Somatic Generation of Antibody Diversity

SUSUMU TONEGAWA
Center for Cancer Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139–4307, USA

By the beginning of the 1970s, immunologists agreed that an individual vertebrate synthesizes many millions of structurally different forms of antibody molecules even before it encounters an antigen. Moreover, Gerald Edelman and Rodney Porter had shown that a typical antibody molecule is composed of two identical light chains and two identical heavy chains (*1*,*2*). It had also been found that each of these two types of chain exhibits great sequence variability in the amino terminal region between one antibody molecule and the next and no sequence variability in the carboxyl terminal regions (*3*). These two regions were then referred to as the variable, or *V*, and the constant, or *C*, regions. However immunologists and geneticists were divided for many years into two schools of thought with respect to the issue of whether the genetic diversity required for the synthesis of these proteins is generated during evolution, and is carried in the germline, or during development, in which case it would be present in somatic but not germline cells. One school of thought held that the germline must include a separate gene for every polypeptide that ultimately appears in an antibody molecule (*4*). In this germline theory, antibody or immunoglobulin genes are expressed in exactly the same way as those for any other protein, and no special gene-processing mechanisms are needed. On the other hand, the model requires an enormous number of immunoglobulin genes inherited from the parents. While the four-chain structure of an immunoglobulin molecule allows diversity to be generated by chain paring, the number of genes required for both light and heavy chains is still very large. One major difficulty for germline theories of antibody diversity was the observation that all antibody polypeptide chains of a given type share a common genetic marker (allotype) that segregates as a single Mendelian gene. If there were many thousands of light and heavy

This article is based on a Nobel Lecture delivered at Kalorinsha Institute on 8 December 1987.

IMMUNOLOGY: THE MAKING OF A MODERN SCIENCE
ISBN 0-12-274020-3

chain genes, how could the same genetic marker in all of these genes have been maintained?

The second theory supposed that there are only a limited number of antibody genes in the germline, and that these genes somehow diversify as the antibody-forming B lymphocytes emerge from their stem cells. In other words, the diversification of antibody gene sequences takes place in specialized somatic, or body, cells rather than being carried from generation to generation by the germ cells (*5–7*). One attraction of this latter theory is that it relieves the host of the need to commit a disproportionately large fraction of the inherited genes to code for antibodies, but the theory demands an unprecedented mechanism for diversifying the inherited genes somatically.

Arguments for and against these contrasting ideas were made both verbally and in written form for many years. However, all of these arguments were based on the interpretation of amino acid sequences of immunoglobulin polypeptide chains or on the generally accepted principles of evolution and genetics. No direct evidence for either view had been obtained. This was because no technique was available that would allow an analysis of the fine structure of specific genes from higher organisms.

Gene Counting

In the early 1970s, the technology for purifying a specific eukaryotic mRNA was just becoming available. Furthermore, a method to determine the number of copies of a specific gene by kinetic analysis of nucleic acid hybridization had already been established (*8,9*). These technical developments led some scientists, including myself, to think that one can experimentally determine the number of immunoglobulin genes contained in a germline genome and thereby decide which of the two major theories of antibody diversity is correct. The validity of this approach is based in part on the fact that the *V* region of a given chain type, while being different, exhibits a high degree of amino acid sequence homology. It was therefore thought that an mRNA coding for a specific immunoglobulin polypeptide chain would hybridize not only with its own gene but also with many other immunoglobulin genes, if they existed in a germline genome.

I thus obtained mouse myeloma cells and put my effort to purifying immunoglobulin mRNA and carrying out the hybridization studies. However, the initial studies focusing on the mouse κ light chain and heavy chain genes gave ambiguous results. The difficulties were threefold: uncertainty about the purity of the mRNA used as the hybridization probe; a lack of knowledge of the extent to which a probe will hybridize with the related but not identical genes; and the precise effect of sequence differences on hybridization kinetics. Thus, it turned out to be nearly impossible to make a convincing interpretation of the data obtained in these early studies in relation to the issue of the evolutionary versus somatic generation of antibody diversity.

One subsequent series of experiments which I carried out on genes coding for the mouse λ light chains, however, was very encouraging (*10*). Using an mRNA preparation that was more than 95% pure, I could show that the mouse λ light chain gene is reiterated no more than the β globin gene. The latter gene had been shown to be essentially unique. Fortunately, Weigert, Cohn and their coworkers had identified at least eight different V_λ region sequences among BALB/c-derived myelomas (*11*). Since the *V* regions were highly homologous, differing by only one, two or three amino acid residues, it was very likely that the corresponding genes would cross-hybridize extensively if they existed separately in the germline genome. Furthermore, statistical analysis of λ light-chain-secreting myelomas strongly suggested that a BALB/c mouse has the capacity to synthesize many more than the eight different V_λ regions identified. Thus, the number of the mouse λ genes determined experimentally (no more than a few) was far smaller than the number of different V_λ regions (at least eight, most probably many more) detected in proteins. On the basis of these results I was convinced that a somatic diversification occurs in this gene system.

Rearrangement

In the meantime, I became aware that some immunologists had been speculating that immunoglobulin polypeptide chains may be encoded by two separate DNA segments, one each for the *V* and *C* regions. Drawing an analogy from the elegant Campbell model (*12*) on the integration and excision of a phage λ genome, Dryer and Bennett had further suggested that one of many '*V* genes' may be excised out from the original chromosomal position and joined with the single '*C* gene' in an immunoglobulin-producing B cell (*13*). This model successfully explained the maintenance of the common genetic marker in all immunoglobulin polypeptide chains of a given type by postulating a single *C* gene for that cell type. Although a somatic recombination between the '*V* and *C* genes' is an inherent aspect of the model, it is clearly a version of the germline theory of antibody diversity because the model assumed that the germline genome carries many '*V* genes', one for every *V* region that an organism can synthesize.

When the Dryer and Bennett model was published in 1965, it was not accepted widely by biologists. This is understandable because the model was built on two hypotheses, both of which violated the then current dogmas of biology. These are the principles of one gene encoding one polypeptide chain, and of the constancy of the genome during ontogeny and cell differentiation. My personal reaction to the model when I learned of it in the early 1970s was also that of scepticism. However, at the same time I thought that the model might be testable if one were to use restriction enzymes. While in Dulbecco's laboratory, I had heard of Daniel Nathans' breakthrough in the analysis of the

SV40 genome by an application of the then newly discovered restriction enzymes (*14*). As one who used to struggle to define the transcriptional units of this DNA virus I was keenly aware of the power of these enzymes for the analysis of DNA structure. However, an extension of the restriction enzyme analysis from a viral genome of 5×10^3 base pairs to the 2×10^9 base pair genome of an eukaryote as complex as a mouse, required the use of an additional trick for the detection of a specific DNA fragment in a vast array of irrelevant fragments. An obvious solution seemed to lie in the combination of an electrophoretic separation of enzyme-digested DNA and the sensitive technique of nucleic acid hybridization. I discussed with Charlie Steinberg the need for developing a method that allows an *in situ* detection of a specific DNA sequence among the electrophoretically fractionated DNA fragments, but we really could not come up with a good idea worthy of exploring. As we all know, a very simple and elegant method ideal for this purpose was later developed by Edward Southern (*15*).

A few weeks passed by before I accidentally saw in one of the Institute's cold rooms a huge plexiglass tray in which someone was fractionating serum proteins by starch gel electrophoresis. I thought one may be able to fractionate a sufficient amount of digested DNA in a gel of such dimensions, so that DNA eluted from gel slices could be used for liquid phase hybridization. A quick calculation seemed to indicate that the experiment was feasible. Nobumichi Hozumi, a postdoctoral fellow in my laboratory, and I therefore decided to give it a try although we were keenly aware of the intense labour required by this type of experiment. As hybridization probes we used purified κ or λ light chain mRNA (*V* + *C* probe) and its 3′-half fragment (*C* probe) that had been iodinated to a high specific activity. The rationale of the experiment was as follows. First, if an immunoglobulin polypeptide chain is encoded by two 'genes', *V* and *C*, in the germline genome, it is highly probable that treatment with a restriction enzyme will separate these DNA sequences into fragments of distinct size, thus allowing their electrophoretic separation. Second, if a somatic rearrangement joins the *V* and *C* 'genes' it is also highly probable that the myeloma DNA digested with the same restriction enzyme will contain a DNA fragment carrying both *V* and *C* 'genes'.

The results obtained were clear-cut. To our pleasant surprise the patterns of hybridization of the embryo (a substitute of germline) DNA and a κ-myeloma DNA were not only drastically different but also perfectly consistent with the occurrence of separate *V* and *C* 'genes' and a joined *V* plus *C* gene, respectively (*16*). We were of course aware of the alternative interpretations of the results, such as fortuitous modification of the enzyme cleavage sites in one of the two types of DNA. However, we considered these alternative explanations of the results unlikely because they all required multiple fortuitous events. Our confidence was fortified soon afterwards as the development of Southern blot techniques allowed us to carry out more extensive analysis using a variety of restriction enzymes and myeloma cells.

Joining of Gene Segments

While the experiments with restriction enzymes were informative, details of the rearrangement were difficult to come by with this approach. Fortunately, recombinant DNA technology was just becoming available and was the ideal means for this purpose. Debates on the possible hazards of this type of research were flaring, initially in the USA and shortly afterwards in European countries. In order to make sure that our research would not become a target of controversy, Charlie and I got in touch with Werner Arber at the University of Basel who was co-ordinating recombinant DNA research activities in Switzerland. A small informal work group was set up by the local researchers interested in this technique. The consensus of the group which was supported by most of the other Swiss researchers was that we should all follow the practices and guidelines being adopted in the USA. We met about once a month and exchanged information regarding both ethical and practical aspects of the technology.

On the basis of the previous experiments attempting to count immunoglobulin genes, I thought that it would be wise to start with the mouse λ light chain system, the simplest of all chain types that had been studied. Our goal was to clone the V_λ and C_λ 'genes' in the germline state from embryonic cells as well as the rearranged *V* plus *C* 'genes' from a λ myeloma, and to determine the relationship between these genomic DNA clones by electron microscopy and DNA sequencing. No precedent existed at that time for cloning 'unique' eukaryotic genes. We therefore had to devise a few tricks as we attempted to clone the first immunoglobulin gene. For instance, our available probe at that time was again 95% pure mRNA rather than a cDNA clone. This situation made the screening of a large number of DNA clones difficult because of the high background. To avoid this problem we pre-enriched the λ gene-containing genomic DNA fragments as much as possible using preparative *R*-loop formation (*17,18*), so that the DNA library constructed would have the clone of interest at a high frequency.

Starting with the embryonic DNA we could isolate a clone that clearly hybridized specifically with the λ mRNA (*18*). When an electronmicroscopist, Christine Brack, who had just joined us from the Biozentrum of the University of Basel, examined the mixture of this clone and λ mRNA that had been annealed under an appropriate condition, she found a beautiful *R* loop from which about a half of the mRNA strand protruded. This and additional analysis convinced us that we had cloned a V_λ 'gene' to which no *C* 'gene' was contiguously attached, thus confirming at the DNA clone level that the *V* and *C* 'genes' are indeed separate in the germline genome. A subsequent DNA sequencing study carried out in collaboration with Allan Maxam and Walter Gilbert of Harvard University revealed that this DNA clone corresponded to the *V* 'gene' for the λ_2 subtype (*19*).

In the meantime Minoru Hirama, another postdoctoral fellow, succeeded in preparing λ and κ cDNA clones. Once these probes became available isolation

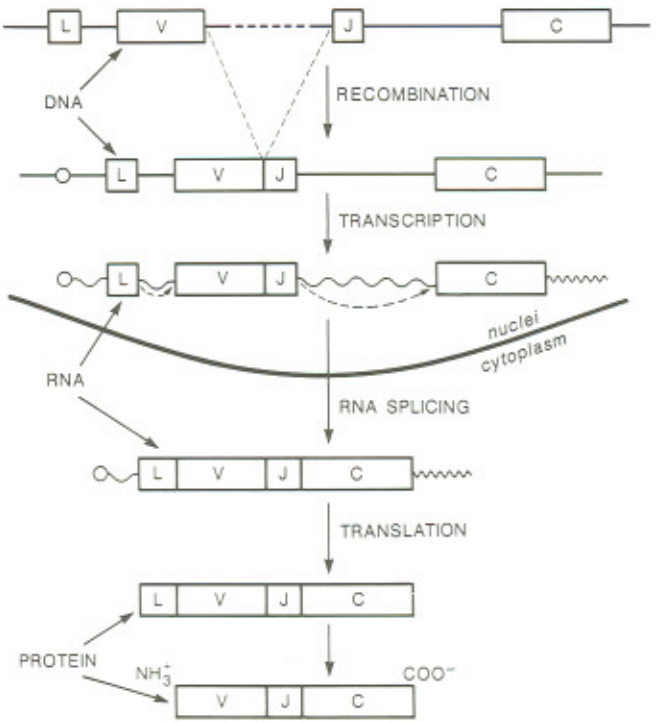

Figure 1 The basic scheme for rearrangement and expression of an immunoglobulin light chain gene. At top is an arrangement of the gene segments on a germline genome. Somatic rearrangement links the *V* and *J* gene segment and generates a complete light chain gene shown just below the germline genome. The entire gene containing the leader exon (*L*), the *V* region (*V* and *J*), the *C* region exon (*C*), and the introns present between these exons are transcribed into a pre-mRNA in the nuclei of the B cell. The pre-mRNA is processed by RNA splicing as it is transported from the nuclei to the cytoplasm. The resulting mRNA, devoid of introns, is translated in the endoplasmic reticulum into a nascent polypeptide chain, from which a mature λ light chain is generated after cleavage of the signal peptide.

of the genomic clones became much easier. My assistant Rita Schuller and I isolated a number of genomic DNA clones from λ and κ chain-synthesizing myelomas as well as from embryos (*20,21*). Analysis of these DNA clones by electron microscopy, by restriction enzyme mapping, and by DNA sequencing not only confirmed the somatic rearrangement of immunoglobulin genes but also revealed some striking features of their arrangement and rearrangement (Fig. 1). These can be summarized as follows:

(1) Although the *V* and *C* 'genes' are rearranged and are much closer to each other in myeloma cells than in embryo cells, they are not contiguous and

are separated by a few kilobases of DNA sequence that does not participate in coding of the polypeptide chain. This untranslated DNA sequence present within the rearranged, complete immunoglobulin gene was unanticipated and was also among the first demonstrations of an intron in eukaryotic genes (*22*).

(2) The *V* 'gene' found in the germline genome is about 13 codons short when it is compared to the length of the conventionally defined *V* region. The missing codons were found in a short stretch of DNA referred to as a *J* or joining) gene segment that is located many kilobases away from the incomplete *V* 'gene' (referred to as a *V* gene segment) and a few kilobases upstream of the *C* 'gene' (also referred to as a *C* gene segment). In myeloma cells the rearrangement event attaches the *J* gene segment to the *V* gene segment and thereby creates a complete *V* region 'gene' (*20,23*).

(3) The signal peptide is encoded in yet another DNA segment referred to as the *L* (or leader) exon that is separated from the *V* gene segment by a short intron (*19,23*).

Finding that the V_λ 'gene' was split into two gene segments, V_λ and J_λ, in the germline genome was completely unexpected. But as soon as this discovery was made its implication for the somatic generation of antibody diversity was obvious. If the germline genome carries multiple copies of different *V* and *J* gene segments, the number of complete *V* 'genes' that can be generated by random joinings between these two types of gene segments would be much greater than the total number of the inherited gene segments. Thus, contrary to the Dryer and Bennett original concept, DNA rearrangement can provide a major means for the somatic diversification of antibody molecules. The amino acid sequence data of the κ light and heavy chains were consistent with this concept (*24,25*). Indeed, the nucleotide sequence analysis of the mouse κ chain gene complex carried out both in my laboratory and in Phillip Leder's laboratory at the United States National Institutes of Health confirmed that a germline genome contains multiple *V* and *J* gene segments and that these gene segments are joined in different combinations in each myeloma cell (*20,26*). Four different J_κ gene segments were found several kilobases upstream of the C_κ gene segment. The exact number of V_κ gene segments is unknown even today, but it is estimated to be 200 to 300 (*27*).

Heavy Chain Genes

Inasmuch as an immunoglobulin heavy chain is also composed of *V* and *C* regions, it was reasonable to expect that its gene also would undergo the type of DNA rearrangement described for the light chain genes. This supposition was confirmed by Leroy Hood and his coworkers at California Institute of Technology and by ourselves (Fig. 2) (*28,29*). As in κ genes, four *J* gene segments were found several kilobases upstream of the *C* gene segments coding for

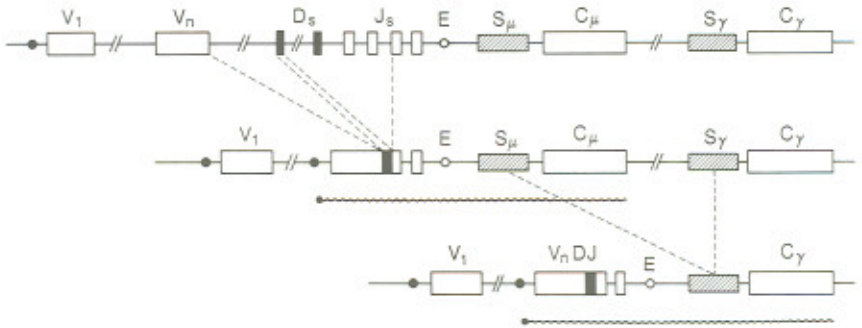

Figure 2 Organization of the immunoglobulin heavy chain family. At the top, middle and bottom are organization in a germline genome, in a genome of B cells synthesizing a μ class heavy chain and in a genome of a plasma cell synthesizing a γ class heavy chain, respectively. A mouse haploid genome carries several hundred different *V* gene segments, about a dozen *D* gene segments, four *J* gene segments, and one copy of *C* gene segment for each of the eight different classes or subclasses of immunoglobulin heavy chains. In a virgin B cell one copy each of the *V*, *D* and *J* gene segment pools have been linked up and the joined *VDJ* DNA sequence is transcribed into a pre-mRNA together with the C_μ gene segment. In different B cells of the same organism, a different set of *V*, *D* and *J* gene segments are usually hooked up and expressed. As the virgin B cell differentiates either to a plasma cell or to a memory B cell (see Fig. 5) the second type of somatic recombination called 'switch recombination' often occurs between a region (S_μ) located upstream of the C_μ gene segment and another region (S_γ) located upstream of the C_γ gene segment.

the *C* region of the μ class heavy chain. Multiple *V* gene segments were also identified.

While these features of the organization of heavy chain genes are essentially the same as those of the light chain genes, one observation made during these studies suggested that the somatic assembly of gene segments plays an even more prominent role in the diversification of heavy chains than of light chains. It was found that from one or two to a dozen amino acid codons, present in the *V–J* junction region of the assembled gene, are not found in either of the corresponding germline *V* or *J* gene segments (*30,31*). This suggested that a third type of short gene segment referred to as *D* (or diversity) might participate in the somatic assembly of a heavy chain gene. Indeed, Hitoshi Sakano and Yoshi Kurosawa, two postdoctoral fellows in my laboratory, soon discovered about a dozen *D* gene segments (*32,33*) which were subsequently mapped in a region upstream of the *J* cluster in the germline genome (*34,35*). Thus, the construction of a complete heavy chain *V* 'gene' requires two DNA recombinational events, one joining a *V* with a *D* gene segment and the other the same *D* with a *J* gene segment.

Recombination Rules

The joining of *V–J* or *V–D–J* involves a site-specific recombination. It might therefore be expected that these gene segments would carry sequences in the vicinity of the joining ends that are recognized by a putative site-specific recombinase. Furthermore, such recognition sequences are likely to be common for all gene segments of a given type (for example V_κ segments), because they all seem to be capable of joining with the common set of gene segments of the appropriate type (for example, J_κ segments). Indeed, a heptamer and a nonamer sequence are conserved in the region immediately downstream of each V_κ gene segment (Fig. 3) (*36,37*). Sequences complementary to the V_κ heptamer and nonamer were also found in the region immediately upstream of each of the

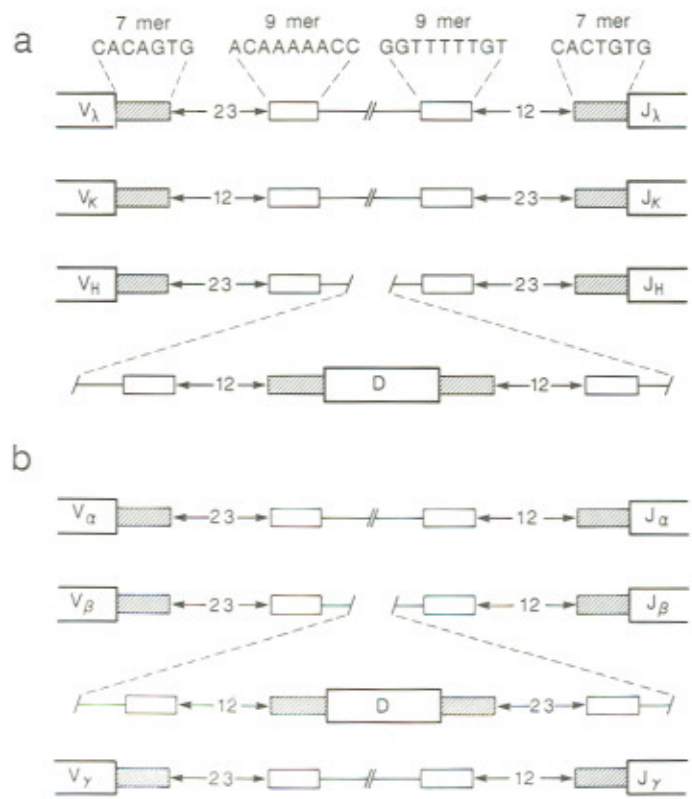

Figure 3 Putative recognition sequences for the rearrangement of immunoglobulin and T cell receptor genes. The conserved heptamer and nonamer sequences and the length of the spacer between these sequences are schematically illustrated for immunoglulubin (a) and for T cell receptor (b) gene families. The sequences shown at the top of (a) and (b) are consensus sequences. Individual sequences may deviate from these consensus sequences by a few nucleotides.

four J_κ gene segments. The same sets of sequences were also found in the corresponding regions of the V_λ and J_λ gene segments (*36*). When the heavy chain *V* and *J* gene segments were analysed subsequently, they too had the common conserved sequences (*30,31*). Furthermore, *D* gene segments carry the heptamer and nonamer sequences both upstream and downstream (*32,33*). Another interesting feature of these putative recognition sequences is the fact that the length of the spacer between the heptamer and nonamer is either 12 or 23 base pairs (*30,31*). In addition, a gene segment carrying a recognition sequence with one type of spacer is able to join only with a gene segment with the spacer of the other type. This 12/23 base pair spacer rule seems to be adhered to strictly. Little is currently known about the recombinase, but proteins with an affinity to the heptamer or nonamer have been identified in the extract of Abelson virus transformed pre-B-cell lines in which the rearrangement occurs *in vitro* at a relatively high frequency (*38,39*). Since then, a pair of genes, RAG-1 and RAG-2, whose protein products are essential for *V*–(*D*)–*J* joining has been identified and cloned in David Baltimore's laboratory (*40,41*).

Diversity Generated at the Joins

The deduced amino acid sequence of a germline J_κ gene segment was compared with the determined amino acid sequences of those κ chains that are encoded in part by that J_κ gene segment. The joining site is not prefixed but rather shifts toward upstream or downstream by several base pairs in different joining events (*36,37*). This flexibility in the precise site of the joining was subsequently found to be characteristic of the joining ends of other gene segments rather than of just J_κ gene segments (*31*). It applies even when the same pair of gene segments were joined in different B cell percursors, such that the completed *V* 'genes' are likely to have slightly different codons in the junction regions.

The *V*–*D* and *D*–*J* junctions exhibit diversity of yet another type. We found that up to a dozen base pairs of essentially random sequence are inserted in these junctions, apparently without a template, during the breakage and reunion of the recombining gene segments (*32,33*). While the precise mechanism is not yet known, the terminal deoxynucleotide transferase which is found in early B lymphatic nuclei, or an enzyme with similar characteristics, is thought to play a role in this phenomenon (*42*).

The part of the *V* region affected by the above two diversification mechanisms is limited. But this does not mean that they do not play a significant role in the determination of antibody specificity. On the contrary, the junctions encode the most variable two of the six loops of polypeptides that make up the antigen binding region of the antibody molecule (Fig. 4). Furthermore, specific cases are known where the affinity of an antibody to a defined antigen is drastically altered by a slight change in one junction sequence (*43*). Thus, the junctional variation also is a potent somatic generator of antibody diversity.

Somatic Mutation

When F. Macfarlane Burnet proposed the clonal selection theory, he recognized the need for some kind of random genetic process in order to generate antibodies able to bind specifically to the vast variety of antigens (*44*). He considered somatic mutations as the most plausible mechanism. Subsequently, this idea was adopted and forcefully presented by many including Joshua Lederberg, Niels Kaj Jerne and Melvin Cohn (*5–7*).

The amino acid sequence data accumulated by Martin Weigert in Melvin Cohn's laboratory at the Salk Institute provided an excellent opportunity to examine directly the role of somatic mutations in antibody diversity (*7,11*). They had analysed the λ_1 light chains derived from 18 myelomas. All the mice were of an inbred strain, BALB/c, and so should have been genetically identical. They found that twelve of the $V_{\lambda 1}$ regions were identical but that the other six differed both from the majority sequence and from one another by only one, two, or three amino acid residues. They proposed that BALB/c mice may carry only one germline $V_{\lambda 1}$ 'gene' which codes for the majority sequence, and that all the other $V_{\lambda 1}$ regions observed are encoded by somatic mutants of this single $V_{\lambda 1}$ 'gene' that arose in B cell development. As I have already mentioned in an earlier section, our gene-counting experiment by hybridization kinetics suggested that the germline BALB/c genome carries no more than a few $V_{\lambda 1}$ 'genes'. This number was reduced to one when we re-evaluated the copy number by the more reliable Southern blotting method (*20*). The final proof of somatic mutation in $V_{\lambda 1}$ came when we cloned and sequenced the sole germline $V_{\lambda 1}$ segment and the rearranged λ_1 genes expressed in a myeloma (*23*). As Weigert and Cohn guessed, the nucleotide sequence of the germline $V_{\lambda 1}$ gene segment corresponded to the major amino acid sequence, while the λ_1 gene expressed in the myeloma had been altered by single base changes.

Since this work, several subsets of κ light and heavy chains and their germline, *V* gene segments have been analysed by cloning and sequencing (*45–48*), and have all confirmed that somatic mutations further amplify the diversity encoded in the germline genome. Particularly revealing was the analysis carried out by Patricia J. Gearhart, Leroy Hood and their coworkers for the V_H regions associated with the binding of phosphorylcholine (PC). They demonstrated that single base changes can be extensive and yet are restricted to the joined *VDJ* sequences and the immediately adjacent regions (*49,50*).

Developmental Control of Rearrangement and Hypermutation

Why have two extraordinary somatic genetic mechanisms, recombination and hypermutation, evolved in the immune system in order to carry out what appears to be one task – namely, to diversify antibodies? The answer may be the differential roles of these two genetic mechanisms. Thanks to the efforts of

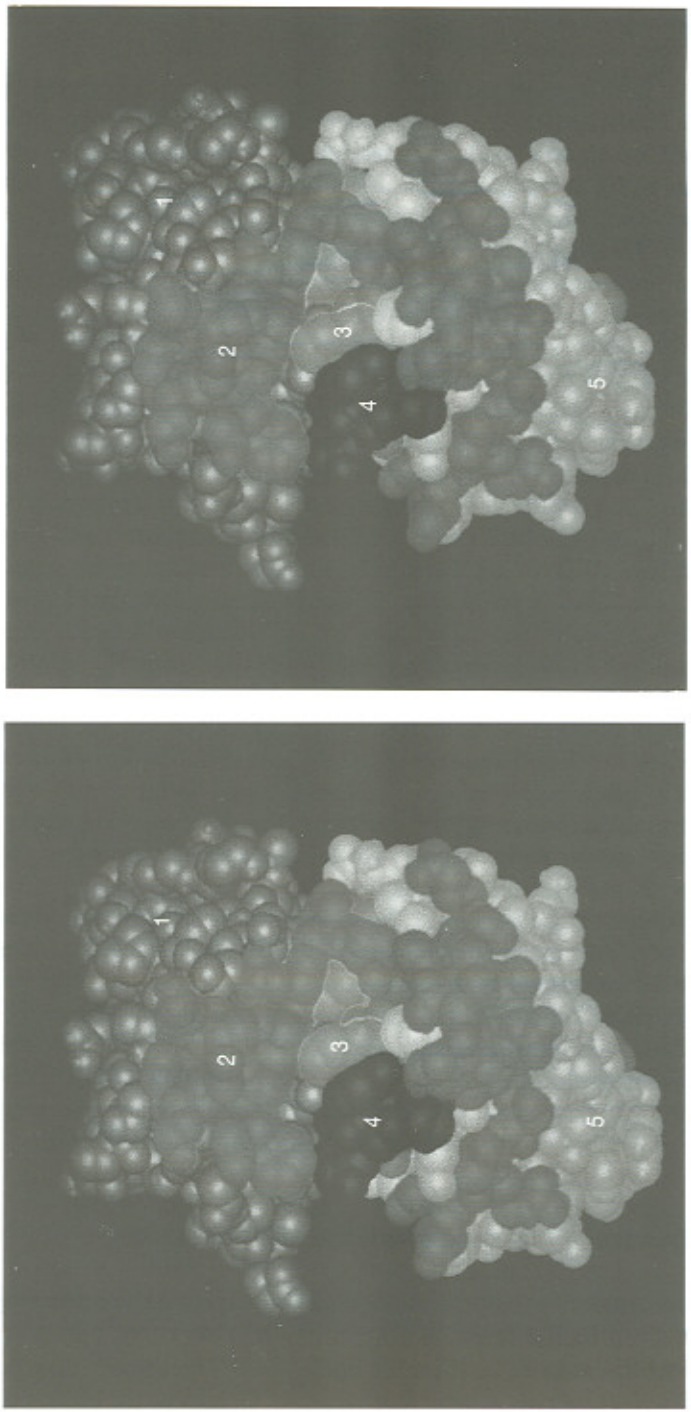

Figure 4 (caption opposite)

several independent groups of cellular and molecular immunologists, a general picture is emerging that describes the relationship between the states of B cell development and the occurrence of somatic recombination or mutation (Fig. 5) (*51–57*). Somatic recombinations contributing to diversity are initiated first for the heavy chain and then for the light chain during the differentiation of progenitor cells, and the completion of somatic recombination is accompanied by the appearance of virgin B cells (*58–60*). These B cells form clones each of which is composed of cells bearing homogenous IgM molecules as surface receptors. Thus, somatic recombination is completed *prior to* any possible interaction of a B cell with antigens.

When an antigen enters lymphatic tissue for the first time, it will be screened by these virgin B cells. The small fraction of these B cells that happen to have sufficient affinity for the antigenic determinants in question will respond and follow either of two pathways: they will produce the primary antibody response, or they will contribute to the generation of memory B cells. In the former pathway, the selected B cells will proliferate and differentiate into antibody-secreting plasma cells. During this process, the *C* region of the heavy chain can switch from μ to another class, but mutation is rare in either the heavy or the light chain *V* region. Consequently, the antibodies secreted by plasma cells in the primary response would largely have the same *V* regions as the immunoglobulin receptors on the virgin B cells from which they derive.

By contrast, immunoglobulin remains in the cell surface receptor form during the other pathway taken by the antigen-activated virgin B cells, namely the generation of memory cells. During this process, the hypermutation

Figure 4 Space-filling, stereo image of antibody combining site. Atomic coordinates of mouse immunoglobulin MOPC 603 (62) were used to produce the picture. The heavy chain variable domain is colour-coded dark grey, and the light chain variable domain light grey. The hypervariable regions (except the V_H third hypervariable region) are blue, the heavy chain segment coded for by the *D* gene is red, and the heavy and light chain segments coded for the *J* genes are yellow. The *D* segment corresponds virtually exactly to the third heavy chain hypervariable region; hypervariable regions were defined as in Novotney *et al.* (63) except for the heavy chain second hypervariable region, which is marked as defined by Kabat *et al.* (25). The antigen of this particular immunoglobulin, phosphoryl choline, binds into the cavity in the middle of the picture between the V_H and V_L domains, making contacts to amino acid residues belonging to the V_H and *J* segments of the heavy chain and the V_L structures of antibodies which bind the protein antigen lysozyme (*64, 65*). There, the contact areas contributed by the *D* segment amount to 50% and 24%, respectively, of the total heavy chain contact area. This image was computer generated by Jiri Novotny using the program SPHERE of Robert Bruccoleri. Because of reproduction difficulties, a black-and-white version of the original colour photograph is shown: 1, heavy chain variable domain; 2, hypervariable regions; 3, heavy and light chain *J* regions; 4, heavy chain *D* region; 5, light chain variable domain.

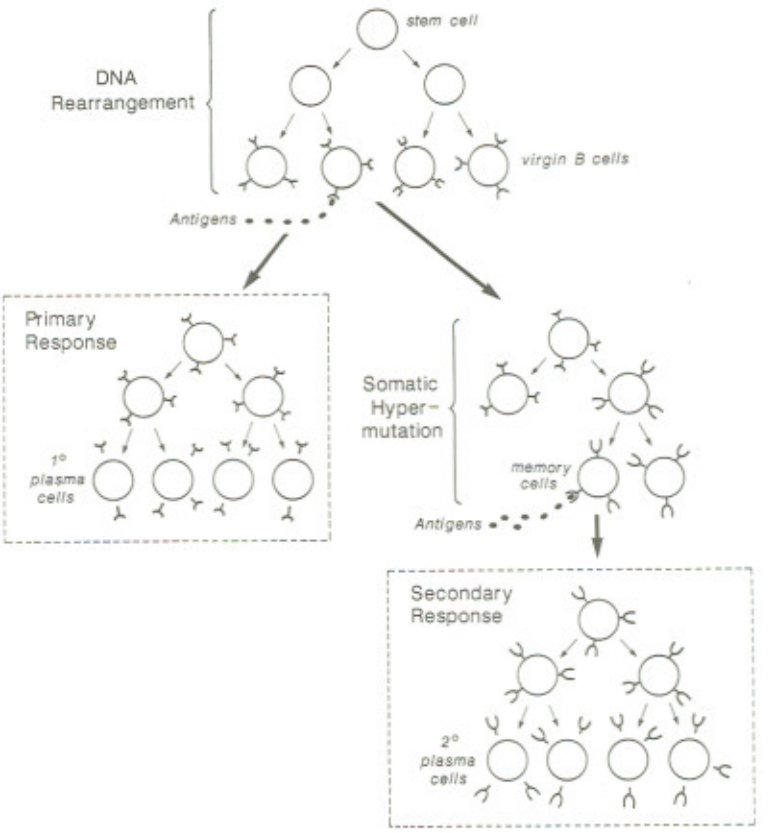

Figure 5 Differentiation of B cells. Note that the receptors present in the memory cells and the antibody molecules secreted by the plasma cells of the secondary response have a tighter fit to the antigen than the receptors on the ancestral virgin B cells or the antibodies secreted by the plasma cells of the primary response. See text for the full explanation.

apparatus appears to be most active and the rate of the mutation approaches 10^{-3} base substitution per cell per generation. Antigen selects, in a stepwise fashion, better and better fitting mutants, so that the immunoglobulins on the surface of memory B cells achieve a substantially higher affinity than the immunoglobulins on the ancestral virgin B cells. Switch recombination also occurs frequently during this process too. When the same antigen as the one that elicited the primary response re-enters the body, the memory B cells are selectively propagated and differentiate into plasma cells. This is the so-called 'secondary antibody response' which, therefore, consists of high affinity antibody of 'mature' isotype; these antibodies show extensive somatic mutation in their *V* regions. Somatic mutations appear to cease after memory cells are

generated, and little or no further mutation takes place during the secondary antibody response.

This scheme of B cell differentiation can be rephrased as follows. An organism is prepared for infection with pathogens bearing virtually any antigens with a large variety of resting B cells. These B cells bear unique immunoglobulin receptors encoded by one copy each of complete light and heavy chain genes that have been constructed by a random or quasirandom assembly of the inherited gene segments. Since the assembly occurs independent of antigens and since the inherited gene segments are not usually selected during evolution for precise fit to particular antigens, the antibody secreted by the plasma cells derived directly from the selected resting virgin B cells during a primary antibody response usually have a relatively low affinity. By contrast, the frequent single base changes that occur during the generation of memory B cells provide the organism with a great variety of finely altered immunoglobulin receptors from which only those with the best fit to the antigen in question will be selected. Since the plasma cells generated during the secondary antibody response are mostly direct descendants of these memory B cells having no further alterations in the antigen-combining sites, these antibodies usually exhibit a much higher affinity for antigen than do primary antibodies. This explains the long-known phenomenon of affinity maturation of antibodies during the course of repeated immunizations (*61*).

Thus, somatic creation of antibody genes can be viewed as a two-step process. In the first step, blocks of gene segments are employed to build, in an antigen-independent fashion, a set of genes coding for antibodies of great diversity but with low affinity. In the second step, once the antigen is defined, a small selected set of B cells bearing low-affinity antibodies as cell surface receptors undergo somatic mutations with the result that a fraction of them develop a higher affinity to that antigen and can be selected for further expansion. This process improves the ability of the immune system to detect a low concentration of antigens. One wonders what happens to those cells in which mutation did not improve affinity. A recent study suggests that at least some of these cells may be set aside for selection by different antigens (*56*). Thus somatic mutation may also contribute to the repertoire of receptors specific for antigens not previously introduced into an immune system.

Concluding Remarks

Use of restriction enzymes and recombinant DNA methods allowed resolution of a long-standing and central issue in immunology, the genetic origins of antibody diversity. It turned out that an organism does not inherit even a single complete gene for antibody polypeptide chains. Rather, the genetic information is transmitted in germline as no more than several hundred gene segments. Through a series of specialized somatic recombinations occurring specifically

during the differentiation of B lymphocytes, these gene segments are assembled into tens of thousands of complete genes. Somatic hypermutation occurring in these assembled genes further diversifies antibody polypeptide chains, so that B cells displaying immunoglobulin receptors having a better fit to a given antigen can be selected in a later phase of B cell differentiation. Thus, in the immune system, organisms have exploited two major processes for modification of DNA, recombination and mutation, as a means to diversify somatically the limited amount of inherited genetic information in order to cope with the vastly diverse antigen universe.

Why has somatic diversification been necessary in the evolution of the immune system? Micro-organisms and substances produced by them are the primary source of biologically relevant antigens against which vertebrates need to produce antibodies for survival. Since the generation time of micro-organisms is several orders of magnitude shorter than that of vertebrates, the former can produce generic variants much faster than the latter. Thus, if genetic alterations in the germline genome were to be the only source of antibody diversity, vertebrates would be unable to deal effectively with the rapidly changing world of antigens. Somatic diversification allows the individual organism to generate a virtually limitless number of lymphocyte variants. Like organisms in an ecosystem, these lymphocytes are subject to selection by antigens and the fittest will survive. Thus, as Jerne and Burnet were aware, the individual immune system can be conceived of as a kind of Darwinian microcosm.

Acknowledgements

The work summarized in this article is the result of collaboration with many colleagues, students and technical assistants. I wish to extend my sincere thanks to every one of them. I also extend my thanks to the Hoffman LaRoche Company which so generously supported my work in Basle. My special thanks are extended to Charles A. Janeway Jr, Nancy Hopkins and Yohtaroh Takagaki for many useful comments on the manuscript, to Jiri Novotny for preparing Fig. 4, and to my secretary, Eleanor Lahey, Basle, for her tireless devotion.

Notes and References

1. R. R. Porter, *Science* **180**, 713 (1973).
2. G. M. Edelman, *Science* **180**, 830 (1973).
3. N. Hilschmann and L. C. Craig, *Proc. Natl Acad. Sci. USA* **53**, 1403 (1965).
4. L. Hood and D. W. Talmage, *Science* **168**, 325 (1970).
5. J. Lederberg, *Science* **129**, 1649 (1959).
6. N. K. Jerne, *Eur. J. Immunol.*, **1**, 1 (1971).
7. M. Cohn, B. Blomberg, W. Geckeler, W. Raschka, R. Riblet and M. Weigert, in *The Immune System: Genes, Receptors, Signals* (E. E. Sercarz, A. R. Williamson and C. F. Fox, eds), Academic Pres, New York (1974), p. 89.

8. A. H. Gelderman, A. V. Rake and R. J. Britten, *Proc. Natl Acad. Sci. USA* **68**, 172 (1971).
9. J. O. Bishop, *Biochem. J.* **126** 171 (1971).
10. S. Tonegawa, *Proc. Natl Acad. Sci. USA* **73**, 203 (1976).
11. M. Weigert, M. Cesari, S. J. Yonkovich and M. Cohn, *Nature* **228**, 1045 (1970).
12. A. Campbell, *Adv. Genet.* **11**, 101 (1962).
13. W. J. Dryer and J. Bennett, *Proc. Natl Acad. Sci. USA* **73**, 3628 (1965).
14. J. J. Danna, G. H. Sack Jr and D. Nathans, *J. Mol. Biol.* **78**, 363 (1973).
15. E. Southern, *J. Mol. Biol.* **98**, 503 (1975).
16. N. Hozumi and S. Tonegawa, *Proc. Natl Acad. Sci. USA* **73**, 3628 (1976).
17. M. Thomas, R. L. White and R. N. Davis (1976). *Proc. Natl Acad. Sci. USA* **73**, 2294 (1976).
18. S. Tonegawa, C. Brack, N. Hozumi and R. Schuller, *Proc. Natl Acad. Sci. USA* **74**, 3518 (1977).
19. S. Tonegawa, A. M. Maxam, R. Tizard, O. Bernard and W. Gilbert, *Proc. Natl Acad. Sci. USA* **75**, 1485 (1978).
20. C. Brack, M. Hirami, R. Lenhard-Schuller and S. Tonegawa, *Cell*, **15**, 1 (1978).
21. R. Lenhard-Schuller, B. Hohn, C. Brack, M. Hirama and S. Tonegawa, *Proc. Natl Acad. Sci. USA* **75**, 4709 (1978).
22. C. Brack and S. Tonegawa, *Proc. Natl Acad. Sci. USA* **74**, 5652 (1977).
23. O. Bernard, N. Hozumi and S. Tonegawa, *Cell* **15**, 1133 (1978).
24. M. Weigert, L. Gatmaitan, E. Loh, J. Schilling and L. Hood, *Nature* **276**, 785 (1978).
25. E. A. Kabat, T. T. Wu, H. Bilofsky, M. Reid-Miller and H. Perry, *Publication No. 91-3242*, US Department of Health and Human Services Publication, Washington, DC (1983).
26. J. G. Seidman, A. Leder, M. Nau, B. Norman and P. Leder, *Science* **208**, 11 (1978).
27. S. Cory, B. M. Tyler and J. M. Adams, *J. Mol. Appl. Genet.* **1**, 103 (1981).
28. M. Davis, K. Calame, P. Early, D. Livant, R. Joho, I. Weismann and L. Hood, *Nature* **283**, 733 (1980).
29. R. Maki, A. Traunecker, H. Sakano, W. Roeder and S. Tonegawa, *Proc. Natl Acad. Sci. USA* **77**, 2138 (1980).
30. P. Early, H. Huang, M. Davis, K. Calame and L. Hood, *Cell* **19**, 981 (1980).
31. H. Sakano, R. Maki, Y. Kurosawa, W. Roeder and S. Tonegawa, *Nature* **286**, 676 (1980).
32. H. Sakano, Y. Kurosawa, M. Weigert and S. Tonegawa, *Nature* **290**, 562 (1981).
33. Y. Kurosawa, H. von Boehmer, W. Haas, H. Sakano, A. Traunecker and S. Tonegawa, *Nature* **290**, 565 (1981).
34. Y. Kurosawa and S. Tonegawa, *J. Exp. Med.* **155**, 201 (1982).
35. C. Wood and S. Tonegawa, *Proc. Natl Acad. Sci. USA* **80**, 3030 (1983).
36. H. Sakano, K. Huppi, G. Heinrich and S. Tonegawa, *Nature* **280**, 288 (1979).
37. E. E. Max, J. G. Seidman and P. Leder, *Proc. Natl Acad. Sci USA* **76**, 3450 (1979).
38. B. D. Halligan and S. V. Desiderio, *Proc. Natl Acad. Sci. USA* **84**, 7019 (1987).
39. R. J. Agutera, S. Akira, K. Okagaki and H. Sakano, *Cell* **51**, 909–917 (1987).
40. D. G. Schatz, M. A. Oettinger and D. Baltimore, *Cell* **59**, 1035 (1989).
41. M. A. Oettinger, D. G. Schatz, C. Gorka and D. Baltimore, *Science* **248**, 1517 (1990).
42. F. Alt and D. Baltimore, *Proc. Natl Acad. Sci. USA* **79**, 4118 (1982).
43. T. Azuma, V. Igras, E. Reilly and H. N. Eisen, *Proc. Natl Acad. Sci. USA* **81**, 6139 (1984).
44. F. M. Burnet, *The Clonal Selection Theory of Acquired Immunity*, Cambridge University Press, London (1959).
45. A. L. M. Bothwell, M. Paskind, M. Reth, T. Imanishi-Kari, K. Rawesky and D. Baltimore, *Cell* **24**, 624 (1981).
46. S. Crews, J. Griffin, H. Huang, K. Calame and L. Hood, *Cell* **25**, 59 (1981).
47. D. Givol, R. Zakut, K. Effron, G. Rechavi, D. Ram and J. B. Cohen, *Nature* **292**, 426 (1981).
48. E. Selsing and U. Storb, *Nucl. Acids Res.* **9**, 5725 (1981).
49. P. J. Gearhart, N. D. Johnson, R. Douglas and L. Hood, *Nature* **291**, 29 (1981).
50. S. Kim, M. Davis, E. Sinn, P. Patten and L. Hood, *Cell* **27**, 573 (1981).
51. B. A. Askonas and A. R. Williamson, *Eur. J. Immunol*, **2**, 487 (1972).

52. C. Berek, G. M. Griffiths and C. Milstein, *Nature* **316**, 412 (1985).
53. D. McKean, K. Huppi, M. Bell, L. Standt, W. Gerhard and M. Weigert, *Proc. Natl Acad. Sci. USA* **81**, 3180 (1984).
54. K. Okumura, M. H. Julius, T. Tsu, L. A. Herzenberg and L. A. Herzenberg, *Eur. J. Immunol.* **6**, 467 (1976).
55. M. Sablitzberg, C. Kocks and K. Rajewsky, *EMBO J.* **4**, 345 (1985).
56. M. Sickevitz, C. Kocks, K. Rajewsky and R. Dildrop, *Cell* **48**, 757 (1987).
57. L. J. Wysocki, T. Manser and M. L. Gefter, *Proc. Natl Acad. Sci. USA* **83**, 1847 (1986).
58. R. Maki, J. Kearney, C. Paige and S. Tonegawa, *Science* **209**, 1366 (1980).
59. P. P. Perry, D. E. Kelley, C. Coleclough and J. F. Kearney, *Proc. Natl Acad. Sci. USA* **78**, 247 (1981).
60. M. G. Reth, P. A. Ammirati, S. J. Jackson and F. W. Alt, *Nature* **317**, 353 (1985).
61. H. N. Eisen and G. W. Siskind, *Biochemistry* **3**, 3996 (1964).
62. Y. Satow, G. H. Cohen, E. A. Padlan and D. R. Davis, *Mol. Biol.* **190**, 593 (1986).
63. J. Novotny, R. E. Bruccolei, J. Newell, D. Murphy, E. Harber and M. J. Karplus, *J. Biol. Chem.* **258**, 14433 (1983).
64. A. G. Amit, R. A. Mariuzza, S. E. V. Phillips and R. Poljak, *Science* **233**, 747 (1986).
65. S. Sheriff, E. W. Silverton, E. A. Padlan, G. H. Cohen, S. J. Smith-Gill, B. C. Finzel and D. R. Davies, *Proc. Natl Acad. Sci USA* **84**, 8075 (1987).