individual V, D, and J pieces is probably random. "Nothing is completely random," Tonegawa explains. "There may be preferential pairings. But with this qualification, if it [the joining] is not random, it is not making use of all the information." The price of using all the information to obtain greater diversity is error, however.

A cell cannot make an antibody until it has assembled genes that code for complete—and functional—variable regions for both light and heavy chains. If the V-J and V-D-J joinings are not accurate, the gene segments thus generated may not be able to direct the synthesis of working antibody chains. Leder, for example, has noted that there is flexibility in V-J joining that may produce missense or nonsense codons instead of codons for one amino acid or another. Similar errors may crop up in the assembly of complete heavy-chain variable genes.

Leder's group and those of Hood, David Baltimore of the Massachusetts Institute of Technology (MIT), and Martin Weigert of the Institute for Cancer Research in Philadelphia have direct evidence that errors frequently occur during both V-J and V-D-J joining. In general, the investigators find that improperly rearranged κ variable genes occur frequently in cells that produce λ chains. Or the κ chain V and J gene segments may be deleted entirely in these cells. But in κ chain-producing cells, inappropriately rearranged λ variable genes are rare; usually the λ chain V and J gene segments remain in the germ-line configuration. Improperly joined heavy chain variable genes are also very common.

The investigators interpret these findings differently. Weigert suggests that the antibody gene rearrangements are governed by the rules of probability. The more gene segments there are, the more opportunities there are for segment joining—and for error. In the mouse, for example, there are only two λ V segments and one λ J segment. Consequently, there would be far fewer opportunities for λ V-J joining to go awry than for κ V-J joining.

According to Leder, strict probability cannot explain the error frequencies he finds in human cells, where λ chains account for about 40 percent of all light

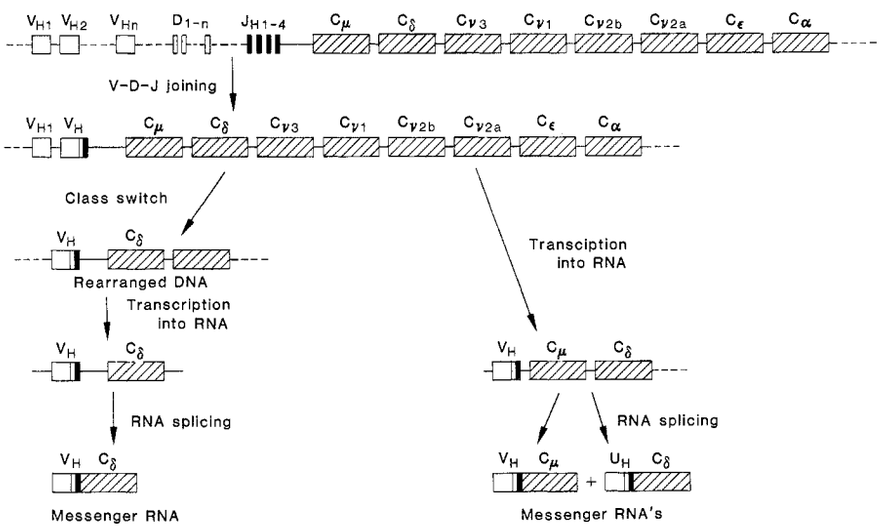

Fig. 3. Heavy chain gene assembly

The upper diagram shows the arrangement in germ-line cells of the gene segments for antibody heavy chains. The assembly of a complete variable region gene (V_H) requires the joining of three gene segments, one of those designated V_{H1} through V_{Hn}, one D, and one J. After V-D-J joining, transcription of the rearranged DNA into RNA, followed by differential splicing of the transcript, allows the simultaneous production of messenger RNA's for two different classes of heavy chain (shown in the right-hand branch of the diagram). The messengers direct the synthesis of either an IgM or IgD heavy chain. The shift to the final class of heavy chain that will be produced by the cell involves a further gene rearrangement (left-hand branch). In this case, the cell will make an IgD heavy chain and the C_μ gene segment has been deleted to bring the completed variable region gene nearer to the C_δ gene segment. If the cell were going to make an IgA heavy chain, gene segments C_μ through C_ϵ would have to be deleted. The rearranged DNA is transcribed and the RNA copy spliced to form the final messenger.

chains produced. He and Hood suggest that the error patterns they have observed reflect the order in which the light chains are synthesized. Baltimore and E. J. Siden, also of MIT, have evidence that the heavy chain is made first, which requires an appropriate V-D-J joining. Once this has happened, V-J joining can proceed with the κ segments rearranging first, according to Leder and Hood. If a productive joining occurs, there is no need for λ V-J joining. But if κ V-J joining does not produce a functional gene, then the cell can go on and try to make a complete λ variable gene. "In this way," Leder points out, "the cell can experiment with flexible joining rules until it finally makes an active gene." Of course, if it fails, it has reached a dead end.

Normally, an antibody-producing cell makes just one species of light chain, even though it carries corresponding information on both members of the chromosome pair and could make two. The need to have a properly rearranged variable-region gene helps to explain this phenomenon, which is called allelic exclusion. The evidence suggests that the rearrangement sequence stops as soon as one complete functional gene is assembled. Genes still in the germ-line configuration are not expressed, and improperly completed variable-region genes either are not expressed or their products cannot function.

The situation with regard to heavy chains is more complicated. Only one variable region is made, but it can be attached to any of five different constant regions. The type of constant region determines the class to which an antibody belongs. The five classes are immunoglobulin (Ig) M, D, G, A, and E, and the corresponding constant-region genes are designated Cμ, Cδ, Cγ, Cα, and Cε, respectively.

The sequence in which the different immunoglobulins are expressed follows a set pattern during development. First the cell makes an IgM; then it may make both an IgM and an IgD at the same time; and finally it becomes committed to producing a single class of heavy chain, which may be any one of the five. This requires that the completed variable-region gene be switched from one type of constant-region gene to another at the appropriate time during development. Hood says, "We know that as B cells [the precursors of the antibody-producing cells] develop, they can replace the Cμ gene with any of the other C_H genes." This change is called "class switching," and it is effected either at the level of RNA or by gene rearrangement.

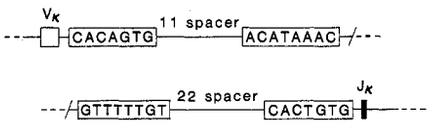

Fig. 4. Signals for V-J joining

The seven-nucleotide sequence CACAGTG occurs just to the right of all the V gene segments that have been sequenced thus far, and is thought to be part of the signal for V-J and V-D-J joining. It is palindromic (meaning it reads the same in both directions) and self-complementary; that is, the DNA sequence on the complementary strand of the DNA molecule reads GTGTCAC. The signals on the right-hand flanks of V gene segments also include a spacer consisting of 11 nucleotides followed by a sequence rich in adenine-containing nucleotides. The signals on the left-hand flanks of J regions consist of a thymine-rich segment, followed by a 22-nucleotide spacer, followed by the sequence CACTGTG. When inverted, as could occur in the looping out of a section of DNA during its excision, the left-hand flank of the J gene segment can base-pair with the right-hand flank of the V gene segment. The rule is that signals having 11-nucleotide spacers pair only with those having 22-nucleotide spacers.

The assorted gene segments that must be rearranged to produce a complete heavy chain are arranged on the chromosome in order of their appearance, with the V, D, and J gene segments followed by the Cμ, Cδ, Cγ, Cε, and Cα genes (Fig. 3). During the early stages of development, before the antibody-producing cell becomes committed to making a single type of heavy chain, processing of RNA transcripts is the most important factor in determining the type of chain to be made. For example, Hood and Randy Wall of the University of California at Los Angeles find that after V-D-J joining, a single RNA transcript, which extends from before the beginning of the completed variable-region gene to beyond the end of the Cδ gene, is made. When the noncoding portions of the transcript are excised, the variable-region gene copy can be attached to either a Cμ or a Cδ gene copy, giving the cell the ability to make an IgM and an IgD at the same time.

In the later stages of development, when the cell makes its final decision about which class of heavy chain to produce, the constant-region genes between the assembled V-D-J region and the desired C gene are apparently removed. Several investigators, notably Tasuku Honjo of Osaka University medical school in Japan, have identified repeated DNA sequences, occurring in the regions flanking the beginnings of the various C genes, that might serve as the switch signals, but the exact mechanism of DNA switching is unclear.

Despite the large degree of variability in antibody structure that can be achieved by these complex gene shifts, gene rearrangement is not the only means of generating the immense repertoire of antibodies. Over the years, three major competing theories sought to explain this great diversity. The germ-line theory held that there was a separate variable gene for every possible antibody chain. The somatic mutation theory held that there need be only one or a few such genes and that diversity was generated by mutations working on these genes. And the minigene theory held that the genes were assembled from the shuffling of small gene segments in varying combinations.

By late 1978, it was becoming clear that each of these theories portrayed elements of the truth (*Science*, 27 October 1978, p. 298). There are at least a few hundred different V gene segments for both light and heavy chains, and their diversity can be further increased by V-J or V-D-J joining. Moreover, several investigators, including Weigert and Melvin Cohn of the Salk Institute, showed that mutations also contribute to antibody diversity.

Recently, Hood and Patricia Gearheart of the Carnegie Institution of Washington in Baltimore, obtained additional evidence in favor of the somatic mutation theory. They compared the amino acid sequences of the variable regions of approximately 30 heavy and 30 light chains with the sequences of the corresponding genes. They found more variation in the protein structures than could be explained by the genes. As Hood describes their conclusions, "The variants unequivocally must arise from somatic mutations. When we looked at the gene sequences, we found that somatic recombination could not explain the diversity."

A second conclusion of Hood and Gearheart may help shed light on the activation of somatic mutation in antibody-producing cells. They found that the light and heavy chains of IgG's and IgA's underwent frequent mutations, whereas those of IgM's rarely did. Gearheart and Hood suggest that somatic mutation is not activated until after the switch from IgM production to the production of the other antibody classes. This finding may make it easier to determine the source of the mutations.

All in all, the exquisite adaptability of antibody synthesis seems well-matched by the complexity of the gene rearrangements and alterations on which it depends. Perhaps not surprisingly, Mother Nature is turning out to be the ultimate producer of designer genes.

—JEAN L. MARX