*SOMATIC RECOMBINATION OF DUPLICATED GENES: AN HYPOTHESIS ON THE ORIGIN OF ANTIBODY DIVERSITY**

By G. M. Edelman and J. A. Gally

THE ROCKEFELLER UNIVERSITY

Communicated by Theodore Shedlovsky, December 15, 1966

The specificity of antibodies for different antigens appears to result from differences in the amino acid sequences of both the heavy and the light polypeptide chains of which antibody molecules are composed.[1-4] Although specific antibodies are too heterogeneous for complete amino acid sequence analysis, homogeneous immunoglobulins may be obtained for this purpose from various plasma cell tumors. Bence-Jones proteins have been shown to be light polypeptide chains[5, 6] and data have accumulated on the amino acid sequences of various samples of both human[7-9] and murine origin.[10, 11]

These amino acid sequence analyses show that light polypeptide chains of type K (κ chains) are composed of 212–214 amino acid residues.[12] From residue 1

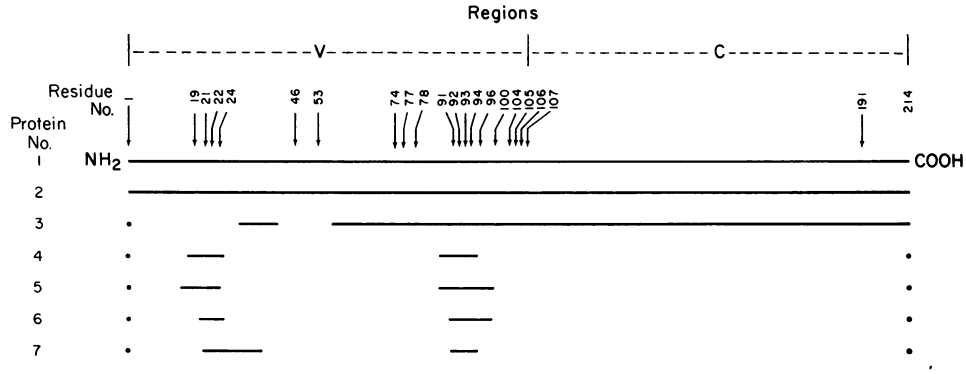

FIG. 1.—Diagram indicating variable and constant regions of light polypeptide chains of human origin (Bence-Jones proteins of antigenic type K); *V*, variable region; *C*, constant region; NH_2, amino terminus; COOH, carboxyl terminus. Incomplete lines indicate proteins only portions of which have been sequenced. Arrows indicate positions where at least one amino acid interchange has been observed among the seven proteins diagrammatically presented here. The numbers are based on the sequence of protein Ag (see refs. 7, 21).

(the amino terminus) to residues 106–107, different Bence-Jones proteins have similar but not identical sequences. A number of positions in the sequence (at least 20 in the human proteins) show substitutions of one amino acid for another (Fig. 1). The majority of the substitutions from protein to protein can be accounted for by single base changes in codons corresponding to the amino acids which have been interchanged.

In contrast to the diversity found in the amino terminal half of Bence-Jones proteins, the carboxyl terminal half (residues 107 (108)–212 (214)) has a constant amino acid sequence except for one position (Fig. 1). This exception is related to an allotypic difference[13, 14] and is represented by a substitution of valine for leucine at position 189 (191).

Although not as much information is available on the sequence of heavy chains,

recent studies[15, 16] suggest that they also consist of a variable portion nearer the amino terminus and a constant portion nearer the carboxyl terminus. The proposal of an antibody model[17] in which amino acid variation in both heavy and light chains is related to differences in specificity and invariance of sequence is related to interchain bonding and other functions is supported by various lines of experimental evidence.[1-4]

An hypothesis to account for the origin of diversity in the amino acid sequences of different antibodies should explain the following facts:

(1) The multiplicity of amino acid replacements in the variable half (V) and the relative invariance of sequence in the constant half (C).

(2) The presence of a majority of amino acid interchanges consistent with single base replacements in the genetic code.

(3) The similarity in lengths of the V and C regions (in the case of the light chains).

(4) The occurrence of invariant segments *within* the V region as well as the finding that certain positions show interchanges among only a few amino acids (e.g., N-terminal glu or asp in κ chains).

(5) The failure to observe a high recombination frequency among allotypes.[13]

(6) The fact that a particular plasma cell tumor can produce a single well-defined protein, the sequence of which does not change from generation to generation. This may be correlated with the finding that the majority of single plasma cells from animals immunized with two unrelated antigens produce antibodies either against one or the other antigen.[18, 19]

Several hypotheses have been advanced to account for some or all of these facts. They may be divided into two general categories: multiple gene hypotheses and somatic variation hypotheses. Multiple gene hypotheses[20-22] propose that each variable region of an antibody molecule is coded for by a different gene in the germ line. Hypotheses of this type face difficulties in explaining how a very large number of similar genes could have evolved or be maintained by natural selection. Somatic variation hypotheses invoke variation of replication in somatic cells (inverted crossing-over,[23] breaks in DNA strands with mistakes in repair[24]), or propose the occurrence of variable translation of the same gene via differences in transfer RNA's or activating enzymes.[25] These hypotheses either encounter difficulties in explaining the presence of multiple amino acid substitutions in well-defined positions or require special mechanisms for control of the synthesis of so many different proteins.

The purpose of the present communication is to suggest an hypothesis to account for the diversity of antibodies and at the same time meet the requirement for strict control and selection (Fig. 2). It is assumed that a number of genes arose in evolution by tandem duplication. The precursor of these genes may have had a length sufficient to specify a region of the same length as the C portion of a light chain.[15, 22] As will be shown below, the number of tandem duplicated genes need not be great; perhaps 50 would suffice. Point mutations are assumed to have accumulated in the duplicated genes so that each differs in at least one codon from the others. It is proposed that these genes recombine at some stage of maturation of the lymphocyte by means of somatic crossing-over. Crossing-over would be favored by the homology of these genes as well as by the fact that they are present in tandem du-

plicated arrays. Such a system might contain instabilities of the sort required to account for the observed amino acid sequences among the immunoglobulins. Because of the close homology of the tandem genes and their similarity or equality in lengths, mispairing between two DNA strands might occur frequently so that nonidentical genes lie side by side. Mispairing could occur either between sister strands or those of homologous chromosomes. Crossing-over at any point along these genes would be expected to generate genes with new sequences of codons. This hypothesis does not require that crossing-over occur within two paired but unlike codons; this type of crossing-over would result only rarely in the appearance of a new point mutation. As shown in Figure 2, each single crossing-over could create

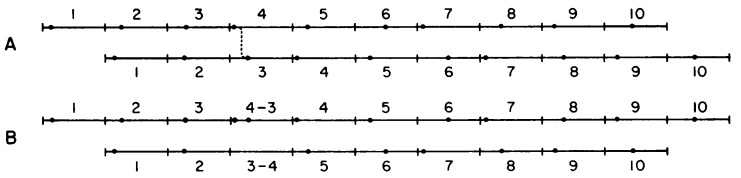

FIG. 2.—Diagram indicating a single crossing-over (4-3) in an hypothetical double array of ten tandem duplicated genes. The dots indicate point mutations at different positions in each gene. Other arrangements are possible. By shifting, for example, gene 6 could cross over with gene 4, etc. (*A*) Before crossing-over; (*B*) after crossing-over between genes 4 and 3.

two new amino acid sequences. Nevertheless, each product would appear to have been generated strictly by point mutation.

This model differs from that of Smithies[23] in that the crossing-over is between tandem duplicated genes and does not invoke inverted duplications. Instead, the multiple amino acid replacements are accounted for on the basis of somatic recombination of a relatively small number of genes that have arisen by mutation and natural selection. This is consistent with the observation that pairs of amino acid interchanges can be accounted for by single base changes in corresponding codons. It is also consistent with the observation that many of the positions within the *V* region appear to be invariable, and accords with the presence of positions in which only a few interchanges of the many possible are observed (e.g., in position 1 of human κ chains only glutamic or aspartic acid has been found).

The present hypothesis does not violate the idea that one gene codes for one distinct polypeptide chain. At the same time, the selective requirements for size and polarity of the variable and constant regions may be achieved via evolutionary pathways. Mutations in cistrons corresponding to the constant regions of antibody polypeptide chains would be subject to stringent selection pressure. The selection would arise because of need to maintain structures capable of interchain bonding and physiologic functions such as complement fixation, opsonization, fixation to cells, etc. In contrast, individual mutations in variable regions are not necessarily selected for by interaction with any particular antigen in the environment. Instead, it is proposed that there is selective advantage for those mutations capable of generating the largest set of possible sequences by somatic crossing-over. This notion is compatible with the suggestion that there is degeneracy in the matching of antigen to antibody.[17] A large variety of sequences of heavy and light chains could

interact to form a degenerate set of combining sites, each of which would combine more or less specifically with a given antigenic determinant. Thus, the heterogeneity of antibodies is considered to be an essential requirement for interaction with a set of antigenic determinants to which the organism may never have been exposed either in evolution or in the lifetime of an individual.[26]

The presence of duplicate genes would increase the probability of random point mutations in this genetic system. As indicated above, those mutations resulting in significant changes in the combining sites of antibodies would be selected for, for they would increase the repertoire of different antibodies the organism can make. Mutations which would not significantly increase this capability would be selected against, because they would decrease the homology between the tandem genes, and lower the probability of crossing-over. The system contains an element of genetic conservatism, which might explain why many positions in the sequence of the variable genes are constant, as well as why different species, such as man and mouse, have similar genes for immunoglobulins.[10, 11, 20, 25]

If a cell contained two genes coding for the amino acid sequence of the light polypeptide chains, and if these genes differed in at least 20 different positions, unlimited somatic crossing-over between these genes could give rise to 2^{20} new genes within the cells of the organism. If a similar source of variability existed for the heavy chains, then somatic crossing-over and subsequent interaction of light and heavy chains[17] could give rise to more than 10^{12} different immunoglobulins from only four original genes, without postulating the occurrence of any somatic point mutations.[27]

Crossing-over among such a small number of genes is a very improbable mechanism for generating the observed variability, inasmuch as genes only very rarely pair with nonidentical partners. Such mispairing is required if the crossing-over is to result in a new nucleotide sequence. In the case of tandem duplication of similar genes, however, the frequency of occurrence of mispairing of nonidentical genes would be expected to be positively correlated with the number of similar genes that are tandemly linked. For this reason, crossing-over between mispaired, nonidentical genes might be *far* more frequent in a system containing a number of tandem duplications than in one without duplicate genes. Indeed, a cumulative effect might be operative, as the occurrence of one nonhomologous crossing-over might make the mispairing of all the genes in the system more probable. Because no genetic system containing more than just a few tandem duplications has been investigated, we do not know the number of duplications that would be required to explain the variability in the amino acid sequences of the immunoglobulins. A system of 20 tandem duplications, each differing from the others by a single nucleotide in different codons, could generate the same number of new genes as two genes differing in 20 codons.

The recombination model encounters several difficulties related to the question of whether one gene codes for all of the C regions and several genes code for V regions. A proposal that the C region is specified by a single gene has been made by Dreyer and Bennett.[20] In this case the low recombination frequency of allotypes related to the C region is no embarrassment. Recombination of V genes in the germ line might have to be restricted by some unknown mechanism in order to prevent instabilities of the type discussed by Thomas.[28] On the other hand, if each of the

tandem duplicated genes specifies both V and C regions, then the C region would be repeated many times and might be expected to show evidence of high recombination frequencies as well as instability. In this case, additional restrictions on recombination in the germ line would have to be invoked.

Although at present there is no direct evidence for somatic recombination of immunoglobulin genes, there have been several observations of crossing-over patterns in chromosomes of lymphocytes in tissue culture.[29, 30] The relationship of these events to immunoglobulin diversity remains to be clarified. The somatic recombination hypothesis[31] leads to several predictions. Bence-Jones proteins should be found to contain combinations of amino acid interchanges that are found singly in other Bence-Jones proteins. The number of sequences examined is still too small to test this prediction. The hypothesis also suggests that there are strict requirements for complementation of genes in the constant regions of both heavy and light chains. Perturbations in the constant regions introduced by chemical modification of amino acid side chains might have dire effects on interchain bonding, physiological activities of antibodies, and the matching of the variable regions of both chains to form a combining site. On the other hand, modifications introduced in many parts of the variable regions should have little or no effect on the over-all structure of the antibody molecule or its physiological function. Finally, according to the somatic recombination hypothesis, a single immunologically competent cell could be at best only pluripotent with respect to antibody production at any one time in its development. It could never be totipotent, i.e., it could produce at most only a small number of different immunoglobulins of a given class. At the same time, however, precursor cells would contain all of the information necessary to generate all antibodies of a given class capable of being made by the organism.

* Supported by NSF grant GB 3920 and USPHS grant AM 04256.

[1] Singer, S. J., and R. F. Doolittle, *Science*, **153**, 13 (1966).

[2] Koshland, M. E., and F. M. Englberger, these PROCEEDINGS, **50**, 61 (1963).

[3] Haber, E., these PROCEEDINGS, **52**, 1099 (1964).

[4] Edelman, G. M., D. E. Olins, J. A. Gally, and N. D. Zinder, these PROCEEDINGS, **50**, 753 (1963).

[5] Edelman, G. M., and J. A. Gally, *J. Exptl. Med.*, **116**, 207 (1962).

[6] Schwartz, J. H., and G. M. Edelman, *J. Exptl. Med.*, **118**, 41 (1963).

[7] Titani, K., E. Whitley, Jr., L. Avogardo, and F. W. Putnam, *Science*, **149**, 1090 (1965).

[8] Hilschmann, N., and L. C. Craig, these PROCEEDINGS, **53**, 1403 (1965).

[9] Milstein, C., *Nature*, **209**, 370 (1966).

[10] Hood, L. E., W. R. Gray, and W. J. Dreyer, these PROCEEDINGS, **55**, 826 (1966).

[11] Perham, R., E. Appella, and M. Potter, *Science*, **154**, 391 (1966).

[12] At present, the data are not sufficiently accurate to justify the conclusion that all Bence-Jones proteins have the same length. The numbering adopted here is based on the findings of Titani *et al.* (ref. 7)

[13] Mårtensson, L., *Vox Sanguinis*, **11**, 521 (1966).

[14] Baglioni, C., L. A. Zonta, D. Cioli, and A. Carbonara, *Science*, **152**, 1517 (1966).

[15] Hill, R. L., R. Delaney, R. E. Fellows, Jr., and H. E. Lebovitz, these PROCEEDINGS, **56**, 1762 (1966).

[16] Press, E. M., P. J. Piggot, and R. R. Porter, *Biochem. J.*, **99**, 356 (1966).

[17] Edelman, G. M., and J. A. Gally, these PROCEEDINGS, **51**, 846 (1964).

[18] Nossal, G. J. V., and O. Mäkelä, *Ann. Rev. Microbiol.*, **16**, 53 (1962).

[19] Attardi, G., M. Cohn, K. Horibata, and E. S. Lennox, *J. Immunol.*, **92**, 335 (1964).

[20] Dreyer, W. J., and J. C. Bennett, these PROCEEDINGS, **54**, 864 (1965).

[21] Titani, K., E. Whitley, Jr., and F. W. Putnam, *Science*, **152**, 1513 (1966).

[22] Edelman, G. M., in *The Neurosciences—A Study Program*, Neurosciences Research Program (The Rockefeller University Press, in press).

[23] Smithies, O., *Nature*, **199**, 1231 (1963).

[24] Brenner, S., and C. Milstein, *Nature*, **211**, 242 (1966).

[25] Potter, M., E. Appella, and S. Geisser, *J. Mol. Biol.*, **14**, 361 (1965).

[26] The selection pressures described here would apply with equal validity to hypotheses which invoke a very large number of genes in the germ line (multiple gene hypotheses, refs. 20–22).

[27] Effective crossing-over is defined as an event which generates a new sequence. The example does not take account of double crossing-over or of looping within the same strand of DNA.

[28] Thomas, C. A., *Progr. Nucleic Acid Res. Mol. Biol.*, **5**, 315 (1966).

[29] German, J., *Science*, **144**, 298 (1964).

[30] Shaw, M. W., and M. M. Cohen, *Genetics*, **51**, 181 (1965).

[31] The hypothesis invoked here to explain antibody variation may be useful in explaining other biological phenomena that require great variability, e.g., proteins involved in specific cell-cell interactions in the nervous system.