Amino Acid Sequence Studies of Immunoglobulins: Implications for the Storage, Processing, and Expression of Genetic Information

Thesis by

Elwyn Yuan Loh

In partial fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1979

(Submitted June 6, 1978)

To my parents

Dr. and Mrs. H. Y. Loh

My thanks go

To my advisor Lee Hood whose enthusiasm, energy, and integrity have made an ideal environment to do science. Whenever an impasse was reached, Lee could provide a new perspective to guide me into a more fruitful line of action or thought,

To all of my friends in the lab Mitch Kronenberg, Jon Fuhrman, Nelson Johnson, Vince Farnsworth, Jeff Hubert, and many others. I could not have made it without their encouragement and their distractions. The hours spent arguing and fantasizing were terrific. I will always be grateful for Mitch's "bad influence," friendship, and help,

To Pehrs Edman, Mike Hunkapiller, and the other sequencing wizards of the world who made sequencing progressively easier from the first set of 160 residues of this thesis to the second set of 1000 to the third set of 2200,

To Bertha Jones for the atmosphere she created and for making the lab function,

To Ray Owen for the model he provided and the immunology he taught,

To those who helped me along the way -- the Alex family, Mr. Ohshima, Paula Samazan, and my family. I hope that I can become more like them,

To this country whose history, freedom, and wealth have created the opportunities which I enjoy,

And thanks to my wife, Louise, with whom I share my life.

Abstract

Antibody molecules form a highly complex set of proteins. A central problem in immunology has been how the information coding for these proteins is stored, processed, and expressed.

The constant region sequences of two rat κ light chain allotypes have been partially sequenced. These allotypes segregate in the Mendelian fashion. In the eighty-one constant region residues compared, 10 amino acid substitutions and one size difference were found. This large number of substitutions raise the possibility that the structural genes for both forms may exist in all rats and the inherited marker is a regulatory gene controlling the expression of one or the other forms.

The diversity of immunoglobulins is reflected in the diversity of myeloma proteins and much of our knowledge of antibodies comes from studies of myeloma proteins. However, the window, created by the myeloma tumors, may be a biased one. By comparing the N-terminal amino acid sequences of myeloma light chains from two inbred strains of mice, BALB/c and NZB, we have found differences which suggest that different populations of lymphocytes are being transformed in the two strains. Thus the true diversity of immunoglobulins may be greater than that seen in myeloma proteins.

By sequencing a set of closely related variable regions, one can ask the question -- what are the protein products coded by a single germ line gene? The nearly complete variable regions of twenty-two κ chains have been sequenced using newly developed automated sequencing technology. This data shows that at least six genes code for this set, assuming that the somatic diversity generating mechanisms cannot produce multiple parallel mutations. Within each subset of these sequences, coded for by at least one gene, additional

variations occur both inside and outside the hypervariable regions, although predominately inside. In addition, the sequence of the approximately twelve residues preceeding the constant region do not correlate with the rest of the variable region. We have termed this region the S or switch region and suggest that it is coded for by a separate segment of DNA that is reorganized during differentiation much in the same way as V and C regions are rearranged.

Table of Contents

Introduction

The immunoglobulin molecule is many things to many people. To the physician, it is the most potent curative agent directed against bacterial and viral invasion. To the immunologist, it is the antigen-binding molecule of the B (and possibly T) lymphocyte. To a chemist, its myriad binding sites offer unlimited possibilities in molecular recognition. To biologists of all fields, it is an experimental reagent of exquisite specificity and flexibility. For the geneticist, amino acid sequence variability of antibodies has raised questions of information storage and evolution.

Originally and ultimately the efforts to understand antibodies come from their importance in maintaining human homeostasis. Within that context I daresay that most human disease lack a molecular understanding, and as a result treatment of these diseases cannot be called strictly rational therapy. A recent catch-all category of so called idiopathic diseases has been "autoimmune disease." This embodies the belief that disorders in the regulation of the immune system may be involved in the pathology of these diseases. A list of autoimmune disease would be several pages long.[1] Self antigens have been determined in some autoimmune diseases -- primary biliary cirrhosis (liver mitochondria), chronic acitve hepatitis (smooth muscle), myasthemia gravis (acetylcholine recpetor), Grave's disease (TSH receptor), bullous pemphgoid (skin basement membrane), and many more. All autoimmune and immunodeficiency diseases will have a genetic component. Some individuals because of their inherited constitution will be more susceptible to the environmental insult that lead to pathology. In most cases which gene products are at fault is unknown. One notable exception is the histocompatibility complex, termed the HL-A complex in man, and H-2 in mouse, to which association of over forty human diseases have been found.[2] Another beginning can be made from the genetics of the

antibody molecule itself, the effector molecule of the immune system. Hopefully, knowledge of antibody genetics and gene regulation can lead to an understanding of diseases of the immune system.

Immunoglobulins form a family of molecules of impressive diversity and complexity. (The first appendix contains a review of the field.) Estimates of the number of different molecules an individual can make, the antibody repetoire problem, range upwards from 10^6 though the difficulties of such guesses are numerous.[3] For the past fifteen years, it has been recognized that the organism does not need 10^6 or more genes to code for this family of molecules. One way to generate diversity is to use two polypeptide chains to construct the binding site. Recent X-ray reconstruction of an antibody has proven this to be so.[4,5] Thus given p light chains and q heavy chains one can theoretically achieve p x q binding specificities. In Chapter VI we introduce data to support the idea that the same approach has been used intramolecularly in that the binding site contribution from each molecule may be constructed from two separate genes so called V for variable region and S for switch region. Thus given P_{VL}, P_{SL}, Q_{VH} and Q_{SH} of each of the variable and switch regions of the light and heavy chain, we can have $P_{VL} \times P_{SL} \times Q_{VH} \times Q_{SH}$ specificities. If each number is 50, then we get over 6×10^6 ultimate different binding specificities. Mechanistically, the intramolecular gene rearrangement has precedent in the antibody molecule in the joining of V and C regions.[6] The V region is responsible for antigen binding and is the N-terminal 110 residues while the C region is responsible for a variety of effector functions; the C region is the molecular link between binding to antigen and the other molecules or cells that are needed to give biological meaning to antigen-antibody binding. For each family of variable regions, of which mammals

have three in all, two for light chains (κ and λ) and one for heavy chains (H), there are from one to ten constant regions which can be eventually joined on a protein level to any one of its family of variable regions. This hypothesis, originally made by Dreyer and Bennett,[7] has now been given direct verification by Tonegawa.[8]

The significance of such a mechanism goes beyond being a clever way to conserve DNA. First, as a developmental mechanism, it may serve in selectively turning on the unique product of a highly differentiated cell and in committing a cell to a developmental pathway. Second, as a well controlled mechanism of mutating DNA the actual molecular events taking place must be fascinating. It must have a way to selectively bring together pieces of DNA which may be millions of base pairs apart (genetically up to 1.5 centimorgans apart) and nicking and splicing them together. One recent experiment suggests that the DNA between the V and C region is deleted.[10] Third, the arrangement and rearrangement of all of these segments of DNA -- the V genes, the S genes, the C genes and all of the intervening DNA -- will undoubtedly raise as many new questions as it resolves. Is the process related to the so called "illegatimate recombination" found in procaryotes?[11] Will insertion sequences[12,13] be necessary to move the DNA? Will the joining process be an orderly one,[3,14] moving sequentially down the chromosome linking each V with an S, then to a C or is it a random process? Will closely related V genes as those described in Chapter VI be scattered or closely linked? Will many silent constant or variable genes exist as suggested in Chapter I and II which form a pool of genetic diversity upon which a species but not an individual can readily draw? The use of recombinant DNA technology

to isolate the entire immunoglobulin coding region is an exciting prospect now underway.

To return to the problem of how antibody diversity is generated, there is yet another way in which the functional diversity is greater than the genetic diversity. The variable region genes may mutate during differentiation to create new sequences not found in the germline. The best evidence for this comes from the mouse λ chains where there appear to be more λ chains sequenced than are counted by hybridization kinetics.[8,15] This is discussed in Chapter VI. The data of Chapter VI addresses the problem of asking what are the somatic protein products of a single gene. The approach taken is to sequence a whole set of molecules which are closely related. This close homology implies that these molecules should be recent derivatives of a common ancestor. Deciding whether the derivation is evolutionary or developmental is of course the difficult central problem. Eventually we will need to compare the genes with the gene products and that comparison should be informative as to what mechanisms, either enzymatic or cellular, we need to search for.

As mentioned above, the diversity of immunoglobulins is not the same when viewed on the germ line or on the somatic levels. One can view antibody sequence information through a variety of windows, starting with the germ line. Even then one must be careful in that structural polymorphisms are frequent and of unknown extent.[9,16] A second level would be to examine so called virgin B cells, those lymphocytes which have not been exposed to antigenic stimulation. If diversity generation depends on antigen presence, then the diversity there would be limited.[17] A third level would be to examine stimulated B-cells, both primary and secondary. The genetic diversity at each level could be examined

at each level by purifying the cells and using isolating the relevant portions of the genome.[18,19] On the third level two additional approaches exist. One is to use myeloma proteins as has been done extensively.[20,21] Chapters II and III discuss the myeloma window and both what one sees and what one does not see. The myeloma transformation process and the preceeding levels of regulation may have limited the view. A statistical method of comparing large sets of sequence data is included as the second appendix. One can also make "hybridomas" which are hybrid cells combining a secreting normal lymphocyte with a myeloma cell.[22] Thus one is able to make immortal cell lines of any antibody producing cell for which one can select.

Finally, it is important to point out that as one progresses along the various levels of antibody diversity, one is not only increasing the amount of diversity by somatic diversification, but one may also be decreasing the diversity. There most likely are genes which cannot be seen on any of the final functional levels of expression because the structural genes are defective, because of regulatory events or because of required somatic changes. An example of this may exist in the complex allotypes discussed in Chapters I and II. Another example may be the λ embryonic sequence which has not been found in myeloma sequences.[18] One would expect that many such variable regions and constant region sequences exist in the germ line since selection against them would be difficult. Mutational events would accumulate in them and either they would be deleted eventually or they would be expressed by changes in regulatory genes, and their fate determined by their selective advantage or disadvantage.

References

[1]Rose, N. R. and Friedman, H., eds. Manual of Clinical Immunology. American Society for Microbiology, Washington, DC (1976).

[2]Saszuki, T., Grumet, F. C., McDevitt, H. O., Ann. Rev. Med. 28:425 (1977).

[3]Klinman, N. R., Sigal, N. H., Metcalf, E. S., Pierce, S. K., and Gearhart, P. J., Cold Spring Harbor Symp. Quant. Biol. 61:165 (1977).

[4]Padlan, E. A., Davies, D. R., Pecht, I., Givol, D., and Wright, C., Cold Spring Harbor Symp. Quant. Biol. 61:627 (1977).

[5]Poljak, R. J., Amzel, L. M., Chen, B. L., Chin, Y. Y., Phizackerley, R. P., Saul, F., and Ysern, X., Cold Spring Harbor Symp. Quant. Biol. 61:639 (1977).

[6]Hood, L., Fed. Proc. 31:177 (1972).

[7]Dreyer, W. J., and Bennett, J. C., Proc. Natl. Acad. Sci. USA 54:864 (1965).

[8]Tonegawa, S., Hozumi, N., Matthyssens, G., and Schuller, R., Cold Spring Harbor Symp. Quant. Biol. 61:877 (1977).

[9]Weigert, M., and Riblet, R., Cold Spring Harbor Symp. Quant. Biol. 61:837 (1977).

[10]Honjo, T., and Kataoka, T., Proc. Natl. Acad. Sci. USA, in press.

[11]Weisberg, R. A., and Adhya, S., Ann. Rev. Genetics 11:451 (1977).

[12]Bukhari, A., Shapiro, J., Adhya, S., eds. DNA Insertions, Plasmids and Episomes. New York: Cold Spring Harbor Lab (1977).

[13]Hu, S., Ohtsubo, E., Davidson, N., Saedler, H., J. Bacteriol. 122:764 (1975).

[14]Huang, H., thesis, California Institute of Technology (1978).

[15]Leder, P. , Honjo, T., Seidman, J., and Swan, D., Cold Spring Harbor Symp. Quant. Biol. 61:855 (1977).

[16]Hood, L., Campbell, J., and Elgin, S., Ann. Rev. Gen. 9:305 (1975).

[17]Cohn, M., Blomberg, B., Geckeler, W., Raschker, W., Riblet, R., Weigert, M., in The Immune System: Genes, Receptors, Signals (E. E. Sercarz, ed.). Academic Press: New York (1974) p. 89.

[18]Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O., and Gilbert, W., Proc. Natl. Acad. Sci. USA 75:1485 (1978).

[19]Benton, W. D., and Davis, R. W. Science 196:180 (1977).

[20]Potter, M., Adv. in Immun. 25:141 (1978).

[21]Kabat, E. A., Wu, T. T., and Bilofsky, H., Variable Regions of Immunoglobulin Chains, Tabulations and Analyses of Amino Acid Sequences, Bolt, Beranek and Newman, Inc., Cambridge, MA (1976).

[22]Kohler, G., and Milstein, C., Eur. J. Immunol. 6:511 (1976).

Chapter 1

This paper was published in Proc. Natl. Acad. Sci. USA.

Reprinted from
Proc. Nat. Acad. Sci. USA
Vol. 72, No. 12, pp. 5046–5050, December 1975
Genetics

10

Structure and regulation of immunoglobulins: Kappa allotypes in the rat have multiple amino-acid differences in the constant region

(complex allotypes/amino-acid sequences/multiple amino-acid substitutions/control gene model)

GEORGE A. GUTMAN*‡, ELWYN LOH†, AND LEROY HOOD†

* Department of Pathology, Stanford University, Stanford, California 94305; and † Division of Biology, California Institute of Technology, Pasadena, Calif. 91125

Communicated by E. B. Lewis, September 15, 1975

ABSTRACT **Immunoglobulin kappa chains from various inbred strains of rats have two serologically detectable forms that segregate in a Mendelian fashion (allotypes *a* and *b* of the *RI-1* locus). Partial amino-acid sequences from the constant regions of these two forms have been compared. Of the 81 residues of the constant region studied, 10 amino-acid substitutions as well as one size difference (sequence gap) were found. This large number of sequence differences among alternative forms of the κ allotype raises provocative questions as to the genetic and evolutionary implications of these light chain allotypes. We designate allotypes whose alternative forms differ at multiple residue positions as complex allotypes. There are basically two genetic models that might explain complex allotypes. First, these allotypes are alleles of a single structural gene with an unusual evolutionary history. Second, all rats have genes that code for each of the light chain allotypes and a control mechanism that permits them to be expressed so that they mimic a Mendelian pattern of segregation. We discuss evidence from other immunoglobulin systems that is compatible with this second model.**

The immune system is one of the most complex physiological systems that has been studied at the molecular, genetic, and cellular levels. The general chemical structure of the immunoglobulin molecule is well understood (1, 2). This molecule is composed of two different polypeptides, light and heavy chains. There are two types of light chains, lambda (λ) and kappa (κ). Each immunoglobulin polypeptide is composed of an NH_2-terminal variable (V) and a COOH-terminal constant (C) region. Some general features are known about the organization of antibody genes (1, 2). Three families (clusters) of genes, unlinked in the mammalian genome, code for the λ, κ, and heavy chain polypeptides. It is generally accepted that the variable and constant regions in these families are coded by separate but closely linked genes. However, very little is known about the genes that regulate the immune response. The immune response genes, linked to the major transplantation locus of the mouse, appear to constitute a family of control genes, unlinked to the structural genes for antibodies, that in some unknown manner regulate the ability of an animal to respond to a variety of different antigens (3). This paper suggests that a new system of control genes may regulate the expression of certain immunoglobulin allotypes.

Inbred rats have a genetic locus, *RI-1*, that controls the expression of two serologically detectable forms of κ light chains, a and b, which segregate as Mendelian codominants (4–6). We report here on the amino-acid sequences of kappa constant (C_κ) regions of the a and b allotypes. The a and b allotypes have multiple amino-acid substitutions, as previously suggested on the basis of differences in peptide maps (7).

Abbreviations: C and V, constant and variable region, respectively, of immunoglobulin polypeptides.

‡ Present address: Walter & Eliza Hall Institute of Medical Research, P.O. Royal Melbourne Hospital, Victoria 3050, Australia.

MATERIALS AND METHODS

Details of the techniques used will be published elsewhere. Briefly, pooled light chains [which are 95% kappa type (8)] were prepared from two rat strains differing at the *RI-1* locus, DA(RI-1^a) and LEW(RI-1^b) (6). Peptide maps of trypsin digests were prepared and the peptides eluted and studied for amino-acid composition and for sequence by manual and automated Edman degradation.

RESULTS AND DISCUSSION

Rat κ Allotypes Exhibit Multiple Amino-Acid Differences. The amino-acid sequences of C_κ regions of normal serum light chains from two inbred strains of rat, DA and LEW, are compared in Fig. 1 with the previously published C_κ sequences of myeloma (Bence-Jones) κ chains from the LOU strain of rat (S211) (9) and the BALB/c strain of mouse (M321) (10). The LEW and LOU strains are of the b serotype, whereas DA belongs to the a serotype. The LEW and DA C_κ regions differ by 10 residues plus one sequence gap, whereas the LEW and S211 C_κ regions differ by only two residues. This is a minimum estimate of the total number of differences for several reasons: (*i*) only 81 of the 108 residues of the C region were compared; (*ii*) the acid and amide forms of aspartic and glutamic acid were not distinguished; and (*iii*) the V regions were not examined.

These allotype-associated differences are distributed in a nonrandom manner. Ten of the 11 differences occur at positions where the LEW sequence differs from that of the mouse, indicating that certain positions are more likely to accumulate changes than others. The fact that the DA sequence is identical to that of the mouse at five of these positions is a puzzling point which may indicate a more rapid accumulation of changes in the LEW gene. Further, the distribution of the substitutions in the tertiary structure of the light chain is not random. Most of the substitutions lie on the external portion of the polypeptide chain (only position 136 is internal) (ref. 11; R. Poljak, personal communication). Two clusters of differences (one including 153 and 155, the other 184, 185, and 188) are external, and both lie very close to one another in a region already known to encompass the serological markers Oz and Inv, and the sequence marker Kern. Since it is known the RI-1 determinants lie exclusively in the C-region (12), it seems likely that one or more of these external substitutions will determine the a and b serological specificities.

The LOU Strain of Rat Appears to Have Two C_κ Genes. Since the LOU and LEW strains of rat are identical at the *RI-1* locus by serological analysis (13), the two sequence differences found between the C_κ regions of the pooled LEW and S211 light chains were surprising. However, the S211 protein seems to be a relatively unusual variant among LOU

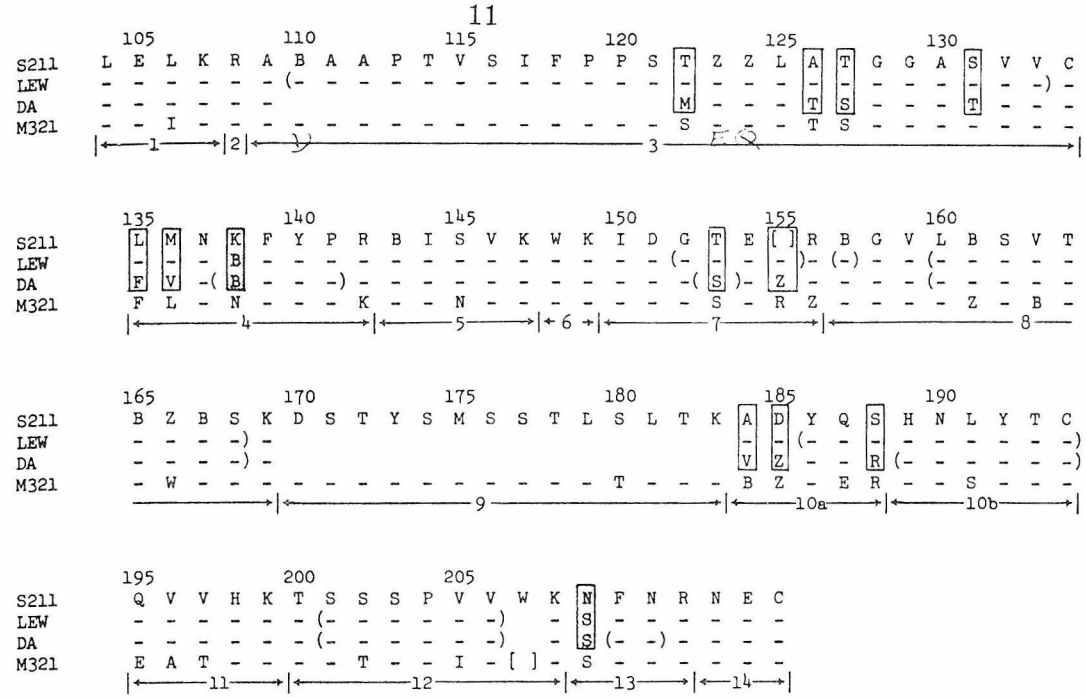


FIG. 1. Amino-acid sequences of rodent κ chain constant regions. S211 is a myeloma κ chain from the LOU rat (11), LEW and DA are sequences from pooled light chains, and M321 is a mouse myeloma κ chain (17). The numbering of peptides is given below the sequences. Differences between DA and LEW are boxed. The two differences between pooled LEW and S211 C_κ regions are boxed and shaded. Dashes indicate an amino acid identical to the S211 sequence. Parentheses indicate that the sequence of the corresponding residues has not been determined. The one letter code of Dayhoff (44) is used for the amino acids.

κ chains, as three other κ myelomas show sequences that are identical at both these positions to the pooled LEW C_κ sequence (P. Querinjean, personal communication). Accordingly, at least two C_κ genes appear to be present in the germ line of the LOU strain of rat. Presumably the same is true of the LEW strain.

Immunoglobulin Allotypes Fall into One of Two Categories, Simple and Complex. It is useful to define two categories of immunoglobulin allotypes, each with very different genetic and evolutionary implications. Alternative forms of *simple allotypes* segregate in a Mendelian fashion in mating studies and differ by one or a few amino-acid substitutions. The InV marker of the human κ chain (14) and the Gm(3) and Gm(17) markers of the human γ1 chain (see ref. 15) are examples of simple allotypes. Simple allotypes are probably coded by alternative alleles at a single structural locus. In contrast, alternative forms of complex allotypes differ by multiple amino-acid residues and *generally* segregate in a Mendelian fashion. The group a allotypes (a1, a2, a3) of the rabbit are V_H markers that differ by multiple amino-acid residues (16, 17). Likewise, the group b allotypes (b4, b5, b6, b9) of the rabbit are C_κ markers that also differ by multiple amino-acid residues (18, 19). Multiple serological specificities have been defined in alternative forms of human γ_3 chains (Gm markers) as well as in certain mouse γ chains. These are designated serologically complex allotypes. If these serological specificities correlate with multiple amino-acid differences, certain human and mouse γ allotypes may represent additional examples of complex allotypes (see ref. 15). Finally, the C_κ regions of the inbred rats described in this paper differ by multiple amino-acid residues. The importance of making a distinction between complex and simple allotypes lies in the very different types of evolutionary or genetic mechanisms they imply. Complex allotypes may be coded by alternative alleles at a single structural locus with an unusual evolutionary history (Fig. 2a and b) or they may result from duplicated genes and the operation of an unusual control mechanism (Fig. 2c). Similar proposals have been made by others (see ref. 15). These three models for complex allotypes will be discussed in subsequent sections using the rat C_κ allotypes as an example.

Complex C_κ Allotypes in the Rat May Evolve by the Divergence of Two Alleles at a Single Genetic Locus. This is the simplest model accounting for the genetics of the RI-1 specificities and it assumes that the two forms have diverged from one another by 10 substitutions and one sequence gap (Fig. 2a). The S211 C_κ gene may represent a very recent duplication in the b strains (LOU and LEW) that may be expressed at low levels in the serum. Two objections can be raised against this model. First, there are a large number of amino-acid differences between the alternative forms. This model assumes that a variant C_κ gene arises and is fixed in the population by natural selection. This new C_κ gene must then incur a second mutation that is once again fixed, and the entire process must be repeated 11 times. Indeed, each new variant gene must be improved in function over its predecessor in order for natural selection to fix (or partially fix) it in the rat population. The question arises as to whether rats as a species have had sufficient evolutionary time for alleles of structural genes to evolve to be so different.

Generally alleles at a single genetic locus are assumed to differ by only one or two residues. Indeed, more than 200 human hemoglobin variants (alleles) have been examined (20). Most differ by one residue, a few differ by two residues, and only one differs by as many as three residues. The allelic forms of the somewhat larger bovine carboxypeptidase A differ by three out of 307 residues (21). Likewise, alternative forms of human κ chains (14), human haptoglobins (22), and a variety of other serum proteins show one or a few amino-acid differences. On the other hand, the "allelic" A

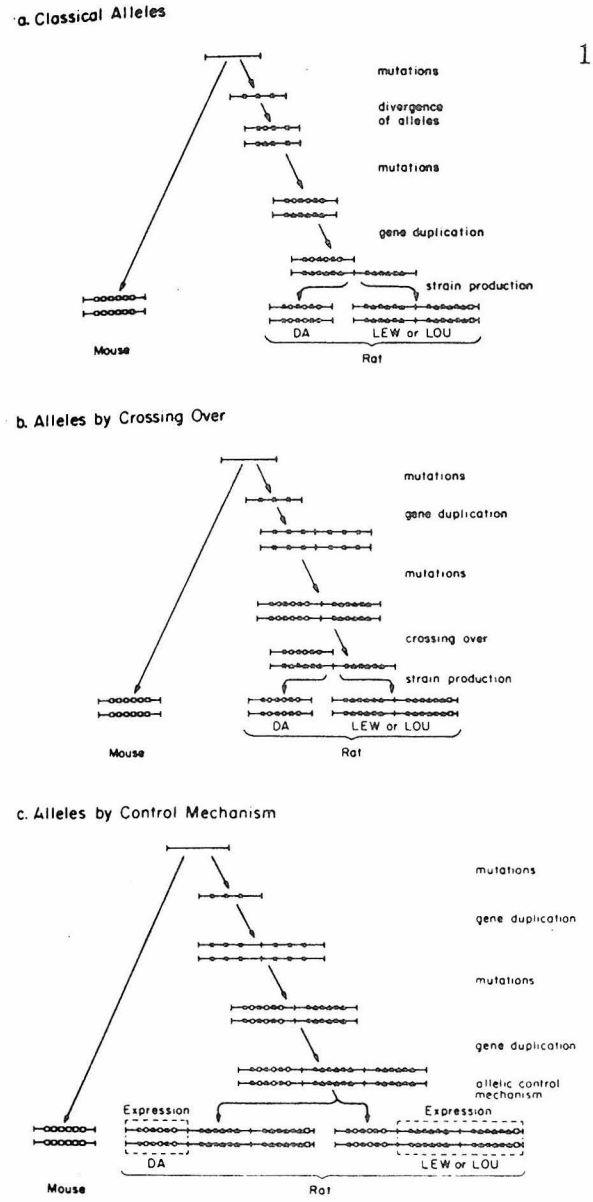


FIG. 2. Genetic models for the expression and evolution of the rat C_κ allotypes. (a) Classical alleles. (b) Alleles by crossing-over. (c) Alleles by gene duplication and a control mechanism. (See *text*.)

and B forms of sheep hemoglobins differ by seven residues out of 145 (23). If the A and B forms are true alleles, which as subsequent discussion will show is often difficult to determine, then alleles with multiple substitutions are possible.

A second objection to this model is the apparent absence (serologically) of any intermediate forms of the rat C_κ region. For example, the white tailed deer has at least seven allelic forms of the β hemoglobin chain that differ by as much as 10% of their amino-acid sequence in the portion of these molecules examined (24). These allelic forms generally express one of two alternative residues at positions that differ. These intermediate forms could arise as (*i*) true intermediates that are maintained in the population between the most extreme alleles and (*ii*) the products of intragenic crossing-over which could scramble the multiply substituted forms. The lack of intermediate forms in κ allotypes of the rat could be explained by hypothesizing that the a and b C_κ gene products have a selective advantage over any of their intermediates.

Complex C_κ Allotypes in the Rat May Evolve by Gene Duplication and Gene Loss Through Crossing-Over. This model suggests that C_κ gene in rat underwent an early gene duplication and that many differences were fixed in these duplicated genes (Fig. 2b). Later in the evolution of the rat line, two unequal but homologous crossing-over events occurred to generate two populations of rats—one (e.g., DA) with a chromosome coding for the *a* form of the C_κ gene and a second (e.g., LEW) with a chromosome coding for the *b* form of the C_κ gene. Once again, the two C_κ forms would be coded by alleles at a single structural locus, but with an evolutionary history that avoids, in part, the need for the highly stringent selective pressures described in the preceding model. For example, one of these C_κ genes may be freed to accept many substitutions while the second is temporarily the functional C_κ gene. Again, the S211 C_κ gene is a recent gene duplication. There are at least two precedents for the evolution via a crossing-over event of alleles that differ by multiple substitutions [β hemoglobin alleles of inbred mice (25) and of Barbary sheep (26)]. This model appears to be a reasonable mechanism for evolving alleles that differ extensively without the necessity of selective pressures to eliminate many intermediate forms. However, intragenic crossing-over might still be expected to scramble these alleles and generate intermediate forms.

Complex C_κ Allotypes in the Rat May Evolve by Gene Duplication and Be Differentially Expressed Via a Control Mechanism. A third model postulates that all rats have genes that code for each of the C_κ allotypes and a control mechanism that permits them to be expressed so that they mimic a Mendelian pattern of genetic segregation (Fig. 2c). Under this model, C_κ gene duplication may have occurred even prior to speciation. Accordingly, complex allotypes could evolve significant differences in their alternative forms. As noted in the previous model, gene duplication can free one gene to accept many substitutions. An important implication of the control gene model is that the serologically detected allotypes follow the inheritance of a control gene(s) that may or may not be closely linked to the corresponding structural (C_κ) genes.

The hypothetical control gene(s) could be operating in a qualitative or quantitative manner. In the latter case small amounts of the "wrong" allotype may be synthesized in homozygous rats. There are precedents for both types of expression in closely linked structural genes. The β-like hemoglobin genes of man (β, δ, γ, and probably ϵ) are closely linked and expressed in a qualitative fashion at different times of development. In the early embryo, the ϵ gene alone is expressed, while later in embryonic life only the γ gene is expressed (27). In the normal adult only the β and δ genes are expressed, at a ratio of about 50 to 1. In addition, an unusual α hemoglobin gene, termed Wazoo, probably present in many primates (e.g., chimpanzee, gorilla), is expressed only in certain individuals of these species (28). Thus, closely linked genes may be expressed in a qualitative manner that varies during development (ϵ, γ), or among individuals (Wazoo), or in a quantitative fashion (adult β and γ). The corresponding control genes may be inherited as Mendelian alleles, as suggested by the inheritance patterns of the ability to express differing ratios of two nonallelic α chains of the stump-tailed macaque (29).

The major objection to the control gene model is its relative complexity. It must mimic with an unknown control

mechanism precisely the same form of expression as the simpler model of allelic structural genes. This is difficult to justify evolutionarily, unless a regulatory mechanism with "allelic" expression has some innate biological advantage. Perhaps it might reflect a strategy to maintain a level of population diversity that may be selectively advantageous in certain environments (30). Two duplicated genes with no control mechanism would result in identical individuals, whereas a control mechanism of the type described in Fig. 2c yields individuals of three kinds (two "homozygotes" and one "heterozygote").

The complex allotypes in rabbits (groups a and b) appear to be coded for by duplicated genes rather than alleles of a single structural locus. Two lines of evidence support this supposition. First, under certain conditions an individual rabbit may express three group a or three group b allotypes (31). If the group a (or group b) allotypes were true alleles, an animal should express at most two alleles. Second, certain rabbits may express low levels of group a allotypes that they should not have according to the genotypes of their parents (32). Two additional examples of serologically complex allotypes may be coded by duplicated genes. A particular congenic strain of mouse, ICR CB-17, has been developed which carries a heavy chain locus homozygous for the C57 Bl/Ka allotype superimposed on a BALB/c background. This strain can express the supposedly absent BALB/c heavy chain allotype under certain conditions (33). This observation implies that the C57 *Bl/Ka* heavy chain locus includes a gene coding for the BALB/c C_H allotype that is ordinarily not expressed. Finally, certain human Gm allotypes not present in donor serum can be found in the supernatant of mixed leukocyte cultures of donor and foreign lymphocytes. Once again this observation implies that humans have C_H genes that are not ordinarily expressed (34). The experiments described above, however, were all carried out using serological methods and will need to be confirmed by structural studies of the "wrong" allotype gene products.

Alternative forms of the complex allotypes of the rabbit, and possibly those of man and mouse, appear to be coded by multiple germ line genes. Complex allotypes may appear on V genes (group *a*, rabbit) or C genes (group *b*, rabbit). In these cases their expression appears to be regulated by a control mechanism that generally causes them to mimic a Mendelian pattern of segregation (Fig. 2c). The obvious parallel between the complex allotypes of the rabbit and the allotypes described in this paper renders the control gene model attractive for the complex κ allotype of rats.

Allelic and Control Gene Models May Be Distinguished by Demonstrating in a Particular Strain "Wrong" Genes or Gene Products. Direct support of the control gene model could be adduced by finding the "wrong" allotype being produced by an inbred rat. A search could be made for the production of a "wrong" allotype in immunologically manipulated situations, as has already been described in rabbits (31), mice (33), and humans (34). Since rat light chains are now fairly well characterized chemically, it will be an easy matter to support serological data with structural studies, even on very small amounts of material. Ultimately, DNA·RNA or DNA·DNA hybridizations of C_κ messenger RNA to somatic or germ cell DNA under very stringent conditions may allow one to determine directly whether the DA and LEW C_κ genes are present in the DNA of all rats.

Rodent C_κ Regions Appear to Be Evolving Rapidly. Mouse and human C_κ regions, which diverged about 75 million years ago, differ by 40 amino-acid residues (35). The rat S211 κ chain differs from its mouse counterpart by 26 substitutions. If mutations were fixed in the rat, human, and mouse evolutionary lines at similar and constant rates, these differences would suggest that the rat and mouse lines diverged more than 40 million years ago [using the naive calculation $(75 \times 26)/40$] rather than 10 million years ago, as is suggested by paleontologic evidence (36). Furthermore, the LEW and DA allotypes are separated by 11 substitutions, suggesting that the divergence of these genes occurred more than 15 million years ago, prior even to the presumed speciation of rat and mouse. There are two possible explanations for these paradoxes. First, the rodent C_κ genes must be diverging considerably more rapidly than their primate counterparts. Recent DNA reassociation studies on primate and rodent single-copy DNA suggest that evolutionary divergence is related to generation time rather than absolute time and, accordingly, rodent genes would be expected to diverge more rapidly than primate genes (37). Second, there is a controversy as to the divergence times of many mammals (38). For example, the mouse and rat evolutionary lines may have diverged from one another much earlier than 10 million years ago. If so, perhaps there is adequate time for the evolution of complex allotypes.

Two types of studies would be valuable in discerning the evolution history of the rat C_κ genes. As examination of the light chain allotypes of wild populations of *Rattus norvegicus* would yield information about the number and range of variation among variants for C_κ chain allotypes. Large sample numbers will be necessary to obtain useful results, however, since it is known that human populations regularly contain rare alleles (with frequencies less than 1%) for many loci (39). The only rat population study reported to date has not demonstrated any new alleles (40). Likewise, the analysis of C_κ regions from other rodents, particularly of the family Cricetidae (lemmings and voles) which is closely related to the Muridae (mice and rats), would answer more general questions relating to the evolution of these light chain genes.

Complex Allotypes of Other Complex Eukaryotic Systems May Also Be Encoded by Duplicated Genes and Expressed by an Unusual Control Mechanism. Serologically complex allotypes have been described in a wide variety of complex eukaryotic systems—the T locus of the mouse (41), the major transplantation locus of mammals (3), the antigens of *Paramecium* (42), the sterility alleles of certain plants (43), etc. Other complex eukaryotic systems may use strategies for the organization, expression, and evolution of genetic information similar to those seen in the vertebrate immune system (1). If so, the presence of complex allotypes may serve as a clue to the presence of multigenic systems with complex regulatory mechanisms.

In summary, alternative forms of complex allotypes may be coded by classical alleles (Fig. 2a), by alleles via gene duplication and crossing-over (Fig. 2b), or by duplicated genes with an unusual regulatory mechanism (Fig. 2c). The latter model can be distinguished from the former two by the presence of the "wrong" genes or gene products in appropriate strains of animals. Two complex allotype systems in the rabbit appear to use the control mechanism model. Perhaps other complex eukaryotic systems will use similar mechanisms for the expression of their information. In any case, the phenotypic expression of complex allotypes raises the possibility that the corresponding genetic system is coded by multiple genes with an unusual control mechanism.

Note Added in Proof. It has come to our attention that W. F. Bod-

mer discussed the concept of complex allotypes in 1973 [W. F. Bodmer (1973) *Transplant. Proc.* V, 1471–1475].

The S211 sequence was kindly provided to us before its publication by Dr. Pierre Querinjean, to whom we are grateful. This work was supported by NSF Grant BMS71-0070 and USPHS Grants AI-10781-04 and AI-09072-06. L.H. has a Research Career Development Award from NIH. This work was carried out while G.A.G. was an NIH Postdoctoral Fellow.

1. Hood, L., Campbell, J. H. & Elgin, S. C. R. (1975) *Annu. Rev. Genet.*, in press.
2. Gally, J. A. & Edelman, G. M. (1972) *Annu. Rev. Genet.* **6**, 1–46.
3. Benacerraf, B. & McDevitt, H. O. (1972) *Science* **175**, 273–279.
4. Rokhlin, O. V., Vengerova, T. I. & Nezlin, R. S. (1971) *Immunochemistry* **8**, 525–538.
5. Armerding, D. (1971) *Eur. J. Immunol.* **1**, 9–45.
6. Gutman, G. A. & Weissman, I. L. (1971) *J. Immunol.* **107**, 1390–1393.
7. Vengerova, T. I., Rokhlin, O. V. & Nezlin, R. S. (1972) *Immunochemistry* **9**, 1239–1245.
8. Hood, L., Gray, W. R., Sanders, B. G. & Dreyer, W. J. (1967) *Cold Spring Harbor Symp. Quant. Biol.* **32**, 133–146.
9. Starace, V. & Querinjean, P. (1975) *J. Immunol.* **115**, 59–62.
10. McKean, D., Potter, M. & Hood, L. (1973) *Biochemistry* **12**, 749–759.
11. Poljak, R., Amzel, L., Avey, H., Chen, B., Phizackerley, R. & Saul, F. (1973) *Proc. Nat. Acad. Sci. USA* **70**, 3305–3310.
12. Nezlin, R. S. Vengerova, T. I., Rokhlin, O. V. & Machulla, H. K. G. (1974) *Immunochemistry* **11**, 517–518.
13. Rohklin, O. V. & Nezlin, R. S. (1974) *Scand. J. Immunol.* **3**, 209–214.
14. Terry, W. D., Hood, L. & Steinberg, A. G. (1969) *Proc. Nat. Acad. Sci. USA* **63**, 71–77.
15. Mage, R., Lieberman, R., Potter, M. & Terry, W. D. (1973) in *The Antigens*, ed. Sela, M. (Academic Press, New York and London), Vol. I., pp. 300–376.
16. Wilkinson, J. M. (1969) *Biochem. J.* **112**, 173–185.
17. Mole, L. E., Jackson, S. A., Porter, R. R. & Wilkinson, J. M. (1971) *Biochem. J.* **124**, 301–318.
18. Appella, E., Rejnek, J. & Reisfeld, R. A. (1969) *J. Mol. Biol.* **41**, 473–477.
19. Goodfleisch, R. (1975) *J. Immunol.* **114**, 910–912.
20. Hunt, L. T., Sochard, M. R. & Dayhoff, M. O. (1972) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, Silver Spring, Md.), pp. 67–88.
21. Petra, P. H., Bradshaw, R. A., Walsh, K. A. & Neurath, H. (1969) *Biochemistry* **8**, 2762–2768.
22. Black, J. & Dixon, G. H. (1968) *Nature* **218**, 736–741.
23. Boyer, S. H., Hathaway, P., Pascasio, F., Bordley, J. & Orton, C. (1967) *J. Biol. Chem.* **242**, 2211–2232.
24. Taylor, W. J. & Easley, C. W. (1975) *Ann. N.Y. Acad. Sci.* **241**, 594–604.
25. Gilman, J. G. (1972) *Science* **178**, 873–874.
26. Huisman, T. H. J. (1975) *Ann. N.Y. Acad. Sci.* **241**, 549–556.
27. Marks, P. A. & Rifkind, R. A. (1972) *Science* **175**, 955–961.
28. Boyer, S. H., Noyes, A. N., Boyer, M. L. & Marr, K. (1973) *J. Biol. Chem.* **248**, 992–1003.
29. Kitchen, H. (1975) *Ann. N.Y. Acad. Sci.* **241**, 12–24.
30. Johnson, G. B. (1973) *Ann. Rev. Ecol. Syst.* **4**, 93–116.
31. Strosberg, A. D., Hamers-Casterman, C., Van der Loo, W. & Hamers, R. (1974) *J. Immunol.* **113**, 1313–1318.
32. Mudgett, M., Fraser, B. A. & Kindt, T. J. (1975) *J. Exp. Med.* **141**, 1448–1452.
33. Bosma, M. J. & Bosma, G. (1974) *J. Exp. Med.* **139**, 512–527.
34. Rivat, L., Gilbert, D. & Ropartz, C. (1973) *Immunology* **24**, 1041–1049.
35. Barker, W. C., McLaughlin, P. J. & Dayhoff, M. O. (1972) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, Silver Spring, Md.), pp. 31–40.
36. Simpson, G. G. (1959) *Cold Spring Harbor Symp. Quant. Biol.* **24**, 255–271.
37. Kohne, D. E. (1970) *Qt. Rev. Biophys.* **3**, 327–375.
38. Sarich, V. M. & Wilson, A. C. (1973) *Science* **179**, 1144–1147.
39. Harris, H., Hopkinson, D. A. & Robson, E. B. (1974) *Ann. Hum. Genet.* **37**, 237–253.
40. Rokhlin, O. V. & Nezlin, R. S. (1974) *Scand. J. Immunol.* **3**, 209–214.
41. Gluecksohn-Waelsch, S. & Erickson, R. P. (1970) *Curr. Top. Dev. Biol.* **5**, 281–316.
42. Beale, G. H. (1954) *The Genetics of Paramecium aurelia* (Cambridge University Press, London and New York).
43. Burnet, F. M. (1971) *Nature* **232**, 230–236.
44. Dayhoff, M. O., Hunt, L. T., McLaughlin, P. J. & Barker, W. D. (1972) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, Silver Spring, Md.), pp. 31–39.

Chapter 2

RAT KAPPA CHAIN ALLOTYPES: PARTIAL CONSTANT REGION AMINO ACID SEQUENCES

George A. Gutman*, Elwyn Loh and Leroy Hood

Department of Pathology, Stanford University, Stanford, California 94305, and Division of Biology, California Institute of Technology, Pasadena, California 91125.

Present address: Walter and Eliza Hall Institute of Medical Research, Post Office, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia.

ABSTRACT

Partial amino acid sequences have been obtained for pooled kappa chains of rats differing at the RI-1 allotype locus. The DA and LEW proteins differ by 10 residues and one sequence gap in the 81 positions compared, raising interesting genetic and evolutionary questions. The usefulness of this genetic marker for immunological experimentation, and in the study of immunoglobulin gene evolution, is discussed.

INTRODUCTION

Allotypic variation among immunoglobulin molecules has provided a great deal of information on the arrangement of antibody genes, and the possible mechanisms for their evolution (see Mage et al., 1973). Most of this information has come from studies of light and heavy chain markers in man and rabbits, and the heavy chain allotypes of mice. More recently, the rat has emerged as a fruitful source of immunoglobulin variants. A rat immunoglobulin alloantigen originally described by Barabas and Kelus (1967) was shown to be localized to immunoglobulin light chains by Wistar (1969). Subsequent work by several groups (Gutman and Weissman, 1971; Armerding, 1971; Rokhlin et al., 1971) showed this to be one of two specificities on Kappa chains inherited as Mendelian co-dominants, and unlinked to the major histocompatibility gene (Ag-B) and several coat color genes. Allotype markers on rat α-chains and γ-2b chains (Bazin et al., 1974; Beckers and Bazin, 1975) have since been described. The work reported here describes the determination of a partial amino acid sequence for pooled Kappa chains from two strains of rats (LEW and DA) differing at this Kappa-chain allotype. The finding of differences at ten residues plus one gap raises interesting genetic and evolutionary questions, which have been extensively discussed elsewhere (Gutman et al., 1975).

MATERIALS AND METHODS

Preparation of Rat Immunoglobulins. Immunoglobulin (Ig) was prepared from two inbred strains that differed in their kappa serological markers - DA ($RI\text{-}1^a$) and LEW ($RI\text{-}1^b$) (Gutman and Weissman, 1971). LEW or DA serum was precipitated with 35% saturated $(NH_4)_2SO_4$, the precipitate redissolved and passed over a DEAE-cellulose column (Whatman DE-52) in 0.07 M sodium potassium phosphate buffer at pH 7.85. The Ig fraction so obtained was concentrated by pressure dialysis to 40 mg/ml.

Preparation of Kappa Chains. Immunoglobulin at 40 mg/ml was partially reduced in 0.5 Tris buffer at pH 8.5 with 28 mM dithiothreitol at 37^o for 1 h. The sample was cooled to 0^o, and iodoacetamide added to a final 80 mM. After 1 hr at 0^o, the partially reduced and alkylated protein was dialyzed against 6 M urea, 0.05 M formic acid, and passed over a Sephadex G-100 column for separation of heavy and light chains. The light chain peak was dialyzed against 1 M acetic acid and lyophilized. It was judged to be 95% light chain by SDS acrylamide gel electrophoresis. Since 95% of rat light chains are of the kappa type (Hood et al., 1967), this preparation is regarded as essentially pure chains.

Aminoethylation and Trypsin Digestion. Aminoethylation was carried out as described by Raftery and Cole (1966). Trypsin digestion was carried out with 2% w/w trypsin-TPCK (Worthington BioChem) at a protein concentration of approximately 10 mg/ml in 0.1 M ammonium bicarbonate for 2 hr at 37^o.

Peptide Maps. Both analytical and preparative peptide maps were prepared. Analytical maps were spotted with the supernatant from a trypsin digest of approximately 1.8 mg of aminoethylated light chain (Fig. 1). The maps were run at pH 3.5 as described by Katz et al. (1959) and the spots detected either by ultraviolet fluorescence or by dipping in collidine ninhydrin stain. The spots of an analytical map were eluted with 10% NH_4OH and hydrolyzed in 6 N HCl at 100° for 12 hr. Preparative maps were prepared as described above except that up to 4 mg of protein were applied in a 0.5 x 3 cm strip parallel to the direction of electrophoresis. The peptides were detected by fluorescence or by spraying lightly with 1% ninhydrin in acetone. The peptides were eluted and characterized.

Peptide Compositions. After hydrolysis in 6 N HCl for 20 hr at 110°, the peptides were dried in vacuo and amino acid compositions were determined either qualitatively by high voltage electrophoresis on paper at pH 1.9 (Dreyer and Bynum, 1967) or by analysis on a Durrum D-500 amino acid analyzer.

Alignment and Numbering. Alignment of the peptides and of the residues within the peptides was accomplished primarily by homology with the sequence of the rat myeloma light chain (Bence-Jones protein) S211 from the LOU inbred strain of rat (Starace and Querinjean, 1975). In the case of one peptide (D9), alignment was determined by homology with a published mouse kappa sequence (Svasti and Milstein, 1972). Tryptic peptides from aminoethylated light chains are numbered from the N-terminus of the S211 constant region. The positioning of the sequence gap at position 155 was chosen to maximize

homology between the rat sequences, at the expense of that between the rat and mouse.

Dansyl-Edman Peptide Degradation. The dansyl-Edman technique was carried out as described by Gray (1972).

Cyanogen Bromide Cleavage and Automated Edman Degradation. The occurrence of methionine residues in the constant region permitted us to succinylate the light chains (10 x excess succinic anhydride by weight at pH 9.5 on 6 mg of polypeptide) and, after desalting and lyophilization, to digest the chains with CNBr in 70% formic acid. After a second lyophilization, the mixture of peptides was analyzed directly on a Beckman 890 automatic sequenator and the resulting thiazolinones were analyzed as previously described (Hood et al., 1973).

RESULTS

Peptide Maps and Peptide Compositions. Using peptide maps, previous workers have suggested that multiple amino acid substitutions distinguish the two rat allotypes (Vengerova et al., 1972). A comparison of the DA and LEW peptide maps (Fig. 1) shows that peptides 7 and 10 have different mobilities in the two strains. Peptide 6 was fluorescent, indicating that it contained one or more tryptophan residues. Amino acid compositions were carried out at least twice on all peptides and the results are shown in Table 1. In some cases the compositions vary from integral values because of slight contamination from variable-region peptides. Peptides D3 and D9 were heavily contaminated or appeared only intermittently on the fingerprints. The staining spots which

are not identified on Fig. 1 gave very low yields and were either variable-region peptides or non-peptide ninhydrin positive material.

Sequence of Succinylated Light Chain Digested with CNBr. A single predominant sequence was obtained from both the DA and LEW CNBr preparations, starting at positions 123 and 137 respectively. The N-terminus of the light chains was blocked by succinylation, and the methionine residue at position 175 probably cleaves in low yield because of the adjacent C-terminal serine (Schroeder et al. 1969). Apparently methionines in the variable region are not highly conserved. For the DA and LEW strains the following sequences were obtained, respectively: Glx-Glx-Leu-Thr-Ser-Gly-Gly-Ala-Thr-Val-Val-Val-( )-Phe-Val-Asx, and Asx-Asx-Phe-Tyr-Pro-( )-Asx-Ile-( )-Val.

Dansyl-Edman Peptide Sequences. Residues were identified for the following residue positions starting from the N-terminus of each peptide: L1-1, 2, 3; L3-1; L4-1; L5-1,2,3; L7-1,2; L8-2,3; L10-1,2; L11-1,2,3; L12-1; L13-1,2; L14-1,2; D1-1,2,3; D3-1; D5-1,2,3; D7-1,2,3,5,6; D8-1,2,3; D9-1,2; D10a-1,2,3; D11-1,2,3; D12-1; D13-1; D14-1,2. A summary of these sequence results and the homology assignments made on compositional data by comparisons with chain S211 are given in Fig. 2.

DISCUSSION

These studies have yielded partial amino acid sequences for the constant region of kappa light chains from DA and LEW rats. They are compared, in Fig. 2, with published sequences of kappa Bence-Jones proteins from LOU rats (S211) and BALB/c mice (M321). It is clear that the DA and LEW sequences differ from each other by at least ten residues and one sequence gap, i.e. in over 13% (11/81) of the positions compared. This value is in fact only a lower limit for the total number of differences, since only 81 of the 108 C-region residues were compared, and the acid and amide forms of aspartic and glutamic acid were not distinguished.

The two differences between the LEW sequence and the LOU S211 myeloma (at positions 138 and 209) were surprising, since these two strains are both serologically RI-1^b. It seems, however, that S211 is an unusual variant among LOU kappa chains, as three other LOU myelomas show sequences identical to that shown here for LEW pooled light chain (P. Querinjean, personal communication). The LOU strain (and perhaps LEW as well) seems to have a kappa-chain gene which is expressed only irregularly or at low levels.

This large number of differences between the DA and LEW proteins, which are presumably the products of allelic genes, is difficult to understand, particularly in the absence of any intermediate forms among many rat strains tested (Gutman and Weissman, 1971; Armerding, 1971; Rokhlin and Nezlin, 1974; Beckers et al., 1974). The genetic and evolutionary implications of these multiple differences have been discussed

elsewhere (Gutman et al., 1975). Briefly, they are taken as support for the idea that the two types of kappa chains are the products of duplicated genes, with an allelically inherited control mechanism allowing the expression of only one or the other gene on each chromosome, an idea which has been discussed as an evolutionary strategy by Bodmer (1973). Such a model allows the two duplicated genes to accumulate changes independently of one another, and without the preservation of intermediate forms. One would otherwise need to hypothesize the elimination of large numbers of such intermediate forms by as yet unknown selective forces.

The lack of intermediate forms of rat kappa chains is based primarily on studies of many inbred lines in various parts of the world. The only published study on wild Rattus norvegicus in Moscow (Rokhlin and Nezlin, 1974) found only the RI-1^a allotype among some 30 rats tested. A more extensive study currently under way (Gutman, unpublished observations) has detected both alleles, in widely varying frequency, among R. norvegicus from Australia and Japan, but has failed to find any rat not bearing one or the other known specificity (among a total of over 100 rats). There remains the possibility that chemical studies will find differences not detected by serological methods, but it seems unlikely that there will turn out to be a great diversity of kappa chain types, at least in modern R. norvegicus. This same study has detected both specificities, expressed in a polymorphic manner, in Rattus rattus from various parts of Asia and California, and the RI-1^b specificity in various other species of Rattus. This suggests that the sequences involved in determining the RI-1 specificities were already present before the divergence of

many modern Rattus species. It remains to be seen, however, if such sequences represent a substantial part of the large number of differences shown here.

The presence of a light-chain marker makes the rat a uniquely useful laboratory animal, since all classes of immunoglobulin share the same light chains (which in the rat are almost exclusively kappa-type). In the mouse, only heavy-chain allotypes have so far been found. There has recently been described an inherited idiotype on rat antibodies to Group A streptococcal polysaccharide, a specificity which has been localized to the light chain (Stankus and Leslie, 1974). Thus, the presence of genetic markers in both the V-region and the C-region of rat immunoglobulin light chains may prove to be an especially valuable experimental asset.

REFERENCES

Armerding, D. (1971). Europ. J. Immunol. 1, 36.

Barabas, A.Z. and Kelus, A.S. (1967). Nature 215, 155.

Bazin, H., Beckers, A., Vaerman, J.-P. and Heremans, J.F. (1974). J. Immunol. 112, 1035.

Beckers, A. and Bazin, H. (1975). Immunochem. 12, 671.

Beckers, A., Querinjean, P. and Bazin, H. (1974). Immunochem. 11, 605.

Bodmer, W.F. (1973). Transplant. Proc. 5, 1471.

Boyer, S.H., Hathaway, P., Pascasio, F., Bordley, J. and Orton, C. (1967). J. Biol. Chem. 242, 2211.

Dreyer, W.J. and Bynum, E. (1967). Methods Enzymol. 11, 32.

Gray, W.R. (1972). Methods. Enzymol. 25, 333.

Gutman, G.A., Loh, E. and Hood, L. (1975). Proc. Nat. Acad. Sci. (U.S.) 72, 5046.

Gutman, G.A. and Weissman, I.C. (1971). J. Immunol. 107, 1390.

Hood, L., Gray, W.R., Sanders, B.G. and Dreyer, W.J. (1967). Cold Spring Harbor Symp. Quant. Biol. 32, 133.

Hood, L., McKean, D., Farnsworth, V. and Potter, M. (1973). Biochem. 12, 741.

Katz, A.M., Dreyer, W.J. and Anfinson, C.B. (1959). J. Biol. Chem. 234, 2897.

Mage, R., Lieberman, R., Potter, M. and Terry, W. (1973). The Antigens, Volume I (edited by M. Sela) p.299, Academic Press, New York.

Raftery, M.A. and Cole, R.D. (1966). J. Biol. Chem. 241, 3457.

Rokhlin, O.V. and Nezlin, R.S. (1974). Scandinav. J. Immunol. 3, 209.

Rokhlin, O.V., Vengerova, T.I. and Nezlin, R.S. (1971). Immunochem. 8, 525.

Schroeder, W.A., Shelton, J.B. and Shelton, J.R. (1969). Arch. Biochem. Biophys. 130, 551.

Stankus, R.P. and Leslie, G.A. (1974). J. Infec. Dis. 130, 169.

Starace, V. and Querinjean, P. (1975). J. Immunol. 115, 59.

Svasti, J. and Milstein, C. (1972). Biochem. J. 128, 427.

Van Hoegaerden, M. and Strosberg, A.D. (1976). FEBS Letters 66, 35.

Vengerova, T.I., Rokhlin, O.V. and Nezlin, R.S. (1972). Immunochem. 9, 1239.

Wistar, R. Jr. (1969). Immunol. 17, 23.

Zeeuws, R. and Strosberg, A.D. (1975). Arch. Int. Physiol. Biochem. 83, 205.

ACKNOWLEDGMENTS

The S211 sequence was kindly provided to us before its publication by Dr. Pierre Querinjean, to whom we are grateful. This work was supported by NSF Grant BMS71-0070 and USPHS Grants AI-10781-04 and AI-09072-06. L. H. has a Research Career Development Award from NIH. This work was carried out while G. A. G. was an NIH Postdoctoral Fellow.

Table 1. Amino acid compositions of the peptides from DA and LEW strains are shown. The numbering of the peptides is as in Figure 2. The molar ratios were calculated by the following formula:

$$\text{molar ratio} = \frac{\sum_{n=1}^{N} R_n}{N}$$

where n = the number of different amino acids in the peptide

R_n = nanomole of the nth amino acid in the peptide

N = total number of residues in the peptide.

The carboxmethyl cysteine and aminoethyl cysteine were not quantitated. All residues below a molar ratio of .2 were not included. Peptides 6 were highly fluorescent, indicating a probable tryptophan. The number in parentheses indicates nearest integral number of the amino acids for each peptide. Peptide D10b has low histidine because the ninhydrin spray partially destroys the amino terminal residue. Peptides 15 have compositions compatible with positions 178-184, but since the peptide was not sequenced and since the cleavage after 177 is unexpected, the peptide was not unequivocally assigned.

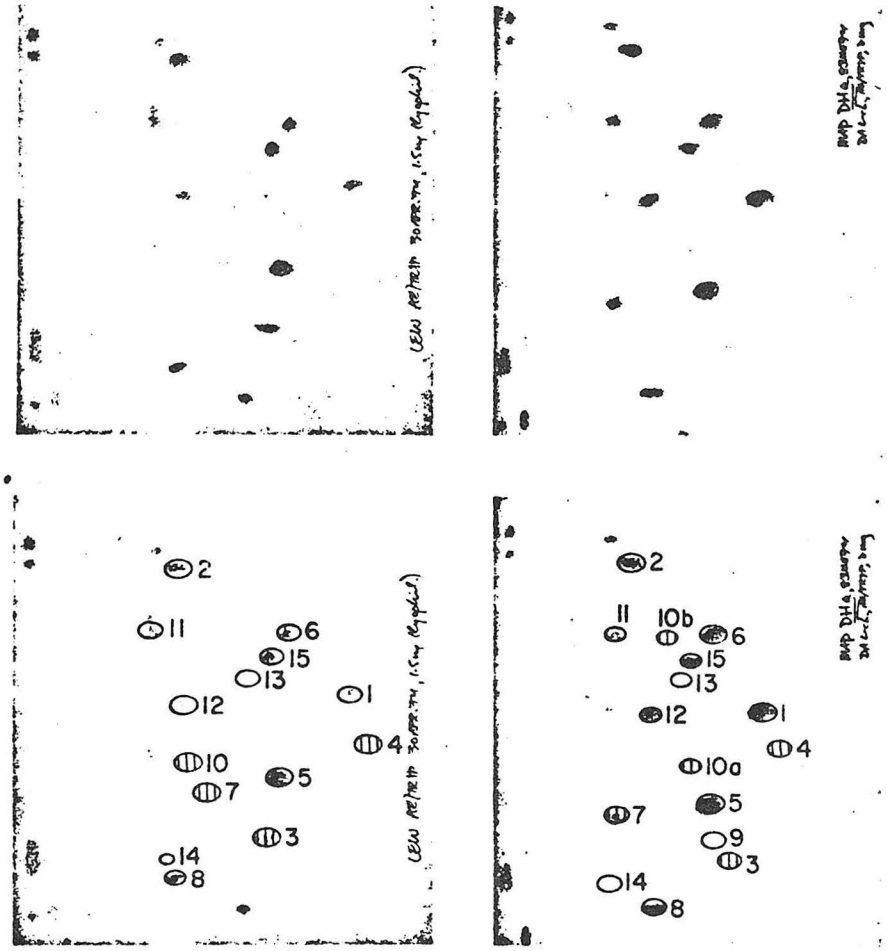
Peptide Strain	1 LEW	1 DA	2 LEW	2 DA	3 LEW	4 LEW	4 DA	5 LEW	5 DA	6 LEW	6 DA	7 LEW	7 DA	8 LEW	8 DA
Asx					1.2(1)	2.1(2)	2.0(2)	.9(1)	1.0(1)			1.1(1)	1.1(1)	3.7(4)	3.9(4)
Thr					2.5(3)							1.0(1)		.9(1)	1.0(1)
Ser					2.8(3)			.9(1)	.9(1)				.8(1)	1.8(2)	1.6(2)
Glx	1.3(1)	1.2(1)			2.2(2)							1.1(1)	2.1(2)	1.2(1)	1.2(1)
Pro					3.0(3)	1.1(1)	1.2(1)								
Gly					2.3(2)							1.2(1)	1.1(1)	1.2(1)	1.1(1)
Ala					4.5(5)										
Val					2.4(3)		1.0(1)	1.0(1)	1.0(1)					2.1(2)	2.2(2)
Met						1.0(1)									
Ile	.4(0)				1.0(1)			1.3(1)	1.3(1)			.9(1)	1.0(1)		
Leu	1.8(2)	1.9(2)			1.3(1)	.9(1)								1.2(1)	1.1(1)
Phe					1.0(1)	1.3(1)	2.2(2)								
Tyr						.9(1)	1.0(1)								
His															
Lys	.9(1)	.9(1)						.8(1)	.8(1)	+(1)	+(1)			.9(1)	.9(1)
Arg			+(1)	+(1)		.7(1)	.7(1)					.8(1)	.9(1)		
CMCys															
Cys															
AcCys					+(1)										

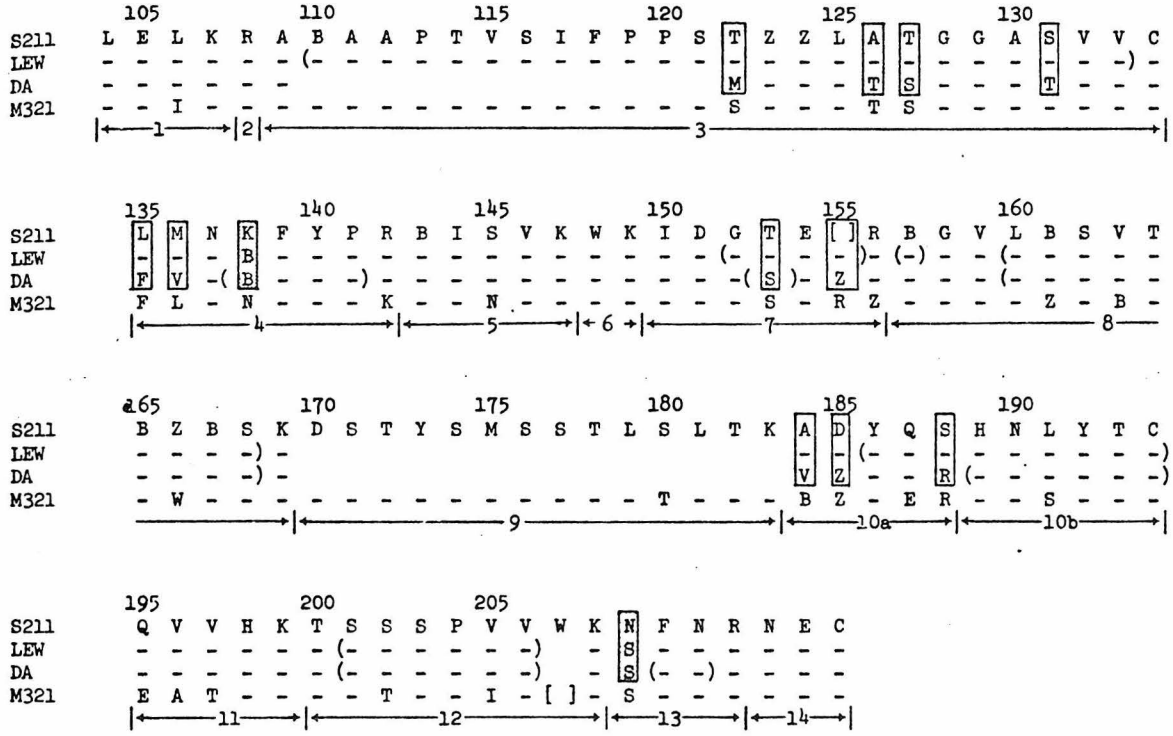
Peptide Strain	10 LEW	10a DA	10b DA	11 LEW	11 DA	12 LEW	12 DA	13 LEW	13 DA	14 LEW	14 DA	15 LEW	15 DA
Asx	2.0(2)		1.0(1)					1.2(1)	1.2(1)	1.0(1)	1.0(1)		
Thr	.9(1)		1.1(1)			.9(1)						1.9(2)	1.6(2)
Ser	1.0(1)					2.8(3)		.7(1)	.6(1)	1.1(1)	1.3(1)	1.2(1)	1.0(1)
Glx	1.3(1)	2.1(2)		1.1(1)	1.0(1)								
Pro						1.3(1)							
Gly													
Ala	.9(1)												
Val		.9(1)		2.0(2)	2.2(2)	2.0(2)							
Met													
Ile													
Leu	1.2(1)		1.1(1)									2.1(2)	2.3(2)
Phe								1.2(1)	1.3(1)				
Tyr	1.6(2)	1.1(1)	1.0(1)										
His	.9(1)		.4(1)	.9(1)	1.0(1)								
Lys				1.0(1)	.9(1)	1.0(1)						.9(1)	1.2(1)
Arg		.9(1)						.9(1)	.9(1)				
CMCys										.6(1)	.4(1)		
Cys										.3	.3		
AcCys	+(1)		+(1)										

FIGURE LEGENDS

Fig. 1. Comparison of peptide maps of trypsin digests of amino-ethylated pooled rat light chain from LEW (left) and DA (right). The lower peptide maps indicate peptides of interest. Hatched peptides are different between the two strains compared. Numbering is according to Fig. 2 except for peptide 15 which is unassigned (see legend to Table 1). Origin is to lower left, and the negative pole is at the top of the map.

Fig. 2. Amino acid sequences of rodent chain constant regions. S211 is a myeloma K chain from the LOU rat (11), LEW and DA are sequences from pooled light chains, and M321 is a mouse myeloma K chain (17). The numbering of peptides is given above the sequences. Differences between DA and LEW are circled and lightly shaded. The two differences between pooled LEW S211 are darkly shaded. Dashes indicate an amino acid identical to the S211 sequence. The one letter code of Dayhoff is used for the amino acids (see Atlas of Protein Sequence and Structure, Dayhoff, M.O., ed.).


2
11
6
15
13
1
12
4
10
5
7
3
14
8
10b
10a
9


105 110 115 120 125 130
S211 L E L K R A B A A P T V S I F P P S T Z Z L A T G G A S V V C
LEW - - - - - - (- -) -
DA - - - - - - - - - - - - - - - - - - M - - - T S - - - T - - -
M321 - - I - - - - - - - - - - - - - - - S - - - T S - - - - - - -
1 2 3
135 140 145 150 155 160
S211 L M N K F Y P R B I S V K W K I D G T E [ ] R B G V L B S V T
LEW - - - B - - - - - - - - - - - - - - (- -) -) - (-) - - (- - - - -
DA F V -(B - - -) - - - - - - - - - - - -(S)- Z - - - - (- - - - -
M321 F L - N - - - K - - N - - - - - - - S - R Z - - - - Z - B -
4 5 6 7 8
165 170 175 180 185 190
S211 B Z B S K D S T Y S M S S T L S L T K A D Y Q S H N L Y T C
LEW - - - -) -)
DA - - - -) - - - - - - - - - - - - - - - - - - V Z - - R (- - - - - -)
M321 - W - - - - - - - - - - - - - - T - - - B Z - E R - - S - - -
9 10a 10b
195 200 205
S211 Q V V H K T S S S P V V W K N F N R N E C
LEW - - - - - - (- - - - - - -) - S - - - - - -
DA - - - - - - (- - - - - - -) - S (- -) - - - -
M321 E A T - - - - T - - I - [ ] - S - - - - - -
11 12 13 14

Chapter 3

This paper has been submitted to Proc. Natl. Acad. Sci. USA

Classification: Immunology

Comparisons of myeloma proteins from NZB and BALB/c mice: structural and functional differences

(protein sequenator/genealogical analysis/NZB myeloma immunoglobulins)

E. Loh*, B. Black*, R. Riblet†, M. Weigert†, J. M. Hood* and L. Hood*

Division of Biology, California Institute of Technology

Pasadena, California 91125

Running title: NZB myeloma proteins

Proofs to be sent to: L. Hood

*Division of Biology, California Institute of Technology, Pasadena, CA 91125

†Institute of Cancer Research, Philadelphia, PA 19111

ABSTRACT Structural and functional analysis of myeloma immunoglobulins from the inbred BALB/c mouse and humans have provided important insights into the structure of the antibody molecule and the expression and evolution of antibody genes. One important question concerning these analyses is whether the myeloma process selects in a nonrandom manner the lymphocytes to be transformed. The availability of myeloma tumors in a second inbred strain of mouse, NZB, permits us to approach this question. In this respect, the N-terminal amino acid sequences of 29 kappa light chains as well as data relating to the antigen-binding properties and immunoglobulin class distribution of NZB myeloma proteins are presented and compared with similar data from the BALB/c mouse. These studies suggest that the myeloma proteins from the BALB/c and NZB mice constitute two populations of immunoglobulins with distinct functional and structural properties. The implications of this observation are discussed.

Myeloma tumors may be artificially induced by injecting mineral oil into the peritoneal cavity of inbred BALB/c mice (1). Chemical analysis of the homogeneous immunoglobulins derived from these myeloma tumors has provided important insights into the structure, genetics, and evolution of antibody molecules and genes (2,3). Moreover, the observations that immunoglobulin variable (V) regions can be divided into related sets or subgroups of sequences and that there are regions of extreme variability termed hypervariable regions have placed important constraints on the modern theories of antibody diversity (4). Indeed, our estimates as to the complexities of the various immunoglobulin gene families have been derived from the nature and extent of the heterogeneity seen in the corresponding myeloma proteins. Thus it is important to determine whether the highly artificial process of myeloma induction provides a random sampling of lymphocyte diversity in the mouse. For example, if only 10% of the lymphocyte population was capable of being transformed by the myeloma process, then diversity estimates would be too low by a factor of ten. The availability of a second inbred mouse, NZB, in which myeloma tumors can be induced gave us an opportunity to examine this question.

It should be noted that three observations suggest that BALB/c myeloma proteins may represent only a selected subset of the antibody V regions. First, approximately 5-10% of the BALB/c myeloma proteins bind a very restricted spectrum of simple haptens, including Dnp, phosphorylcholine, and various simple carbohydrate moieties (5). This frequency is much greater than would be expected from the frequency of lymphocytes binding the same haptens in normal individuals (6). Accordingly, the BALB/c myeloma proteins appear to represent a restricted sample of the potential functional repertoire of the BALB/c mouse. Second, nearly all of the BALB/c myeloma

heavy chains screened thus far have unblocked N-terminal groups, whereas only about 20% of the serum heavy chains are unblocked (7). The presence of blocked and unblocked N-terminal groups is indicative of different V-region subgroups (8). Thus the BALB/c myeloma proteins and serum immunoglobulins reflect different distributions of V subgroups and, presumably, different antibody specificities. Finally, if the residue alternatives at certain positions of myeloma sequences are compared to their counterparts from the normal serum pool, it is clear that some residues in the myeloma pool are not found in the normal pool (9). Conversely, residue alternatives found in the normal pool are missing in the myeloma pool. These differences again suggest that the normal and myeloma pools of sequences express somewhat different subsets of the antibody repertoire.

Selection also appears to operate at the level of myeloma constant (C) regions. Indeed, two-thirds of BALB/c myeloma immunoglobulins are of the IgA class whereas less than 10% of the circulating immunoglobulins are of the IgA class (10). Thus there is evidence from both the V and C regions that the BALB/c myeloma proteins do not represent a random selection of normal serum antibody molecules.

The observation that myeloma tumors can be induced in a second inbred strain of mice, NZB, permits us to compare two populations of myeloma immunoglobulins induced within a single species, the mouse (11). In this paper we report on the antigen-binding properties and immunoglobulin class distribution of NZB myeloma immunoglobulins as well as the N-terminal sequences from 29 kappa chains. These studies suggest that the myeloma proteins from the BALB/c and NZB mice constitute two populations of immunoglobulins with distinct functional and structural properties.

MATERIALS AND METHODS

Myeloma tumors. The plasmacytomas employed for these studies were induced in NZB/NIH mice (Weigert and Riblet, in preparation).

Immunoglobulin class identification and antigen-binding assays. The serological techniques for immunoglobulin class identification and the antigen-binding assays are described in a separate publication (H. C. Morse III, R. Riblet, R. Asofsky and M. Weigert, in preparation).

Immunoglobulin production. Plasmacytomas were passaged subcutaneously in (NZB x BALB/c)F_1 hybrid mice. These F_1 mice are produced at the Institute for Cancer Research (ICR) from matings of NZB/NIH and BALB/c ICR mice. The tumors secreting myeloma proteins were passaged intraperitoneally to 30-50 (NZB x BALB/c)F_1 mice primed one week prior to passage with 0.5 ml pristane (2, 6, 10, 14 tetramethylpentadecane, Aldrich Chemical Co., Milwaukee). The ascites fluid from these mice was collected and pooled.

Immunoglobulin purification. The ascites fluids obtained from tumor bearing (NZB x BALB/c)F_1 were clarified by centrifugation at 15,000 for 10-18 min. An equal volume of phosphate buffered saline with azide and EDTA (PBSAE) (.15 M NaCl, .01 M PO_4, 1 mM Na azide, 1 mM EDTA, pH 7.4) was added and diluted to a 50% concentration with a neutral solution of saturated ammonium sulfate. The ammonium sulfate solution was dissolved in PBSAE and volumes containing about 200 mg of protein were applied to G-200 column equilibrated PBSAE.

Reduction/alkylation. The peak corresponding to the myelomas protein was concentrated to 20 mg/ml and dialyzed against a tris-saline-EDTA solution (0.15 M Tris-Cl, pH 8; 0.15 M NaCl; .002 M EDTA, pH 7.0). One molar

dithiotheitol (DTT) was added to a concentration of .01 M and the solution was stirred at 37° for 1.5 hr. The sample was then placed in an ice bath and 1 M iodoacetamide was added to reach a concentration of 0.022 M. The alkylation reaction was terminated by the addition of DTT to a molarity equal to that of the iodoacetamide added. The solution was dialyzed against 8 M urea in propionic acid for 2 hr.

Separation of heavy and light chains. Reduced and alkylated proteins were fractionated on G-150 columns equilibrated in urea-propionic. Pooled light chains were dialyzed against water and lyophilized.

Automated sequence analysis. The amino acid sequence analysis of the light chains was carried out on a Beckman 890A Sequencer using a dimethylbenzylamine program as previously described (12,13,14). Briefly, aliquots of the phenylthiohydantoin (PTH) amino acids were identified by gas chromatography and thin layer chromatography. The PTH amino acids were also hydrolyzed to free amino acids and analyzed by a Durrum D-500 amino acid analyzer as previously described (13). A high pressure liquid chromatography system (Waters Associates, Inc.) using a pH 4.27 buffer system with sodium acetate-water-methanol was used to identify the PTH amino acids for some of the proteins. Generally each protein was analyzed a single time. Repetitive yields ranged from 85-92%.

Mathematical analyses. There are 50 published N-terminal sequences of kappa chains from BALB/c mice that can be compared with the 29 NZB kappa chains presented in this paper. Mathematical techniques have been developed which allow us to ask several questions of these two populations of kappa sequences (15). First, given one set of proteins (e.g., the BALB/c κ chains), what is the likelihood that a second set of proteins (e.g., the NZB κ chains)

were drawn from the same population of sequences? Second, in quantitative terms how much diversity exists within each population of sequences? These mathematical techniques can be summarized briefly as follows.

To determine whether the BALB/c and NZB sequences come from the same pool, one calculates a diversity distance index [D(B,N)] where D is the diversity distance, B represents the BALB/c sequences and N the NZB sequences. The diversity distance index is a measure of the similarity of the amino acid distributions in the two sets of sequences. The significance of the distance index is estimated by creating a statistical model which employs a randomly chosen second set of 29 BALB/c sequences drawn from the complete set of BALB/c sequences. Then one calculates the most likely distance [$\mu(B,29)$] this second set will have from the original BALB/c set and the standard deviation [$\delta(3,29)$] of the distance index from $\mu(B,29)$ if one repeats the random selection many times. Thus, to ask if BALB/c and NZB kappa chains are drawn from the same pool, one calculates the distance index D(B,N) and compares this to $\mu(B,29)$ and $\delta(B,29)$. If the difference of D(B,N) and $\mu(B,29)$ is greater than 2 $\delta(B,29)$ then one can conclude that the BALB/c and NZB are drawn from different pools.

The diversity within each set of proteins can be determined by calculating the distance index for each sequence of a population to the population as a whole and averaging all of the individual distances. This gives a quantitative measure of the diversity within a population which is designated the variation index of W(B) for the case of the BALB/c sequences. These mathematical techniques are fully described in reference (15).

Genealogical trees. Amino acid sequence data can be visually compared by drawing genealogical trees which depict a theoretical ancestry of the sequences based on the assumption of minimizing mutational events. Distances

on the tree are proportional to the number of genetic events (substitutions, deletions, insertions) that are needed to account for sequence differences. Where the tree branches (nodal points) occur, a gene duplication is represented with the sequences below the branchpoint sharing a common ancestor.

Genealogical trees were prepared by a variation of the method of Fitch and Margoliash (16). Data for the BALB/c κ chain sequences are taken from Hood et al. (9). Because of the large number of sequences, the proteins were assigned to one of four subpopulations according to their relatedness and a detailed tree was constructed for each subdivision. The four trees were joined by making a master tree of selected members of each separate tree.

RESULTS AND DISCUSSION

The nature and extent of V region diversity is similar in NZB and BALB/c myeloma κ light chains. By several criteria the diversity of NZB and BALB/c myeloma κ chains is similar even though they appear to be drawn from distinct sets of proteins. i) The distribution of variation in the first 23 residues is similar (Table 1). For example, residues 5, 6 and 16 show very little variation while 9-12, 15, 17 and 18 vary extensively. ii) The degree of variation as measured by the variation indices, W(B) and W(N), for 23 residues is very similar for the BALB/c and NZB pools of κ chains (Table 5). iii) The NZB light chains have a minimum of four different sizes in their first hypervariable region. Table 2 shows that the four size classes differ by 1, 2 and 5 residues, having, respectively, 33, 34, 35 and 38 residues before the tryptophan at position 36. One of these, the 35 residue size class, has not been seen in BALB/c sequences (17). BALB/c κ chains have three other size classes which have not been seen in NZB sequences thus far. Thus, the great diversity of

the NZB κ chains in the first 23 residues extends through the first hypervariable region and, accordingly, will be reflected in the diversity of the corresponding antigen-binding sites. iv) As has been found with BALB/c variable regions, a correlation exists between the size of the first κ chain hypervariable region and the N-terminal sequence (18). Indeed, light chains which are identical in their first 23 amino acids invariably have first hypervariable regions of identical size. For example, the first six proteins in Table 2 constitute three pairs with identical N-terminal sequence and each pair has the same hypervariable region size. This correlation is true of other sequences in the literature including several sets of κ chains from antigen-binding proteins and the set of closely related BALB/c kappa chains including M321, M70, M653, and T124.

The pool of NZB myeloma proteins appears to be distinct from the pool of BALB/c myeloma proteins by several criteria. The NZB myeloma proteins differ from their BALB/c counterparts by three criteria: immunoglobulin class distribution, antigen-binding properties, and amino acid sequence. First, constant region differences occur in that NZB myeloma proteins are predominantly of the IgG class, whereas BALB/c proteins are predominantly of the IgA class (Table 3). Thus the myeloma process amplifies different ratios of immunoglobulin classes in these two inbred strains even though the same method of induction was used. Second, the profile of simple haptens to which the NZB myeloma proteins bind appears to be distinct from that of their BALB/c counterparts (Table 4). For example, it is striking that no NZB myeloma proteins in over 200 screened bind dinitrophenol or phosphorylcholine--the two most common haptens bound by BALB/c myeloma proteins. Indeed, 12 NZB myeloma proteins from those screened bind DNA (Weigert and Morse,

unpublished data) whereas few BALB/c proteins bind this antigen. Third, the V_κ regions from the NZB and BALB/c myeloma proteins appear to form distinct populations of amino acid sequences.

The third supposition is supported by three observations. i) Twenty-nine V sequences of NZB myeloma κ chains have been examined over their N-terminal 23 residues (Table 5). Twenty-one different V_κ sequences are found among the 29 NZB-κ chains examined, and only <u>one</u> of these sequences is identical to any of the 43 different BALB/c V_κ sequences available for comparison over this region. The identity of but a single V_κ sequence between NZB and BALB/c populations would be expected if the pool of possible V region sequences in mice is so large that few repetitions would be seen. However, both the BALB/c and NZB pools of sequences show multiple repeats of V regions examined over their N-terminal 23 residues. For example, within the NZB pool of V_κ regions two identical pairs of sequences appear (e.g., PC2367 and PC2316) and three probably identical triplets appear (e.g., PC144, PC2419, PC2454). In addition, the BALB/c sequences demonstrate at least 12 pairs of identical V_κ regions (9). Thus if the NZB and BALB/c myelomas were drawn from the same pool, one would expect more V region sequences to be shared by the NZB and BALB/c myeloma kappa chains. ii) A genealogical analysis of the NZB and BALB/c κ sequences also suggests that they constitute, at least in part, distinct populations. Figure 1 is a genealogic tree which graphically displays the sequence relationships of the NZB and BALB/c myeloma κ chains in their first 23 residues. Individual immunoglobulin chains are represented by number designations at the terminal twigs of the tree. The NZB V regions are boxed to distinguish them from the BALB/c V_κ counterparts. Many NZB sequences are three or more base changes from their nearest BALB/c counterpart. For example, the PC2316,

PC2367, and PC2200 cluster in the upper right and the PC144, PC2419, PC2454 set in the lower left constitute groups of related NZB sequences which have no BALB/c counterparts. Conversely, sets of related κ chains of the BALB/c set such as the TEPC-15, HOPC-8 set have no NZB counterpart. iii) We have found a diversity distance index D(B,N) of 4.63 which is a measure of the difference between the distribution of amino acids in the first 23 residues of the BALB/c and NZB light chains (Table 3). Randomly chosen sets of sequences taken from the BALB/c population would give a diversity distance index [D(B,29)] of only 3.50 with a standard deviation of .56. Thus the NZB sequences deviate from their BALB/c counterparts by more than two standard deviations and, accordingly, the probability these two populations of proteins were chosen from the same pool is small (<.025).

Two models may account for the observation that the myeloma process in BALB/c and NZB mice appears to be transforming distinct populations of lymphocytes. The BALB/c and NZB mice may have i) different V genes or, ii) genetic differences, outside the V region structural genes, which either produce different antigenic exposure (e.g., susceptibility to viral infections) or operate as distinct control elements to modulate V gene expression. Obviously both of these possibilities could be true.

If the V genes for immunoglobulins are different in the two strains, different V regions would be expressed in these strains. Indeed, different structural genes may have been fixed in each strain from polymorphisms that preexisted in their ancestors. Consistent with this hypothesis is the observation that the NZB and BALB/c strains differ in their heavy chain (C region) allotype. Thus some of the sequences found in one strain but not in the second (Fig. 1) BALB/c may constitute structural gene polymorphisms.

However, this model does not account for the fact that the myeloma proteins in the two strains bind distinct sets of antigens (Table 4) since a response to any of these haptens can be obtainable in immunized animals from either strain. Neither does this hypothesis account for the fact that the class distribution is different in the two strains (Table 3).

A seemingly more likely explanation is that because of genetic differences other than in the V region genes, either the antigenic exposures are distinct in these two strains due to physiological differences, or control genes differ and these differences lead to the clonal expansion of (or induction of tolerance in) different sets of lymphocytes in the two strains. Accordingly, the myeloma process may transform different populations of peritoneal lymphocytes or plasma cells. This supposition is attractive in view of the propensity of the NZB mice to develop autoimmune disease with the concomitant expansion of clones of lymphocytes to self antigens (20). Consistent with this hypothesis is the fact that twelve different NZB myeloma proteins bind DNA whereas few of their BALB/c counterparts do (Table 4). Also consistent with this model is the observation that in NZB hybrids with other strains, anti-red blood cell (RBC) antibodies of the other strain appeared as frequently as anti-red blood cell antibodies of the NZB strain thus showing that NZB genes (antigens?) brought about the expression of large amounts of antibodies normally unexpressed in the non-NZB strain (21). This model would propose that the same structural genes for both sets of variable region sequences exist in both strains but that other genetic differences create distinct internal conditions which lead to the expression of different <u>subsets</u> of V genes. A possible way to differentiate between the models would be through nucleic acid chemistry, by determining whether the identical structural genes for

individual κ chains exist in both strains. The final answer may be that both models are partially correct.

The myeloma process appears to transform only a subset of the total lymphocyte repertoire the mouse can express. The spectra of binding specificities, the V_H-subgroup distribution, and the sequence data demonstrate that the myeloma process transforms different populations of lymphocytes in these two inbred strains of mice. It provides windows through which we can glimpse the antibody repertoire, but the windows probably view just a fraction of the total diversity. Since it is difficult to say how small this fraction is, it is impossible to estimate what fraction of the antibody repertoire is expressed by the myeloma process in each of the inbred strains. Estimates of the number of V-region subgroups in the antibody families, based on the sequence analysis of myeloma proteins, are therefore minimal estimates and could be far too low. Thus the number of V-region genes in many antibody families may be substantially higher than previously estimated.

This work was supported by grant AI-10781 of the National Institutes of Health. E.L. was supported by an NIH Graduate Research Training Grant.

REFERENCES

1. Potter, M. (1972) Physiol. Rev. 52, 631-719.
2. Gally, J. A. & Edelman, G. M. (1970) Nature 227, 341-348.
3. Hood, L., Campbell, J. H. & Elgin, S. C. R. (1975) Ann. Rev. Genet. 9, 305-353.
4. Hood, L., Loh, E., Hubert, J., Barstad, P., Eaton, B., Early, P., Fuhrman, J., Johnson, N., Kronenberg, M. & Schilling, J. (1977) Cold Spring Harbor Symp. Quant. Biol. 41, 817-835.
5. Potter, M. (1970) Ann. N.Y. Acad. Sci. 190, 306-321.
6. Press, J. L. & Klinman, N. (1974) Eur. J. Immunol. 4, 155-159.
7. Capra, J. Donald, Wasserman, R. L. & Kehoe, J. M. (1973) J. Exp. Med. 138, 410-427.
8. Capra, J. D. & Kehoe, J. M. (1975) J. Immunol. 114, 678-681.
9. Hood, L., Barstad, P., Loh, E. & Nottenburg, C. (1974) in The Immune System: Genes, Receptors, Signals (Academic Press, New York).
10. Barth, W. F., McLaughlin, C. L. & Fahey, J. (1965) J. Immunol. 95, 781-790.
11. Warner, N. L. & Varry, Z. (1970) J. Immunol. 105, 812-817.
12. Hood, L., McKean, D., Farnsworth, V. & Potter, M. (1973) Biochemistry 12, 741-749.
13. Smithies, O., Gibson, D., Fanning, E. M., Goodfliesh, R. M., Gilman, J. G. & Ballantyne, D. M. (1971) Biochemistry 10, 4912-4921.
14. Pisano, J. & Bronzert, T. (1972) Analytical Biochem. 45, 43-59.
15. Hood, J. M., Loh, E. & Hood, L. (1976) Biochem. Genet. 14, 467-479.
16. Fitch, W. & Margoliash, E. (1967) Science 155, 279-
17. Potter, M., Padlan, E. & Rudikoff, S. (1976) J. Immunol. 117, 626-629.

18. Barstad, P., Hubert, J., Black, B., Eaton, B., Weigert, M. & Hood, L.

19. Capra, J. D. & Kindt, T. J. (1975) Immunogenetics 1, 417-427.

20. Howie, J. B. & Helyer, B. J. (1968) Adv. Immunol. 9, 215-266.

21. Warner, N. L. (1974) Clinical Immunol. Immunopathol. 2, 556-562.

FIGURE LEGEND

Fig. 1. Genealogical tree of mouse light chains. Boxes enclose NZB light chains. Examples of BALB/c light chain sequences with multiple members are not indicated. Lengths of lines connecting proteins and nodes indicate the minimum number of base substitutions.

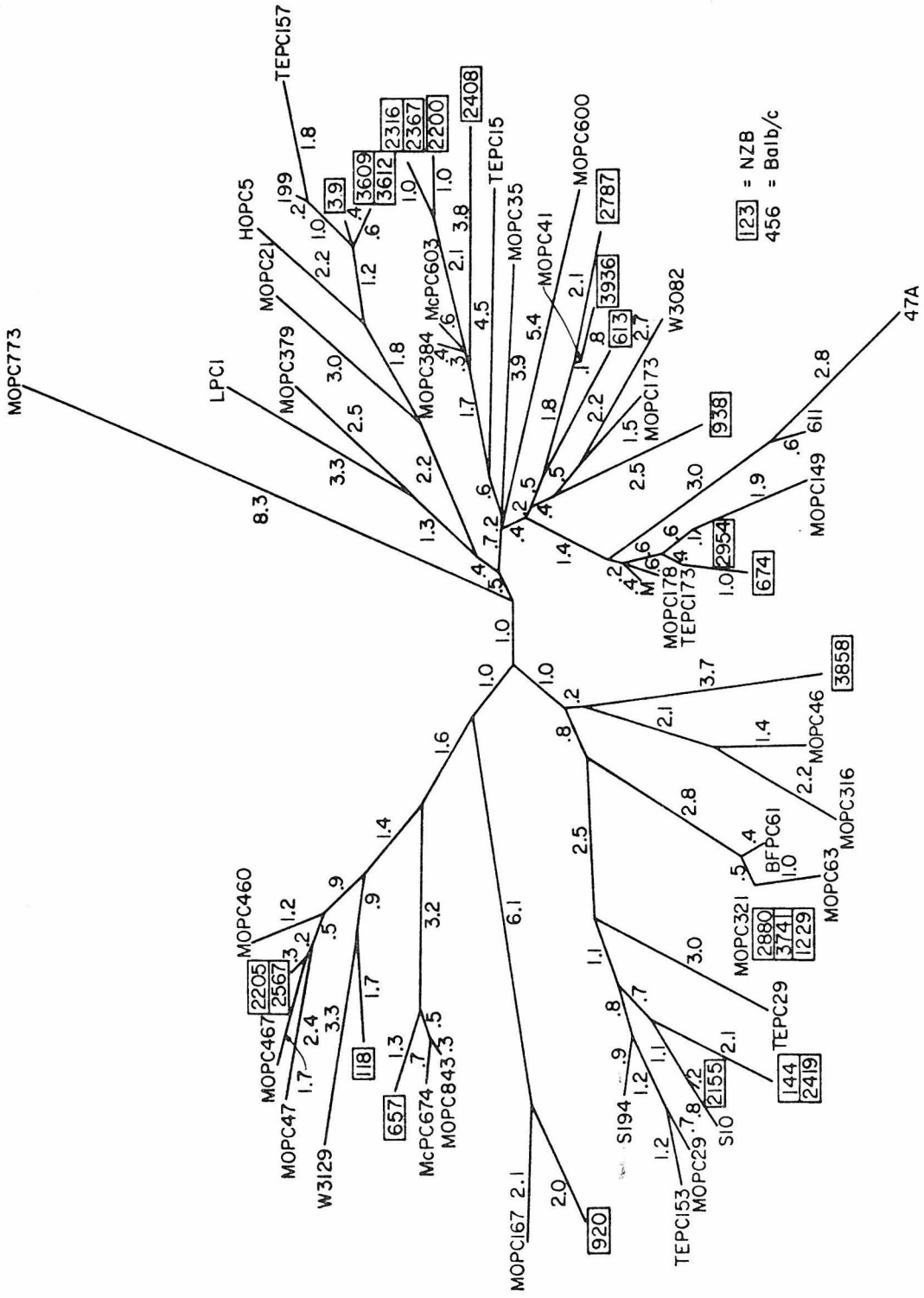

MOPC773
LPC1
HOPC5
MOPC21
TEPC157
MOPC379
MOPC384
McPC603
TEPC15
MOPC35
MOPC41
MOPC600
W3082
MOPC173
MOPC78
TEPC173
MOPC149
47A
MOPC460
MOPC467
MOPC47
W3129
MOPC843
McPC674
MOPC167
S194
S10
MOPC29
TEPC153
TEPC29
MOPC321
BFPC61
MOPC63
MOPC316
MOPC46
2205
2567
118
657
920
144
2419
2155
2880
3741
1229
3858
3609
3612
2316
2367
2200
2408
3936
2787
613
938
2954
674
611
123 = NZB
456 = Balb/c

Table 1. Amino acid sequences of NZB kappa chains.[+]

Light Chain	Position 1	5	10	15	20	Heavy Chain Class
PC2408	D I V M T Q F P S S L N V S A G E S V T M S C					IgG2b
PC2367	D I V M S Q S P S S L A V S V G Z K V T M S C					IgG2a
PC2316	———					IgA
PC3612	D I V M T Q S H K F M S T S V G B R V S I T C					IgG2b
PC3609	———			D X ———		IgG2b
PC3249	——— (H)(K) ———			D X —— X X X		IgG1
PC2200	D I V M S Q S P S S L A V S A G Z K V T M B C					IgG2a
PC118	D V V V T Q T P L S L P V S F G D Q V S I S C					IgA
PC39	D I V M T Q S Q K F M S T S V G D R V S I T C					IgG2a
PC657	D V V L T Q T P L I L S V T I G Z P A X X X X					IgA
PC920	D I V L T Q D E I S N P V T S G E R V X I S C					
PC938	D I Q M T Q S (S) S S F (S) V S L G B R V T I T C					
PC144	E I V L T Q S P A L M A A X P G Z K V T I T C					IgA
PC2419	———			S —— E ———		IgG2a
PC2454	———			S —— E ———		IgG2a
PC613	D I Q M T Q S P S X L S A S L G K G V T I T C					
PC674	D I Q M T Q S P A S L S V S V G E T V T I T C					IgA
PC2880	D I V L T Q S P A S L A V S L G Q R A T I S C					IgG2b
PC3741	———					IgM
PC1229	———			Z ———		IgG2a
PC2155	E I V L T Q S P A I T A A S L G Z K V T I T C					IgG2b
PC2787	D I Q M T Q S P S S L A A S L G Z R I S L T C					IgM
PC2205	D V V M T Q T P L S L P V S L G D Q A S I S C					IgG2b
PC2567	———					Ig3
PC2954	D I Q M T Q S M A S L S A S V G Z T V T I T C					IgA
PC3858	D I L L T Q S P A T L S V S P G E R V S F S C					IgA
PC3936	D I Q M T Q S P S N L S A S L G E R V S L T C					IgA
PC3660	D I L M T Q S P S S M S V X L G X X X X X X X					IgA
PC2279	D V Q I T Q S P S Y L A A S P G E (T) I T I N C					IgG2b
PC373	Blocked					
PC2426	Blocked					IgG2b

[+]The one letter amino acid code is used. X denotes an unknown residue. Brackets indicate an uncertainty of sequence assignment. Blocked denotes that the α amino group is covalently blocked and not free for the Edman reaction.

Table 2. Hypervariable region sequences of κ light chains. Table uses one letter amino acid code as in Table 1.

Light Chain	Position																			
				25			30		a	b	c	d				35				
PC3741	R	A	S	Z	S	V	B	X	Y	G	B	S	F	M	(D)	(W)	Y	Z	Z	K
PC2880	R	A	S	Z	S	V	B	B	Y	S	I	S	F	M	N	W	F	Z	Z	K
PC144	X	V	X	(S)	(S)	I	X	X	X	B	L	Y				X	Y	Z	Z	K
PC2419	X	V	S	(S)	(S)	I	X	X	X	B	L	X				X	Y	Z	Z	K
PC3609	K	A	S	Q	D	V	S	T	A	V	A					W	Y	Z	Z	K
PC3612	K	X	S	Z	B	V	S	T	A	X	V					X	Y	Z	Z	K
PC39	K	A	S	Q	B	V	G	S	S	V	A					X	Y	Z	Z	K
PC938	K	X	X	Z	B	I	Y	X	X	L	X					X	Y	Z	Z	(K)
PC2954	R	A	S	Z	B	I	Y	S	Y	L	A					X	Y	Z	Z	K
PC2787	X	A	S	Z	B	I	Y	G	K	L	(B)					X	F	Z	Z	K
PC3858	R	A	S	Q	N	I	G	T	S	I	H					W	Y	Q	Q	R
PC3936	R	A	S	Q	E	I	X	G	Y	L	S					X	L	Q	Q	K
PC2279	R	A	S	K	X	I	X	K	Y	L	A					X	Y	Q	Q	K
PC674	X	A	S	Z	B	I	Y	X	B	L						X	Y	Z	Z	K
PC2155	S	A	M	S	X	V	X	Y	X	X						X	Y	Z	Z	K
PC2367	K	X	X	Z	X	L	B	Y	S	X	T	Z	K	B	Y	L				
PC118	R	S	S	Z	S	L	A	B	X	Y	G	X	T	Y	L	S	P			
PC920	R	X	S	K	S	L	L	Y	T	B	X	K	T	Y	L	B				

X indicates an uncertainty in residue assignment.

Table 3. Class distribution of NZB and BALB/c myelomas. NZB and BALB/c data from Morse and Weigert (unpublished data). BALB/c normal serum data from ref. 10 . Columns do not sum to 100% because of tumors that contain multiple classes or give ambiguous results.

Constant Region	NZB	BALB/c	BALB/c Normal Serum
	%	%	mgm/ml
γA	20.9	45.5	.7
γM	1.7	1.1	.8
γ1	9.1	6.4	2.4
γ2a	15.2	6.4	1.4
γ2b	25.7	8.1	
γ3	2.2	0.7	.1

Table 4. Antigen binding spectrum of NZB and BALB/c myelomas

	NZB	BALB/c
DNA	++	-
DNP	-	++
Phosphorylcholine	-	++
α-1,3 Dextran	-	+
α-1,6 Galactan	-	+
α-1,6 Dextran	+	+
α-1,2 Levan	+	+

+ indicates from 1-10 tumors of that specificity have been found.

++ indicates greater than 10.

Chapter 4

This paper will be submitted to J. Immunol.

Running title: NZB Myeloma Heavy Chains

Comparisons of Myeloma Proteins from NZB and BALB/c Mice:

Structural and Functional Differences of Heavy Chains

E. Loh[*†], J. M. Hood[*‡], R. Riblet[+], M. Weigert[+], and L. Hood[*]

[*]Division of Biology, California Institute of Technology

Pasadena, California 91125

[+]Institute for Cancer Research

Philadelphia, Pennsylvania 19111

[‡]Present Address: Department of Mathematics, California State Polytechnic University, San Luis Obispo, California 93407

[†]Present Address: Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305.

Abstract

The N-terminal sequences from heavy variable regions of 47 myeloma proteins of the NZB mouse have been analyzed. Sixteen of these V_H regions have unblocked α amino groups and have been analyzed over their N-terminal 20 residues by automatic sequence analysis. These sequence data along with the antigen-binding profiles and immunoglobulin class distribution are compared with comparable data from BALB/c myeloma proteins. These comparisons suggest that the NZB and BALB/c populations of myeloma proteins are distinct from one another. The genetic implications of this conclusion are discussed.

Comparative studies have been carried out on the structure of immunoglobulins from a variety of mammalian and nonmammalian species (see 1, 2). In many cases, the homogeneous immunoglobulins derived from myeloma tumors have been employed for these comparative studies. Amino acid sequence studies on myeloma immunoglobulins have provided important insights into the structure and function of antibody molecules and the organization, diversity, regulation and evolution of antibody genes (3-7).

There are two general possibilities regarding the expression of myeloma immunoglobulins. i) The myeloma population in a particular animal may represent a random selection from the normal immunoglobulin repertoire of that animal. ii) The population of myeloma proteins or the "myeloma window" may express only a fraction of the diversity the animal is capable of expressing. The existence of large libraries of myeloma tumors in two inbred strains of mice, BALB/c and NZB, allows us to begin to distinguish between these two possibilities. One approach to the analysis of this problem is to carry out comparative structural, functional and immunoglobulin class analyses of the myeloma proteins from these two inbred strains. Our laboratory as well as many others have carried out detailed structural and functional analyses of the BALB/c myeloma proteins (see 1, 2). An earlier analysis of the N-terminal structures of 29 light chains from NZB mice demonstrated that the BALB/c and NZB mice appear to express distinct populations of light chains (8). In addition, the antibody class distribution and antigen-binding properties of myeloma immunoglobulins from these two strains are distinct. These observations suggest that the process of myeloma induction leads to the expression of different populations of myeloma immunoglobulins in the BALB/c and NZB mouse strains.

In this paper we analyze the N-terminal sequences of the heavy variable regions (V_H) from 47 myeloma proteins of the NZB mouse. Sixteen of these V_H regions

have free α amino groups and, accordingly, have been analyzed over their N-terminal 20 residues by automatic sequence analysis. These partial N-terminal sequences have allowed us to estimate the range and nature of diversity among the NZB V_H regions and to compare it with the comparable BALB/c data. Indeed, we have developed special statistical methods for comparing the diversity and relatedness in the populations of NZB and BALB/c myeloma immunoglobulins (9). These results confirm and extend the previous conclusion that the majority of the myeloma immunoglobulins sampled from the BALB/c and NZB mouse are generally distinct from one another.

MATERIALS AND METHODS

Myeloma induction and chain isolation. Myeloma tumors were induced as previously described using NZB/NIH mice (10, 11, M. Weigert and R. Riblet, in preparation). The myeloma immunoglobulins were purified and the light and heavy chains separated as described previously (8).

Automated sequence analysis. The amino acid sequence analysis of these proteins was carried out on an updated Beckman 890A sequencer as previously described (12). Briefly, aliquots of the PTH (phenylthiohydantoin) amino acid derivatives were analyzed by gas chromatography, thin layer chromatography, and high pressure liquid chromatography. The PTH amino acids also were hydrolyzed to their respective amino acids and amino acid analysis was done on a Durrum D-500 amino acid analyzer.

Assay for blocked proteins. Many heavy chains of immunoglobulins have a "blocked" N-terminal amino acid (i.e., a pyrrolidone-2 carboxylic acid) and cannot be analyzed by the Edman degradation procedure. To screen for heavy chains with blocked N-termini, approximately one mg of the heavy chain was analyzed on the

automatic sequenator for four steps and the resulting PTH amino acids, if any, were assayed by gas chromatography. If no identifiable PTH residues were seen, the protein is assumed to be blocked. All proteins were assayed at least twice.

Data analysis. Relatedness or genealogical trees were constructed by the method of Fitch and Margoliash (13).

Because two populations of sequences, NZB and BALB/c V_H regions, must be compared, it was necessary to develop statistical methods for comparing distinct sets of proteins. Two general questions were analyzed. First, given one set of V regions (the V_H regions of BALB/c mice) what is the likelihood that a second set of V regions (the V_H regions of NZB mice) was drawn from the same population of sequences as the first? Second, how much diversity exists within each population of sequences in quantitative terms? The detailed analyses of these questions is considered in a separate paper (9). Briefly, our approach was as follows.

To determine whether the BALB/c (designated C) and NZB (denoted N) sequences come from the same pool, a diversity distance index [D(C,N)] is calculated which measures how different the amino acid distributions of the two sets of sequences are. The significance of the distance index is estimated by creating a statistical model which assumes that a randomly chosen set N of 16 sequences (i.e., the number of analyzed NZB sequences) is drawn from the C set. Then one calculates the most likely distance this second set will have from the original BALB/c set [μ(C,16)]. This random selection process is repeated many times to determine the standard deviation of the distance index [σ(C,16)]. Thus, to ask if heavy chains from NZB and BALB/c myeloma tumors are drawn from the same pool, one calculates the distance index D(C,N) to compare it with μ(B,16) and σ(B,16). If the difference of D(C,N) and μ(C,16) is greater than 2σ(B,16), then it is highly probable that the heavy chains of BALB/c and NZB myeloma proteins are drawn from different pools.

The diversity within each set of proteins is quantified as follows. The distance index between each individual sequence of a population and the population as a whole is determined. Then all of the individual distances are averaged. This average is designated the variation index or W(C) in the case of BALB/c sequences. Thus, the diversity within the BALB/c and NZB V regions can be compared.

RESULTS AND DISCUSSION

The unblocked N-terminal sequences of NZB heavy chains demonstrate a diversity that is comparable to that of their BALB/c counterparts. Forty-seven V_H regions from NZB immunoglobulins were analyzed by automatic sequence analysis. Thirty-one of these heavy chains had blocked N-termini and were inaccessible to Edman degradation. The remaining 16 V_H regions were analyzed over their N-terminal 20 residues (Fig. 1). Thirteen of 16 V_H sequences are different from one another. These V_H regions fall into at least two distinct subgroups which are for historical reasons designated II and III (2, 12). In subgroup III, PC674 and PC2567 share several residues that distinguish them from their other counterparts and, accordingly, may constitute an additional subgroup.

The 16 V_H sequences from NZB mice (boxed) are compared with 17 V_H sequences from BALB/c mice in a relatedness or genealogical tree in Figure 2. A relatedness tree is an attempt to construct the minimum mutational pathway by which each of the sequences of a particular set can be derived from a single common ancestor. The lengths of the lines between branch points on the tree and V_H regions indicate the number of base substitutions that are required to account for the observed sequences. Thus, V_H regions that are on adjacent twigs are more similar to one another than to those more distantly placed. If proteins share common residues or substitutions, then they will share a branch, with further splitting of the terminal

twigs indicating their differences. Several interesting diversity patterns emerge from an analysis of this tree. i) Three major branches corresponding to three subgroups (i.e., I, II, and II) are evident on the tree. As more extensive V region data are accumulated on these and additional V_H sequences, it appears likely that additional subgroups will be defined. Indeed, as noted above, PC674 and PC2567 as well as W3082 may constitute a distinct subgroup (Figures 1 and 2). ii) Within each major branch or subgroup considerable diversity even in the amino terminal 20 residues is evident. For example, the seven NZB VHIII sequences reported here differ from a prototype sequence by two to six residues and from each other by up to seven residues. In contrast V_H regions of the homologous human subgroup III differs from a prototype sequence by only one to three residues in the first 20 residues (2). Qualitatively, the relatedness tree demonstrates that the NZB and BALB/c V_H region exhibit comparable diversity. The genetic interpretation of these sequence data is that if a single V gene encodes each of these major branches of the tree, then the genetic mechanism for producing somatic diversity must be capable of extensive amino acid substitutions. Of course, the existence of multiple germ line V_H genes in each subgroup could reduce the magnitude of somatic mutation require. iii) If the BALB/c and NZB mice share very similar, if not virtually identical V genes, then the diversity in mouse V regions is quite extensive. Only two of the 13 different NZB sequences were identical to their BALB/c counterparts. Indeed, in 33 different mouse V_H regions analyzed in Figure 2, 29 different V_H regions were noted. The extent of this diversity begins to approach that of the highly diverse mouse K family. It also should be stressed that the diversity analyzed (i.e., residues 1-20) is in a framework portion of the V_H region and hence is not directly associated with the antigen-binding site.

The diversity in the NZB and BALB/c populations of V_H regions can be analyzed by statistic methods described earlier in this paper and elsewhere (9). The variation

indices for both populations fall within one standard deviation of one another (Table 1). Thus, the diversity in the N-terminal region of the NZB V_H regions is comparable to that of the BALB/c V_H regions.

Thirty-one of the V_H regions from NZB myeloma tumors were blocked at their N-terminus and unavailable for direct automated sequence analysis. This is nearly two thirds of the sampled population. The generalizations made for the unblocked heavy chains cannot be extended to this group of proteins. It obviously will be important to analyze the nature and extent of diversity present in the blocked heavy chains. The unblocked heavy chains are listed in Fig. 1; the remainder of the proteins in Table II are blocked.

The myeloma immunoglobulins of inbred NZB and BALB/c mice appear to represent distinct populations of proteins by several criteria. These criteria include the expression of different populations of V_H regions, the binding to different profiles of antigens and distinct immunoglobulin class distributions.

i) The unblocked V_H regions from NZB and BALB/c myeloma proteins have distinct subgroup distributions. No NZB V_H regions fall into subgroup I. Nine of 11 V_H regions in subgroup II are of NZB origin. The vast majority of the unblocked BALB/c sequences fall into subgroup III, whereas less than half of the NZB sequences currently analyzed fall into this subgroup.

ii) More than 60% of the V_H regions from NZB myeloma immunoglobulins have blocked N-termini, whereas less than 10% of their BALB/c counterparts exhibit blocked N-termini. This is a significant observation because the V_H regions with blocked N-termini probably belong to different subgroups than those with free N-termini. For example, human heavy chains can be divided into at least three subgroups at least one of which, III, is homologous to its mouse counterpart (14, 17). The V_H regions in human immunoglobulins with a blocked N-terminus always belong

to subgroups I or II, but never III. If the same is true of the mouse V_H regions, this means that 7 out of 47 NZB V_H regions belong to subgroup III, whereas 13 out of 17 BALB/c V_H regions belong to subgroup III. Thus, the subgroup distribution among blocked as well as unblocked V_H regions appears quite different in these two inbred strains.

iii) A statistical method which compares sets of sequences rather than individual sequences, suggests that the V_H regions of NZB and BALB/c mice are selected from different populations. In comparing the amino acid distribution of the 16 unblocked NZB and 28 unblocked BALB/c heavy chains, a diversity distance of 11.6 is calculated (Table I). Comparing the amino acid distribution of any 16 randomly chosen BALB/c V_H regions to the total collection of BALB/c V_H regions gives an expected distance of 2.5 with a standard deviation from that distance of 0.4. Thus the calculated diversity distance between BALB/c and NZB V_H regions is many standard deviations from that expected if NZB and BALB/c heavy chains are drawn from the same pool of possible sequences. Thus, by several different analyses, the populations of V_H regions from BALB/c and NZB mice appear distinct.

iv) If different populations of V_H regions are expressed in BALB/c and NZB mice, these sequence differences should be reflected in distinct antigen-binding profiles. The myeloma proteins of NZB and BALB/c mice appear to have distinct antigen-binding profiles, as previously reported (8). For example, no NZB myeloma proteins in over 200 screened bind dinitrophenol or phosphorylcholine - the two most common haptens bound by BALB/c myeloma proteins. Twelve NZB myeloma proteins bind DNA but not DNP, whereas no BALB/c myeloma proteins have a similar specificity (15). The BALB/c myeloma proteins that bind α1,6-dextran are quite different from the NZB myeloma proteins binding this same antigen (Fig. 3). The NZB myeloma protein binding levan has a blocked N-terminus and presumably does not belong to subgroup III, whereas all the BALB/c myeloma proteins binding levan

fall into subgroup III. Thus, the antigen-binding spectra and even the V_H sequences from myeloma proteins binding haptens are distinct in these two strains.

v) The NZB and BALB/c myeloma proteins express different ratios of immunoglobulin classes. The distribution of heavy chain classes and subclasses among the NZB myeloma proteins are presented in Table II. About 70% of the NZB heavy chains examined fall into the IgG class. In contrast, about 65% of the BALB/c heavy chains are of the IgA class (10). The explanation for these differing class ratios in these two inbred strains is not obvious.

One may raise the possibility that particular classes of immunoglobulins are associated with particular subgroups of V regions. Indeed, heavy chains from human IgA and IgE myeloma proteins are predominately associated with subgroup III sequences, whereas other variable sequences predominate in IgG, IgM and IgD myeloma proteins (16). This striking association of V region subgroup and immunoglobulin class is not seen in the NZB myeloma proteins examined to date. In Table III are compiled the class and subgroup associations among the NZB myeloma proteins. If one assumes that the blocked heavy chains are not of subgroup III, as is the usual case for human myeloma proteins (17) then the sum of the two righthand columns gives the total number of sequences not of subgroup III for each immunoglobulin class. While the proportion of IgG heavy chains associated with subgroup III sequences is similar in NZB and human myelomas, 4/24 and 5/28 respectively, the percentage of IgA heavy chains associated with subgroup III sequences is much lower than for humans, 2/10 and 20/30, respectively (16). Thus, NZB myeloma proteins of the IgA class do not appear to be predominately associated with subgroup III sequences.

Two models may account for the observation that the myeloma process in BALB/c and NZB mice appears to be transforming distinct populations of lymphocytes. The BALB/c and NZB mice may have 1) different V genes, or 2) genetic differences

outside the V region structural genes which either produce different antigenic exposure (e.g., susceptibility to viral infections) or operate as regulatory elements to modulate V gene expression. Obviously, both of these possibilities could be true.

It appears likely that the time span since the separation of the strains is probably too short, perhaps on the order of hundreds of years, for significant divergence of structural V genes. For example, the most rapidly known evolving polypeptide, the fibrinopeptides, takes 10^6 years to fix one substitution per 100 amino acid residues (1). Thus it appears likely that most of the V genes in NZB and BALB/c mice will be identical apart from the possibility that the inbreeding process may have incorporated different structural alleles in the two strains, such as different C_H region allotypes. However, mere allelic differences do not explain the differences in class distribution nor the distinct profiles of antigens bound by each of the populations of myeloma proteins.

The expression of immunoglobulin V regions can be regulated by a wide variety of different factors. For example, tolerance to self antigens regulates the expression of antigen-binding sites. The immune response genes control the ability of an animal to respond in a quantitative fashion to certain antigens. The expression of certain polymorphisms of immunoglobulin genes, termed complex allotypes, may be regulated by control genes (18). The fact that NZB mice have a propensity for autoimmune disease also may be reflected in the repertoire of immunoglobulin specificities that are expressed in the serum. Hence there are a variety of factors that may lead to the expression of distinct immunoglobulin repertoires in these two inbred strains.

Summary. Partial N-terminal sequence analyses are useful in quickly delineating general properties about the nature of populations of V regions. The V region subgroups were first defined in precisely this fashion (3, 4). Obviously these studies

must be extended by appropriate complete V region analysis. Nevertheless, we believe that these N-terminal analyses, along with the antigen-binding properties and subclass distributions, suggest that the populations of myeloma proteins from the inbred NZB and BALB/c mice are distinct. One important implication relating to the nature of antibody diversity emerges from these observations. Myeloma proteins provide an incomplete "window" of unknown size for viewing the true extent of immunoglobulin diversity in any particular species. Thus generalizations about the extent and nature of antibody diversity derived from myeloma proteins must be tempered with the realization that the myeloma window provides only a glimpse of unknown magnitude of true immunoglobulin diversity.

Acknowledgments. This research has been supported by NIH grant AI-10781 and NSF grant PCM 76-81542.

REFERENCES

1. Dayhoff, M. O. 1972. In Atlas of Protein Sequence and Structure. Vol. 5. Biomedical Research Foundation, Washington, D. C.
2. Kabat, E. A., T. T. Wu, and H. Bilofsky. 1976. In Variable Region of Immunoglobulin Chains. Bolt, Beranek and Newman, Cambridge, Mass.
3. Milstein, C. 1967. Linked groups of residues in immunoglobulin kappa chains. Nature, 216:330.
4. Hood, L., W. Gray, E. Sanders, and W. Dreyer. 1967. Light chain evolution. Cold Spring Harbor Symp. Quant. Biol. 32:133.
5. Wu, T. T., and E. A. Kabat. 1970. An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma chains and their implications for antibody complementarity. J. Exp. Med. 132:211-258.
6. Edelman, G. M., B. A. Cunningham, W. E. Gall, P. D. Gottlieb, V. Rutishauser, and M. Waxdal. 1969. The covalent structure of an entire γG immunoglobulin molecule. Proc. Nat. Acad. Sci. USA 63:78.
7. Hood, L., E. Loh, J. Hubert, P. Barstad, B. Eaton, P. Early, J. Fuhrman, N. Johnson, M. Kronenberg, and J. Schilling. 1977. The structure and genetics of mouse immunoglobulins: An analysis of NZB myeloma proteins and sets of BALB/c myeloma proteins binding particular haptens. Cold Spring Harbor Symp. Quant. Biol. 41:817.
8. Loh, E., B. Black, R. Riblet, M. Weigert, J. M. Hood, and L. Hood. Comparisons of myeloma proteins from NZB and BALB/c mice: structural and functional differences. Proc. Nat. Acad. Sci. (submitted for publication).
9. Hood, J. M., E. Loh, and L. Hood 1976. A mathematical approach to the analysis of diversity in antibody gene families. Biochem. Genet. 14:467.

10. Potter, M. 1972. Immunoglobulin-producing tumors and myeloma proteins of mice. Physiol. Rev. 52:631-719.

11. Warner, N. L. 1971. Autoimmunity and the origin of plasma cell tumors. J. Immunol. 107:936-937.

12. Barstad, P., V. Farnsworth, M. Weigert, M. Cohn, and L. Hood. 1974. Mouse immunoglobulin heavy chains are coded by multiple germ line variable region genes. Proc. Nat. Acad. Sci. USA 71:4096-4100.

13. Fitch, W., and E. Margoliash. 1967. Construction of phylogenetic trees. Science, 155:279.

14. Wang, A. C., J. R. L. Pink, H. H. Fudenberg, and J. J. Ohms. 1970. A variable region subclass of heavy chains common to immunoglobulin G, A, and M and characterized by an unblocked amino terminal residue. Proc. Nat. Acad. Sci. USA 66:657-663.

15. Potter, M. 1971. Antigen-binding myeloma proteins in mice. Annals of N.Y. Acad. of Sci. 190:306.

16. Capra, J. D., and J. M. Kehoe. 1975. Distribution and association of heavy and light chain variable region subgroups among human IgA immunoglobulins. J. Immunol. 114:678.

17. Kehoe, J. M., and J. D. Capra. 1972. Sequence relationship among the variable regions of immunoglobulin heavy chains from various mammalian species. Proc. Nat. Acad. Sci. USA 69:2052.

18. Gutman, G., E. Loh, and L. Hood. 1975. Structure and regulation of immunoglobulins: Kappa allotypes in the rat have multiple amino acid differences in the constant region. Proc. Nat. Acad. Sci. USA 72:5046.

19. Barstad, P., J. Hubert, M. Hunkapillar, A. Goetze, J. Schilling, B. Black, B. Eaton, M. Weigert, and L. Hood. Immunoglobulins with hapten-binding activity: structure function correlations and genetic implications. Eur. J. Immunol. (submitted).

TABLE I

Statistical Comparison of NZB and BALB/c Heavy Chains

C = 28 BALB/c V_H Regions	N = 16 NZB V_H Sequences	
Diversity distance	D(C,N)	= 11.6
Expected distance	D(C,16)	= 2.5
Standard deviation	λ (C,16)	= 0.4
Internal diversity NZB	W(N)	= 12 ± 3
Internal diversity BALB/c	W(C)	= 9 ± 3

The indices are explained in the text and in reference (9).

TABLE II

Class Distribution of NZB Heavy Chains

A	M	G1	G2a	G2b	G3	Unknown*
PC2316	PC2787	PC3581	PC2200	PC2880	PC2567	PC2167
PC657	PC3741	PC3048	PC1229	PC2155		PC373
PC3660		PC3249	PC613	PC3612		PC920
PC144			PC2485	PC2205		PC938
PC3746			PC2997	PC3609		PC3519
PC3858			PC2454	PC2419		PC3941
PC3936			PC3698	PC3635		PC674
PC2954			PC2367	PC2154		
PC2193			PC39	PC2899		
PC118			PC3061	PC2426		
				PC3798		
				PC3678		
				PC3808		
				PC3815		

*These proteins were either not done or yielded ambiguous results with regard to class assignment.

TABLE III

Class and Variable Region Subgroup Association

for the NZB Heavy Chains

	Unblocked			
Class	Subgroup III	Other Subgroups	Blocked	Total
IgG1	0	1	2	3
IgG2a	1	2	7	10
IgG2b	2	3	9	14
IgG3	1	0	0	1
IgM	0	0	2	2
IgA	2	3	5	10
Untyped	1	0	6	7
TOTALS	7	9	31	47

FIGURE LEGENDS

Figure 1. N-terminal sequences of heavy chains from NZB mouse myeloma proteins. The most common residue is given as the prototype. Unassigned residues are given as question marks. Uncertainty is indicated by parentheses. The grouping is discussed in the text.

Figure 2. Relatedness tree of BALB/c and NZB heavy chains, derived from N-terminal 20 residues. The NZB sequences are boxed. The VHI sequences are circled to the right, the VHIII in the middle and the VHII are to the left.

Figure 3. N-terminal sequences of myeloma heavy chains binding α1,6 dextran.

Figure 1. N-Terminal Sequences of Heavy Chains from NZB Myeloma Proteins

Subgroup	Tumor Number	1	5	10	15	20
	Prototype	Glu Val Gln Leu Gln Gln Ser Gly Pro Glu Leu Val Lys Pro Gly Ala Ser Val Lys Ile				
II	PC3249	—	—	—	—	Met
	PC2426	—	—	—	—	—
	PC3858	—	—	—	—	—
	PC3936	—	—	—	—	—
	PC2367	—	—	Ala —	—	? Leu
	PC2954	— Ile —	—	Ala —	—	—
	PC3798	—	—	Val —	? —	? ?
	PC3678	—	—	Thr — (Glu) —	—	? ?
	PC3061	—	—	Thr Val — Ala ? —	—	Met
III	PC674	— Lys — Glu Glu —	—	Gly Gly — Gln —	Gly — Met —	Leu
	PC2567	— Lys — Glu Glu —	—	Gly Gly — Gln —	Gly — Met —	Leu
	PC2193	— Lys — Val Glu —	—	Gly Ser — Gln —	Gly ? Leu —	Leu
	PC3808	— Val Glu —	—	Gly Ala —	Gly — Leu —	Leu
	PC3815	— Val Glu —	—	Gly Gly — Ser —	Gly — Leu —	Leu
	PC118	— Met — Val Glu —	—	Gly Gly —	Gly — Leu —	Leu
	PC39	— Met — Val Glu —	—	Gly Ala Ser —	Glu — Leu —	Leu

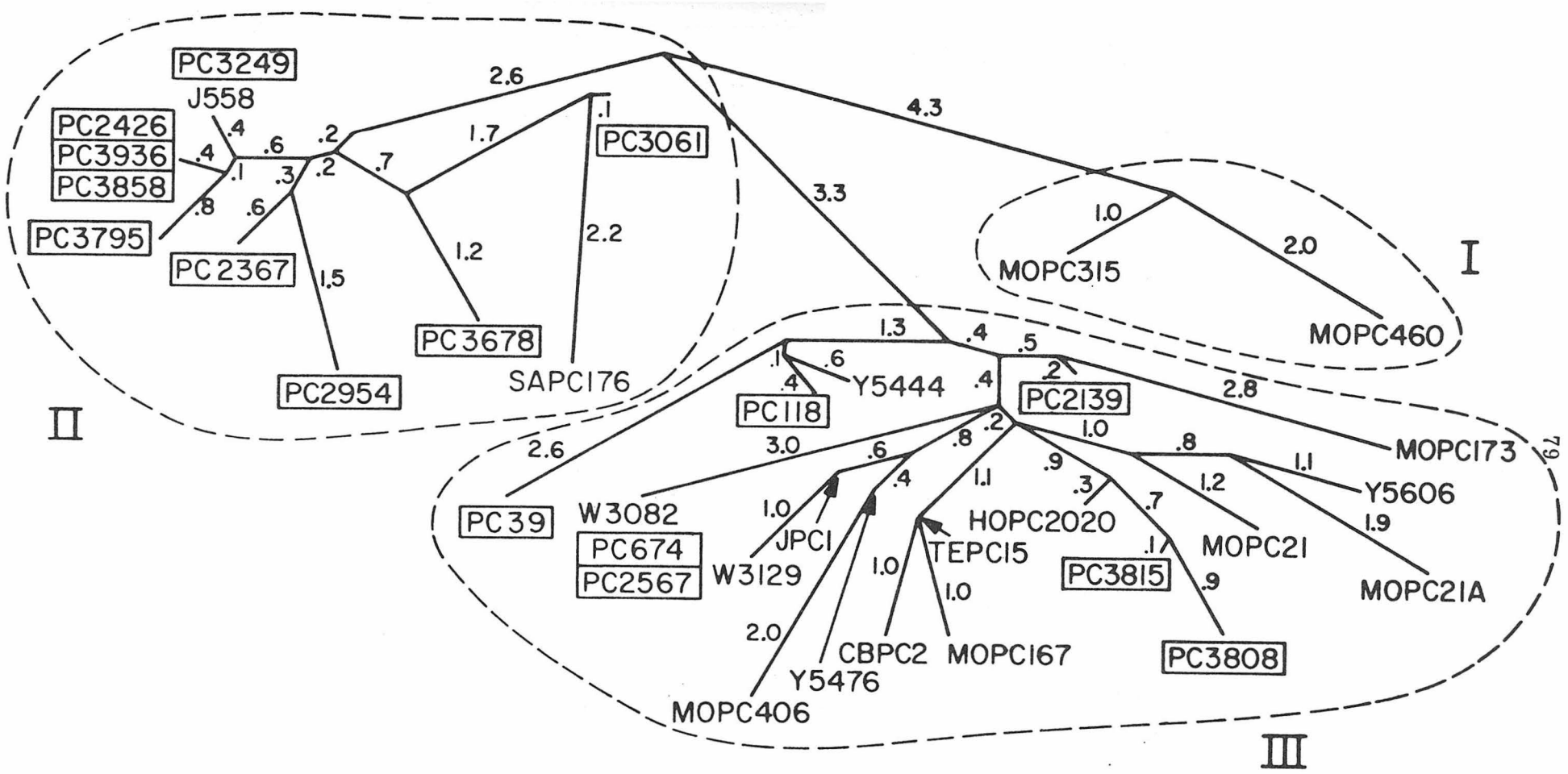

PC3249
J558
PC2426
PC3936
PC3858
PC3795
PC2367
PC2954
PC3678
SAPC176
PC3061
MOPC315
MOPC460
I
II
III
PC118
Y5444
PC39
W3082
PC674
PC2567
W3129
JPC1
PC2139
MOPC173
Y5606
MOPC21
MOPC21A
HOPC2020
TEPC15
PC3815
PC3808
CBPC2
MOPC167
Y5476
MOPC406

Figure 3. **N-terminal Sequences of Myeloma Heavy Chains Binding α1,6 dextran**

Tumor Number	1	5	10	15	20	
W3434*	Glu Val Lys Leu	Leu Glu Ser Gly	Gly Gly Leu Val	Gln Pro Gly Gly	Ser Leu Lys Leu	BALB/c
W3129*	——	Val Ile	——	——	——	BALB/c
PC3936	——	Gln —— Gln Gln ——	Pro Glu ——	Lys ——	Ala —— Val —— Ile	NZB
PC3585	——	Gln —— Gln Gln ——	Pro Glu ——	Lys ——	Ala —— Val —— Ile	NZB

*Data from (19).

Additional Heavy Chain Amino Acid Sequence Data[1]

	21	2	3	4	5	6	7	8	9	30	1	2	3	4	5	6	7	8	9	40
PC118	S	(G)	A	A	S	G	F	T	F	S	Z	Y	A	M	S	W	V	R	Z	
PC39	S	(C)	A	A	S	G	F	T	F	K	N	Y	V	M	(S)	W	V	R	Z	
PC2426	(	)(C)	K	A	S	G	Y	T	F	T	B	Y	Y	L	(S)(	)	V	(	)	K
PC2954	S	(C)	K	A	S	G	Y	S	F	T	G	Y	(B)	M	(	)(	)	V	K	Z
PC199	S	(C)	A	A	S	G	F	T	F	S	(S)	F	A	M	(	)(	)	V	(	) Z
PC3061	(	)(C)	K	A	(S)	G	Y	T	F	T	(	)	Y							
PC2567	(	)(C)	V	A	(	) G	F	T	F	S	(	)	Y	(	)	M	T	(	)	V
PC674	S	(C)	V	A	(	) G	F	(	)	F										
PC3936	(	)(C)	K	A	(	) G	Y	T	F	T	(	)	Y							
PC960+	S	(C)	K	A	(S)	G	Y	T	F	T	B	Y	Y	M	K					

+This is an IgA protein from a myeloma induced in a BALB/c x NZB F1. It has ability to bind 1,3-dextran.

PC3936 and PC3858 are proteins that bind 1,6-dextran. Bruce Black and I have data that indicates that the variable regions of the two heavy chains have two methionines. Thus the CNBr cleaved native proteins give two large peaks at OD 280 after separation on G-100. The second peak includes a variable region peptide beginning at residue 36. The first peak after reduction and albylation separates into 4 peaks. The third peak is the peptide that begins after the second methionine and for 3858 has the sequence ELHSLTSEDSAVYYCAREGPFVYWGQG()LVT. PC3936 has a similar sequence for the first five residues.

Chapter 5

This paper will be submitted to Biochemistry

Mouse Immunoglobulin Kappa Chains: The Amino Acid Sequence of Four Closely Related Variable Regions of Subgroup $V_{\kappa}21A$

E. Loh[‡¶], J. Schilling[‡], L. Gatmaitan[§], M. Weigert[§], and L. Hood[‡*]

Running Title: Closely Related κ Immunoglobulin Chains

Footnotes

From the Division of Biology, California Institute of Technology, Pasadena, California 91125. This study was supported by NIH Grant AI-10781 and NSF Grant PCM 76-81542.

[+]Division of Biology, California Institute of Technology, Pasadena, California 91125.

[§]Institute for Cancer Research, Philadelphia, Pennsylvania.

[*]To whom reprints should be addressed.

[¶]Present address: Biochemistry Department, Stanford University School of Medicine, Stanford, California 94305.

[1]The numbering of this paper is to count each residue including the insertion at 100 in PC7132. This differs from previous reports (Gray et al., 1967; Potter, 1977) who number residues 28-31 as an insertion, 27a,b,c,d. Thus our numbers are four higher than theirs from 32-99 and five higher from 101-219.

ABSTRACT: The variable (V) region amino acid sequences of four kappa (κ) immunoglobulins of the V_κ 21 subgroup from the inbred NZB mouse strain are presented. Two of these variable regions are identical. The third differs from the first two by one amino acid substitution and one amino acid insertion, both of which are located in the third hypervariable region. The fourth V region differs from the identical pair by 12 amino acid substitutions, both within and outside of the hypervariable regions. The implications of these data with regard to the genetic mechanism of antibody diversity are discussed.

Immunoglobulins constitute a highly complex family of proteins (Gally, 1973). The storage and expression of the genetic information coding for immunoglobulins is complicated by at least two ways in which germ line and somatic information differ: 1) Mechanisms exist which rearrange existing germ line nucleic acid sequences during development, and 2) germ line nucleic acid sequences may mutate during differentiation to create novel sequences which increase the antibody antigen-binding capabilities of the individual. The antibody molecule can be divided into two regions, the N-terminal portion of approximately 110 residues termed the variable (V) region and the C-terminal portion of varying size termed the constant (C) region. The V and C regions are encoded by separate germ line genes. Rearrangement of V and C genes during development has been shown to occur (Tonegawa et al., 1977). The origin of V region diversity has been a central problem in immunology for the past 60 years (Edelman and Gall, 1969; Hood et al., 1975).

The study of a closely related set of V regions, the mouse λ immunoglobulin family, has had a major impact on the immunologist's view of antibody diversity. The variable regions of 18 $_{\lambda}$ light chains have been sequenced; 12 are identical and the remaining 6 differ from the others by one to three amino acid substitutions (Weigert et al., 1977; Appella, 1971). The substitutions in the variant V_{κ} regions all fall in one of three hypervariable regions which fold to comprise the walls of the antigen-binding site. These studies are consistent with a somatic theory of antibody diversity in which the 12 identical V_{λ} regions

are directly encoded by a single germ line V gene and the six variants arise by somatic mutation of this gene (Cohn et al., 1974). This view is supported by DNA hybridization studies that suggested the mouse has relatively few (1-5) V_λ genes (Leder et al., 1976; Tonegawa et al., 1976). This hypothesis is now being directly tested by the isolation and sequence analysis of V_λ genes from embryonic (undifferentiated) (Tonegawa et al., 1978) and plasma (differentiated) cells. The mouse λ family has one peculiarity that forces immunologists to view generalizations emerging from these studies with caution. The mouse λ family is expressed infrequently in the serum--only 1-3% of the serum immunoglobulins are of the λ type, whereas the remainder are of the second light chain type, κ (McIntire and Rouse, 1970). This observation has several consequences. First, since the type of myeloma tumor roughly reflects the serum distribution of immunoglobulins, λ myeloma tumors are relatively infrequent. This means that it is difficult to obtain multiple homogeneous V_λ variants for protein or nucleic acid analysis. Second, because of their low serum levels, it is difficult to study normal λ chains to determine whether they show similar patterns of variation. Third, perhaps the mouse λ family does have very few V genes and employs a somatic mechanism that is not ordinarily required in immunoglobulin families encoded by larger numbers of V genes. In contrast, mouse κ chains constitute a highly diverse family of immunoglobulin genes (Hood et al., 1973) that are associated with more than 95% of the serum immunoglobulins (McIntire and Rouse, 1970).

We set out to find a closely related set of mouse κ chains. An earlier sequence analysis of 44 κ chains from the inbred BALB/c mouse revealed four chains that were virtually identical over their N-terminal 23 residues (Hood et al., 1973). Complete V region analysis revealed these κ chains differed by 3 to 22 residues from one another (Grey et al., 1967; McKean et al., 1973a; McKean et al., 1973b). These V_κ regions are designated the V_κ 21 subgroup (Potter, 1978). Subsequently we screened 29 myeloma κ chains from the NZB mouse and found three light chains with the V_κ 21 sequence over their N-terminal 23 residues (Loh et al., 1978). Using two of these NZB light chains as immunogens, antisera were raised which only react with κ chains closely related to the V_κ 21 sequence (Julius, M., Gatmaitan, L., and Weigert, M., unpublished data). Roughly 10% of the NZB myeloma κ chains react with one or the other of these antisera. Moreover, about 10% of serum immunoglobulins react with these screening antisera. Accordingly, the V_κ 21 subgroup appears to constitute a set of closely-related mouse κ chains that are present at the 10% level as normal serum immunoglobulins and as myeloma proteins.

We have begun to determine the amino acid sequence of approximately 22 of these closely related κ chains from the NZB mouse. The large scope of the task (2500 residues) has been made possible by the use of improved automated sequencing technology (Hunkapiller and Hood, 1978). In this report we present the sequence of four closely-related V_κ 21 regions from NZB myeloma immunoglobulins that constitute a closely-related subset of the V_κ 21 subgroup.

Experimental Procedure

Myeloma Tumors. Plasmacytomas were passaged subcutaneously in (NZB x BALB/c)F_1 hybrid mice. These mice are produced at the Institute for Cancer Research from matings of NZB/NIH and BALB/c mice. The tumors secreting the κ chains analyzed here were passaged intraperitoneally into 30-50 hybrid mice primed one week prior to passage with 0.5 ml pristane (Aldrich Chemical Co., Milwaukee). The ascites fluid from these mice was collected and pooled.

Protein Purification. The ascites fluid obtained from tumor-bearing mice was clarified by centrifugation at 15,000 rpm for 10-18 min. An equal volume of PBSAE (0.15 M NaCl, 0.01 M PO_4, 1 mM Na azide, and 1 mM EDTA at pH 7.4) was added. This solution was made 50% saturated with a neutral, saturated solution of ammonium sulfate. The resulting protein precipitate was dissolved in PBSAE and about 200 mg of protein was applied to G200 column (5 cm x 120 cm) equilibrated in PBSAE.

Partial Reduction and Alkylation of Proteins. The immunoglobulin peak was concentrated to 20 mg/ml and dialyzed against TSE (0.15 M Tris-Cl, 0.15 M NaCl, and 0.002 M EDTA at pH 7.0). One molar dithiothreitol (DTT) was added to a concentration of 0.01 M (01 M/10 ml) and the solution was stirred at 37°C for 1.5 h. Then the sample was placed in an ice bath and 1 M iodoacetic acid was added to reach a concentration of 0.022 M. The alkylation reaction was

terminated by the addition of DTT to a molarity of 0.022. The solution was dialyzed against 8 M urea in propionic acid for 2 h.

Separation of Heavy and Light Chains. Reduced and alkylated immunoglobulins were fractionated on G150 columns (2.5 cm x 120 cm) equilibrated in 8 M urea and propionic acid. The light chain pool was dialyzed against water and lyophilized.

Cyanogen Bromide Cleavage and Fragment Separation. Thirty to 50 mg of light chain were dissolved in 70% formic acid (0.5 μM/ml). Cyanogen bromide was then added to reach a 1.5-fold excess over protein (w/w) and the reaction allowed to proceed overnight at room temperature. The reaction was terminated by the addition of distilled water to the solution and the solution was lyophilized. The cleaved light chains were dissolved in 1-2 ml of 0.1 M acetic acid and fractionated on G50 column (1.5 cm x 80 cm) equilibrated in 0.1 M acetic acid.

Amino Acid Analyses. Proteins and peptides were hydrolyzed in evacuated tubes in 5.7 N (constant-boiling) HCl at 110°C for 24 h. The amino acid analyses were performed on a Durrum D500 Amino Acid Analyzer.

Complete Reduction and Aminoethylation or Carboxymethylation of Peptides. Peptides were dissolved in 8 M urea and 2 M Tris-HCl at pH at a concentration of 5-10 mg/ml. This solution was made 0.1 M in DTT and stirred for 60 min

at 37°C. Ethyleneimine was added three times at 5 min intervals (50 µl/ml) and the solution was agitated vigorously after each addition. The solution was dialyzed against 0.05 M ammonium bicarbonate and directly applied to a G50 column (1.5 cm x 80 cm). Alternatively, 0.022 M iodoacetic acid was added to the 0.1 M DTT solution and the sample was applied directly to a G50 column (1.5 cm x 80 cm) equilibrated with 0.1 M ammonium hydroxide.

Chemical Reagents. Dithiothreitol was purchased from California Biochemical Corp. and used without further purification. Iodoacetamide (Pierce Chemical Corp.) was recrystallized three times from ethanol. ^{14}C iodoacetamide was purchased from New England Nuclear Corp.

Ethyleneimine was purchased from Matheson Coleman and Bell Co. Tris (Ultra Pure) was obtained from Schwarz Bio-Research. Guanidine-HCl (Ultra Pure) was obtained from Mann Research. Reagent grade cyanogen bromide was obtained from Eastman Chemical Co.

Enzymes. Trypsin treated with L-(tosylamido-2-phenyl)-ethyl chloromethyl ketone and chymotrypsin (three times recrystallized) were purchased from Worthington Biochem. Corp., and stored as 1% solutions in 0.001 N HCl at -20°C. Thermolysin (three times recrystallized) was purchased from California Biochemical Corp. and freshly prepared as 1% solution in distilled water before each digestion.

Enzyme Digestions. Peptides were dissolved in 0.05 M ammonium bicarbonate and digested with trypsin (2% w/w) or chymotrypsin (20 µl/0.5 ml in 0.5 M ammonium bicarbonate at pH 8) or thermolysin (0.1 mg/ml or 2% w/w) and incubated at 37°C for 2-6 h.

Preparative Fingerprints. Peptide mixtures were applied as a narrow 2-inch band on a 36 x 12-inch sheet of Whatman 3 MM paper. Electrophoresis was carried out in a pH 4.7 buffer (25 ml pyridine, 25 ml glacial acetic acid, 25 ml n-butanol, 900 ml water) at 3000 V for 4 h under cooled Varsol. Ascending chromatography was done at room temperature for 12 h in a mixture of 244 ml n-butanol, 378 ml pyridine, and 76 ml glacial acetic acid diluted to 1000 ml with water. After drying, the paper was dipped in a solution of 0.02% ninhydrin in acetone and allowed to develop at room temperature.

Peptides were cut out and eluted with either 5.7 N (constant-boiling) HCl or 50% pyridine.

Automated Sequenator Analysis. Sequence analysis was carried out using a Beckman 120A sequencer modified according to Wittmann-Liebold (1973) and Wittmann-Liebold et al. (1976). These modifications included a new vacuum system, a new valving and delivery system, and a chamber for the automatic conversion of the thiazolinone derivatives to the more stable phenylthiohydantoin (PTH) derivatives. The nature of these changes as well as the purification

of the Edman reagents and solvents, the use of Polybrene to completely sequence small peptides in the spinning cup instrument and the utilization of high pressure liquid chromatography to separate, identify and quantitate all 20 PTH-amino acid derivatives is thoroughly documented in a previous publication (Hunkapiller and Hood, 1978). Some peptide sequences were determined on a solid phase sequenator of our own design employing (^{35}S)phenylisothiocyanate (W. J. Dreyer, D. Helphrey, and L. Hood, in preparation). The (^{35}S)phenylthiohydantoin amino acids were identified by thin layer chromatography and autoradiography.

Results

Sequence Strategy. The four NZB chains whose sequences are presented here, PC2880, PC1229, PC7132, and PC2413, have generally similar amino acid compositions and share four methionine residues. One previously sequenced BALB/c κ chain of the $V_{\kappa}21$ subgroup, MOPC 70, also has four methionine residues (Gray et al., 1967). Automated sequence analysis of the isolated cyanogen bromide fragments from each of these NZB chains reveals that they are homologous to their MOPC 70 counterparts. The methionine residues fall at positions 37, 82, 89, and 179 and the corresponding cyanogen bromide peptides are designated CN1, CN2, CN3, CN4, and CN5 respectively.[1] Those portions of these peptides corresponding to the V region were sequenced on the sequenator and where necessary smaller peptide fragments were prepared and analyzed.

Isolation of Cyanogen Bromide Peptides. Native PC1229 light chain, cleaved with cyanogen bromide, gave the elution profile on G50 Sephadex shown in Figure 1a. Peak I includes three cyanogen bromide fragments, 1-37 (CN1), 90-180 (CN4), and 181-219 (CN5), covalently linked to one another as well as intact and partially cleaved light chain. Peak II contained the peptide extending from 38-82 (CN2). Peak III was the peptide extending from 83-89 (CN3). Similar fraction patterns were obtained for cyanogen bromide digests of PC2880, PC7132, and PC2413.

After reduction and aminoethylation, peak I from PC1229 was fractionated on Sephadex G50 (Figure 1b). Peak Ia contained intact and partially cleaved light chain. Peak Ib contained residues 90-180 (CN4) while peak Ic contained two peptides, 1-38 (CN1) and 181-219 (CN5). Similar gel filtration patterns were obtained for the peak I of PC2880, PC2413, and PC7132 after reduction and aminoethylation.

Enzymatic Cleavage of Cyanogen Bromide Peptides. Peptides from peaks Ia, Ib, II, and III were cleaved with trypsin and/or thermolysin and the resulting peptides were purified using preparative fingerprint techniques. The amino acid compositions of certain of these peptides are presented in Table I.

Arginine Cleavage of PC2413. Totally reduced and alkylated PC2413 was succinylated to block the lysine residues and digested with trypsin to cleave

at the arginine residues. The resulting peptides were fractionated on Pharmacia AcA-54 (1.5 cm x 80 cm) in 3 M guanidine and 0.2 M ammonium bicarbonate (Figure 2). Peak ST contained two peptides, 66-112 and 113-160.

Amino Acid Sequence Analysis of Four V_K Regions. A schematic diagram of the strategies employed to determine the V region sequences of PC2413, PC1229, PC2880, and PC7132 is presented in Figure 3. In general the automatic sequenator was used to sequence the N-terminal 38 residues of the light chains, thus providing an overlap between cyanogen bromide peptides 1 and 2. Then cyanogen bromide peptide 2 (residues 38-82) was sequenced but for its last few residues. The composition of these residues was determined from the amino acid analysis of smaller tryptic peptides. Cyanogen bromide peptide 3 was sequenced except for the last two or three residues. These were determined by composition and homology. Compositional data is shown in Table I for the relevant peptides. Cyanogen bromide peptide 4 (residues 90-178) was sequenced into the constant region. Representative data for the sequence analysis of intact light chains and cyanogen bromide peptide CN2 are presented in Figures 4 and 5, respectively. The repetitive yields for these sequenator runs ranged between 92 and 95%. The deviations from this sequence strategy for individual variable regions are indicated below.

PC2413. CN2 was sequenced from residues 38-66. An arginine peptide was sequenced from residues 66-94. This variable-region-arginine peptide

was simultaneously sequenced with a constant-region-arginine peptide (residuces 113-160) whose previously established sequence (Gray et al., 1967) could readily be subtracted from that of the V region peptide (Figure 6). The simultaneous analysis of two peptides is greatly facilitated by the precise quantitation of the PTH-derivatives that is made possible by our modified sequenators (see Hunkapiller and Hood, 1978). Thus overlap sequences were obtained for each of the four V-region methionine residues and the complete V region sequence of PC2413 was unambiguously determined.

PC2880. The N-terminus of this light chain was sequenced for 37 residues with some ambiguity (Figure 3). To resolve these ambiguities, CN1 was succinylated, digested with trypsin and two arginine peptides with free α amino groups (residues 19-24 and 25-37) were sequenced simultaneously as a mixture to provide an unambiguous sequence for residues 19-37. CN3 was sequenced on the solid phase sequenator. The other CN peptides were sequenced as described above. Thus the V region of PC2880 is completely sequenced but for the compositional data at positions 80 and 81 (Figure 3).

PC1229. The sequence analysis of this protein was as described in the introduction to this section (Figure 3), except that CN3 was not sequenced. The CN3 peptides from each of these four light chains had identical electrophoretic mobilities on preparative fingerprints. Moreover, PC1229 and PC2880

had identical electrophoretic mobilities by isoelectric focusing (D. Gibson, personal communication). Since these proteins are otherwise identical throughout their V-region amino acid sequences, their amide distribution must be identical and it is highly likely their amino acid sequences are identical as well. Accordingly, the V region of PC2880 is completely sequenced except for the compositional data at positions 83-88.

PC7132. This V region is completely sequenced except for residues 87 and 88 which were ordered by homology using the approach outlined above (Figure 3).

Discussion

The V-region sequences of four NZB κ chains, PC2880, PC1229, PC7132, and PC2413, and one BALB/c κ chain, MOPC 70, are presented in Figure 7. PC2880 and PC1229 are identical while PC7132 is identical to PC2880 but for the insertion of a proline residue at position 100 and the substitution of a tyrosine for a tryptophan at position 101. The BALB/c sequence, MOPC 70, is identical to PC2280 except for two substitutions in the first hypervariable region. In contrast, PC2413 exhibits 12 amino acid substitutions when compared with PC2880. Four are in the framework regions (i.e., position 78, 87, 108, and 112) whereas four fall in the first hypervariable region and two fall in each of the second and third hypervariable regions.

Several interesting points may be made from these data. Two V regions in this set of five are identical. Immunologists have generally argued that two or more identical V region sequences probably represent the products of germ line V genes (Cohn et al., 1974; Hood et al., 1977), although somatic mechanisms can be postulated which give identical V regions. Thus the PC2880 sequence may represent a germ line V gene.

If the PC2880 sequence does represent a germ line V gene and those sequences differing from PC2880 are produced by somatic mutation, three of the five V sequences in this subgroup are somatic variants. Three interesting constraints may be placed on this somatic mutation mechanism. i) The putative somatic mechanism must be capable of generating sequence insertions (e.g., PC7132). ii) The somatic mutation mechanism must be capable of generating as many as 13 nucleotide substitutions. One of the 12 amino acid substitutions in PC2413 is a two base substitution. iii) Somatic mutation must occur in framework as well as hypervariable regions. These last two constraints would be eliminated if PC2413 is encoded by a germ line V gene distinct from that coding for PC2880.

The BALB/c sequence MOPC 70 is very similar to the NZB sequence PC2280. It is unlikely the two amino acid substitutions that occur at positions 27 and 28 are strain specific, because the NZB residues are seen in other BALB/c κ chains at these positions (D. McKean, personal communication). Thus we

conclude that very similar, if not identical, V genes will encode the V_K 21 subgroup of variable regions in the inbred NZB and BALB/c mice.

The relationship of these four V_K 21 sequences to 15 additional V_K sequences of the NZB mouse will be considered in separate papers where a more detailed summary of the implications of these data are presented (Weigert, M., L. Gatmaitan, E. Loh, J. Schilling, and L. Hood, in preparation; and J. Schilling, E. Loh, L. Gatmaitan, M. Weigert, and L. Hood, in preparation).

References

Appella, E. (1971) Proc. Natl. Acad. Sci. U.S.A. 68, 590-594.

Cohn, M., Blomberg, B., Geckeler, W., Raschke, W., Riblet, R., & Weigert, M. (1974) in The Immune System: Genes, Receptors, Signals (Sercarz, E. E. et al., Eds.) pp. 89-117, Academic Press, New York.

Edelman, G. M., & Gall, W. E. (1969) Ann. Rev. Biochem. 38, 415-466.

Gally, I. (1973) in Antigens (Sela, M., Ed.) Vol. I, pp. 162-299, Academic Press, New York and London.

Gray, W. R., Dreyer, W. J., & Hood, L. (1967) Science 155, 465-467.

Hood, L., Campbell, J. H., & Elgin, S. C. R. (1975) Ann. Rev. Genet. 9, 305-353.

Hood, L., Loh, E., Hubert, J., Barstad, P., Eaton, B., Early, P., Fuhrman, J., Johnson, N., Kronenberg, M., & Schilling, J. (1977) Cold Spring Harbor Symp. Quant. Biol. 41, 817-836.

Hood, L., McKean, D., Farnsworth, V., & Potter, M. (1973) Biochemistry 12, 741-749.

Hunkapiller, M., & Hood, L. (1978) Biochemistry, in press.

Leder, P., Honjo, T., Seidman, J., & Swan, D. (1977) Cold Spring Harbor Symp. Quant. Biol. 61, 855-862.

Loh, E., Black, B., Riblet, R., Weigert, M., Hood, J. M., & Hood, L. (1978) Proc. Natl. Acad. Sci. U.S.A., submitted.

McIntire, K. R., & Rouse, A. M. (1970) Fed. Proc. Fed. Am. Soc. Exp. Biol. 29, 704 (Abstract).

McKean, D., Potter, M., & Hood, L. (1973a) Biochemistry 12, 760-771.

McKean, D., Potter, M., & Hood, L. (1973b) Biochemistry 12, 749-759.

Potter, M. (1978) Adv. Immunol. 20, 1-40.

Summers, M. R., Smythers, G. W., & Oroszlan, S. (1973) Anal. Biochem. 53, 624-628.

Tonegawa, S., Hozumi, N., Matthyssens, G., & Schuller, R. (1977) Cold Spring Harbor Symp. Quant. Biol. 61, 877-889.

Tonegawa, S., Maxam, A, M., Tizard, R., Bernard, O., & Gilbert, W. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 1485-1489.

Weigert, M., and Riblet, R. (1977) Cold Spring Harbor Symp. Quant. Biol. 61, 837-846.

Wittmann-Liebold, B. (1973) Hoppe-Seyler's J. Physiol. Chemie 354, 1415-1431.

Wittmann-Liebold, B., Graffunder, H., & Kohls, H. (1976) Anal. Biochem. 75, 621-633.

Table I. Amino acid compositions of selected peptides. Expected number of amino acids from sequences data or homology with other members of set shown in parentheses.

	PC2413		PC2880		PC1229		PC7132	
	CN2 83-89	CN3 66-82	CN2 66-82	CN3 83-89	CN2 66-82	CN3 83-89	CN2 66-82	CN3 83-89
Asp	1.0(1)	1.9(2)	2.3(s)	1.6(s)	2.0(2)	2.0(2)	2.2(2)	(2)
Thr	1.1(1)		1.0(1)	1.0(1)	1.0(1)	1.2(1)	1.0(1)	(1)
Ser	4.1(4)	1.0(1)	4.3(4)		4.1(4)		3.7(4)	
Glu		2.1(2)		1.9(2)		1.5(2)		(2)
Pro+	0.2(1)		1.2(1)		0.7(1)		0.7(1)	
Gly	3.3(3)		3.6(3)		3.3(3)		3.9(3)	
Ala		1.1(1)		1.4(1)		1.2(1)		(1)
Val								
Met++	0.7(1)	0.6(1)	(1)*	1.2(1)	0.4(1)	0.3(1)	(1)*	(1)
Ile	1.4(2)		0.7(1)		0.6(1)		1.1(1)	
Leu	0.9(1)		1.0(1)		0.8(1)		1.1(1)	
Tyr								
Phe	1.4(2)		1.0(2)		1.2(2)		1.9(2)	
His	1.0(1)		0.6(1)		0.5(1)		1.0(1)	

+Proline is difficult to normalize to the same standard because it is measured at a different and more noisy wavelength.

++Methionine is converted to homoserine lactone in the cyanogen bromide cleavage.

*Homoserine lactone values were not quantitated.

Figure Legends

FIGURE 1a: Elution profile of PC1229 light chain from Sephadex G50 after cyanogen bromide cleavage. 1b. Elution profile of CN1 from Sephadex G50 after full reduction and aminoethylation.

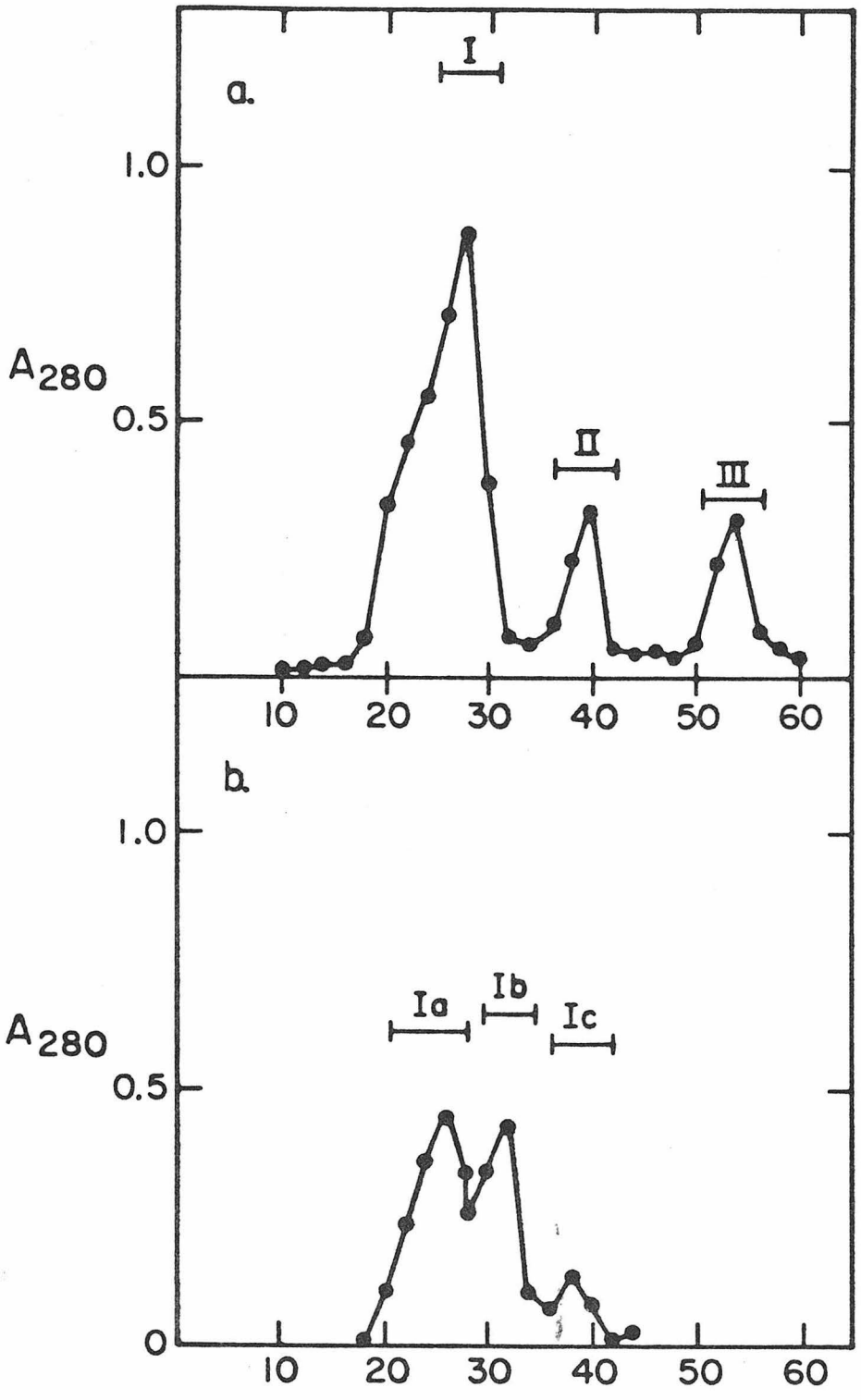

a.
I
II
III
1.0
0.5
A_{280}
10
20
30
40
50
60
b.
Ia
Ib
Ic
1.0
0.5
0
A_{280}
10
20
30
40
50
60

Figure 1

FIGURE 2: Elution profile of PC2413 from Pharmacia AcA-54 after total reduction, alkylation, succinylation, and trypsin digestion. The peak denoted ST contains a V region peptide extending from residues 66-112 and a C region peptide extending from residues 113-160.

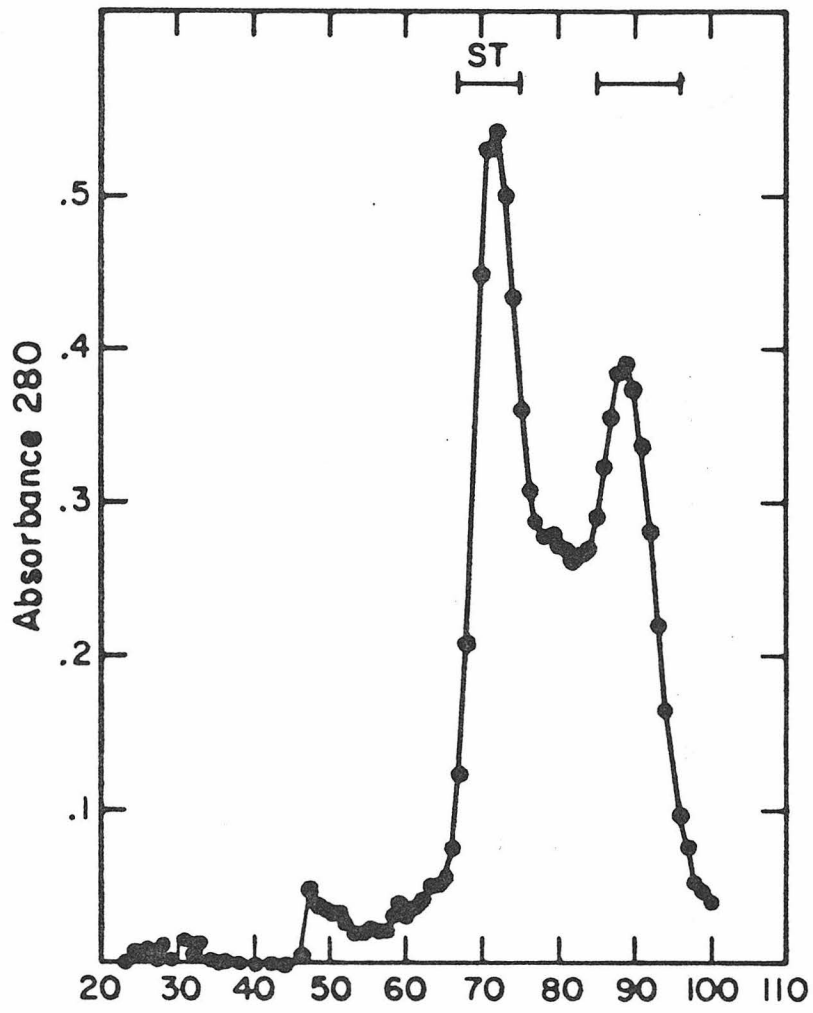


Figure 2

FIGURE 3: Strategies employed to sequence four V_K21 regions from NZB myeloma light chains. The prototype sequence is given at the top and deviations from this sequence are indicated by appropriate letters on the corresponding lines. L indicates intact light chain; CN denotes cyanogen bromide peptides; ST indicates a fragment derived from a specific arginine cleavage; and Compositions indicates the amino acid composition of certain peptides whose identity is indicated. ( ) indicates a deletion of one residue. Blank regions indicate areas of the corresponding peptide whose sequence was not determined on the sequenator.

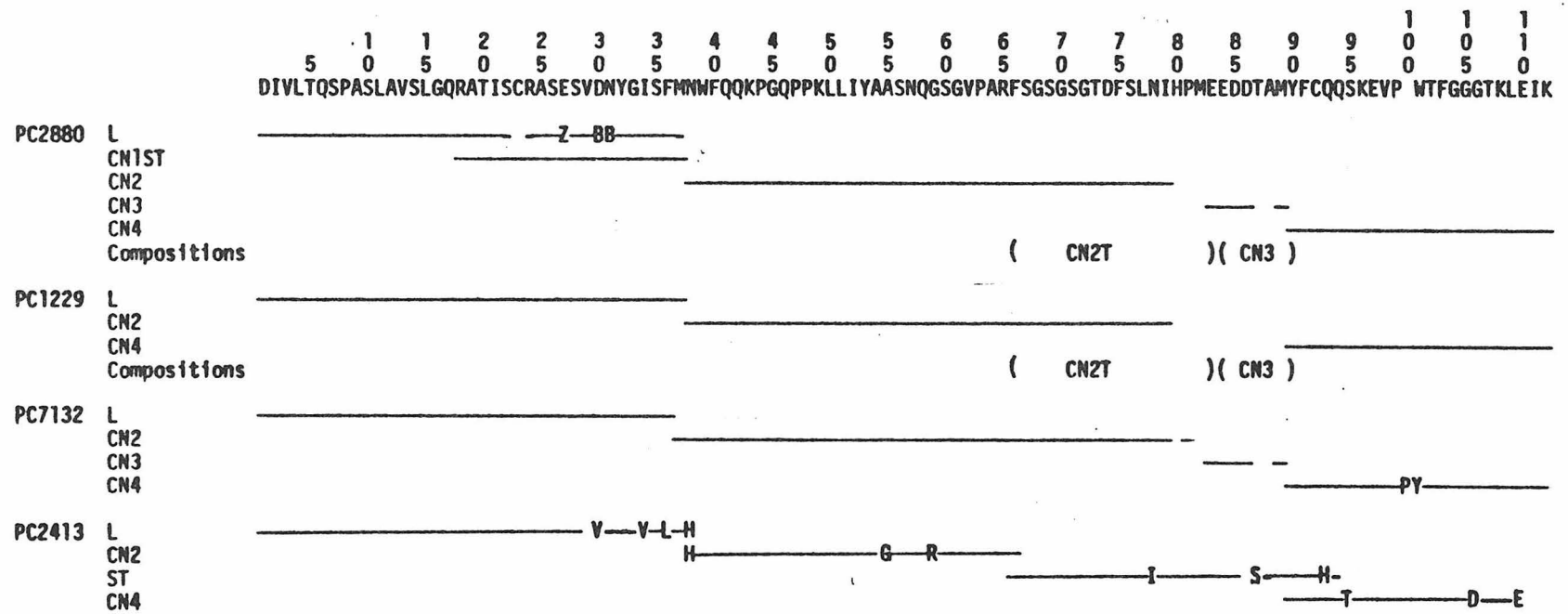

5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
DIVLTQSPASLAVSLGQRATISCRASESVDNYGISFMNWFQQKPGQPPKLLIYAASNQGSGVPARFSGSGSGTDFSLNIHPMEEDDTAMYFCQQSKEVP WTFGGGTKLEIK
PC2880
L
CN1ST
CN2
CN3
CN4
Compositions
(CN2T)(CN3)
PC1229
L
CN2
CN4
Compositions
(CN2T)(CN3)
PC7132
L
CN2
CN3
CN4
PY
PC2413
L
CN2
ST
CN4

FIGURE 4: Peak heights of amino acid phenylisothiohydantoins from an N-terminal run of PC2413 taken by high pressure liquid chromatograms. The identified residue ast each position are indicated by solid circles. A scale change at position 21 amplifies the signal by a factor of two.

```
1                   5                  10                  15                  20
Asp-Ile-Val-Leu-Thr-Gln-Ser-Pro-Ala-Ser-Leu-Ala-Val-Ser-Leu-Gly-Gln-Arg-Ala-Thr
                   25                  30                  35
Ile-Ser-Cys-Arg-Ala-Ser-Glu-Ser-Val-Val-Asn-Tyr-Gly-Val-Ser-Leu-Met- ? -Trp
```

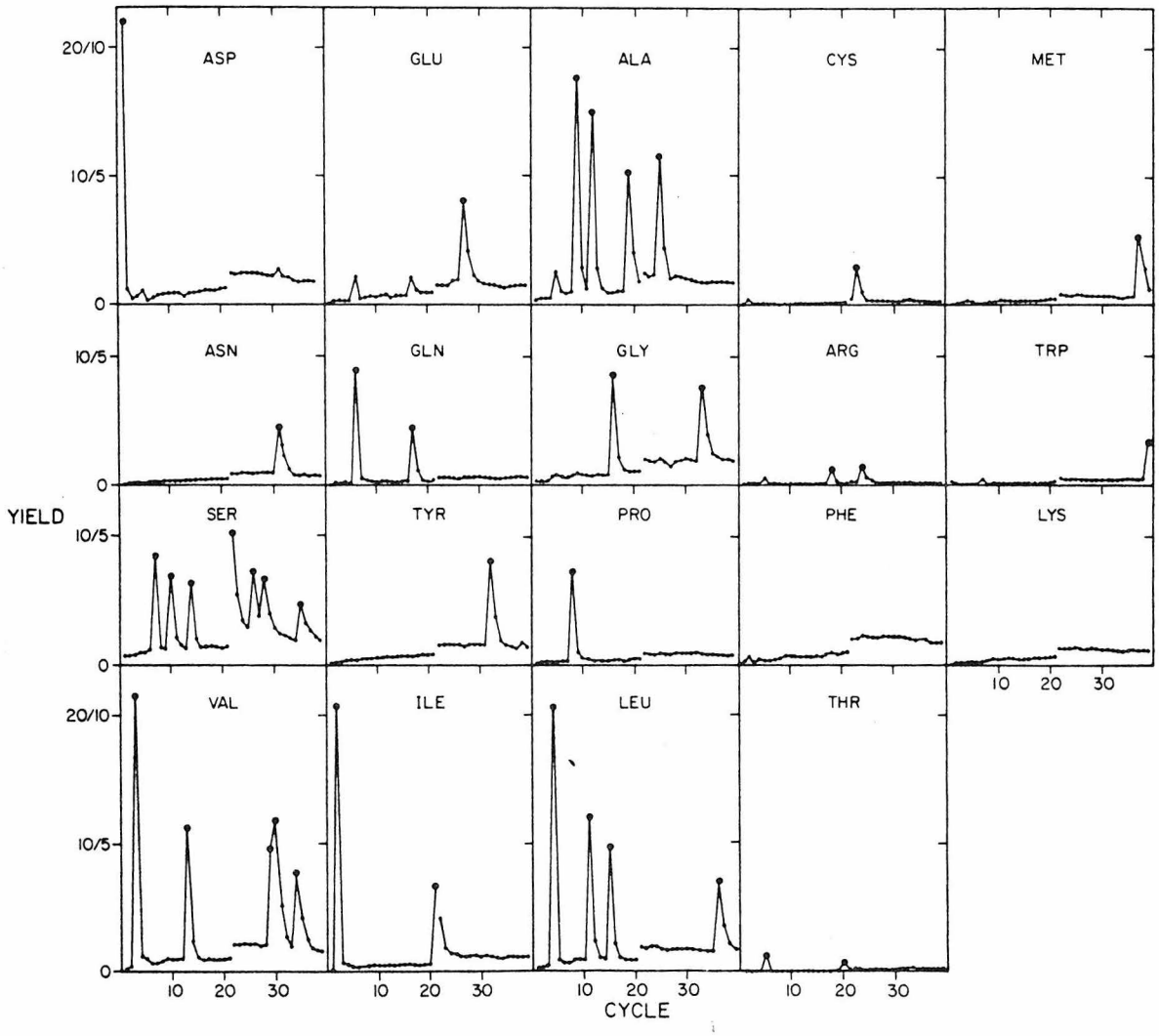


Figure 4

FIGURE 5: Peak heights of amino acid phenylthiohydantoins from PC7132 CN2. See legend to Figure 4. The scale is increased at position 25. The peptide was contaminated by CN3 and the sequence of this latter peptide, Glu-Glu-Asp-Asp was subtracted to determine the sequence of PC7132 CN2. Small amounts of the intact light chain can also be seen. Pro and Trp at 2, 7, 10, 11 were identified on thin layer chromatography as described by Summers et al. (1973). The Pro at 26 was determined from the composition of a peptide extending from 54-65.

10

Asn-Trp-Phe-Gln-Gln-Lys-Pro-Gly-Gln-Pro-Pro-Lys-Leu-Leu-Ile-Tyr-Ala

20 30

Ala-Ser-Asn-Gln-Gly-Ser-Gly-Val-Pro-Ala-Arg-Phe-Ser-Gly-Ser-Gly-Ser-Gly-Thr

Asp-Phe-Ser-?-Asn-Ile

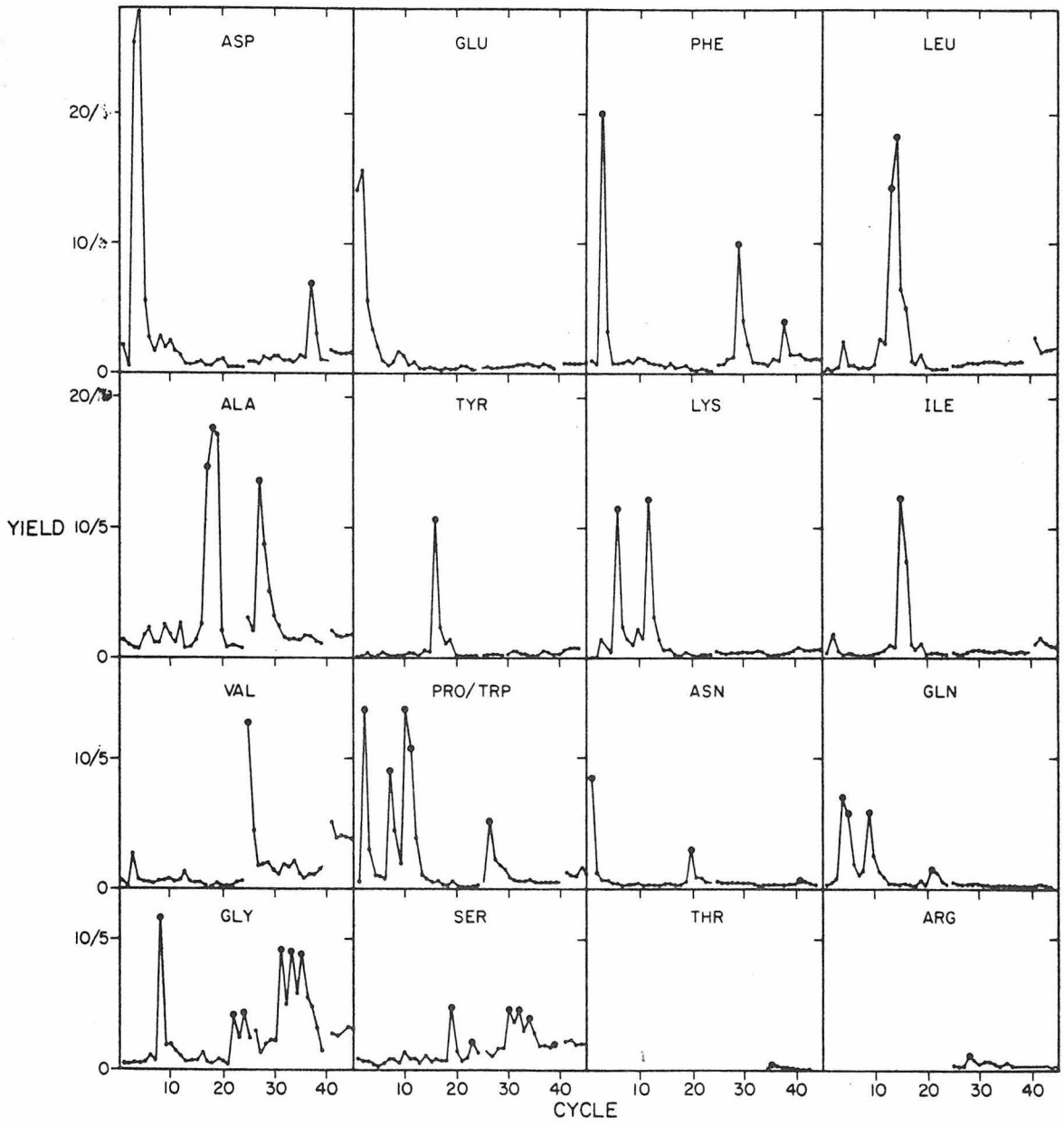


Figure 5

FIGURE 6: Peak heights of amino acid phenylthiohydantoins from PC2413 ST. See legend to Figure 4. The scale is doubled at position 24. Two sequences are present in roughly equal molar amounts. One, the constant region sequence extending from residues 113-159, is indicated by x's. The V region sequence extending from residues 66-112 is indicated by solid circles. At residue 10, both sequences have phenylalanine.

C region peptide: Ala-Asp-Ala-Ala-Pro-Thr-Val-Ser-Ile-Phe-Pro-Pro-Ser-Ser-Glu-Gln
Leu-Thr-Gly-Gly-Ser-Ala-Ser-Val-Val-Cys-Phe-Leu-Asn-Asn-Phe-Tyr

V region peptide: Phe-Ser-Gly-Ser-Gly-Ser-Gly-Thr-Asp-Phe-Ser-Leu-Ile-Ile-His-Pro
Met-Glu-Glu-Asp-Asp-Ser-Ala-Met-Tyr-Phe-Cys-His-Gln

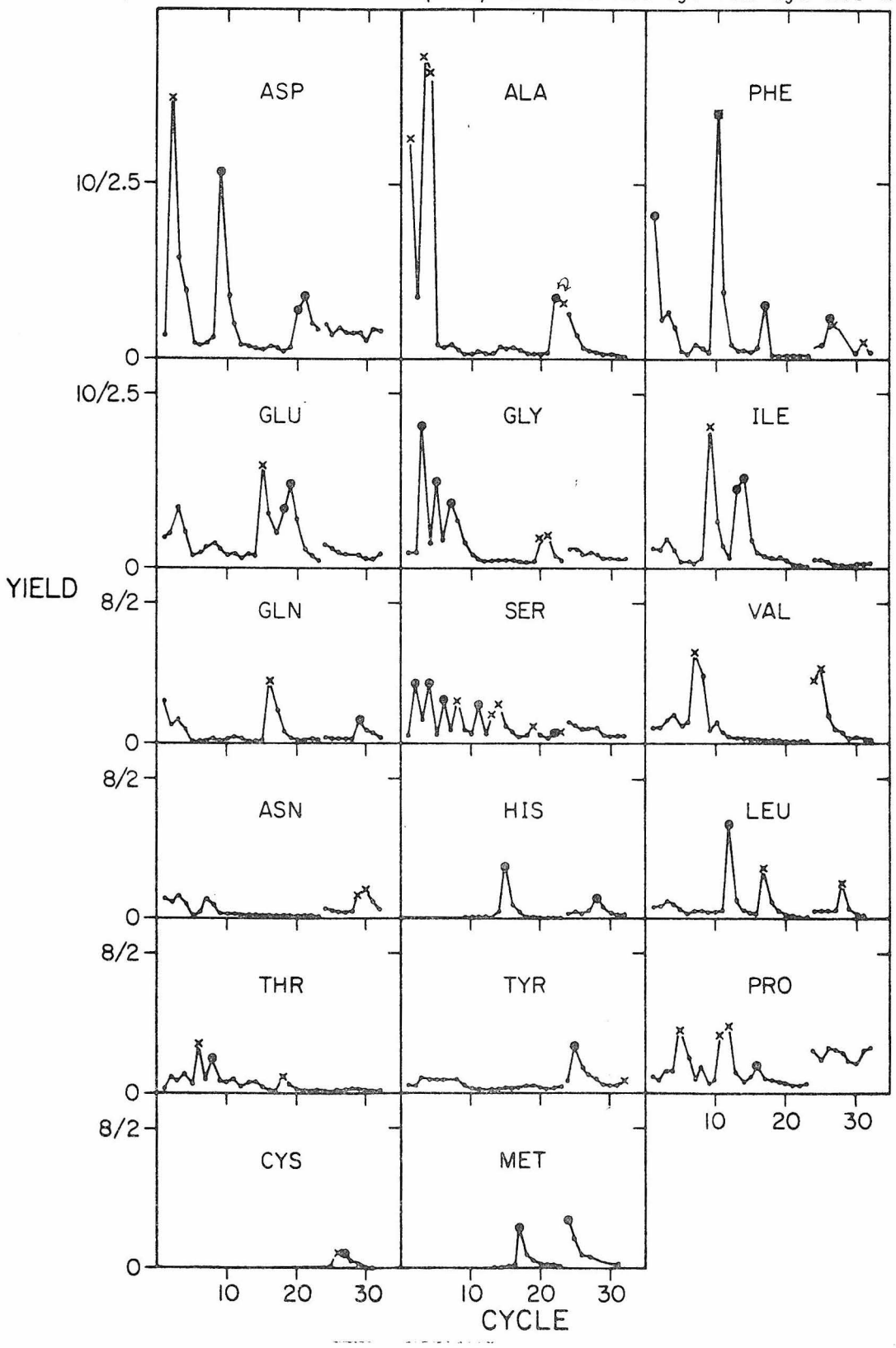
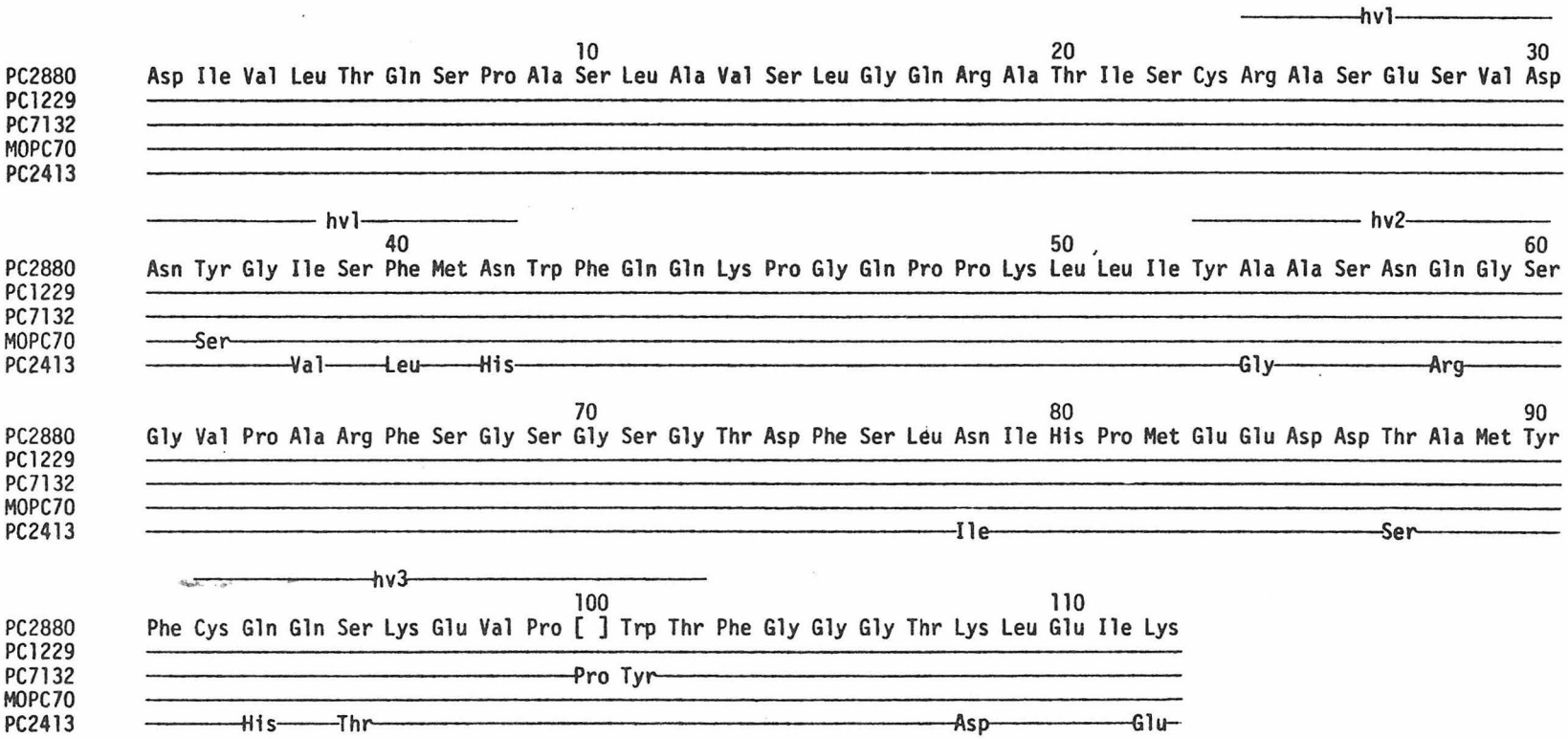


Figure 6

FIGURE 7: Amino acid sequence of five V_K 21 regions. MOPC70, a BALB/c myeloma protein is from Gray et al., 1967. Lines indicate identity to PC2880. Hypervariable regions are as defined by Kabat et al. (1976) and are indicated as arrows above corresponding sequences.


hv1
10
20
30
PC2880 Asp Ile Val Leu Thr Gln Ser Pro Ala Ser Leu Ala Val Ser Leu Gly Gln Arg Ala Thr Ile Ser Cys Arg Ala Ser Glu Ser Val Asp
PC1229
PC7132
MOPC70
PC2413
hv1
hv2
40
50
60
PC2880 Asn Tyr Gly Ile Ser Phe Met Asn Trp Phe Gln Gln Lys Pro Gly Gln Pro Pro Lys Leu Leu Ile Tyr Ala Ala Ser Asn Gln Gly Ser
PC1229
PC7132
MOPC70 Ser
PC2413 Val Leu His Gly Arg
70
80
90
PC2880 Gly Val Pro Ala Arg Phe Ser Gly Ser Gly Ser Gly Thr Asp Phe Ser Leu Asn Ile His Pro Met Glu Glu Asp Asp Thr Ala Met Tyr
PC1229
PC7132
MOPC70
PC2413 Ile Ser
hv3
100
110
PC2880 Phe Cys Gln Gln Ser Lys Glu Val Pro [ ] Trp Thr Phe Gly Gly Gly Thr Lys Leu Glu Ile Lys
PC1229
PC7132 Pro Tyr
MOPC70
PC2413 His Thr Asp Glu

Chapter 6

This paper is to be submitted to Nature

REARRANGEMENT OF GENETIC INFORMATION AND THE ORIGINS OF IMMUNOGLOBULIN DIVERSITY

M. Weigert

Institute for Cancer Research

Philadelphia, Pennsylvania 19111

E. Loh*, L. Hood, J. Schilling

Division of Biology

California Institute of Technology

Pasadena, California 91125

*Present address: Biochemistry Department, Stanford Medical School, Stanford, CA 94305

Amino acid sequences of 22 closely related κ chains are compared. Subsets sharing multiple substitutions can be defined. The distribution of variability shows that the mouse κ immunoglobulin chain may be encoded by three noncontiguous nucleotide segments, V, S and C, which are rearranged during differentiation.

Antibody polypeptides are divided into two regions, an N-terminal variable (V) region and a C-terminal constant (C) region. The infrequent amino acid sequence variation within the constant region is encoded by alleles of a single structural gene[1] or by a small number of closely linked duplicated genes.[2,3] The diversity in the highly variable V regions may arise from allelic polymorphisms, from multiple genes in the germ line, termed isotypes, or from somatic mutational mechanisms that amplify genetic information during development. Variation due to allelic polymorphisms are eliminated in the extensive analyses of homogeneous immunoglobulins and antibody genes from inbred strains of mice. With allelic variation not present in inbred mice, it is clear that each individual carries multiple variable region genes for κ light chains[4] and heavy chains.[5] The association of these multiple variable regions with a single constant region to constitute a single polypeptide chain made probable that separate "genes" coded for V and C regions and that a rearrangement of genetic material occurred during development.[6,7] Much evidence has been found for this process[7] including the recent direct demonstration of rearrangement or change in context of genetic information during differentiation at the nucleic acid level.[8]

In this report, we analyze the variable regions that form a closely related set of 22 mouse κ chains. These data suggest that κ chains are encoded by three noncontiguous nucleotide sequences, V (residues 1-99),*

---

*A four-residue insertion occurs at residues 28-31 in these κ chains as compared to others in the literature.[9] Since other investigators number the insertion as 27A, 27B, 27C, 27D, our contiguous numbering is four residues greater after residue 27, and 5 greater after the insertion in PC7132 at 100.

S or switch region (residues 100-112), and C (residues 113-219), which are rearranged to eventually code for a single polypeptide chain. Furthermore, these 22 variable region sequences comprise a large set of somatic cell sequences which have been produced by the antibody diversity generating mechanisms. An attempt to reconstruct the genetic history of these molecules provides additional insights into these mechanisms. Ultimately, a comparison of this somatic cell information (differentiated) with the corresponding germ line (undifferentiated) will give a clear picture as to these processes of genetic rearrangement and modification, particularly as to what diversity is present in the germ line and what is somatically derived. Before analyzing these data, let us discuss several general features of antibody diversity.

Patterns of variability

The variable regions of the three immunoglobulin families, λ, κ and H fall into subgroups that are defined by linked amino acid residues and shared sequence deletions or insertions.[10,11] This pattern of amino acid diversity in V regions has permitted immunologists to begin to determine whether variation is germ line or somatic in origin. Each subgroup consists of a set of V regions from independently induced myeloma tumors that share extensive amino acid sequence homology and yet differ as a group from other subgroups. Because the V regions of different subgroups can differ by up to 50% of their amino acid sequence, each subgroup appears to be encoded by a separate germ line gene so as to avoid extensive identical or parallel mutations.

Hence the minimum number of germ line genes can be estimated by counting V region subgroups. Nucleic acid hybridization data support this general approach in that each subgroup appears to be encoded by one to a few genes[8,12] and different subgroups are encoded by distinct V genes.[12]

A second sequence pattern has been described that has implications for the distinction between germ line genes and somatic mutation. When many variable regions are compared, certain polypeptide segments are hypervariable.[13,19] These hypervariable segments fold to form the walls of the antigen binding crevice[14] whereas the remainder of the V region constitutes a scaffold or framework. Three such hypervariable regions have been defined for light chains[9] (refer to Figure 4). The hypervariable regions have been given special genetic interpretation by some investigators.[15,16] For example, one model of somatically derived diversity, the antigen driven model, proposes that somatic mutations in the hypervariable region will create variants that bind antigen molecules more effectively and thus the mutant cells will be selected and differentially expanded.

Mouse λ chain variability

Observations on mouse λ chains and their genes have convinced many immunologists that somatic mutation plays an important role in generating diversity. Amino acid sequence analyses have established that 12 of 18 λ chain V regions are identical and that the remaining six differ by one to three amino acid substitutions, all of which fall

in the hypervariable regions.[17] Thus these V regions of mouse λ chains form a single subgroup. It has been proposed that the 12 identical V_λ regions are encoded by a single germ line gene and that the variants arise through antigen driven mutations in the hypervariable regions.[17,18] The interpretation that only one or a few genes code for the λ subgroup is supported by nucleic acid hybridization data.[8,12]

The V_κ21 subgroup as a model of κ chain diversity

The pattern of variability in the mouse κ family is quite different from that of the λ family. When the myeloma sequences from the inbred strains of mouse, BALB/c and NZB, are pooled, approximately 60 different sequences have been found for the N-terminal 23 residues.[4,19] Figure 1, a relatedness tree for the N-terminal 23 residues of 55 κ myeloma sequences, shows the large diversity of the mouse κ family. Using the N-terminal data, minimum statistical estimates of the number of subgroups and hence the number of germ line genes have been made. With the understanding that numerous assumptions need to be made in arriving at these statistical estimates, a minimum number of 100 germ line genes has been determine for the mouse κ family.[17]

The precise definition of a subgroup is difficult to make because often a one-to-one correlation of subgroup to germ line gene is assumed in defining a subgroup. However, we can not be certain about the nature of somatic mutation mechanisms. In this paper we define a subgroup to be a closely related set of proteins which are encoded by at least one germ line gene. When enough data exist to

say that more than one germ line gene codes for a set of sequences, then we subdivide it into two subgroups. One recent attempt has been made to divide all BALB/c κ chains into subgroups+.[20] Each subgroup thus defined has been specifically named and we will follow this nomenclature.

To more carefully analyze what constitutes the protein products of a single gene, we screened by automatic sequence analysis a large number of BALB/c[4] and NZB[19] myeloma light chains. One closely related set was found relatively frequently and this set shared virtually identical N-terminal sequences.[19] This set, designated $V_{\kappa}21$, is circled in Figure 1. Four $V_{\kappa}21$ examples from BALB/c have already been published.[21,22] Based on their differences, these four have been subdivided into three separate subgroups or subsets $V_{\kappa}21A$ (MOPC 70), $V_{\kappa}21B$ (MOPC 63) and $V_{\kappa}21C$ (MOPC 21 and TEPC 124).[20]

Using PC3741 and PC2880, two $V_{\kappa}21$ chains found in the initial NZB screening, antisera were made which could specifically identify $V_{\kappa}21$ light chains and which could subdivide $V_{\kappa}21$ into anti-PC2880 positive and anti-PC3741 positive groups (Julius, M., Potter, M., and Weigert, M., in preparation). Of the approximately 700 NZB myelomas screened, 5.7% fell into the first group and 3.1% into the second. In addition, the normal light chains from the sera of several inbred strains of mice cross react to the level of 7.2% and 3.2%, respectively. Thus

---

+The author suggests the term isotype replace that of subgroup. We do not see any advantage in this new term and have not used it.

the $V_\kappa 21$ closely related set constitutes about 10% of normal and myeloma light chains. This set affords an ideal model system to study the nature of diversity within the κ family of mouse because many members of the set are available due to the antisera screening and because the appearance of these molecules in normal sera suggests this set is biologically relevant.

Subdividing $V_\kappa 21$ into six subsets

In Figure 2, 22 NZB $V_\kappa 21$ V regions are compared to four BALB/c V_κ taken from the literature. A summary of the methodology employed to sequence these V regions is given in the legend to the figure and the data is presented in detail elsewhere[22,23] (Schilling, J., Loh, E., Weigert, M., Hood, L., in preparation). Twenty-two of these V_κ regions fall into six subsets which are defined by two or more V_κ regions sharing linked groups of subset-associated residues which are boxed if they are present in two or fewer subsets. Four of these V_κ regions do not fall into one of the six subsets, thus raising the possibility that they represent additional subsets for which two or more members have not yet been identified. We have named only those subsets which have multiple linked substitutions in two or more V_κ regions, and these names are as given in Figure 2.‡ Note that subset-associated

---

‡We have not designated the sequences without partners as separate subsets because our assumption is that any somatic diversity mechanism probably does not create parallel mutations, but there is no obvious limit as to the number of differences it would generate. Thus the odd sequences may be the most diverse products of a somatic mechanism, but if they occur twice, then they identify at least one separate germ line gene.

residues often occur in hypervariable regions. This illustrates that germ line diversity is an important element in generating a vast repertoire of antigen-binding sites.

A nucleotide difference matrix of the various subsets is given in Table 1 comparing the predominant sequences of each subset. Table 1 shows clearly the V_κ21D and V_κ21F are closely related as are V_κ21B and V_κ21C, and, accordingly, are perhaps of more recent origin by gene duplication and divergence. A visual display of all of the sequence information is shown in Figure 2 as a relatedness tree which attempts to derive a set of contemporary sequences from a single ancestral sequence using the minimal number of nucleotide substitutions.[24] Both Table 1 and Figure 2 analyze only the V region sequences through residue 99, for reasons that will become obvious later.

We would like to ask how many genes code for the V_κ21 sequences or conversely, which subset or set of sequences constitutes the products of a single germ line gene. Saturation hybridization studies using homogeneous DNA probes derived from several V_κ21 mRNAs suggest that five to six germ line genes encode the V_κ21 regions.[25] This study assumes that all of the V_κ21 genes cross hybridize, an assumption tested for only several members. Thus these results are in correspondence with the interpretation that each of the six subsets we have defined is derived from separate germ line gene. If new sequences are found which pair up with the unmatched sequences shown in Figure 2, then one of two possibilities must be considered: there is an error in the saturation hybridization approach in counting V_κ21 genes and more than six genes exist; alternatively, a somatic mechanism can permit a single germ line

V gene to generate two or more subsets of $V_\kappa 21$ regions. The latter possibility requires that in different lymphocyte lines identical mutations must occur repeatedly at the subset-associated positions. This requirement for multiple identical substitutions is unattractive to many immunologists and geneticists.

Intrasubset variation

If each subset is encoded by a single germ line V gene, the substitutions within each subset should reflect the mechanism of somatic mutation. Thus the nature and patterns of intrasubset variation should reflect the nature of the somatic mechanism.

The 23 intrasubset substitutions are shown in Figure 2 by circles. A summary of their properties is shown in Table 2. One should note that the assignment of intrasubset substitutions is somewhat arbitrary in that certain positions differ in all the members of a subset (e.g., position 27 in $V_\kappa 21C$). In addition, there is the possibility that one variant differing by three or so residues from the prototype suequence will become a distinct subset when an identical sequence is determined. For example, PC2485 was a member of the $V_\kappa 21D$ subset with three intrasubgroup substitutions until the sequence of PC4039 was determined. These two identical sequences then became subset $V_\kappa 21F$ and three intrasubset substitutions were eliminated. The possibility must be considered that additional variants will fall into new subsets as more $V_\kappa 21$ regions are sequenced.

Several features distinguish these V_κ substitutions from their V_λ counterparts discussed earlier (Table 2). i) Four of 23 substitutions

are found outside the hypervariable regions. Thus a putative somatic mechanism must operate outside as well as within hypervariable segments. ii) Variant sequences are found more frequently than identical or prototype sequences. Indeed, of the 22 sequences that can be assigned to one of the six subsets, 11 are sequences with variant residues. iii) Six positions exhibit two to four amino acid substitutions (i.e., 27, 29, 31, 32, 91 and 98) and, indeed, more than half the intrasubset substitutions are found at these positions. Of the 15 variant sequences, one has four substitutions, three have three substitutions, three have two substitutions and only four have one substitution. Accordingly, there appears to be a tendency to accumulate multiple mutations in the variant sequences at certain positions. As discussed above, perhaps certain of these variant sequences represent new subsets. It is also remarkable that all 23 substitutions are single base changes.

Distribution of $V_\kappa 21$ variability

There is a fourth set of hypervariability in the $V_\kappa 21$ set of light chains. One measure of the variability of this set of 26 V_κ regions is given by a Wu-Kabat plot which measures the variability at each position (Figure 4). The variable region can be divided into seven segments: four framework (FR) regions denoted FR 1, 2, 3 and 4 encompass residues 1-23, 39-52, 61-92 and 102-112, respectively; three hypervariable (hv) regions or complementarity determining residues denoted hv 1, 2 and 3 include residues 24-38, 53-60 and 93-102, respectively.[9] The pattern of variability for the $V_\kappa 21$ subgroup as

compared with that of all mouse V_κ regions shows the diversity in FR 1 and 2 to be much lower than for all κ chains. This diversity in the pool of V_κ regions presumably reflects subgroup specific differences that is lacking in the proteins of the V_κ21 subgroup. The diversity in the first half of FR3 is low in both populations and the diversity in hv segments 1, 2 and 3 is correspondingly high in both. The last half of FR3 in the V 21 regions shows a diversity that is comparable in the pool of all myeloma light chains. This is surprising because of the restriction in the diversity of the other V_κ21 FR regions 1 and 2.

Positions 76-91 appear to consitute a fourth hypervariable segment in that they exhibit 12 subset-associated residues and three intrasubset substitutions (Fig. 2). This region of hypervariability is interesting in several regards. i) The X-ray analysis of immunoglobulin molecules reveals that it lies well outside the antigen-binding site.[14] This raises a question as to how intrasubgroup variability generated by a somatic mechanism could be selected if not on the basis of generating a better antigen-binding site for the relevant antigen. Alternatively, perhaps this intrasubgroup variability reflects multiple germ line genes. ii) A fourth hypervariable segment has been noted in heavy chains in an analogous position.[25,26] Once again this region is quite distant from the antigen-binding site. iii) The antiserum to PC2880 recognizes κ chains of the V_κ21A, V_κ21D, V_κ21F and V_κ21E subsets. Conversely, the antiserum to PC3741 recognizes κ chains of the V_κ21B and V_κ21C subsets. From the available data a correlation between sequences and serology can be made. The only

residues uniquely shared by V_κ21 subsets B and C are at positions 72 (arginine), 79 (threonine), and 84 (alanine). At these positions, V_κ21 subsets A, D, E and F have glycine, asparagine and glutamic acid, respectively. The corresponding sites in the McPC603 light chain as revealed by X-ray analysis of the M603 Fab fragment occur in a hydrophilic loop region that lies well outside the combining sites and does not interact with V_H residues.[14] A comparable loop region also is found in the V_H region. In addition, side chains of the amino acids at these sites in the McPC603 light chain face to the outside of the molecule. Assuming that the V_κ21 L chains in association with their respective H chains take on the same conformation as the McPC603 light chain, it is reasonable to suppose that residues at 72, 79 and 84 could comprise these V_κ21-specific antigenic determinants. It is interesting to note that the group A serological markers of the rabbit heavy chain also are found in a homologous position.[40] Whether this region has a functional significance is unknown. These may, however, be segments of the V_L and V_H regions that can vary without affecting the antigen-binding specificity of the V_L and V_H domains. Accordingly, perhaps diversity in this region reflects the existence of multiple germ line genes.

Genetic rearrangement during differentiation

The V 21 polypeptides appear to be coded by three noncontiguous segments V, S and C of DNA that are rearranged during differentiation. The V segment codes for residues 1-99, the S or switch segment codes

for residues 100-112, and the C segment codes for residues 113-219. The hypothesis of an independent S segment, which was formerly considered a part of the V region, is based on our observations of the $V_\kappa 21$ subgroup at the protein level and observations of others at the nucleic acid level.

The $V_\kappa 21$ sequences have several features that distinguish the S region from the rest of the molecule. i) Subgroup-associated residues end at position 98. Indeed, residue 101 is one of the most variable residues seen (6 alternatives). It marks the beginning of the S region for all but one of the sequences. ii) There are nine distinct sequences for the S region (Fig. 5). Thus the region is quite variable, certainly not a part of the C region. iii) A single S sequence can be associated with the V_κ regions from two or more subsets (e.g., in Fig. 5, sequences 1 of the S region is found associated with three subsets and one "other" while sequence 5 is found associated with four subsets and one "other"). iv) Alternatively, 3 subset pairs identical over the first 99 residues have different S regions (PC6684 and PC7175, PC7013 and PC7183, and PC4050 and PC9245). Thus identical V regions can be associated with different S regions. v) Furthermore, one subset can be associated with as many as three different S regions (V 21D). vi) From the literature[9,28] seven other non-$V_\kappa 21$ V regions (MOPC41, MOPC173, MOPC21, J606, W3082, MOPC511 and MOPC167) have four different S regions, each of which has been seen in the $V_\kappa 21$ sequences. Thus identical S regions are associated with V regions from widely different subgroups. In these ways the V and S regions of the $V_\kappa 21$ subgroup associate independently

with one another, just as do the V_H and C_H genes. Let us consider the genetic implications of this independent behavior.

Three possible genetic explanations exist for these observations on the nature of the S region. i) Each V gene codes for residues 1-112 with frequent mutations occurring in that position which we have designated the S region. This explanation appears unlikely because many parallel mutations must occur to explain the occurrence of the same S region on many different V regions. ii) Each $V_\kappa 21$ gene may encode residues 1-112 with a frequent propensity for recombination to occur in the S region. This would explain different V regions associated with the same S region. However, to explain different S regions associated with the same V region, one must hypothesize a duplication of the entire V gene (residues 1-112) for each S region. This postulate would further increase the number of $V_\kappa 21$ germ line genes required above the number that is acceptable by the hybridization diata as discussed earlier. iii) Alternatively, the immunogloublin light chain gene may be encoded by three noncontiguous gene segments, V, S and C as illustrated in Figure 6. These three gene segments are rearranged at the DNA level during the differentiation of the antibody-producing cell[29] so that the three segments are juxtaposed (Fig. 6). In the differentiated cell these three gene segments appear to be separated by untranslated intervening DNA,[29,30] that undergoes processing at the mRNA level.[30] Support for this latter statement comes from the following observations. 1) A mouse V_λ gene isolated from embryonic DNA ends its coding portion at position 98.[32] This would correspond to our V segment. 2) Several

mouse V genes isolated from myeloma DNA terminate at approximately position 98.[31] 3) the C region appears to be encoded by one or a few C genes.[33] Certainly there are not enough C_κ genes to account for all of the different S region sequences. These points suggest that the V, S and C segments are probably separated by intervening DNA sequences in the differentiated DNA. These intervening sequences have been found in many other genes.[34,35,36,37] While there are noncoding stretches of DNA between V and S genes and between S and C genes, the remarkable point which has not been demonstrated in other systems is that rearrangement of DNA occurs during differentiation, probably to select out one V and one S to express with the C. That is the most likely possibility. One cannot eliminate the less likely possibility that the multiple V and multiple S are transcribed on the same RNA precursor and the processing enzymes provide the specificity to express only one protein.

Based on a computer analysis of all available V region sequences, Kabat and his coworkers have recently suggested the mini-gene model which proposes that each of the three hypervariable segments and the four framework regions are encoded by mini-genes which are assembled during differentiation.[16] Our data argue rather compellingly against this model as stated above. All of the hypervariable and framework segments of a particular $V_\kappa 21$ subset are always uniquely associated only in V regions of that subset. That is, the hypervariable and framework segments of a particular subset are always linked to one another. The mini-gene model would predict that the hypervariable segments of

one $V_\kappa 21$ subset should be associated with the framework regions of other subsets. The extensive data we have gathered on the $V_\kappa 21$ subgroup make this constraint unequivocal and require that the mini-gene model make special and unattractive postulates to explain the linked association of certain hypervariable segments with certain framework regions. However, we do agree with Kabat and his coworkers that FR4 (approximately our S region) is independent in its association with the rest of the V region.

Several interesting points may be raised about the S region. First, it has been defined based on amino acid sequence data which are incapable of determining precisely the initiation and termination of the S region. Furthermore, the S region coding sequence may be flanked by noncoding DNA in which the DNA joining actually occurs. Precise definition of this region will have to await the DNA sequence analysis of κ light chain genes from differentiated and nondifferentiated DNA. Second, one may ask how diverse the S region is. There are seven additional V_κ sequences complete in the S region which are associated with four S sequences, all of which were seen in the $V_\kappa 21$ subgroup. Thus in examining 31 complete V_κ regions, nine different S regions were found. It is obvious that a great deal more data must be accumulated in the S region before the extent of its diversity can be determined. Third, will the S region be found in the λ and H chain families? The sequence data at the protein level are insufficient to convincingly answer this question. As noted earlier, the V_λ gene isolated from embryonic DNA terminates at position 98.[32] An intervening sequence

between V and S regions then could be a sign that the mouse λ family will have DNA rearrangements between V, S and C. Sequence studies at the DNA level on V coding sequences isolated from differentiated and undifferentiated DNA should resolve this question.

The function of the S region is, of course, unknown. However, several interesting possibilities suggest themselves. i) The different S regions may be joined in a combinatorial fashion to each V region to provide another stage in the amplification of antibody diversity. For example, the combinatorial association of 100 V genes and 20 S genes could generate 2000 different light chains. Moreover, the joining of S and V regions occurs in the third hypervariable region, providing diversity directly in the antigen-binding site. This diversity would include sequence gaps (insertions or deletions) as well as nucleotide substitutions. The sequence gaps would result from S regions of differing length. Indeed, this may be the explanation for the insertion of proline at position 100 in PC7132. ii) Perhaps the joining process itself induces somatic mutations in the third hypervariable region if, for example, the joining process is error-prone. Thus joining a particular S region to a V region could generate many new sequences not found in the germ line. This assumes that the joining process occurs next to the S coding region. Both of these mechanisms suggest that the joining of S and V region provides one or more mechanisms for increasing antibody diversity from a limited number of germ line genes.

Acknowledgement

This work was supported by NIH grant AI-10781 and NSF grant PCM76-81542.

REFERENCES

[1]Terry, W. D., Hood, L. E., and Steinberg, A. G., Proc. Nat. Acad. Sci. USA 63, 71-75 (1969).

[2]Gibson, D., Levanson, M., and Smithies, O., Biochemistry 10, 3114 (1971).

[3]Ein, D., Proc. Nat. Acad. Sci. USA 60, 982-985 (1968).

[4]Hood, L., McKean, D., Fransworth, V., and Potter, M., Biochemistry 12, 741-749 (1973).

[5]Barstad, P., Farnsworth, V., Weigert, M., Cohn, M., and Hood, L., Proc. Nat. Acad. Sci. USA 71, 4096-4100 (1974).

[6]Dreyer, W. J., and Bennett, J. C., Proc. Nat. Acad. Sci. USA 54, 864-869 (1965).

[7]Hood, L. E., Fed. Proceedings 31, 177-187 (1972).

[8]Tonegawa, S., Hozumi, N., Mathyssens, G., and Schuller, R., Cold Spring Harbor Symp. Quant. Biol. 61, 877-889 (1977).

[9]Kabat, E. A., Wu, T. T., and Bilofsky, H., Variable Regions of Immunoglobulin Chains, Tabulations and Analyses of Amino Acid Sequences Bolt, Beranek and Newman, Inc., Cambridge, MA 1976).

[10]Hood, L., Gray, W., Sanders, E., and Dreyer, W., Cold Spring Harbor Symp. Quant. Biol. 32, 133-146 (1967).

[11]Milstein, C., Nature 216, 330 (1967).

[12]Leder, P., Honjo, T., Seidman, J., and Swan, D., Cold Spring Harbor Symp. Quant. Biol. 41, 855-862 (1977).

[13]Wu, T., and Kabat, E., J. Exp. Med. 132, 211-250 (1970).

[14]Amzel, L., Polyjak, R., Saul, F., Varga, J., and Richard, F., Proc. Nat. Acad. Sci. USA 71, 1427-1430 (1974).

[15]Capra, J. D., and Kindt, T. J., Immunogenetics 1, 417-427 (1975).

[16]Kabat, E. A., Wu, T. T., and Bilofsky, H., Proc. Nat. Acad. Sci. USA (in press).

[17]Weigert, M., and Riblet, R., Cold Spring Harbor Symp. Quant. Biol. 61, 837-846 (1977).

[18]Cohn, M., Blomberg, B., Geckeler, W., Raschke, W., Riblet, R., and Weigert, M., in The Immune System, Genes, Receptors, Signals (Edit. by Sercarz, E. E., Williamson, A. R., and Fox, C. F.), 89-118 (Academic Press, New York, 1974).

[19]Loh, E., Black, B., Riblet, R., Weigert, M., Hood, J. M., and Hood, L., Proc. Nat. Acad. Sci. USA (submitted.

[20]Potter, M., Adv. in Immunol. 25, 141-211 (1978).

[21]McKean, D., Potter, M., and Hood, L., Biochemistry 12, 760-771 (1973).

[22]McKean, D., Potter, M., and Hood, L., Biochemistry 12, 749-759 (1973).

[23]Loh, E., Schilling, J., Weigert, M., and Hood, L., Biochemistry (submitted).

[24]Fitch, W. M., and Margoliash, E., Science 155, 279-284 (1967).

[25]Perry, R., personal communication.

[26]Capra, J. D., and Kehoe, J. M., Proc. Nat. Acad. Sci. USA 71, 845-848 (1974).

[27]Capra, J. D., and Kehoe, J. M., Cont. Topics Mol. Immunol.

[28]Johnson, N., and Hood, L., personal communication.

[29]Hozumi, N., and Tonegawa, S., Proc. Nat. Acad. Sci. USA 73, 3628-3632 (1976).

[30]Gilmore-Herbert, M., and Wall, R., Proc. Nat. Acad. Sci. USA 75, 342-345 (1978).

[31]Leder, P., personal communication.

[32]Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O., and Gilbert, W., Proc. Nat. Acad. Sci. USA 75, 1485-1489 (1978).

[33]Leder, P., Honjo, T., Swan, D., Peckman, S., Nau, M., and Norman, B., in Molecular Approaches to Immunology, Miami Winter Symposia (Edit. by Smith, E. E., and Ribbons, D. W.) Vol. 9, 173-188 (Academic Press, New York, 1975).

[34]Goodman, A. M., Olson, M. V., and Hall, B. D., Proc. Nat. Acad. Sci. USA 74, 5453-5457 (1977).

[35]Jeffreys, A. J., and Flavell, R. A., Cell 12, 1097-1108 (1977).

[36]Tilghman, S. M., Tiemeier, D. C., Seidman, J. G., Peterlin, B. M., Sullivan, M., Maizel, J. V., and Leder, P., Proc. Nat. Acad. Sci. USA 75, 725-729 (1978).

[37]Tilghman, S. M., Curtis, P. J., Tiemier, D. C., Leder, P., and Weissman, C., Proc. Nat. Acad. Sci. USA 75, 1309-1313 (1978).

[38]Schilling, J., Loh, E., Weigert, M., Gatmaitan, L., and Hood, L. (in preparation).

[39]Hunkapiller, N., and Hood, L., Biochemistry (in press) (1978).

[40]Mage, R. G., Young-Cooper, G. O., Rejnek, J., Ansari, A. A., Alexander, C. B., Apella, E., Carta-Sorcini, M., Landucci-Tosi, S., and Tosi, R. M., Cold Spring Harbor Symp. Quant. Biol. 61, 677-686, 1977.

Table 1. A minimal nucleotide difference matrix of the substitutions separating various subsets within the $V_\kappa 21$ subgroup*

	$V_\kappa 21D$	$V_\kappa 21F$	$V_\kappa 21E$	$V_\kappa 21B$	$V_\kappa 21C$	$V_\kappa 21A$
$V_\kappa 21D$		3	18	19	19	23
$V_\kappa 21F$			21	22	22	25
$V_\kappa 21E$				18	15	19
$V_\kappa 21B$					7	21
$V_\kappa 21C$						20

*At positions where substitutions occur within a subset the most frequent residue is compared. If no residue is most frequent (e.g., position 27 in the $V_\kappa 21C$ subset), the prototype is assumed to be any one of the possible alternatives at that position.

Table 2. A comparison of the properties of the intrasubset (subgroup) substitutions of the mouse λ_I and V_K21 subgroups

Property		M70	λ_I
Single base substitution		23/23	8/9
Substitutions within hv regions		19/23	9/9
Variants and their number of			
substitutions:	1 substitution	4	4
	2 substitutions	3	1
	3 substitutions	3	1
	4 substitutions	1	0
Number of variant V regions		11/22	6/18
Number of positions with multiple substitutions		6	0
Number of identical substitutions		1	0

Figure 1. A relatedness tree for N-terminal 23 residues of 55 V_κ sequences of the mouse. These V_κ regions are from myeloma proteins from the inbred BALB/c and NZB strains of mice. Boxed V_κ regions indicate NZB V_κ regions. The $V_\kappa 21$ set is indicated by a dotted circle. The numbers on the lines indicate the number of nucleotide substitutions that separate two successive junction (node) points on the tree. Thus the extend to which individual chains differ from one another at the nucleotide level can be determined by summing the lengths of the lines between the two V_κ regions. (Adapted from reference 19.)

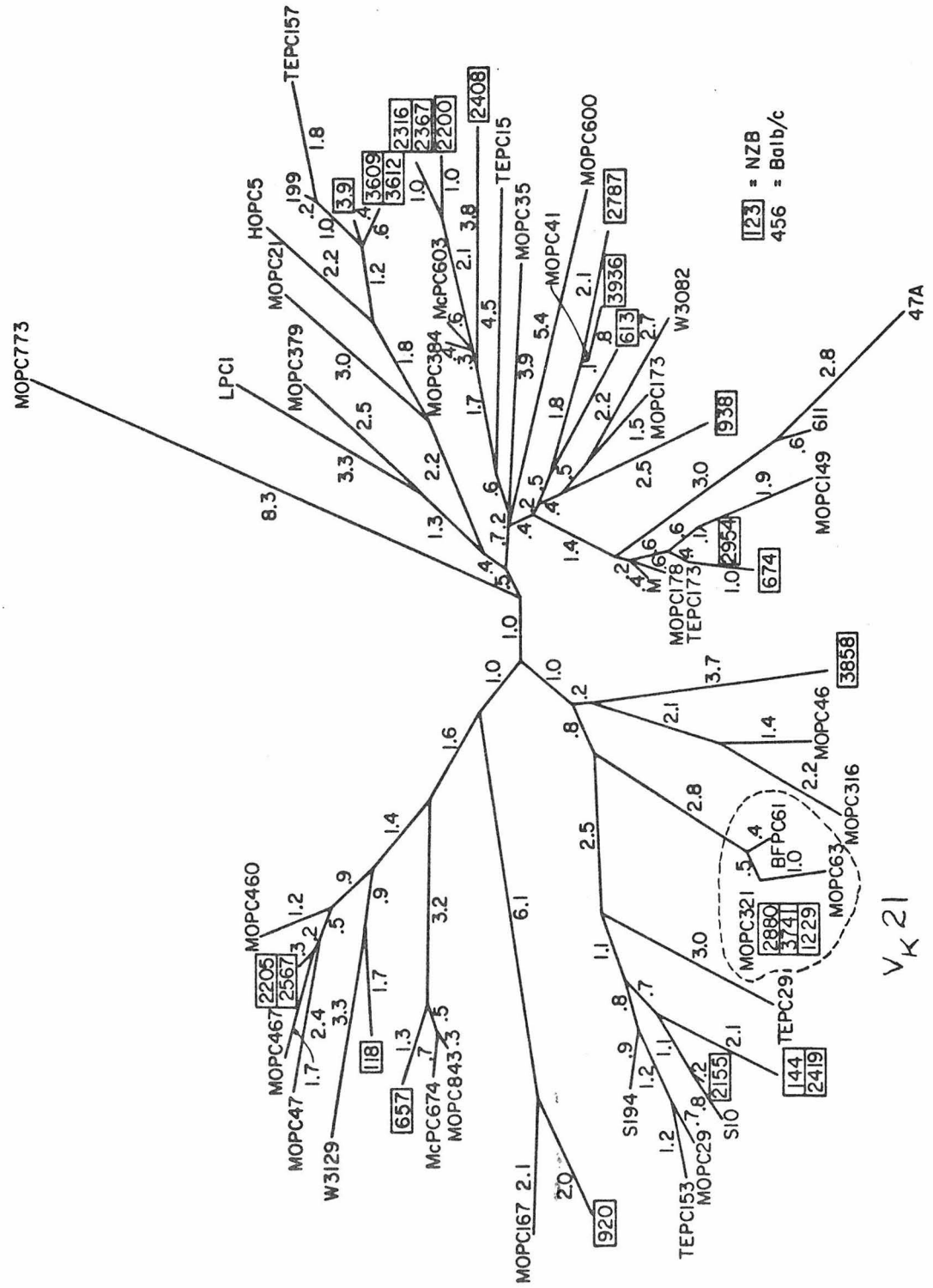

MOPC773
LPC1
MOPC379
MOPC21
HOPC5
TEPC157
MOPC384
McPC603
TEPC15
MOPC35
MOPC41
MOPC600
2787
3936
613
W3082
MOPCI73
938
MOPCI49
611
47A
2408
2316
2367
2200
3609
3612
3.9
199
MOPCI78
TEPCI73
2954
674
3858
MOPC46
MOPC316
MOPC63
BFPC61
MOPC321
2880
3741
1229
TEPC29
144
2419
2155
S10
S194
MOPC29
TEPCI53
MOPCI67
920
MOPC460
2205
2567
MOPC467
MOPC47
W3129
118
657
McPC674
MOPC843
123 = NZB
456 = Balb/c
Vκ21

Figure 2. The amino acid sequences of V_κ regions from the mouse $V_\kappa 21$ set. The sequences are grouped by subsets which are designated $V_\kappa 21A$ through $V_\kappa 21F$. The subset-associated residues unique to two or fewer subsets are boxed. The intrasubset substitutions are circled. The three hypervariable regions are indicated by arrows and hv1, hv2 and hv3. The S region extends from residue 100-112 and is boxed. The serological cross reactivity of individual proteins to antisera directed against two screening κ chains is given at the right. The three residues associated with the serological specificity to these two antisera are starred.

The following general strategy was employed for sequencing these V regions.[23,38] The intact light chain was analyzed on the automatic sequenator for 40 residues. The procedures for automatic sequence analysis are given in ref. 39. Methionine fragments were prepared; all V_κ regions but one have a methionine at position 37. Methionine fragments starting at position 38 were sequenced for 40 to 65 residues. Only the $V_\kappa 21A$ subset has additional methionine residues and these fragments were isolated and sequenced. Arginine fragments extending from positions 65 to 113 were prepared for certain κ chains and sequenced for 35 to 40 residues. Thermolysin and chymotryptic peptides were isolated from certain of these methionine or arginine fragments and compositions were determined. In some cases these peptides were sequenced in the presence of polybrene.[39]

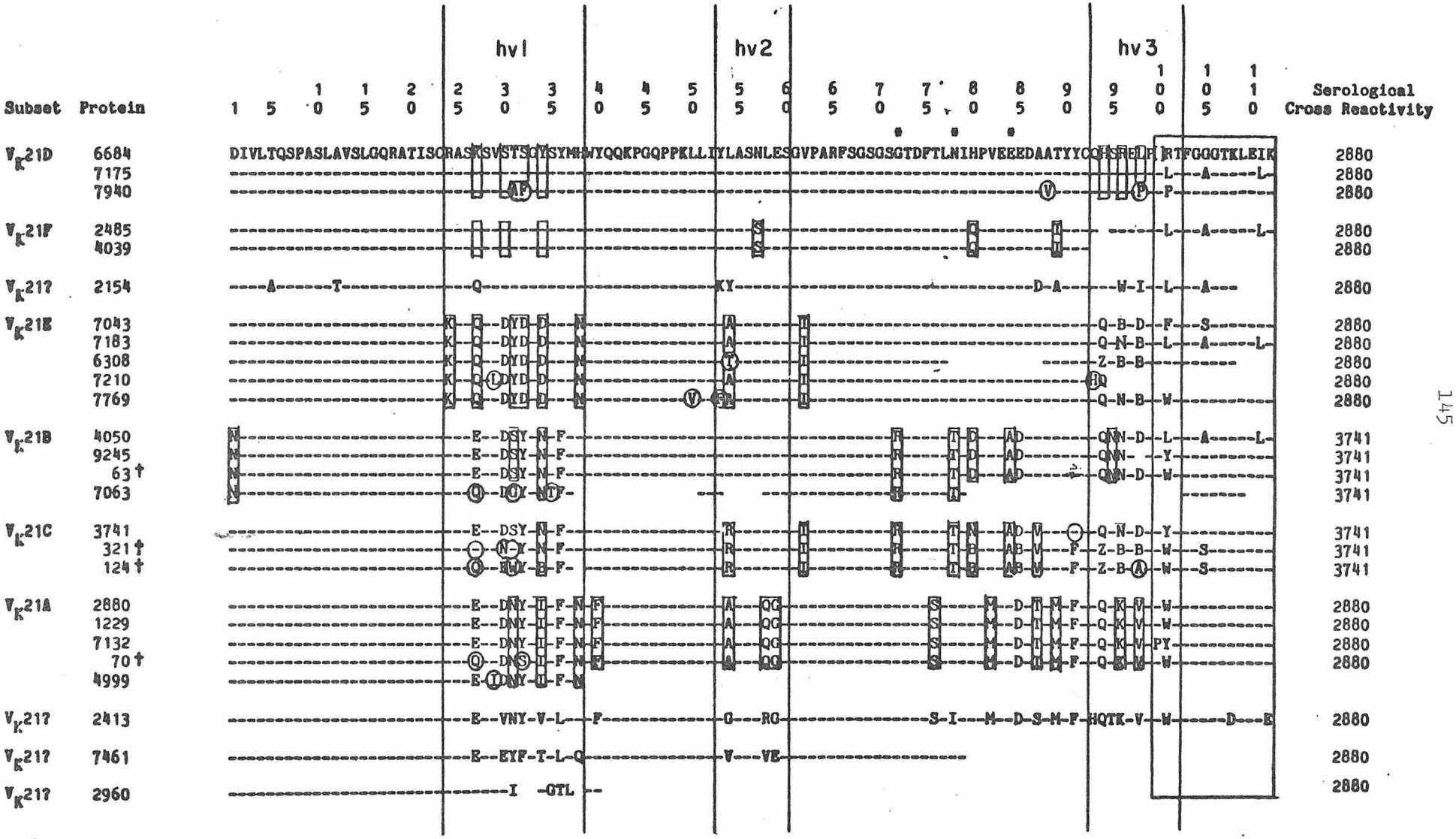

hv1
hv2
hv3
Subset
Protein
Serological Cross Reactivity
DIVLTQSPASLAVSLGQRATISCRASKSVSTSGYSYMHWYQQKPGQPPKLLIYLASNLESGVPARFSGSGSGTDFTLNIHPVEEEDAATYYCQHSRELPRTFGGGTKLEIK
V_K21D 6684 2880
7175 2880
7940 2880
V_K21F 2485 2880
4039 2880
V_K21? 2154 2880
V_K21E 7043 2880
7183 2880
6308 2880
7210 2880
7769 2880
V_K21B 4050 3741
9245 3741
63† 3741
7063 3741
V_K21C 3741 3741
321† 3741
124† 3741
V_K21A 2880 2880
1229 2880
7132 2880
70† 2880
4999
V_K21? 2413 2880
V_K21? 7461 2880
V_K21? 2960 2880

Fig. 2.

Figure 3. A relatedness tree of all $V_{\kappa}21$ sequences. Each subset is circled and named.

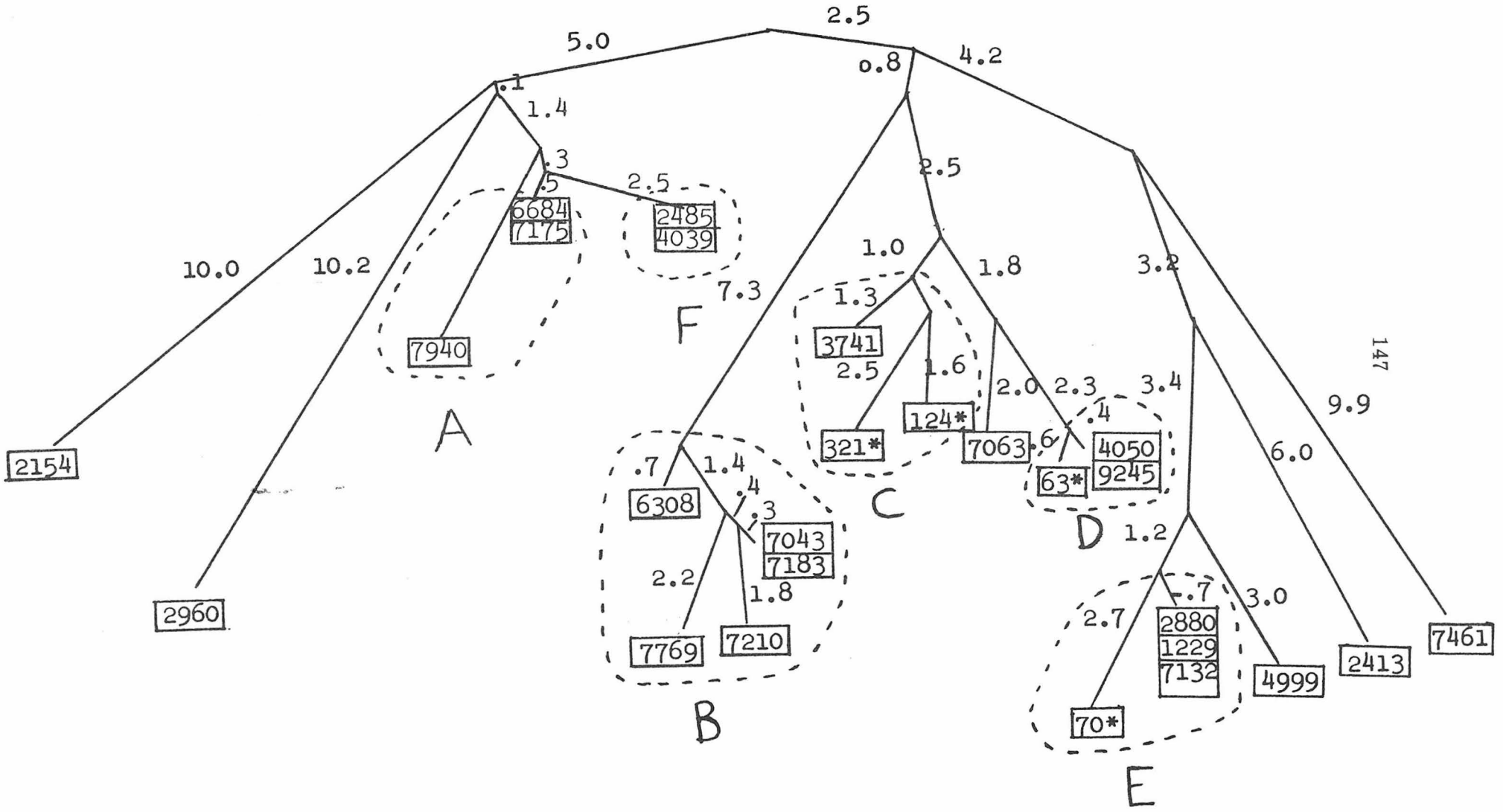

2.5
5.0
0.8
4.2
.1
1.4
.3
.5
2.5
2.5
10.0
10.2
7.3
1.0
1.8
3.2
1.3
2.5
1.6
2.0
2.3
3.4
.4
.6
.7
1.4
.4
.3
2.2
1.8
1.2
.7
2.7
3.0
6.0
9.9
6684
7175
2485
4039
7940
2154
2960
3741
321*
124*
7063
4050
9245
63*
6308
7043
7183
7769
7210
2880
1229
7132
4999
2413
7461
70*
A
F
C
D
B
E
147

Figure 4. Variability distribution for all V_{κ}21 κ chains. These data were calculated according to ref. 13.

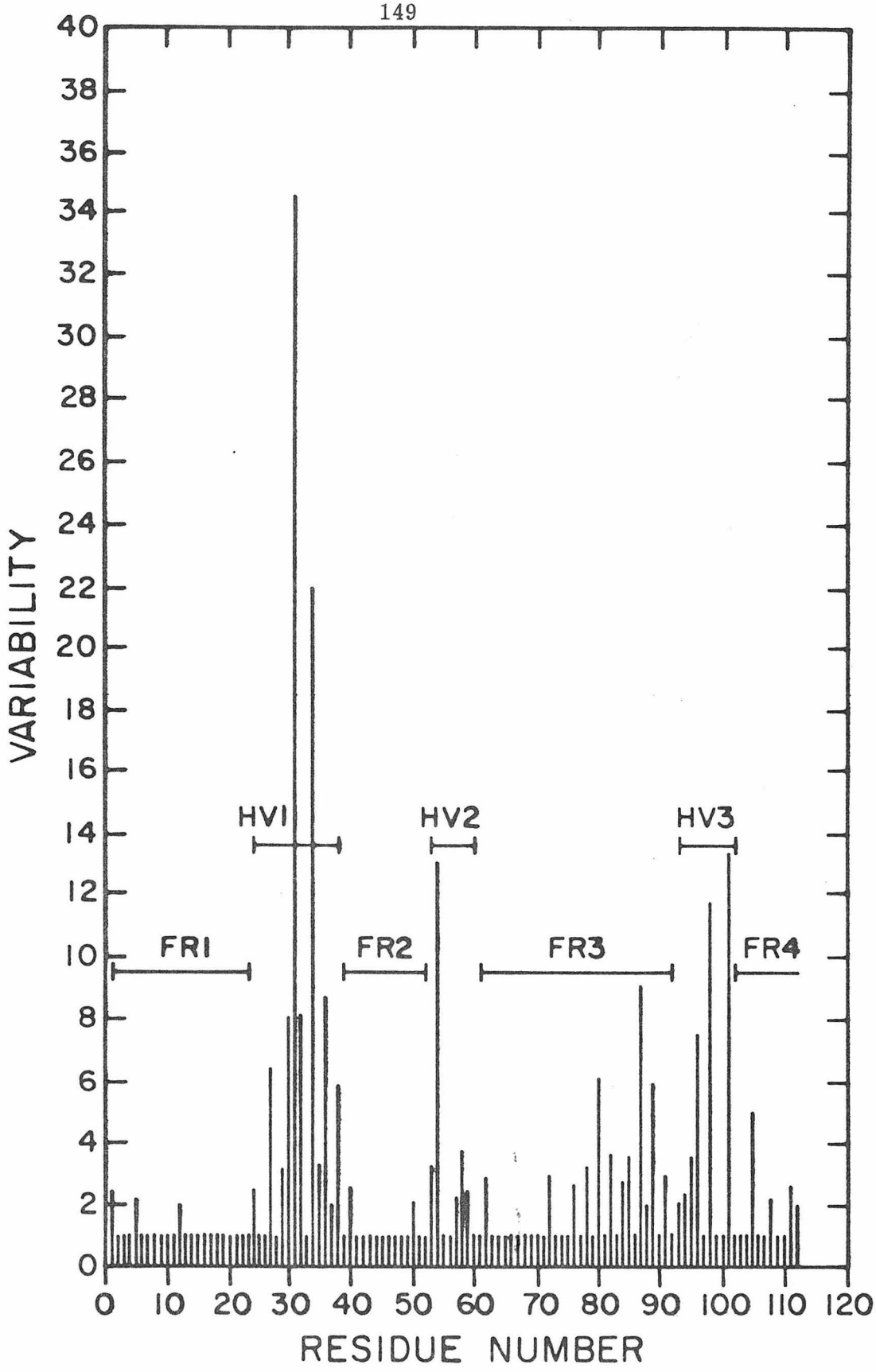

VARIABILITY
RESIDUE NUMBER
HV1
HV2
HV3
FR1
FR2
FR3
FR4
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
0
10
20
30
40
50
60
70
80
90
100
110
120

Figure 5. A diagram illustrating the independent association of the V region subsets (residues 1-99) and the S region (residues 100-112).

Figure 5

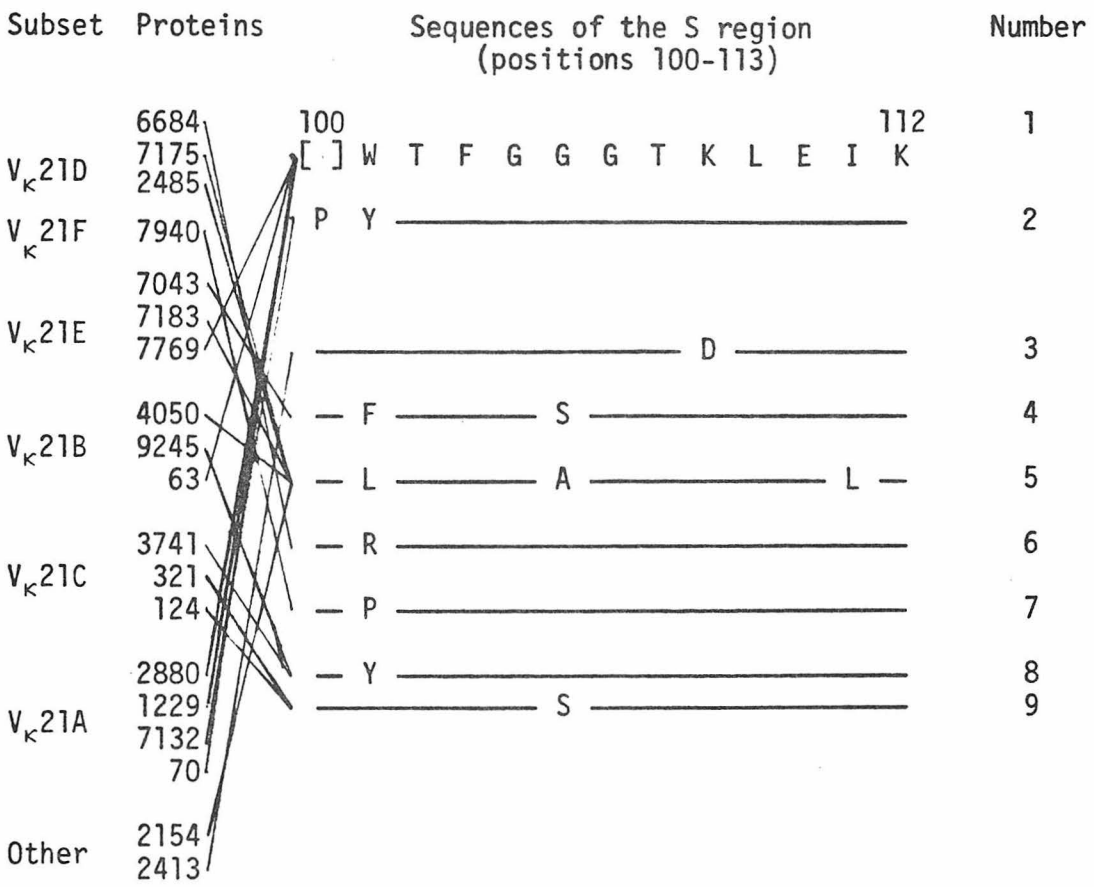
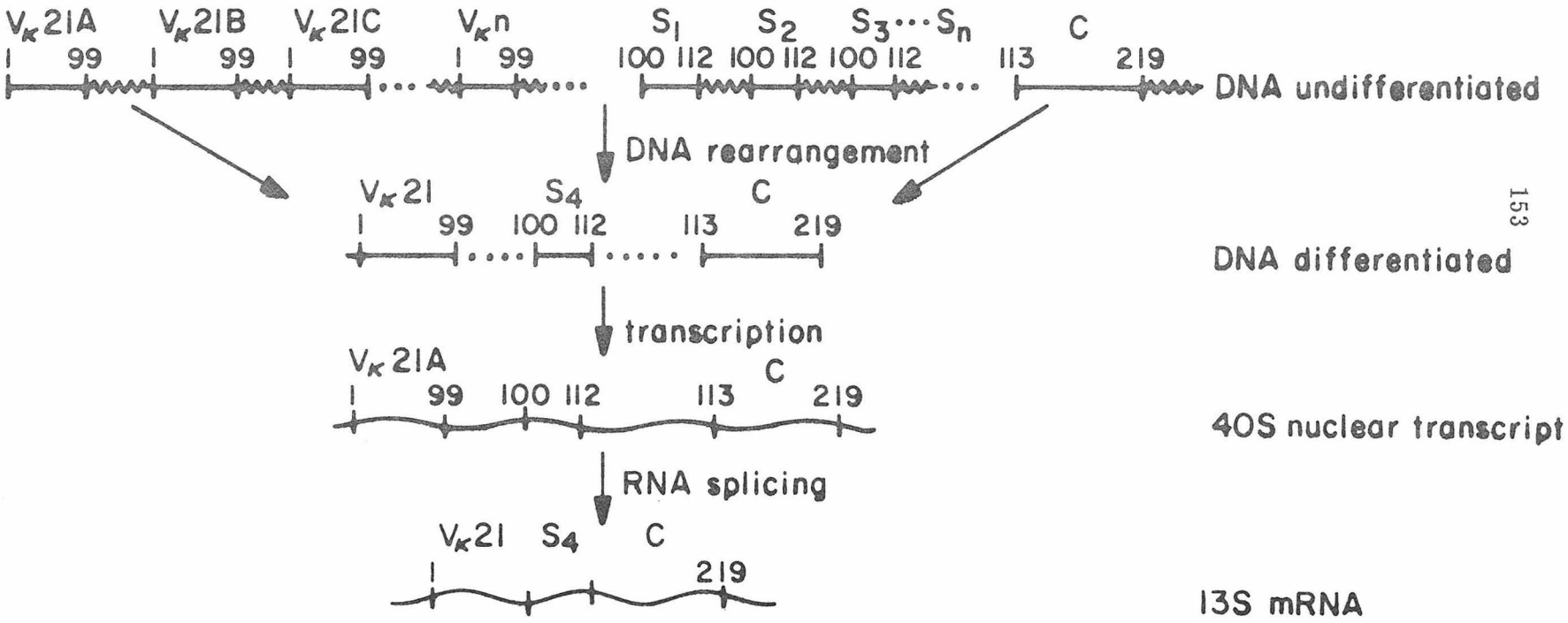

Subset
Proteins
Sequences of the S region (positions 100-113)
Number
100
112
[ ] W T F G G G T K L E I K
V_K21D
V_K21F
V_K21E
V_K21B
V_K21C
V_K21A
Other
6684
7175
2485
7940
7043
7183
7769
4050
9245
63
3741
321
124
2880
1229
7132
70
2154
2413
P Y
D
F
S
L
A
L
R
P
Y
S
1
2
3
4
5
6
7
8
9

Figure 6. A hypothetical model of the DNA segments for the V_{κ} genes. These segments are rearranged at the DNA level, presumably as the lymphocyte differentiates. The rearranged (but not joined) gene segments are transcribed, presumably as a 40S nuclear transcript[30] which undergoes several RNA splicing events to yield a 13S cytoplasmic mRNA with the V, S and C regions joined together.


Vκ21A
Vκ21B
Vκ21C
Vκn
S1
S2
S3···Sn
C
1 99 1 99 1 99 1 99 100 112 100 112 100 112 113 219
DNA undifferentiated
DNA rearrangement
Vκ21
S4
C
1 99 100 112 113 219
DNA differentiated
transcription
Vκ21A
C
1 99 100 112 113 219
40S nuclear transcript
RNA splicing
Vκ21 S4 C
1 219
13S mRNA

Appendix A

This paper was published in Cold Spring Harbor Symp. Quant. Biol.

Reprinted from
Cold Spring Harbor Symposia on Quantitative Biology, Volume XLI

The Structure and Genetics of Mouse Immunoglobulins: An Analysis of NZB Myeloma Proteins and Sets of BALB/c Myeloma Proteins Binding Particular Haptens

L. HOOD, E. LOH, J. HUBERT, P. BARSTAD,* B. EATON, P. EARLY, J. FUHRMAN, N. JOHNSON, M. KRONENBERG AND J. SCHILLING

California Institute of Technology, Division of Biology, Pasadena, California 91125

Speculations about the origins of antibody diversity have intrigued immunologists for the past 75 years. How can the vertebrate organism generate perhaps as many as 10^5 to 10^7 different antibody molecules? Two points of view have developed in response to this question, the germ-line theory and the somatic theory. Does an organism inherit all the genes required for the myriad antibody molecules it will produce, or do most antibody genes arise during differentiation by a process of somatic mutation?

At the first Cold Spring Harbor Symposium dealing with immunology, partial amino acid sequence data from myeloma proteins led some to infer that relatively few genes encoded antibody diversity (Smithies 1968; Edelman 1968). At this symposium, based primarily on nucleic acid hybridization data, others have emphasized the importance of somatic mutation in generating antibody diversity (Jerne; Tonegawa et al.; both this volume). But enthusiasm for this simple solution, just as for the simple solutions in 1967, should not obscure the fact that the antibody gene families generally are multigenic in nature and that germ-line genes in some cases do directly encode antibody molecules. Accordingly, we should not merely ask what are the fundamental mechanisms that contribute to antibody diversity, but what are the relative contributions of each of these mechanisms to the total antibody repertoire?

The tools of molecular immunology have revealed important constraints on contemporary theories of antibody diversity. This paper will review the modern theories of antibody diversity and possible constraints placed on those theories, particularly by sequence analysis of myeloma proteins and by serologic, nucleic acid hybridization, and ontogenetic data. We will focus on recent protein sequence data from our laboratory analyzing: (1) N-terminal sequences of NZB myeloma proteins, (2) N-terminal sequences of sets of BALB/c myeloma proteins binding particular haptens, and (3) the complete variable-region (V_H) amino acid sequences of nine heavy chains from myeloma proteins binding the hapten phosphorylcholine. Finally, we will discuss four fundamental mechanisms for generating antibody diversity.

There Are Two Major Theories of Antibody Diversity

The enormous variability of antibody molecules has led to two general theories of diversity. The *germ-line theory* postulates that most, if not all, antibody genes are encoded separately in the zygote or germ line of the organism and that these genes arose by gene duplication and changed by mutation and selection during vertebrate evolution. This theory suggests that the diversity of antibody genes exists prior to the somatic differentiation of each individual and that antibody synthesis merely requires the activation of distinct antibody genes in each individual lymphocyte. In contrast, the *somatic mutation theories* postulate that antibody diversity is encoded by a more limited number of germ-line genes which diversify by some type of somatic mutational process during the differentiation of each individual. Thus somatic mutation generates antibody diversity anew in each individual. The germ-line theories include the classical formulation as well as theories which scramble germ-line information during differentiation. The somatic theories are categorized by the various genetic mechanisms they employ to explain antibody diversity. The contemporary forms of these general theories are outlined in Table 1. Let us consider the constraints imposed on contemporary theories of antibody diversity by molecular immunology.

Organizational Features and Diversity Patterns of the Antibody Gene Families Have Placed Constraints on Theories of Antibody Diversity

Protein chemistry and serologic analyses have illuminated the outlines of the organizational features and diversity patterns of the antibody gene families.

(1) Three genetically unlinked clusters or families of antibody genes, λ, κ, and H, are present in all mammals that have been studied to date and they display gene products that are homologous to one

* Present address: Department of Microbiology, University of Alabama in Birmingham, Birmingham, Alabama 35274.

Table 1. Contemporary Theories of Antibody Diversity

Theory	Category	Comments	Definition of germ-line V genes	Estimated number of mouse V_κ germ-line genes
Classical germ line[a]	germ line	most antibody variable (V) regions are encoded by distinct germ-line genes	most immunoglobulins with a distinct V sequence	1000's
Combinatorial germ line[b]	germ line	each framework and each hypervariable sequence is encoded by a distinct germ-line gene; in one simple formulation of this model, a single framework and three hypervariable genes are joined during differentiation for each V gene	each distinct framework and each hypervariable-region sequence	?
Predetermined permutation[c]	germ line	contains elements of the classical germ-line and episomal insertion models	uncertain	?
Ordinary somatic mutation[d]	somatic	ordinary spontaneous mutants are selected for clonal expansion	each distinct framework sequence and each distinct sequence gap (Cohn)	100's
Somatic recombination[e]	somatic	somatic recombination among germ-line V genes generates diversity	uncertain	100's to 1000's
Special mutation[f]	somatic	these theories may evoke any one of a variety of special somatic mutational mechanisms	depends on the specific form of this theory	few to 100's

[a] Hood (1973). [b] Capra and Kindt (1975). [c] Klinman and Press (1975a). [d] Cohn et al. (1974); Jerne (1971). [e] Gally and Edelman (1970). [f] Brenner and Milstein (1966); Baltimore (1974); Leder et al. (this volume).

another (Dayhoff 1972; Mage et al. 1973). These observations imply that the antibody gene families descended from a common ancestral family and that they arose prior to the mammalian divergence. Thus similar mechanisms for generating antibody diversity probably exist for all mammals.

(2) Antibody polypeptides are divided into variable (V) and constant (C) regions which are encoded by separate germ-line genes (Tonegawa et al., this volume). These regions reflect a fundamental functional dichotomy of antibody molecules: The V regions comprise the antigen-binding site (Amzel et al. 1974), whereas the C regions encode the effector functions such as complement fixation (Gally 1973). The fundamental functional duality is also reflected in the organization of the antibody gene families. The C-region genes are all encoded directly in the germ line (Mage et al. 1973). Since the antigen-combining site is encoded in the V regions, proposed mechanisms of antibody diversity must explain the patterns of variability observed in the V regions. These patterns will be discussed further on in detail.

(3) A comparative amino acid sequence analysis of many V_L and V_H regions reveals three and four segments, respectively, of extreme diversity. These are termed *hypervariable regions* (Wu and Kabat 1970). Five or six of these seven hypervariable regions comprise walls of the antigen-binding site of the antibody molecule (Amzel et al. 1974; Padlan et al. 1974). The nonhypervariable portions of the V region are remarkably constant in their three-dimensional structure (Poljak et al. 1973) and these have been termed the *framework regions*.

Diversity in the active-site-associated hypervariable regions presumably arises as a consequence of the need to generate many different antigen-binding sites. The V_λ sequences of the mouse are particularly interesting in this regard; 12 of 18 complete V regions studied are identical, and the six variant sequences differ by one to three residues from the most common sequence (Cohn et al. 1974). The striking observation is that all nine of these amino acid substitutions are in the hypervariable regions. This restriction of mutations to the hypervariable regions could arise by random mutation and selection in the germ line or soma or by a special mutation mechanism. Indeed, some immunologists conclude that fundamentally different mechanisms may be required to explain hypervariable and framework diversity (e.g., separate genes for or special mutational mechanisms in hypervariable regions) (Wu and Kabat 1970; Capra and Kindt 1975). Hypervariable regions outside the antigen-binding site might arise as a result of (1) acceptable neutral mutations or (2) the presence of a site with another function which requires diversity (e.g., modulation of V_H and V_L interactions).

(4) A comparison of the V regions from a particular antibody family reveals related sets or *subgroups* of V regions that have similar amino acid sequences and sequence gaps (insertions or dele-

tions) placed at homologous positions (Hood et al. 1968; Milstein 1967). Furthermore, every individual appears capable of producing these subgroups (Grant and Hood 1971). The implications of these observations can best be understood in terms of a graphic display of sequence information termed a *genealogic tree.*

(5) The accumulation of nearly a thousand partial or complete V-region sequences (which constitute more than 22,000 individual amino acid residues) has created a problem of data analysis. How can patterns be extracted from the extensive amino acid sequence data which may tell immunologists more about the genetic organization of antibody families? The genealogic analysis permits sequence data to be compressed and visualized in a form suitable for abstracting patterns. This mode of analysis will be described in detail as it is central to an understanding of much of what we discuss subsequently. (For a more detailed description of this approach, see the Appendix to Chapter 7 in Hood et al. 1975b).

(a) *The rationale.* All immunologists agree that V-region diversity should be explained by the simplest genetic mechanism that is compatible with the experimental data. The genealogic approach assumes that V genes are related to one another by divergent evolution (germ line or soma). Furthermore, their relationships can be depicted by a genealogic tree which attempts to retrace the events of divergent evolution and to determine how a given set of V regions (i.e., V genes) can be derived from a single ancestral V gene using the minimum number of genetic events (i.e., gene duplication, single base substitutions, and the insertion or deletion of one or more codons [sequence gaps]). The genetic events required to generate a hypothetical genealogic tree for V_κ genes are depicted in Figure 1. A single primordial V gene is the ancestor to all contemporary V_κ genes (level E). This ancestral gene undergoes a gene duplication to produce two daughter genes ($V_{\kappa A1}$ and $V_{\kappa A2}$), each of which diverges in nucleotide sequence from the other to become the primordial V_κ genes for their respective branches of the genealogic tree. The number on a line joining two nodal points in the tree designates the number of nucleotide substitutions that have occurred during this divergence. For example, the $V_{\kappa A1}$ gene has diverged from the primordial V_κ gene by five substitutions and from the $V_{\kappa A2}$ gene by ten substitutions (5 + 5). In addition, the $V_{\kappa A1}$ gene has incurred a deletion of codons 31 through 33, which is indicated by brackets. All the V_κ genes above level A on the $V_{\kappa A1}$ branch will share these genetic events (i.e., deletions and base substitutions) with the primordial $V_{\kappa A1}$ gene. It is these shared genetic events in contemporary proteins (genes) which allow their ancestry to be traced by a genealogic tree. Gene duplication followed by mutation and occasional codon deletion occurs at each of the successively higher levels of the genealogic tree (B, C, and D) until the diversity of contemporary genes (level E) is produced. In actual practice, the genealogic tree is generated by reversing this process with the aid of a computer.

How does this genealogic analysis relate to the two major theories of antibody diversity? The genealogic tree permits an analysis of the *number* and *types* of genetic events that might reasonably be expected to occur during somatic differentiation. Two points are of interest in this regard. First, how many mutational events would be required during somatic differentiation to generate most antibody V genes? To take an extreme case, suppose in Figure 1 that only the primordial V_κ gene is in the germ line. Then all higher levels of diversity are generated by somatic mutation. If so, 25 somatic nucleotide substitutions and in some cases a deletion must occur in each lymphocyte line to generate the V_κ regions shown at level E. In addition, some type of selective pressure must act on each of the 25 consecutive mutations so that the appropriate lymphocytes will

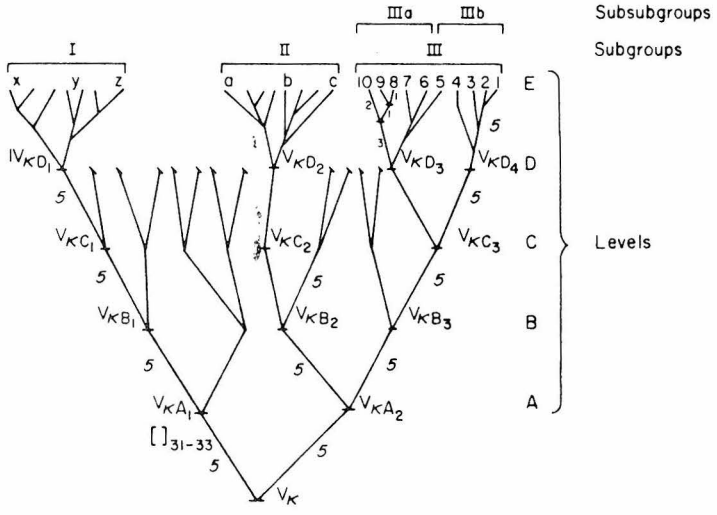


Figure 1. A hypothetical genealogic tree for human V_κ regions. The tree is constructed from a hypothetical set of proteins (i.e., 1, 2, 3, 4, etc.), such as human V_κ regions, by generating a series of ancestral or nodal sequences (levels D, C, B, etc.) using the minimum possible number of "genetic events" (base substitutions, sequence insertions or deletions, and gene duplications). The genetic events responsible for generating this genealogic pattern could occur, in part, during somatic differentiation (somatic theory) or entirely during the evolution of the species (germ-line theory). The V_κ branch is divided into three sub-branches (subgroups) designated I, II, and III. The fact that each of the sub-branches can be further subdivided is indicated by IIIa and IIIb. The fine details of one region of the genealogic tree are represented by V_κ regions 1–10.

be selected and clonally expanded at each stage of this process in order that the next mutation can occur in the expanded clone. The various contemporary theories of antibody diversity have different views on how many somatic mutational events might occur during the differentiation of each individual (see Table 1). A second but related question is how much *parallel* (identical) mutation is likely to occur with a somatic mutational mechanism? For example, assume that the V_κ-region sequences 8 and 9 in Figure 1 were derived from *different* individuals. Once again, if we assume that only the primordial V_κ gene is in the germ line, then 25 mutations are required to generate V genes 8 and 9. However, in this case, 24 mutations must be identical (i.e., parallel mutations) in two separate individuals. What type of selective forces could produce such parallel mutations? None are known. Accordingly, most contemporary immunologists would feel uncomfortable postulating even one or two parallel mutations. The number of mutational events and parallel mutations required by somatic mechanisms can be reduced by moving the level of germ-line V genes up the genealogic tree (to level B, C, or D). Thus the critical question with regard to theories of antibody diversity is "What level on the V-region genealogic tree represents germ-line genes?"

(b) *Constraints of the genealogic pattern.* The earliest analysis of the amino acid sequences of immunoglobulin V regions demonstrated that they fit into a genealogic pattern (Smith et al. 1971). Subsequent sequence analysis has revealed more of the fine structure of the genealogic trees of various immunoglobulin families and permitted the genealogic analysis of diversity in new antibody families (see Hood et al. 1975a). These analyses place three constraints on theories of antibody diversity.

First, the V-region subgroups represent the terminal divergences of particular branches on the genealogic tree. For example, if the V-region sequences had been determined on proteins 1, 4, and 8, a, b, and c, and x, y, and z in Figure 1, then these proteins would have defined three subgroups designated I, II, and III. To avoid excessive parallel mutations, at least one germ-line gene is necessary for each subgroup. However, with additional sequence data, subsubgroups have been defined (e.g., IIIa and IIIb in Fig. 1), and once again the level of germ-line V genes must be raised to avoid parallel mutation. This points out the difficulty of precisely defining a subgroup. The important observation, however, is that most antibody V-gene families have multiple branches on their genealogic tree and, accordingly, must be encoded by multiple germ-line genes.

Second, the V-region genealogic trees of some immunoglobulin families are extremely diverse; others are not. For example, mouse V_κ regions appear to be extremely diverse, whereas the V_λ regions are not. This implies, but does not prove, that the diverse families are encoded by many germ-line V genes, whereas the restricted families are encoded by few germ-line V genes.

Third, identical V regions have been observed in myeloma tumors that arose independently in different individuals. To avoid parallel mutation in different individuals, a genealogic analysis suggests that identical V regions are coded directly by a germ-line V gene that has not been altered by somatic mutation. For example, in Figure 1 if two separate individuals produce V-region 1, then the only way to avoid parallel mutation is to move the germ line to level E. If one accepts this argument, the V region is directly encoded by a V germ-line gene.

By the reasoning given above, subgroups have become a convenient means of counting the minimal number of germ-line genes required to encode a particular group of V regions.

We will use subgroups to designate a set of closely related V-region sequences that have two properties: (1) they show extensive sequence homology to one another and therefore constitute a distinct branch on the genealogic tree; and (2) most immunologists agree that they are encoded by *at least* one germ-line gene. Confusion in the literature occurs when the term "subgroup" is used interchangeably with "germ-line gene." For example, if one equates the term "subgroup" to "germ-line gene," the germ-line theory would contend that every separate sequence is a new subgroup. However, since the correlation of each sequence to a germ-line gene is controversial, we could not use the term "subgroup" in this situation. One can avoid this confusion by using subgroup to refer only to sets of proteins, with closely related V regions, that are clearly encoded by separate genes—for example, the three human V_κ sets of sequences. The crux of the argument among the various theories of diversity then becomes a question as to the number of V genes in each subgroup. Where the germ-line theory would contend that each distinct V-region sequence within a subgroup reflects a separate germ-line gene, one form of ordinary somatic mutation would argue that only new framework sequences within subgroups need define a new germ-line gene. All theories of diversity, however, agree that most antibody families have multiple V-region subgroups and, accordingly, are multigenic in nature.

(6) The idiotypic analysis of certain myeloma proteins and homogeneous antibodies suggests that there are multiple germ-line V_H genes. Theoretically, an idiotype is the collection of antigenic determinants that distinguishes one V domain from another. In practice, antibodies directed against an idiotype (i.e., anti-idiotypic antibodies) may be

capable of identifying a single V domain or a group of very closely related V domains. Anti-idiotypic antibodies have been prepared against certain murine myeloma proteins with defined hapten-binding activity and against homogeneous antibodies (Weigert et al. 1975). Studies on the expression of these idiotypes in various inbred strains of mice permit several conclusions to be drawn.

(a) These idiotypes segregate in a Mendelian fashion (Weigert et al. 1975). This suggests that the idiotypes are coded by germ-line V genes.

(b) Twelve different idiotypes appear to be linked to genetic markers on the C_H genes (i.e., allotypes) (M. Weigert, pers. comm.). This suggests that multiple germ-line V_H genes are closely linked to their corresponding C_H genes.

(c) Recombinational analysis permits a preliminary grouping of certain of the V_H idiotypes on a genetic map (Weigert et al. 1975). With more complete studies of this type, it will be possible to order the V genes with respect to one another.

(d) Recombinational analysis suggests that the V_H chromosomal map distance is three map units (Eichmann 1975; R. Riblet, pers. comm.). Though this number appears unreasonably large (i.e., with some simplistic genetic assumptions this region could code for $\sim$12,000 V_H genes), it does stress that the V_H segment of the mouse chromosome potentially has sufficient genetic information for many germ-line V_H genes.

(7) The diversity of the λ and κ families may vary markedly from one mammal to the next. For example, the human V_λ sequences are highly diverse (Hood and Talmage 1970), whereas their mouse counterparts are highly restricted (Cohn et al. 1974). If the expressed diversity of an antibody family is proportional to its genetic diversity, then some mammals have many V_λ genes and others have few. This suggests that the antibody gene families can readily be expanded by gene duplication or contracted by gene deletion over the time period of mammalian divergence (75×10^6 years). If some mammalian species have predominately V_κ genes and others predominately V_λ genes (Hood et al. 1968), the light-chain gene families can apparently perform equivalent functions.

(8) Experiments by several laboratories have demonstrated that the time sequence of the appearance of immunocompetent cells during development is nonrandom with respect to specificity (Press and Klinman 1974; Sherwin and Rowlands 1975), immunoglobulin isoelectric point (Klinman and Press 1975b), and idiotype (Sigal et al. 1976). Since the same order of expression occurs in different individual mice, such experiments imply the existence of multiple germ-line V genes which can be expressed in an ordered fashion during development. These observations render less likely "big bang" theories which in their simpler forms predict that all the diversity is generated rapidly during early development, with no predetermined order (Edelman 1974).

In summary, the antibody gene families and their diversity-generating mechanisms are ancient adaptations for the storage and generation of protein diversity (Hood et al. 1975a). This diversity is concentrated in the hypervariable segments which are the walls of the antigen-binding site. This pattern of diversity must be produced by mutation and selection in the germ line or soma. The facts that the individuals of a species have identical subgroups and idiotypes and that the programmed expression of certain antibody molecules occurs during development both suggest that multiple germ-line V genes exist. Indeed, all theories of antibody diversity now accept the existence of multiple germ-line genes rather than requiring extensive parallel mutations in many individuals. Table 2 summarizes the constraints mentioned above as well as those deduced from sequence data to be presented subsequently and nucleic acid hybridization data.

Conceptually, we can divide the diversity problem into two important questions: (1) How many V-region subgroups (i.e., branches on the genealogic tree) exist within a particular antibody gene family? That is, what is the minimal contribution of the germ line and its evolutionary selection and mutation to diversity? Contemporary estimates of the number of V-region subgroups come almost entirely from the amino acid sequence analysis of myeloma proteins. We shall argue from data on NZB myeloma proteins that this approach may be seriously underestimating the number of such germ-line genes. (2) How many germ-line V genes are present within individual subgroups? This is a point on which somatic and germ-line theories clearly differ. This has been approached by nucleic acid hybridization studies, which suggest that some somatic mutation must occur in the mouse V_λ family (Tonegawa 1976). We will return to this argument shortly.

BALB/c and NZB Myeloma Proteins Differ in Their Structural and Functional Properties

BALB/c myeloma tumors may express only a subset of the total antibody repertoire in the mouse. Myeloma tumors may be artificially induced by injecting mineral oil into the peritoneal cavity of inbred BALB/c mice (Potter 1972). Does this highly artificial induction process provide a random sampling of the lymphocyte diversity the mouse is capable of producing? Three observations suggest that these myeloma proteins may represent only a selected subset of the antibody repertoire. First, approximately 5% of the BALB/c myeloma proteins bind a very restricted spectrum of simple haptens, including DNP, phosphorylcholine, and various

Table 2. Organizational Features and Diversity Patterns of Antibody Gene Families

Type	Observation	Conclusion
Structural	λ, κ, and H immunoglobulins	three families of antibody genes
	V and C regions	separate V and C genes
	hypervariable regions	(a) antigen-binding-site–associated selection (b) site accepting neutral mutations
	mouse λ chains exhibit few substitutions limited to hypervariable regions	mutation followed by selection (germ line or soma)
	V subgroups or genealogic trees	multiple germ-line V genes
	identical V regions	some antibodies are coded directly by germ-line genes
	the diversity of a particular light family may vary markedly from one mammal to the next	expansion (duplication) and contraction (deletion) of antibody gene families
	normal mouse κ chains exhibit residue alternatives not seen in the myeloma pool of sequences	the myeloma systems are windows which express only a subset of the total V-gene diversity
	NZB and BALB/c myelomas differ in structural and functional properties	the myeloma systems are windows which express only a subset of the total V-gene diversity
	myeloma proteins binding common haptens have identical or closely related sets of V_L and V_H regions	(a) multiple germ-line V genes (b) some antibodies are coded directly by germ-line genes
Serologic	idiotypes behave as Mendelian markers linked to C_H	the V_H genes are closely linked to C_H genes
	approximately 12 idiotypes map to the C_H genes	multiple germ-line V genes
	the V_H genes span 3 genetic map units	there is sufficient DNA to code for thousands of V genes
Nucleic acid	mRNA or cDNA from different V_κ myelomas hybridizes with non-myeloma DNA with single-copy kinetics	multiple germ-line V genes
	V_λ mRNA or cDNA hybridizes with single-copy kinetics to excess myeloma or liver DNA	somatic mutation may generate variants
Ontogenetic	ordered expression of individual antibody molecules	programmed readout of germ-line genes

simple carbohydrate moieties (Potter 1970). This frequency is much greater than would be expected from the frequency of lymphocytes binding the same haptens in normal individuals (Press and Klinman 1974). Accordingly, the BALB/c myeloma proteins appear to represent a restricted sample of the potential functional repertoire of the BALB/c mouse. Second, approximately 80% of the BALB/c myeloma heavy chains have unblocked N-terminal groups, whereas only about 20% of the serum heavy chains are unblocked (Capra et al. 1973). The presence of blocked and unblocked N-terminal groups is indicative of different V-region subgroups (Capra and Kehoe 1975). Thus the BALB/c myeloma proteins and serum immunoglobulins reflect quite a different distribution of V_H subgroups and, presumably, antibody specificities. Finally, if the residue alternatives at certain positions of myeloma sequences are compared to their counterparts from the normal serum pool, it is clear that some residues in the myeloma pool are not found in the normal pool (Hood et al. 1974). Conversely, residue alternatives found in the normal pool are missing in the myeloma pool. These differences again suggest that the normal and myeloma pools of sequences express somewhat different subsets of the antibody repertoire. We wondered whether a similar phenomenon would be reflected in the comparison of BALB/c myelomas with the only other inducible myeloma system available in mice – the NZB system.

The NZB myeloma proteins appear to be distinct from BALB/c myeloma proteins by several criteria. The NZB myeloma proteins differ from their BALB/c counterparts by three criteria (E. Loh et al., in prep.). First, NZB myeloma proteins are predominately of the IgG class, whereas BALB/c proteins are predominately of the IgA class (Table 3). Second, the profile of simple haptens to which the NZB myeloma proteins bind appears to be distinct from that of their BALB/c counterparts (Table 4). For example, 12 NZB proteins have been

Table 3. Comparison of Heavy-chain Isotype for BALB/c and NZB Myeloma Proteins

Class of Ig		NZB (%)[a]	BALB/c (%)[a]
A		20.9	45.5
M		1.7	1.1
G	1	9.1	6.4
G	2a	15.2	6.4
G	2b	25.7	8.1
G	3	2.2	0.7
Number of proteins screened		230	558

These data were collected by H. C. Morse III (pers. comm.).

[a] The percentages do not add up to 100% because Bence-Jones proteins (urinary light chains) and unclassified immunoglobulins are included in the total.

Table 4. Haptens Bound by NZB and BALB/c Myeloma Proteins

	NZB	BALB/c
DNA	++	*
DNP	−	++
Phosphorylcholine	−	++
1,3 Dextran	−	+
1,6 Galactan	−	+
1,6 Dextran	+	+
2,1 Levan	+	++

−Indicates that no proteins bind; + indicates that 1 to 10 proteins bind; ++ indicates that > 10 proteins bind; * denotes that a number of BALB/c proteins bind DNA and DNP, whereas the NZB proteins bind DNA. (Data from E. Loh et al., in prep.)

found which bind DNA. Indeed, it is striking that no NZB myeloma proteins bind DNP or phosphorylcholine – the two most common haptens bound by BALB/c myeloma proteins. Furthermore, in one case where pairs of NZB and BALB/c myeloma proteins bind a similar hapten, sequence analysis of the corresponding V regions shows that the NZB and the BALB/c sequences are similar within the strain yet quite distinct between strains (Table 5). Third, the V_κ and V_H regions from the NZB and BALB/c myeloma proteins appear to be quite distinct from one another by amino acid sequence analysis. Twenty-five V sequences of NZB myeloma κ chains have been examined over their N-terminal 23 residues. Figure 2 is a genealogic tree which graphically displays the sequence relationships of the NZB and BALB/c myeloma κ chains. Individual immunoglobulin chains are represented by number designations at the terminal twigs of the tree. The NZB V_κ regions are boxed to distinguish them from their BALB/c V_κ counterparts. This figure illustrates that the diversity of the NZB κ chains is comparable to the diversity of BALB/c κ chains and that very little overlap exists between the two sets of proteins. Eighteen different V_κ sequences are found among the 25 NZB-κ chains examined, and only one of these sequences is identical to one of the 43 different BALB/c V_κ sequences available for comparison. Accordingly, new mouse V_κ subgroups can be defined. Sequence data from the heavy chains also show differences between the NZB and BALB/c V_H regions. Approximately 60% of the V_H regions from the NZB heavy chains have a blocked α-amino group, whereas very few BALB/c heavy chains are blocked (Table 6). This observation suggests that a different ratio of V_H subgroups is being expressed in the myeloma pools of the two strains. This supposition is supported by the N-terminal sequence of the unblocked chains (Fig. 3). Twelve NZB heavy chains were examined over their N-terminal 20 residues. Ten different V_H sequences were noted. Nine out of ten NZB V_H sequences are different from each of more than 30 BALB/c V_H regions available for comparison (Fig. 3). The main cluster of sequences in the center of Figure 3 is homologous to the unblocked V_HIII subgroup (Barstad et al. 1974). Non-V_HIII sequences, representing different subgroups, constitute a much higher proportion of the NZB heavy chains than of the BALB/c chains. Thus the V-region sequences of both the κ family and the heavy-chain family appear to be distinct in these two myeloma populations from inbred strains of mice.

Only one V_κ sequence and only one V_H sequence are seen in both the NZB and BALB/c myeloma proteins examined. One could argue that this slight degree of overlap between the NZB and BALB/c sequences is expected if the pool of possible sequences is so large that repetitions would not be seen. Although it is true that the pool of sequences is large, more NZB and BALB/c identities would be expected than are found, since within the NZB

Table 5. The N-terminal Sequences of V_H Regions Derived from BALB/c and NZB Myeloma Proteins That Bind α1,6 Dextran

Strain		Position 1 5 10 15 20	Specificity
BALB/c	W3434	E V K L L E S G G G L V Q P G G S L K L	α1,6 dextran
BALB/c	W3129	——V-I—( )——	α1,6 dextran
NZB	3936	——Q——Q-Q——P E——K——A—V—I	α1,6 dextran
NZB	3858	——Q——Q-Q——P E——K——A—V—I	α1,6 dextran

The one-letter code of Dayhoff (1972) is used. A line indicates that the corresponding sequence is identical to W3434. Substitutions are indicated by letters. Parentheses indicate that the residue assignment is uncertain. The BALB/c sequences are from Barstad et al. (1974) and the NZB sequences from E. Loh et al. (in prep.).

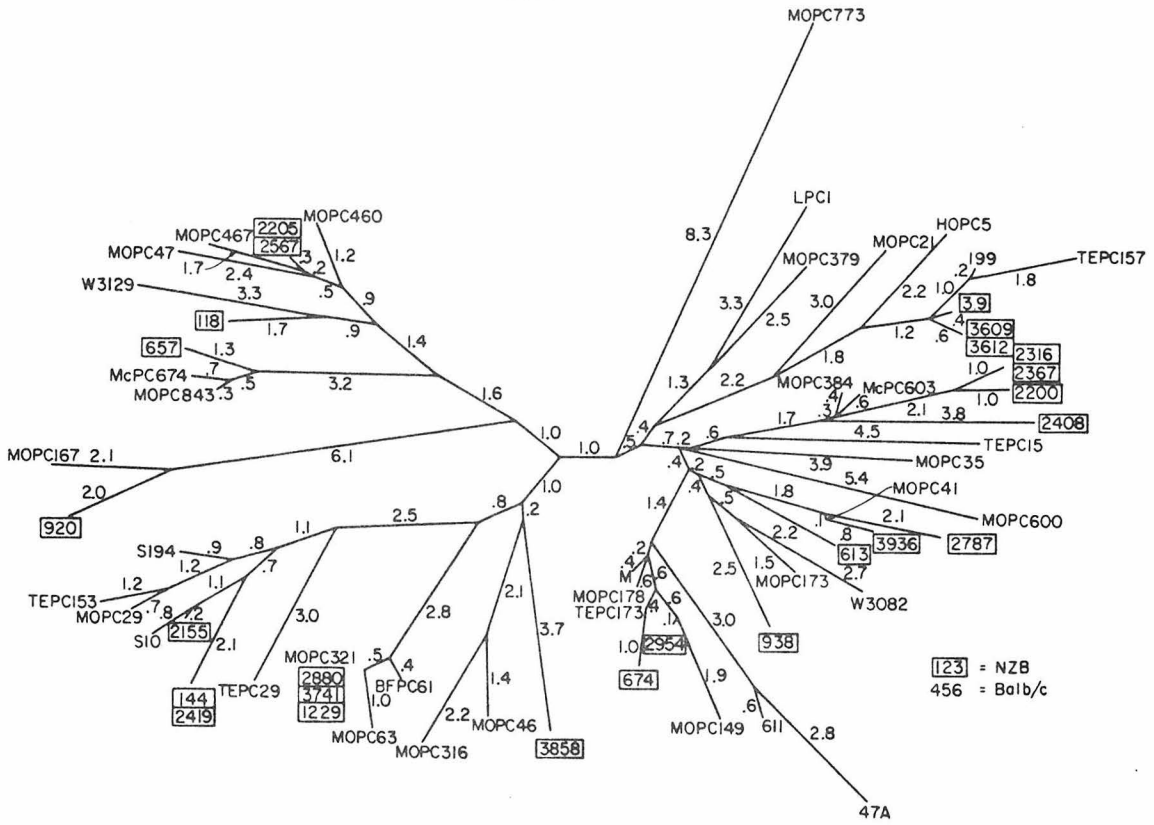


Figure 2. A genealogic tree of the N-terminal 23 residues of BALB/c and NZB V_κ regions. The NZB V_κ regions are boxed and the BALB/c V_κ regions are not. The fractional numbers designating nucleotide substitutions arise from the averaging process that is necessary to generate these trees (see Hood et al. 1975b, Appendix to Chapter 7).

κ chains themselves five sets of identities exist, and at least twelve exist within the BALB/c set (Hood et al. 1974).

The one overlap that exists between NZB and BALB/c κ sequences is a special case which may be of particular interest. Three NZB V_κ regions (2880, 3741, 1229) and four BALB/c V_κ regions (T124, M70, B32, M321) have the same N-terminal sequence for 23 residues. This set of proteins is interesting for three reasons. First, approximately 10% of the NZB and BALB/c κ myeloma proteins examined thus far have this N-terminal sequence. Up to 10% of the normal serum light chains also appear to be of this same subgroup by serologic analysis (M. Weigert, pers. comm.). Second, these BALB/c κ chains can be split into two distinct sets after the first hypervariable region – the M70-like and the T124-like. The NZB V_κ regions seem to split into similar sets – 2880 and 1229 being M70-like and 3741 being T124-like (Table 7). We would argue that these proteins form two distinct subgroups and are therefore encoded by at least two germ-line genes present in both strains. Third, with the possible exception of 1229 and 2880, no two members of this set of eight proteins have identical V regions (McKean et al. 1973). The patterns of differences within this set provide some constraints on possible mechanisms of diversity that will be discussed later.

Table 6. Comparison of V_H Subgroup Distributions of BALB/c and NZB Myeloma Proteins

	NZB	BALB/c
"Blocked" N termini	34/49	5/30
"Unblocked" V_HIII subgroup	6/49	21/30

Data from E. Loh et al. (in prep.).

On the basis of heavy-chain class distribution, antigen-binding properties, and sequence analysis, we suggest that NZB and BALB/c myelomas are distinct populations of proteins.

Two models may account for the observation that the myeloma process in BALB/c and NZB mice appears to be transforming distinct populations of lymphocytes. The BALB/c and NZB mice may have (1) different V genes or (2) genetic differences outside the V-region structural genes which cause different internal environments.

If the V genes for immunoglobulins are different in the two strains, different V regions would be expressed in these strains. Different structural genes may have been fixed in each strain from polymorphisms that existed in their ancestors.

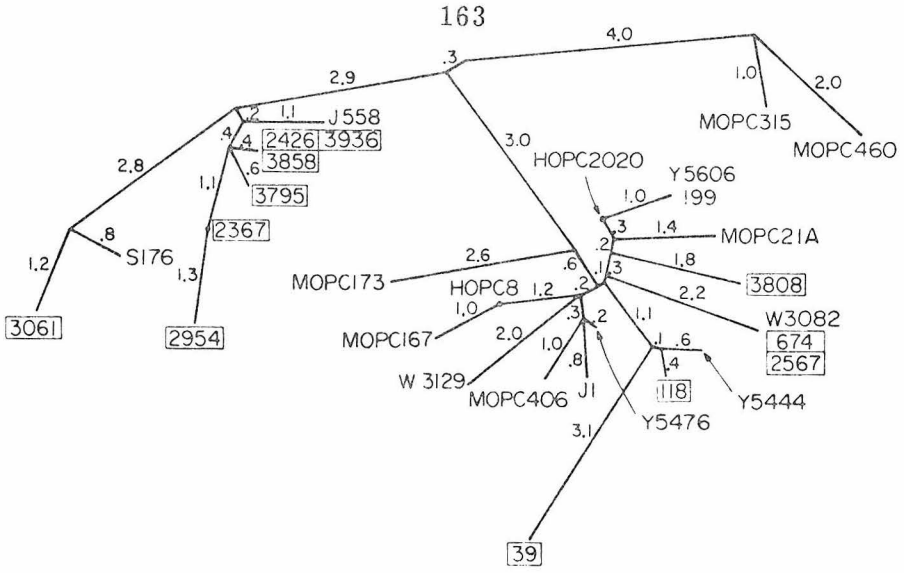


Figure 3. A genealogic tree of the N-terminal 20 residues of BALB/c and NZB V_H regions. The NZB V_H regions are boxed and the BALB/c V_H regions are not.

Indeed, the two strains differ in their heavy-chain allotypes. However, this model does not account for the fact that the myeloma proteins in the two strains bind distinct sets of antigens (Table 4) or that the class distribution is different in the two strains (Table 3).

The more likely explanation is that because of genetic differences outside the V-region genes, the internal antigenic environments are distinct in these two strains, and this difference leads to the clonal expansion of (or induction of tolerance in) different sets of lymphocytes in the two strains. Accordingly, the myeloma process may transform different populations of peritoneal lymphocytes or plasma cells. This supposition is attractive in view of the propensity of the NZB mice to develop autoimmune disease, with the concomitant expansion of clones of lymphocytes to self antigens. Consistent with this hypothesis is the fact that 12 different NZB myeloma proteins bind DNA but not DNP, whereas none of their BALB/c counterparts bind only DNA (Table 4). This model would propose that the structural genes for both sets of sequences exist in both strains but that other genetic differences create distinct internal environments which lead to the expression of different *subsets* of V genes.

The myeloma process appears to transform only a subset of the total lymphocyte repertoire the mouse can express. The spectra of binding specificities, the V_H-subgroup distributions, and the sequence data all demonstrate that the myeloma process is selective. It provides windows through which we can glimpse the antibody repertoire, but the windows probably view just a fraction of the total diversity. Since it is difficult to say how small this fraction is, it is impossible to estimate what fraction of the antibody repertoire is expressed by the myeloma process in each of the inbred strains. Estimates of the number of V-region subgroups in the antibody families, based on the sequence analysis of myeloma proteins, are therefore minimal estimates and could be far too low. Thus the number of V-region subgroups in many antibody families may be substantially higher than previously estimated. Let us now consider the sets of V regions

Table 7. The N-terminal Sequences of Mouse V_κ Regions from Two Closely Related Subgroups

Subgroup	κ Chain	Position	Strain
		10 20 30 40	
T124-like	TEPC-124	D I V L T Q S P A S L A V S L G Q R A T I S C R A S Q S V N W Y G N S F M Q W Y Z Z K	BALB/c
	MOPC-321	—— K —— T ——	BALB/c
	MOPC-63	N —— S ——	BALB/c
	3741	—— ? —— D ——	NZB
M70-like	MOPC-70	—— B S — I —— B — F ——	BALB/c
	2880	—— B —— I —— B — F ——	NZB
	1229	—— B —— I —— ? — F ——	NZB

Residues 27–32 in 1229 are determined by homology from a peptide fragment. See Table 5. TEPC-124, MOPC-321, MOPC-63 are from McKean et al. (1973). MOPC-70 is from Gray et al. (1967). 3741, 2880, and 1229 are from E. Loh et al. (in prep.).

164

derived from myeloma proteins binding the same hapten.

Variable Regions from Myeloma Proteins Binding a Common Hapten Are Closely Related

Many myeloma proteins from BALB/c mice have been screened for binding activity to a variety of haptens. Approximately 5% bind specifically to one of a spectrum of haptens, including phosphorylcholine, $\beta(2 \rightarrow 1)$ levan, $\beta(2 \rightarrow 6)$ levan, $\alpha(1 \rightarrow 3)$ dextran, $\alpha(1 \rightarrow 6)$ dextran, $\alpha(1 \rightarrow 6)$ galactan, and DNP (Potter 1970). Two lines of evidence indicate that these myeloma proteins are members of the normal antibody population that bind these antigens. First, many of the haptens that bind to the myeloma proteins are found on the normal intestinal flora in BALB/c (Potter 1972). BALB/c myeloma tumors arise in the peritoneal cavity and presumably represent transformed lymphocytes derived from expanded clones of antibody-producing cells directed against intestinal antigens. Second, the idiotypes of certain myeloma proteins that bind phosphorylcholine and $\alpha(1 \rightarrow 3)$ dextran are indistinguishable from the idiotypes of the antibodies normally induced in BALB/c mice against these antigens (Sher and Cohn 1972; Carson and Weigert 1973).

The fact that certain of these myeloma proteins may be very similar, if not identical, to normal antibody molecules makes them attractive candidates for analysis vis-a-vis theories of antibody diversity. These proteins allow us to search for patterns of sequence diversity among biologically relevant sets of V regions which bind the same antigen. In this section, we will examine the N-terminal portions of V regions from myeloma proteins binding seven different haptens.

The hapten-binding properties of myeloma proteins appear to correlate with the V_H sequences. Table 8 gives the N termini of 23 V_H sequences, 21 of which are derived from immunoglobulins able to bind to one of eight different hapten specificities (P. Barstad et al., in prep.). One of these heavy chains (Ars) is derived from normally induced antibody to the arsenate moiety (Capra et al. 1975). These sequences are compared against three prototype V_H sequences to facilitate the analysis of related V_H regions (see Barstad et al. 1974). In most cases, there is a striking correlation between the amino acid sequence of a particular V_H region and the hapten with which the parent immunoglobulin combines. Note that this structure–function correlation holds for the framework residues (i.e., 1–27) as well as for the first hypervariable region (i.e., residues 28–36) which folds to comprise a portion of the antigen-binding site. Figure 4 is a genealogic tree of these V_H sequences illustrating the differences as well as the similarities among these sequences. The V_H regions of a particular specificity (e.g., phosphorylcholine) are tightly clustered together on the tree and generally are widely separated from the V_H regions derived from immunoglobulins with different binding specificities.

The V_L regions from myeloma proteins binding different haptens fall into distinct sets. Table 9 gives the N termini of 21 V_L sequences, 20 of which are derived from myeloma proteins with hapten-binding activity (P. Barstad et al., in prep.). For the myeloma proteins binding each carbohydrate specificity ($\beta[2 \rightarrow 1]$ levan, $\alpha[1 \rightarrow 3]$ dextran, $\beta[2 \rightarrow 6]$ levan, and $\beta[1 \rightarrow 6]$ galactan), the V_L regions are generally identical in amino acid sequence (Fig. 5). The V_L regions derived from proteins binding phosphorylcholine fall into three distinct subgroups. Light chains derived from the two myeloma proteins binding DNP are of the λ and κ type, respectively, and thus are quite different in amino acid sequence. However, within the limits of this study, there appears to be a correlation of light-chain sequences with hapten-binding specificity. Indeed, the three subgroups of phosphorylcholine-binding light chains may have some functional importance in that they appear to correlate with three distinct subspecificities of phosphorylcholine-binding myeloma proteins (P. Barstad et al., in prep.).

These amino acid sequence data have important implications for antibody diversity. The data suggest that the V regions which bind particular haptens are generally from different subgroups and hence derived from different germ-line genes. For example, two or more V_L regions show sequence gaps of one residue, two residues, three residues, six residues, and seven residues in the first hypervariable region when these V_κ regions are compared against the T15 V_κ region (Table 9). Indeed, each of the major specificities listed here is associated with a distinct sequence length in the first hypervariable region, except for two related levan specificities. In addition, the V_H region shows two different sequence lengths in the first hypervariable region. It is generally felt that sets of V regions of differing lengths are probably encoded by separate germ-line genes in order to avoid the repeated (parallel) generation of identical gaps in different individual mice. The existence of three sets of identical V_H sequences for the first 40 residues under the V_HIII prototype (e.g., 1,6G; 2,1L; and PC–Table 8) demonstrates that additional sequence data can subdivide one subgroup into multiple new subgroups. Most immunologists would agree that these three sets of proteins which each differ by three residues over their N-terminal 27 residues (framework residues) constitute three distinct germ-line genes (i.e., subgroups). Thus the set of mouse heavy chains analogous to human V_HIII, originally defined as a single subgroup, is

Table 8. The N-terminal Amino Acid Sequences of V_H Regions Derived from Myeloma Proteins Binding Various Haptens

Source	Ig class	Light-chain type	Position	Specificity	Ref.
			←——HV$_1$——→		
			1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38		
V_HI prototype			E V Q L Q E S G P S L V K P S Q T L S L T C S V T G S D I T N G Y F W B W I		
M460	IgA	κ	——————————————————————(Z)——————————————————————M-[ ]——	DNP	1
M315	IgA	λ	D————————G————————S————————Y-S————S————N——V	DNP	2
V_HII prototype			E V Q L Q Q S G P E L V K P G A S V K M S C K A S G Y T F T B Y Y M K [ ] W V		
558	IgA	λ	——	1,3D	1
M104E	IgM	λ	——————————————————————A————V————?————T——	1,3D	1
α Ars	IgG	κ	————————A————A——S————S——T————S—S——G-L—Y——	Ars	3
V_HIII prototype			E V K L L E S G G G L V Q P G G S L K L S C A A S G F D F S R Y W M S [ ] W V		
S117	IgA	κ	——————————————————————————————B————————E——	Ga	1
M173	IgG	κ	————————P————L————————————————————————————	unknown	4
S10	IgA	κ	——	1,6G	5
T191	IgA	κ	——	1,6G	5
X44	IgA	κ	——	1,6G	5
J1	IgA	κ	————I———————————————————————————————————————	1,6G	5
Y5476	IgA	κ	——————————————————————————————A-?————?——	2,6L*	1
U10	IgA	κ	——————————————————————————————G——?——————	2,6L*	S.R.
W3082	IgA	κ	————E————————M————V————T——B————B——?——	2,1L	1
J606	IgG3	κ	————E————————M————V————T——B——?——B——?——	2,1L	1
T15	IgA	κ	————V————————R————T————T——B-F-Y——E——	PC	6
H8	IgA	κ	————V————————R————T————T——B-F-Y——E——	PC	6
S107	IgA	κ	————V————————R————T————T——B-F-Y——E——	PC	6
M511	IgA	κ	————V————————R————T————T——B-F-Y——E——	PC	1
W3207	IgA	κ	————V————————R————T————T——B-F-Y——E——	PC	1
M603	IgA	κ	————V————————R————T————T——B-F-Y——E——	PC	6
M167	IgA	κ	———V-V————————R————T————T——B-F-Y——E——	PC	6
M21	IgG	κ	D—Q——V————————R————————T——S-F-G——H——	unknown	

Ars indicates the sequence of V_H regions derived from normal antibody (Capra et al. 1975). A line indicates that the sequence is identical to the nearest prototype sequence above it. The prototype sequences are taken from Barstad et al. (1974). Brackets indicate a deletion. Question mark indicates that the residue assignment is uncertain or that the amide assignment has not been determined. HV_1 indicates the extent of the first hypervariable region. Specificity denotes the hapten-binding activity of the intact immunoglobulin: DNP, dinitrophenyl; 1,3D, α(1 → 3) dextran; 1,6G, β(1 → 6) galactan; 2,1L, β(2 → 1) levan; 2,6L, β(2 → 6) levan; PC, phosphorylcholine; and G, glucosamine. Asterisk indicates a probable but not established hapten-binding assignment. The references noted are as follows: (1) P. Barstad et al. (in prep.); (2) Francis et al. (1974); (3) Capra et al. (1975); (4) Bourgois and Fougereau (1970); (5) Rudikoff et al. (1973); (6) Barstad et al. (1974); (7) Schulenberg et al. (1971); (8) Cesari and Weigert (1973). S.R. indicates personal communication from S. Rudikoff.

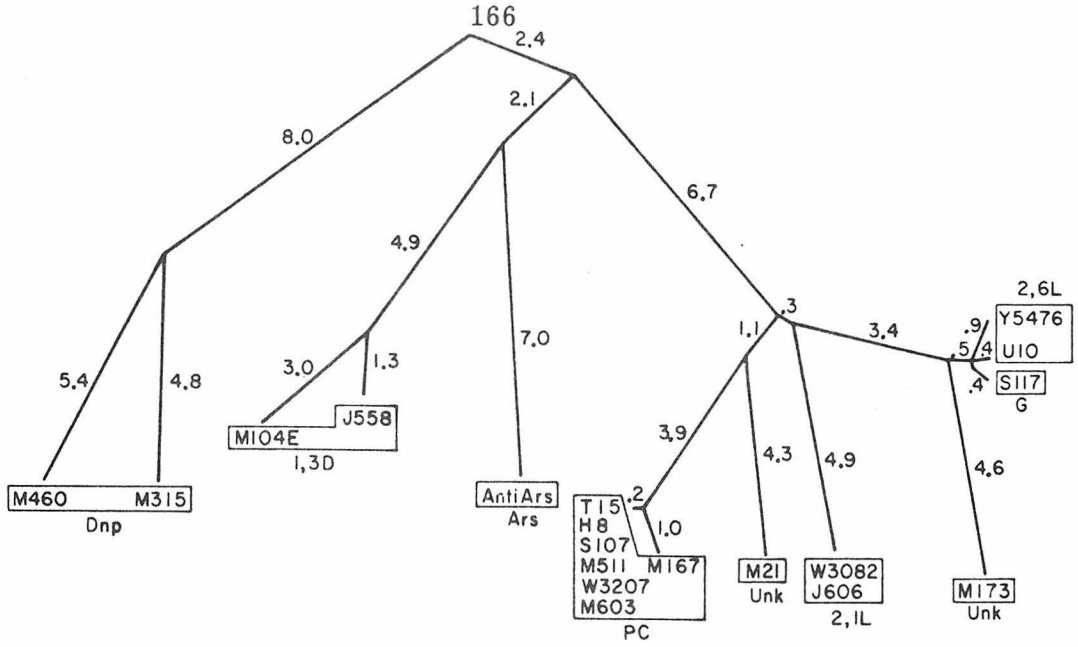


Figure 4. A genealogic tree generated from the N-terminal 38 residues of heavy chains from 17 BALB/c proteins. Individual immunoglobulin chains are represented by number designations at the terminal twigs of the tree. Boxes indicate myeloma proteins that share a hapten-binding specificity. See notes to Table 8 for hapten abbreviations.

actually composed of a series of related sets of V_H regions coded for by at least three genes. It will be important to obtain the complete V_H sequences of these sets to determine whether these relationships hold throughout the entire V_H region. It is interesting to note that in the closely related V_H (Fig. 6) and V_κ (Fig. 8) sets of sequences, as discussed in the next section, differences in the N-terminal 23 residues do indeed presage multiple differences throughout the remainder of the variable region.

One may increase the number of V genes further by arguing that more than one gene codes for each related set. This is a more controversial point. These data demonstrate that, within related sets of V_L or V_H sequences, minor variants are seen that lie outside the active-site–associated hypervariable region (Tables 8 and 9). For example, two sets of V_H regions that show variants are the α1,3 dextran binders (positions 21 and 25) and the phosphorylcholine binders (position 4) (Table 8). The antiphosphorylcholine V_L regions show variants at position 18 (M603) and 11 (M511). This observation is of particular importance for theories of antibody diversity that require a new germ-line V gene for each framework substitution (e.g., classical germ-line, combinatorial germ-line, and ordinary somatic mutation) because the number of known framework substitutions is rapidly increasing (see Haber et al., this volume). For example, Table 8 shows 15 different V_H sequences. If each framework substitution requires an additional germ-line gene, then 12 germ-line V_H genes would be required (for the V_L sequences, 12 germ-line genes would be required for 12 different sequences [Table 9]). Again, an important question is whether one substitution in the first 27 residues generally signals multiple substitutions throughout the remainder of the V region.

The correlation of amino acid sequence in framework and hypervariable regions renders unlikely the combinatorial germ-line theory. A simple form of this theory proposes that the framework and hypervariable regions of each immunoglobulin chain are encoded by separate genes in the germ line and that the hypervariable regions are inserted into a framework gene during the differentiation of antibody-producing cells to generate a complete V gene – much as the V and C genes join to produce a complete light or heavy gene (Wu and Kabat 1970; Capra and Kindt 1975; Capra et al., this volume). This view is based in part on the belief that only the hypervariable regions determine the antigen-binding specificity since the framework regions are remarkably invariant in their tertiary structure (Davies et al. 1975; Poljak et al., this volume). If this theory is correct, a given set of hypervariable regions (e.g., those coding for phosphorylcholine binding) should be able to insert into any framework sequence to produce a phosphorylcholine-binding antibody. Yet particular V_H and V_L framework sequences (i.e., residues 1–27) always appear associated with the same hypervariable sequences (Tables 8 and 9). Accordingly, the combinatorial germ-line model appears to require that each of the antibody specificities examined to date employ distinct hypervariable genes associated with distinct framework genes. This additional ad hoc association renders this simple version of the theory unlikely. An alternative possibility for this theory is that the frame-

Table 9. The N-terminal Amino Acid Sequences of V_L Regions Derived from Myeloma Proteins Binding Various Haptens

Protein	Chain type	Position	Specificity (subspecificity)	Ref.[a]
		←HV$_1$→		
		1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 a b c d e f 31 32 33 34 35		
T15 prototype		D I V M T Q S P T F L A V T A S K K V T I S C T A S Z S L Y S S K H K V H Y L A W		
T15	κ	———	PC(A)	6
H8	κ	———	PC(A)	6
S107	κ	———	PC(A)	6
M603	κ	——S—S——S——S——G—E—R———M———K—S———L—B——G—B—Z—K—B—F———	PC(C)	6
W3207	κ	——S—S——S——S——G—E———M———K—S———L—B——	PC(?)	1
M511	κ	——I———D—E—L—S—K—P———S—G—E—S——S———R—S——K———L—Y—K—D—G——T—[ ]———N—	PC(B)	1
M167	κ	——I———D—E—L—S—N—P———S—G—E—S——S———R—S——K———L—Y—K—B—G——T—[ ]———B—	PC(B)	6,1
M460 prototype		D V V M T Q T P L S L T V S L G D R A S I S C R S S Q L V H S T B G B [ ] Y L H W		
M460	κ	———[ ]———	DNP	1
M315	λ	Z—A——V———Q—S—A—[ ]———T——P——G—T—T—V—L—T———T—G—A—V—T—[ ]—T—S—N——A—N—	DNP	3
W3082 prototype		D V Q M I Q S P S S L S A S L G D I V T M T C Z A S Z G T B I B [ ] L B W		
W3082	κ	———[ ]———	2,1L	1
J606	κ	———(K)———(S)———[ ]———	2,1L	1
M104E prototype		Z A V V T Q Q S A [ ] L T T S P G E T T V L T C R S S T G A V T [ ] T S N Y A N W		
M104E	λ	———[ ]———	1,3D	8
J558	λ	———[ ]———	1,3D	8
U10 prototype		D I Q M T Q T T S S L S A S L G D R V T I S C R A S Z B I S B [ ] Y L B W		
U10	κ	———[ ]———	2,6L*	S.R.
Y5476	κ	——?———[ ]———	2,6L*	1
M173	κ	———	Unknown	1
S117 prototype		E I V L T Q S P A I T A A S L G Q K V T I T C A A ? ? S V S [ ] Y M B W		
117	κ	———	Ga	1
S10	κ	———	1,6G	5
X44	κ	———	1,6G	5
T191	κ	———	1,6G	5
J1	κ	———	1,6G	5

Subspecificities, given in parentheses, are those which will be discussed in P. Barstad et al. (in prep.). The prototype sequences are those of one V_L region in the related set so as to facilitate the comparison of similar sequences.
* See note to Table 8.

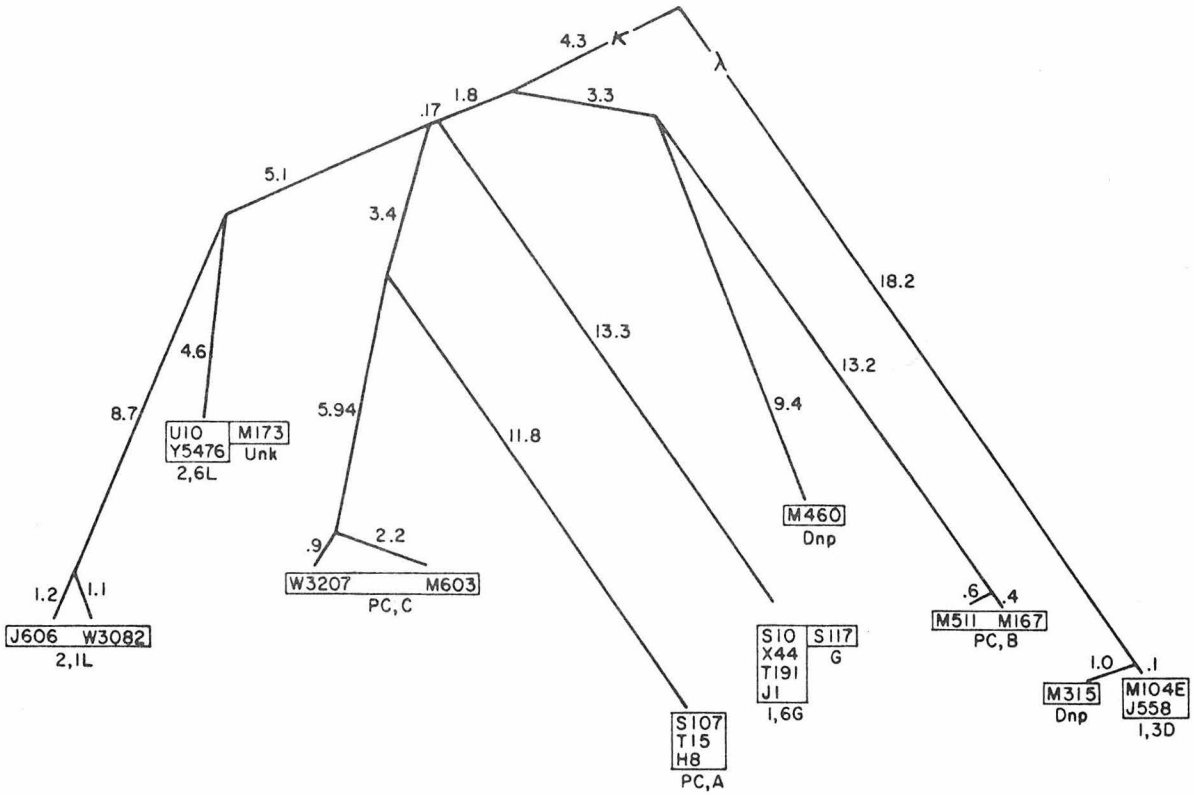


Figure 5. A genealogical tree for the N-terminal portion of κ chains from BALB/c myeloma proteins binding various haptens. See legend to Fig. 4.

work residues *do* influence the structure of the antigen-binding site, and, accordingly, selection can occur for the framework as well as the hypervariable regions.

Closely Related Sets of Complete V-region Sequences Place Constraints on Theories of Antibody Diversity

The V_H regions from myeloma proteins binding phosphorylcholine constitute a set of closely related immunoglobulin V regions. Figure 6 gives the complete V-region sequences for nine heavy chains derived from myeloma proteins binding phosphorylcholine. Six of these sequences have been determined in our laboratory. Several interesting features emerge from these data. First, four of these V_H regions are identical (S107, S63, T15, Y5236), whereas the variants differ by 1 to 11 amino acid substitutions and in some cases sequence insertions or deletions. The nine proteins contain a total of 27 substitutions from the T15 V_H region, requiring 19 one-base and 8 two-base changes. Second, sequence gaps occur in four of these six sequences, all localized in the third hypervariable region. Compared to T15, M603 has a deletion, M511 an insertion, and M167 two insertions. Third, the amino acid substitutions are concentrated in the second and third hypervariable regions with only about 30% occurring in the

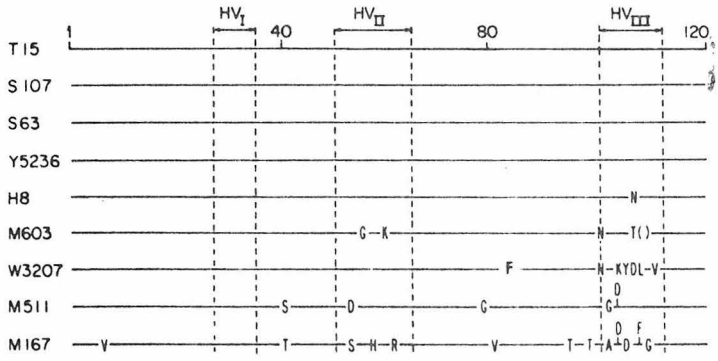


Figure 6. The complete amino acid sequences of nine mouse V_H regions from myeloma proteins binding phosphorylcholine. The three active-site-associated hypervariable regions are indicated by HV_I, HV_{II}, and HV_{III}. Sequence gaps are indicated by parentheses. Sequence insertions are designated by a vertical bar which lifts the residue above the main sequence. The various residues are indicated by the one-letter code. T15 and S107 are from Rudikoff and Potter (1976). S63 and Y5236 are from J. Hubert et al. (in prep.). M603 and M167 are from Rudikoff and Potter (1974, 1976). M511 is from E. Appella (pers. comm.).

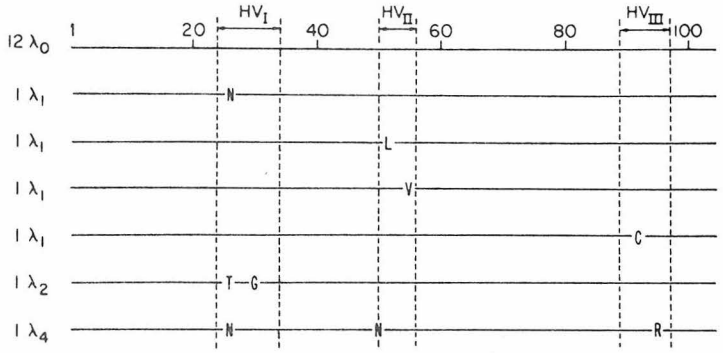


Figure 7. The amino acid sequences of 18 mouse V_λ regions. See legend to Fig. 6. (Reprinted, with permission, from Cohn et al. 1974.)

framework region. Seven of the 19 hypervariable substitutions and one of the eight framework substitutions are two-base changes. The six different V_H regions shown in Figure 6 contain four different framework sequences. Finally, there is a block of six amino acid substitutions that occurs over ten residues in the third hypervariable region of W3207, and half of these are two-base substitutions from the T15 codons. The implications of these features of diversity are discussed in subsequent sections.

One set of closely related V regions has been described for each immunoglobulin family in the mouse. In addition to the V_H regions derived from myeloma proteins binding phosphorylcholine, two other sets of closely related V regions have been completely sequenced—the mouse λ chains (Fig. 7) and the mouse T124-like κ chains (Fig. 8). As mentioned earlier, 18 λ chains have been examined, and 12 appear to be identical to one another (λ_0) (Cohn et al. 1974). Of the six variants, four exhibit a single amino acid substitution (λ_1), a fifth shows two amino acid substitutions (λ_2), and a sixth has three amino acid substitutions (λ_4), one of which requires two nucleotide substitutions (Fig. 7). All substitutions occur in the hypervariable regions.

Three mouse κ chains fall into the T124 subgroup (McKean et al. 1973). A prototype sequence can be determined for all positions except 27d by determining what residue a majority of the V regions have at each position (Fig. 8). Two V_κ regions differ from the prototype by a single residue, and the third differs by six residues. All eight substitutions may be accounted for by single nucleotide substitutions. Four out of six substitutions occur outside the hypervariable regions. We noted earlier that one of the NZB V_κ sequences may also fall into this subgroup (see Table 7). The nature of the substitutions from these three sets of V regions is summarized in Table 10.

The sets of closely related V genes place constraints on certain theories of antibody diversity. Taken together, these three closely related sets (subgroups) of sequences provide definite patterns of variability which the various theories of diversity must explain. A number of interesting observations can be made which constrain the theories. First, from the number of different substitutions and their locations, one can calculate the number of germ-line genes the various theories need to explain these data (Table 11). The general guidelines for counting germ-line V genes are given in Table 1. The combinatorial germ-line, classical germ-line, and ordinary somatic mutation theories require, respectively, 35, 16, and 9 germ-line genes for 16 different V sequences. Thus all of these theories need large numbers of V genes. The additional diversity introduced by ordinary somatic mutation appears to amplify the V-region sequence diversity by less than a factor of 2 (i.e., 9 genes for 16 V regions). However, special mutational mechanisms probably will require fewer germ-line genes. Second, 30% of the substitutions occur in framework regions (Table 10). Thus any special mutational mechanism must operate on framework regions as well as hypervariable regions (unless one postulates that the special mutational mechanism starts from a large number of germ-line genes which differ in size). Third, the frequency of gaps is very high in the V_H regions (four in six different sequences) (Table 10). This implies that all mutational mechanisms, somatic or germ line, must be

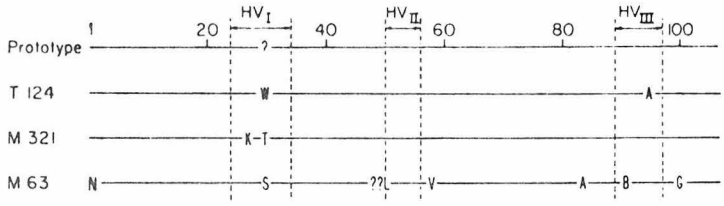


Figure 8. The amino acid sequences of three T124-like mouse V_κ regions. See legend to Fig. 6.

Table 10. Nucleotide Substitutions Occurring in the Three Sets of Closely Related V Regions

	V_H	V_κ	V_λ	Total
Total amino acid substitutions	27	10	9	46
Hypervariable amino acid substitutions	19	7	9	35
Framework amino acid substitutions	8	3	0	11
Two-base substitutions	7	0	1	8
Sequence gaps	4	0	0	4

Table 11. Number of Germ-line V Genes Required to Encode the Three Closely Related Sets of V Genes by Various Theories of Antibody Diversity

	V_H		V_κ		V_λ		Total	
Theory	no. of sequences	minimum no. of genes	no. of sequences	minimum no. of genes	no. of sequences	minimum no. of genes	no. of sequences	minimum no. of genes
Classical germ line	6	6	3	3	7	7	16	16
Combinatorial germ line	6	4 (Fr) 11 (Hv)[a]	3	2 (Fr) 8 (Hv)	7	1 (Fr) 10 (Hv)	16	7 (Fr) + 29 (Hv) = 36
Ordinary mutation (Cohn)	6	6	3	2	7	1	16	9
Recombination	6	3	3	2	7	3	16	8
Special mutation	6	?	3	1–?	7	1	16	?

The rationale for counting germ-line V genes is given in Table 1.
[a] Fr designates framework genes; Hv denotes hypervariable genes.

capable of generating frequent sequence gaps. Fourth, identical V_H regions can occur (e.g., V_H T15) in many individuals, so at least some germ-line genes are expressed without modification. Fifth, one can note that in the closely related V_H subgroup (Fig. 6), the V region that is most different from the remainder of the set (V_H M167) differs by 11 single-base substitutions and two sequence gaps. This V region constitutes an interesting test for somatic mutation mechanisms. Can the putative somatic mechanism generate 13 mutational events in the V_H T15 gene during differentiation to produce a V_H M167 gene? Sixth, it has been suggested that special mutation mechanisms may be revealed by careful examination of patterns of nucleotide changes. An analysis of the types of nucleotide substitutions that occur in these closely related sets of V genes (Table 12) reveals a tendency to favor the $X \rightarrow$ purine substitutions, where X is any base (29/38 in the framework region and 9/12 in the hypervariable regions). However, this tendency does not appear to reflect any special mutational mechanism in V genes; when 78 mutants of human hemoglobin α chains and 113 mutants of human hemoglobin β chains (Dayhoff 1972) are analyzed in a similar manner, the same tendency is noted. A statistical analysis confirms the supposition that the $X \rightarrow$ purine nucleotide substitutions in the globins and immunoglobulins are similar (J. M. Hood, pers. comm.).

An important question can be asked of these sets of closely related V regions. Namely, do they exhibit a tree-like genealogic structure (e.g., the $V_{\kappa D3}$ branch in Fig. 1) or do they all appear to be derived from one or a few prototype sequences? If these sets show a genealogic structure, somatic mutation models would require extensive parallel mutation. If these sets show divergence from one or a few prototype genes, the germ-line theory would require that a prototype gene be duplicated many times (i.e., saltatory replication) and that each of the duplicated genes diverge independently from the others. With the paucity of variants available in the V_λ and V_H sets, it is difficult to determine whether or not they exhibit a genealogic structure. Because large numbers of variants are potentially available in the M70 and T124 sets, these appear to be an ideal system for testing whether or not closely related sets of V regions have a genealogic structure.

Nucleic Acid Hybridization Studies Suggest That There Are Multiple Germ-line V Genes and That Somatic Mutation May Occur

Clearly, the number of germ-line V genes within a subgroup can be determined most directly by

Table 12. Types of Nucleotide Substitutions in Sets of Closely Related Proteins

Type of change	Hypervariable	Framework	Total
U → X	8	2	10
C → X	9	4	13
A → X	11	4	15
G → X	13	2	15
X → U	4	1	5
X → C	5	2	7
X → A	17	3	20
X → G	12	6	18

Substitutions from the V_H, V_λ, and V_κ sets given in Figs. 6, 7, and 8 are summed for this table.

studies at the nucleic acid level. Messenger RNA has been isolated from a number of mouse κ and λ myeloma tumors (Swan et al. 1972; Milstein et al. 1974; Tonegawa et al. 1974). These mRNAs or their complementary (c)DNA copies can be hybridized to myeloma or other somatic or germ-line DNA. Under conditions of excess genome DNA, the kinetics of the hybridization reaction are theoretically related to the number of genes in the DNA complementary to the mRNA or cDNA probe (see Hood et al. 1975b, Appendix to Chapter 5). These studies have concluded that BALB/c mouse V_λ regions and individual V_κ subgroups appear to be encoded by distinct germ-line genes. These data are consistent with one to a few identical or closely related V genes for each subgroup (Rabbitts et al. 1975; Tonegawa 1976). Statistical calculations suggest that for V_λ this is fewer genes than would be predicted by a germ-line theory. Accordingly, it is postulated that somatic mutation generates most of the V_λ variants, although the functional significance of these variants has not been demonstrated. There is one important reason for being cautious in accepting this conclusion. There are serious limitations on the accuracy of hybridization kinetics in determining the number of related (but not identical) germ-line V genes in conditions of vast DNA excess (for a more thorough discussion of this point, see Smith, this volume). Indeed, Smith contends that it is difficult to distinguish between 1 and 20–30 V_λ genes based on the hybridization data now available.

Techniques other than hybridization kinetics may prove more suitable for resolving the presence of specific V-region genes in the germ line. Nucleic acid probes can be used to detect particular sequences in the various DNA fragments produced by restriction endonuclease digestion of genome DNA. Recently, this technique was used (Tonegawa et al., this volume) to demonstrate V–C joining at the DNA level in lymphocyte differentiation. If related genes from a single subgroup are present in the germ line, it is possible that they would be detected in different sized restriction fragments. Although the nucleic acid probe may not distinguish between these sequences, presumably only one would join to the C region in DNA from the appropriate myeloma. It will soon be possible to obtain quantities of germ-line V gene DNA sufficient for direct nucleic acid sequencing. This promises to detect whatever heterogeneity exists in DNA hybridizing to probes from a particular subgroup.

Four General Strategies Exist for the Storage and Generation of Information in the Antibody Gene Families

The immune system of vertebrate organisms is capable of storing or generating a great deal of information—the information required to produce antibody molecules recognizing many different antigens. The basic strategies for storage and generation of this information fall into two general categories (Table 13). Genetic strategies amplify information by producing multiple antibody V genes, whereas molecular strategies amplify information by employing certain fundamental characteristics of antibody molecules themselves. Unfortunately, molecular immunology has focused on the genetic mechanisms discussed above almost to the exclusion of the molecular mechanisms. There are two molecular mechanisms for expanding the number of functional antibody molecules (Table 13).

Combinatorial association is possible because antibody molecules are made up of nonidentical subunits (light and heavy chains). This mechanism suggests that a single light chain may associate with many different heavy chains to produce many different antibody molecules. For example, if 1000 light and 1000 heavy chains can each associate freely with one another, 10^6 different antibody molecules would be produced. X-ray crystallographic studies suggest that appropriate residues are highly conserved at the sites of V_L and V_H interaction so that virtually any light chain might associate with virtually any heavy chain (Poljak et al. 1975).

There is evidence to suggest that when separated heavy and light chains are reassociated to form whole antibody molecules, homologous recombination (the recombination of heavy and light chains originally from the same molecule) is preferred to heterologous recombination (Olander and Little 1975). This does not necessarily argue against combinatorial association, since the preference may be kinetic. That is, certain light and heavy chains may associate more rapidly, but since the antibody-producing cell expresses a single light and a single heavy chain, perhaps the kinetics of association are unimportant. Somatic cell hybrids formed between two different myeloma cells or myeloma cells and antibody-producing cells suggest that many different light–heavy combinations are permissible (Schwaber and Cohen 1974). For example, 9 out of a possible 11 light–heavy combinations were seen in various hybrids created at Cambridge (C. Milstein, pers. comm.). Thus combinatorial association may be a powerful strategy for amplifying a given amount of germ-line information.

Table 13. Mechanisms for Storing and Generating Information

Strategy	Category
Multiple germ-line genes	genetic
Somatic mutation	genetic
Combinatorial association	molecular
Multispecificity	molecular

Multispecificity indicates that a single antibody molecule may interact with a variety of different antigens, some presumably related in tertiary structure and others possibly unrelated (Inman 1974; Richards et al. 1975). The antigen-binding crevice is a shallow trough ~15 Å × 10 Å × 7 Å (Amzel et al. 1974; Davies et al. 1975). Accordingly, there is room in this active site for many different molecular interactions. Indeed, the fact that many common enzymes interact through their active sites with a variety of structurally dissimilar hydrophobic probes supports this hypothesis (Glazer 1970). If individual antibody molecules may combine with 10 to 100 different antigenic patterns, a fixed number of germ-line V genes may interact with a disproportionately large fraction of the foreign antigenic universe. Surprisingly, few serious experiments have been directed at assessing the relative contributions of combinatorial association and multispecificity to the antibody repertoire. These molecular strategies appear to offer extremely important mechanisms for information generation in multigene families.

Clearly, multispecificity and combinatorial association will have to be studied at the level of antibody molecules since their diversity arises as a result of the interaction of two nonidentical polypeptides—a relationship which does not exist at the nucleic acid level. In contrast, questions of antibody gene numbers and organization can be approached most effectively by the techniques of nucleic acid chemistry.

SUMMARY

The mouse myeloma data presented in this paper add to our understanding of the organizational features and diversity patterns of the antibody gene families: (1) The NZB and BALB/c myeloma proteins differ in their structural and functional properties. The most likely explanation is that different subsets of lymphocytes are transformed by the myeloma process in these two strains. If so, the myeloma process may sample only a small subset of the antibody repertoire. Thus V-gene estimates based on myeloma data may be low. (2) The examination, through the first hypervariable region, of sets of myeloma proteins binding various haptens demonstrates that the corresponding V_L and V_H regions generally fall into subgroups characteristic for each specificity. Thus new subgroups (V genes) are defined for the variable regions of myeloma proteins binding most haptens examined. Indeed, the V_HIII subgroup of the mouse has now been divided into at least three separate subgroups by this analysis. (3) The closely related sets of completely sequenced V regions suggest that certain V regions are encoded directly by germline genes. These observations also place constraints on the putative somatic mutational mechanism.

The nucleic acid hybridization studies on the mouse λ family suggest that somatic mutation may generate the λ variants. Various studies also suggest that bona fide antibodies can be directly encoded by germ-line genes. Thus both of these genetic mechanisms, multiple germ-line genes and somatic mutation, may contribute to antibody diversity. Two other potential mechanisms for generating functional diversity—combinatorial association and multispecificity—have not been effectively assessed.

All of these strategies—multiple germ-line genes, somatic mutation, combinatorial association, and multispecificity—might be used by any multigenic family. The cardinal question for the future is, "What is the relative contribution of each mechanism to the total *functional* diversity of vertebrate antibody molecules?"

Acknowledgments

This work was supported by grants from the National Institutes of Health and the National Science Foundation. L. H. has a Research Career Development Award from the National Institutes of Health.

REFERENCES

AMZEL, L., R. POLJAK, F. SAUL, J. VARGA and F. RICHARDS. 1974. The three-dimensional structure of a combining-site ligand complex of immunoglobulin NEW at 3.5-Å resolution. *Proc. Nat. Acad. Sci.* **71:** 1427.

BALTIMORE, D. 1974. Is terminal deoxynucleotidyl transferase a somatic mutagen in lymphocytes? *Nature* **248:** 409.

BARSTAD, P., V. FARNSWORTH, M. WEIGERT, M. COHN and L. HOOD. 1974. Mouse immunoglobulin heavy chains are coded by multiple germ line variable region genes. *Proc. Nat. Acad. Sci.* **71:** 4096.

BOURGOIS, A. and M. FOUGEREAU. 1970. Partial amino acid sequence of the variable region of a mouse G2a immunoglobulin heavy chain. Evidence for the existence of a third subgroup of variability for the heavy chain pool. *FEBS Letters* **8:** 265.

BRENNER, S. and C. MILSTEIN. 1966. Origin of antibody variation. *Nature* **211:** 242.

CAPRA, J. D. and J. M. KEHOE. 1975. Hypervariable regions, idiotypy and the antibody combining site. *Adv. Immunol.* **20:** 1.

CAPRA, J. D. and T. J. KINDT. 1975. Antibody diversity: Can more than one gene encode each variable region? *Immunogenetics* **1:** 417.

CAPRA, J. D., A. S. TUNG and A. NISONOFF. 1975. The heavy chains of anti-*p*-azophenyl arsonate antibodies from A/J mice bearing a cross-reactive idiotype. *J. Immunol.* **114:** 1548.

CAPRA, J. D., R. W. WASSERMAN and J. M. KEHOE. 1973. Phylogenetically associated residues within the V_{HIII} subgroup of several mammalian species: Evidence for a "pauci-gene" basis for antibody diversity. *J. Exp. Med.* **138:** 410.

CARSON, D. and M. WEIGERT. 1973. Immunochemical analysis of the cross-reacting idiotypes of mouse myeloma proteins with anti-dextran activity and normal anti-dextran antibody. *Proc. Nat. Acad. Sci.* **70:** 235.

CESARI, I. and M. WEIGERT. 1973. Mouse lambda-chain sequences. *Proc. Nat. Acad. Sci.* **70:** 2112.

COHN, M., B. BLOMBERG, W. GECKELER, W. RASCHKE, R. RIBLET and M. WEIGERT. 1974. First order considerations in analyzing the generation of diversity. In *The immune system: Genes, receptors, signals* (ed. E. Sercarz et al.), p. 89. Academic Press, New York.

DAVIES, D. R., E. A. PADLAN and D. M. SEGAL. 1975. Three-dimensional structure of immunoglobulins. *Ann. Rev. Biochem.* **44:** 639.

DAYHOFF, M. O. 1972. *Atlas of protein sequence and structure*, vol. 5. National Biomedical Research Foundation. Silver Spring, Maryland.

EDELMAN, G. M. 1974. Origins and mechanisms of specificity in clonal selection. In *Cellular selection and regulation in the immune response.* (ed. G. M. Edelman), p. 1. Raven Press, New York.

———. 1968. General discussion on theories of antibody variability. *Cold Spring Harbor Symp. Quant. Biol.* **32:** 169.

EICHMANN, K. 1975. Genetic control of antibody specificity in the mouse. *Immunogenetics* **2:** 491.

FRANCIS, S. H., R. G. Q. LESLIE, L. HOOD and H. N. EISEN. 1974. Amino-acid sequence of the variable region of the heavy (alpha) chain of a mouse myeloma protein with anti-hapten activity. *Proc. Nat. Acad. Sci.* **71:** 1123.

GALLY, J. A. 1973. Structure of immunoglobulins. In *The antigens.* (ed. M. Sela), vol. I, p. 162. Academic Press, New York.

GALLY, J. A. and G. M. EDELMAN. 1970. Somatic translocation of antibody genes. *Nature* **227:** 341.

GLAZER, A. N. 1970. On the prevalence of "non-specific" binding at the specific binding sites of globular proteins. *Proc. Nat. Acad. Sci.* **65:** 1057.

GRANT, J. A. and L. HOOD. 1971. N-terminal analysis of normal immunoglobulin light chains. I. A study of thirteen individual humans. *Immunochemistry* **8:** 63.

GRAY, W. R., W. J. DREYER and L. HOOD. 1967. Mechanisms of antibody synthesis: Size difference between mouse kappa chains. *Science* **155:** 465.

HOOD, L. 1973. The genetics, evolution and expression of antibody molecules. *Stadler Genet. Symp.* **5:** 73.

HOOD, L. and D. TALMAGE. 1970. Mechanism of antibody diversity: Germ line basis for variability. *Science* **168:** 325.

HOOD, L., J. H. CAMPBELL and S. C. R. ELGIN. 1975a. The organization, expression and evolution of antibody gene families and other multigene families. *Annu. Rev. Genet.* **9:** 305.

HOOD, L., J. H. WILSON and W. B. WOOD. 1975b. *Molecular biology of eucaryotic cells,* pp. 286–291. W. A. Benjamin, Menlo Park, California.

HOOD, L., P. BARSTAD, E. LOH and C. NOTTENBURG. 1974. Antibody diversity: An assessment. In *The immune system: Genes, receptors, signals* (ed. E. E. Sercarz et al.), p. 119. Academic Press, New York.

HOOD, L., W. GRAY, E. SANDERS and W. DREYER. 1968. Light chain evolution. *Cold Spring Harbor Symp. Quant. Biol.* **32:** 133.

INMAN, J. 1974. Multispecificity of the antibody combining region and antibody diversity. In *The immune system: Genes, receptors, signals* (ed. E. E. Sercarz et al.), p. 37. Academic Press, New York.

JERNE, N. K. 1971. The somatic generation of immune recognition. *Eur. J. Immunol.* **1:** 1.

KLINMAN, N. R. and J. L. PRESS. 1975a. The B cell specificity repertoire: Its relationship to definable subpopulations. *Transplant. Rev.* **24:** 41.

———. 1975b. The characterization of the B-cell repertoire specific for the 2, 4-dinitrophenol and 2, 4, 6-trinitrophenol determinants in neonatal BALB/c mice. *J. Exp. Med.* **141:** 1133.

MAGE, R., R. LIEBERMAN, M. POTTER and W. D. TERRY. 1973. Immunoglobulin allotypes. In *The antigens* (ed. M. Sela), vol. I, p. 300. Academic Press, New York.

MCKEAN, D., M. POTTER and L. HOOD. 1973. Mouse immunoglobulin chains. Pattern of sequence variation among κ chains with limited sequence differences. *Biochemistry* **12:** 760.

MILSTEIN, C. 1967. Linked groups of residues in immunoglobulin kappa chains. *Nature* **216:** 330.

MILSTEIN, C., G. G. BROWNLEE, E. M. CARTWRIGHT, J. M. JARVIS and N. J. PROUDFOOT. 1974. Sequence analysis of immunoglobulin light chain messenger RNA. *Nature* **252:** 354.

OLANDER, J. and J. R. LITTLE. 1975. Preferential homologous recombination of H and L chains from mouse myeloma proteins which bind Dnp ligands. *Immunochemistry* **12:** 383.

PADLAN, E. A., D. M. SEGAL, G. H. COHEN and D. R. DAVIES. 1974. The three-dimensional structure of the antigen-binding site of McPc 603 protein. In *The immune system: Genes, receptors, signals* (ed. E. E. Sercarz et al.), p. 7. Academic Press, New York.

POLJAK, R. J., L. M. AMZEL, B. L. CHEN, R. P. PHIZACKERLEY and F. SAUL. 1975. Structural basis for the association of heavy and light chains and the relation of subgroups to the conformation of the active site of immunoglobulins. *Immunogenetics* **2:** 393.

POLJAK, R., L. AMZEL, H. AVEY, B. CHEN, R. PHIZACKERLEY and F. SAUL. 1973. Three-dimensional structure of the Fab' fragment of a human immunoglobulin at 2.8-Å resolution. *Proc. Nat. Acad. Sci.* **70:** 3305.

POTTER, M. 1970. Antigen-binding myeloma proteins in mice. *Ann. N.Y. Acad. Sci.* **190:** 306.

———. 1972. Immunoglobulin-producing tumors and myeloma proteins of mice. *Physiol. Rev.* **52:** 631.

PRESS, J. L. and N. KLINMAN. 1974. Frequency of hapten-specific B cells in neonatal and adult murine spleens. *Eur. J. Immunol.* **4:** 155.

RABBITTS, T. H., J. M. JARVIS and C. MILSTEIN. 1975. Demonstration that a mouse immunoglobulin light chain messenger RNA hybridizes exclusively with unique DNA. *Cell* **6:** 5.

RICHARDS, F., W. KONIGSBERG, R. ROSENSTEIN and J. VARGA. 1975. On the specificity of antibodies. *Science* **187:** 130.

RUDIKOFF, S. and M. POTTER. 1974. Variable region sequence of the heavy chain from a phosphorylcholine binding myeloma protein. *Biochemistry* **13:** 4033.

———. 1976. Size differences among immunoglobulin heavy chains from phosphorylcholine-binding proteins. *Proc. Nat. Acad. Sci.* **73:** 2109.

RUDIKOFF, S., E. B. MUSHINSKI, M. POTTER, C. P. J. GLAUDEMANS and M. E. JOLLEY. 1973. Six BALB/c IgA myeloma proteins that bind $\beta(1\rightarrow6)$ galactan. *J. Exp. Med.* **138:** 1095.

SCHULENBERG, E. P., E. S. SIMMS, R. G. LYNCH, R. A. BRADSHAW and H. N. EISEN. 1971. Amino acid sequence of the light chain from a mouse myeloma protein with anti-hapten activity: Evidence for a third type of light chain. *Proc. Nat. Acad. Sci.* **68:** 2623.

SCHWABER, J. and E. P. COHEN. 1974. Pattern of immunoglobulin synthesis and assembly in a human-mouse somatic cell hybrid clone. *Proc. Nat. Acad. Sci.* **71:** 2203.

SHER, A. and M. COHN. 1972. Inheritance of an idiotype associated with the immune response of inbred mice to phosphorylcholine. *Eur. J. Immunol.* **2:** 319.

SHERWIN, W. K. and D. T. ROWLANDS, JR. 1975. Determinants of the hierarchy of humoral immune responsiveness during ontogeny. *J. Immunol.* **115:** 1549.

SIGAL, N., P. J. GEARHART, J. L. PRESS AND N. KLINMAN.

1976. Late acquisition of a germ line antibody specificity. *Nature* **259:** 51.

SMITH, G. P., W. FITCH and L. HOOD. 1971. Antibody diversity. *Annu. Rev. Biochem.* **40:** 969.

SMITHIES, O. 1968. The genetic basis of antibody variability. *Cold Spring Harbor Symp. Quant. Biol.* **32:** 161.

SWAN, D., H. AVIV and P. LEDER. 1972. Purification and properties of biologically active messenger RNA for a myeloma light chain. *Proc. Nat. Acad. Sci.* **69:** 1967.

TONEGAWA, S. 1976. Reiteration frequency of immunoglobulin light chain genes: Further evidence for somatic generation of antibody diversity. *Proc. Nat. Acad. Sci.* **73:** 203.

TONEGAWA, S., C. STEINBERG, S. DUBE and A. BERNARDINI. 1974. Evidence for somatic generation of antibody diversity. *Proc. Nat. Acad. Sci.* **71:** 4027.

WEIGERT, M., M. POTTER and D. SACHS. 1975. Genetics of the immunoglobulin V region. *Immunogenetics* **1:** 511.

WU, T. and E. KABAT. 1970. An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma chains and their implications for antibody complementarity. *J. Exp. Med.* **132:** 250.

Appendix B

This paper was published in Biochemical Genetics

A Mathematical Approach to the Analysis of Diversity in Antibody Gene Families

J. M. Hood,[1] E. Y. Loh,[2] and L. Hood[2]

Received 13 Oct. 1975—Final 24 Nov. 1975

In this article, we develop a mathematical approach for the analysis of diversity in antibody gene families. This approach is arrived at by examining two general questions about protein populations: (1) What is a relative measure of the diversity exhibited by one protein family when compared with a second? (2) What is the probability that two protein populations were derived from a single common population? These quantitative approaches permit a variety of precise evolutionary, genetic, and developmental questions to be asked of antibody gene families. Using this methodology, we demonstrate that the diversity in mouse κ-immunoglobulin chains is considerably greater than in their human κ counterparts. We also show that the variable (V_L) regions of light chains associated with IgG and IgA immunoglobulins in the mouse appear to have been derived from a common population of V_L genes. This approach also can be used to analyse sequence data from other informational multigene families.

KEY WORDS: antibody gene families; probability matrix; diversity distance; pool variation.

INTRODUCTION

The analysis of amino acid sequence patterns of immunoglobulins has led to important insights into the structure, genetics, differentiation, control, and evolution of antibody molecules (Hood *et al.*, 1975; Gally and Edelman, 1972; Hood, 1973). Each antibody molecule is composed of light and heavy

This work was supported by grants from the National Science Foundation (BMS 71-00770) and the National Institutes of Health (AI 10781). L. H. has a Research Career Development Award from NIH (AI 203 88).

[1] Department of Mathematics, Occidental College, Los Angeles, California.
[2] Division of Biology, California Institute of Technology, Pasadena, California.

polypeptide chains (Smith *et al.*, 1971). Antibody polypeptides can be divided into an *N*-terminal variable (V) region that codes for the antigen-binding function and a *C*-terminal constant (C) region that codes for various effector functions (Edelman *et al.*, 1969). The V and C regions of each antibody polypeptide are coded by separate genes (see Hood, 1972). Thus the V and C genes must be joined at some stage of the differentiation of the antibody-producing cell in order to be translated into individual protein molecules containing both the V and C regions (Milstein *et. al.*, 1974). There are three families of genes unlinked in the mammalian genome that code for antibody molecules; two code for light chains, λ and κ, and the third for heavy chains (see Mage *et al.*, 1973). Each antibody family has an unknown number of V genes and approximately one to ten C genes (Hood, 1973). The amino acid sequence of the constant region of the heavy chain (C_H) determines the class of immunoglobulin to which a particular antibody molecule belongs (e.g., IgM-C_μ, IgG-C_γ, IgA-C_α) (Gally, 1973).

An analysis of the patterns of diversity in antibody molecules has provided important insights into the evolution of antibody gene families (Smith, 1973; Hood *et al.*, 1975; Gutman *et al.*, 1975), the generation of antibody diversity (Gally and Edelman, 1970; Cohn, 1974, Hood, 1973), and structure–function correlations for antigen binding (Wu and Kabat, 1970; Barstad *et al.*, 1974, 1975). Initially, diversity patterns were analyzed by visual inspection of the amino acid sequence data. This approach led to the discovery of V and C regions (Hilschmann and Craig, 1965; Putnam, 1969), V-region subgroups (Hood *et al.*, 1967; Milstein, 1967), species-specific residues (Doolittle, 1966; Hood *et al.*, 1970), and the three families of antibody polypeptides (Gally and Edelman, 1970). Later the computer was used to analyze immunoglobulin sequences for regions of homology and extreme variability and for their genealogical relationships. These approaches demonstrated that immunoglobulins are composed of homology units (Edelman *et al.*, 1969) and that the V regions of immunoglobulins have hypervariable segments which X-ray crystallography later established fold to form the walls of the antigen-binding site (Wu and Kabat, 1970; Amzel *et al.*, 1974). The genealogical analysis led to a detailed analysis of antibody diversity within individual antibody gene families (Smith *et al.*, 1971; Smith, 1973; Hood *et al.*, 1975).

With the advent of the protein sequenator, enormous numbers of immunoglobulin sequences have been determined from a variety of different species (see Gally, 1973). Accordingly, it has become important to develop quantitative approaches for analyzing individual families of gene products (e.g., mouse V_κ regions) and for comparing them with other families of gene products (e.g., human V_κ regions). Indeed, Fig. 1, which depicts the *N*-terminal sequences of V_κ regions from the human and the mouse, illustrates the magnitude of this problem.

In this article, we develop techniques which are (1) capable of quantifying the diversity exhibited in a particular gene family and (2) capable of determining whether one population of proteins may have been derived from a second population of proteins. We then illustrate the use of these techniques on two independent problems: (1) a comparison of the diversity of human and mouse V_κ regions and (2) the determination as to whether the mouse V_L regions associated with α heavy chains might have been drawn from the same population as those associated with γ heavy chains.

Our basic approach is to display the diversity of a given protein population as a probability matrix which gives the likelihood that any random protein from the population will have a particular amino acid at a specific residue position. From these probability matrices, we define the "diversity distance index" or simply "distance index" between two different protein populations. Next we find the average distance between a given protein pool and each of its members. This average distance is designated the "variation index" of the pool. Finally, we employ statistical methods to analyze the significance of a specific distance between two populations and compare their individual variation indices.

THE MATHEMATICAL ANALYSIS

Distance and Variation Indices

We begin our analysis by first quantifying a single protein chain. This is done by assigning a matrix to each individual protein in the following manner: The 20 amino acids found in proteins are numbered 1 through 20. If a protein chain has T residues, then the (i,j) entry of a $20 \times T$ matrix is assigned the number 1 if the protein has the ith amino acid in the jth residue position. Otherwise, the (i,j) entry is given the value 0. Gaps and unknown residues are also given the value 0. Thus each protein corresponds to a $20 \times T$ matrix, all of whose entries are either 0 or 1 (Fig. 2).

This quantification can be extended from a single protein to n proteins in a given population. We will denote this population by R. Define n_{ij} to be the number of proteins in R that have the ith amino acid in the jth residue position. Let $p_{ij} = n_{ij}/n$. Thus p_{ij} is the probability that if a protein is drawn at random from R it will have the ith amino acid in the jth residue position. The $20 \times T$ matrix (p_{ij}) will be called the "amino acid vs. position probability matrix for R," and it will be noted by APP(R) (Fig. 3).

We now detail a procedure for comparing the APP matrices for two populations, R and S. We define the "distance index" or "degree of diversity" between these two populations as $D(R,S) = \sum_{j=1}^{T} D_j(R,S)$, where $D_j(R,S) = \sum_{i=1}^{20} |p_{ij} - q_{ij}|$, and p_{ij} and q_{ij} are the entries of APP(R) and APP(S),

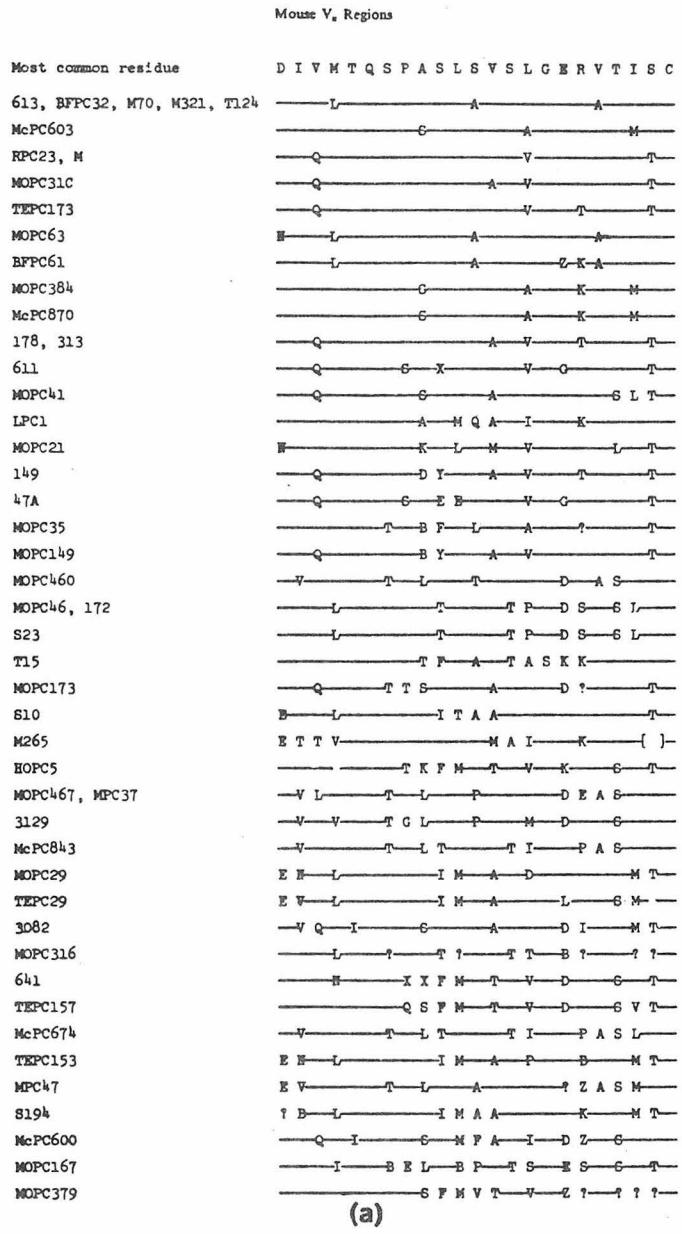
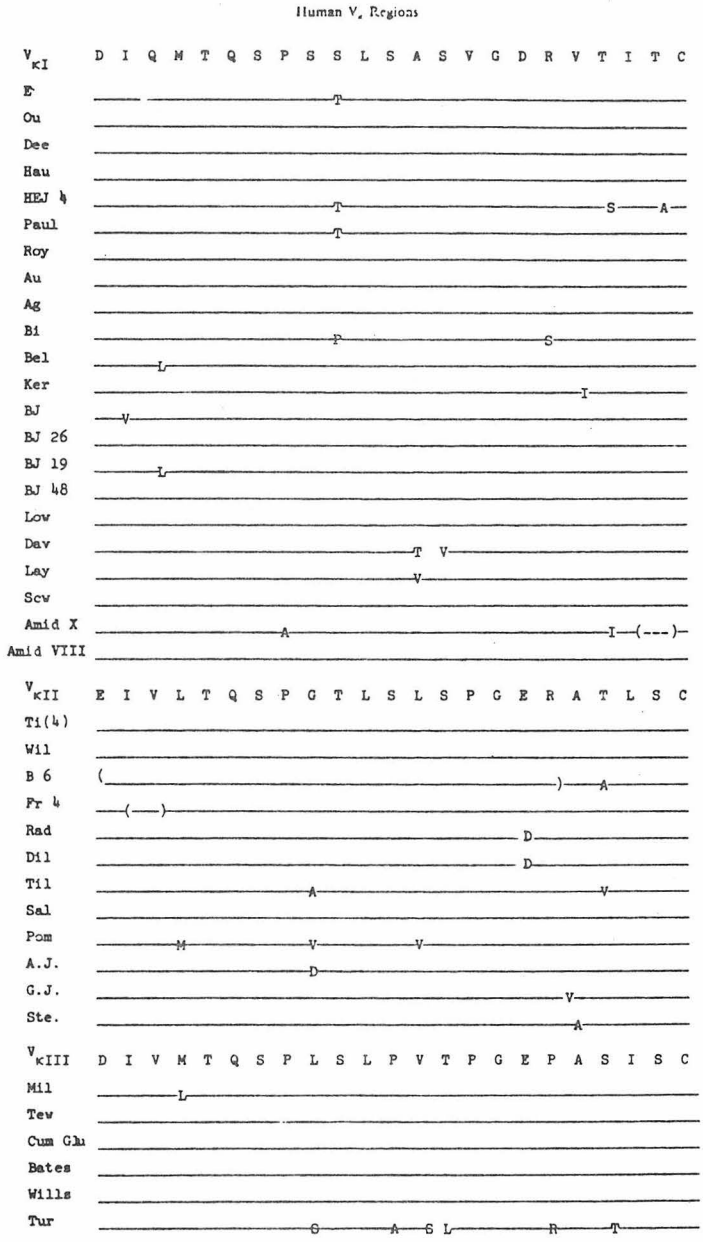


Fig. 1. *N*-terminal sequences of mouse and human V_κ regions. The one-letter code of Dayhoff is used. A horizontal line indicates that the particular sequence is identical to the prototype sequence above. The mouse sequences are compared against a prototype sequence of the most common residues. The human



(b)

sequences are compared against three prototype (subgroup) sequences. The mouse data (a) were taken from Hood *et al.* (1974) and the human data (b) from Gally (1973).

	1	2	3	4	5	.	.	.	T
Sequence	D	I	V	M	T	.	.	.	
Amino Acids (i)	Positions (j)								
	1	2	3	4	5	.	.	.	T
Alanine	0	0	0	0	0	.	.	.	
Aspartic Acid	1	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Isoleucine	0	1	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Methionine	0	0	0	1	0	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Threonine	0	0	0	0	1	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Valine	0	0	1	0	0	.	.	.	

Fig. 2. The matrix of an individual protein (see text).

respectively. $D_j(R,S)$ is designated as the distance between populations R and S at the jth residue position and $D(R,S)$ is the total distance between the two populations R and S. When R and S are understood, $D_j(R,S)$ and $D(R,S)$ will be denoted by D_j and D, respectively. It can be seen that $0 \leq D_j \leq 2$ and $0 \leq D \leq 2T$. If p_{ij} and q_{ij} are nearly identical for all is and js, then the distance between R and S will be small, which indicates that the two protein populations are very similar to one another. In contrast, if for some j, D_j is near 2, then the two populations exhibit extensive diversity in their amino acids at the jth residue position.

This concept of diversity can be used to quantify the variation within a single population of proteins in the following manner: Let R be a protein population and let S be a subpool of R. Assume that S contains m proteins and denote S by $\{S_1, S_2, \ldots, S_m\}$, where each S_i represents a protein of the subpool S. If we consider each protein in S as a "pool" of size 1, we can compute the distance between R and each S_k; that is, we can find $D(R,\{S_k\})$ for $k = 1, 2, \ldots, m$. Define $W(R,S) = (1/m) \sum_{k=1}^{m} D(R,\{S_k\})$. $W(R,S)$ is called the "variation index" of S in R. $W(R,S)$ is a measure of the average

Population of Proteins	1	2	3	4	5	.	.	.	T
	D	I	V	M	T	.	.	.	
	D	I	L	M	T	.	.	.	
	D	V	V	M	T	.	.	.	
	D	I	V	M	T	.	.	.	
	E	I	V	M	T	.	.	.	

Amino Acids (i)	Positions (j)								
	1	2	3	4	5	.	.	.	T
Alanine	0	0	0	0	0	.	.	.	
Aspartic Acid	.8	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Glutamic Acid	.2	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Isoleucine	0	.8	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Leucine	0	0	.2	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Methionine	0	0	0	1	0	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Threonine	0	0	0	0	1	.	.	.	
.	0	0	0	0	0	.	.	.	
.	0	0	0	0	0	.	.	.	
Valine	0	.2	.8	0	0	.	.	.	

Fig. 3. The matrix of a population of proteins (see text).

distance between each element of S and the entire pool R. Again, we have that $0 \leq W(R,S) \leq 2T$.

If $R = S$, then $W(R,S)$ gives the average variation of an element of R. In this sense, $W(R,R)$ can be considered as measuring the "width" of R. $W(R,R)$ will be used to determine the internal variation of single populations of proteins. Thus a comparison of $W(R,R)$ and $W(R',R')$ for two distinct populations R and R' yields information as to which pool has the greater variation. Further, a small number for $W(R,R)$ would indicate that most of the proteins are very similar to one another.

Interpretation of the Indices

If two sets of proteins are being compared at T different residues positions, the distance index can vary from 0 to $2T$. While the values of 0 and $2T$ have an obvious biological meaning, the significance of specific intermediate values is not so clear. In interpreting intermediate values, we found it convenient to use a statistical approach. Given two protein pools, R and S, we determine the likelihood that the smaller pool is similar to a subset of the larger pool. Let R and S have n and m sequences, respectively, where m is less than n. This likelihood is found by comparing the distance index, $D(R,S)$ with the distribution of the indices $D(R,X)$, where X runs through all possible subsets of size m contained in R.

The actual calculation of all such $D(R,X)$'s is not feasible unless n and m are both small. However, we can calculate the mean and standard deviation of this distribution. Thus if a value $D(R,S)$ were several standard deviations greater than the mean, then there would be only the smallest probability that the pool S is similar to some subpopulation of R.

We denote the mean and standard deviation of the distribution of $D(R,X)$'s by $\mu(R,m)$ and $\sigma(R,m)$, respectively. In calculating these, we begin by first finding the mean, $\mu_j(R,m)$, and the standard deviation, $\sigma_j(R,m)$, of the distribution of $D_j(R,X)$s for the jth position. The formulas for all these expressions are given in the Appendix.

It can be shown that, for m and n large enough, the distribution of $D(R,X)$ for all subpools X of size m approximates a log normal distribution. This information, together with $\mu(R,m)$ and $\sigma(R,m)$, allows us to analyze the diversity index, $D(R,S)$, and determine precisely the probability that any set of m proteins, S, is similar to a subset of m proteins taken from the R population, i.e., the probability that S is drawn from R.

The variation index $W(R,R)$ is identical to $\mu(R,1)$. If S is a subpool of R, then $\sigma(R,1)$ can be used to analyze the significance of $W(R,S)$ and $W(S,S)$ by comparing them to $W(R,R)$. If $W(R,S)$ and $W(S,S)$ are very different from

Table I. Diversity Indices for a Population of Human (H) and BALB/c (B)

V_κ regions[a]
1. $W(H,H) = 10.95$
2. $W(B,B) = 18.62$
3. $D(B,H) = 10.45$
4. $\mu(B,40) = 2.4$
5. $\sigma(B,40) = 0.38$

[a] The B and H populations are given in Fig. 1.

$W(R,R)$ relative to $\sigma(R,1)$, then it is unlikely that S comes from the same pool as R.

RESULTS AND DISCUSSION

To illustrate the application of these diversity indices, we have chosen two examples. They illustrate the diversity comparison and the determination of whether two pools can be drawn from a common pool. The first is a comparison of the N-terminal 23 residues of V_κ regions of the myeloma immunoglobulins derived from 40 humans and 50 BALB/c mice (an inbred strain) (see Fig. 1). The data were tabulated using a computer and are summarized in Table I. Undetermined acid or amide residues (Glx and Asx) were counted as unknowns. Identical sequences were counted only once to avoid bias in selection of identical proteins for sequencing. Lines 1 and 2 in Table I indicate that the average diversity with the population of mouse V_κ regions is considerably greater than that in their human counterparts. This information is obvious from visual inspection of these data or from visual inspection of the corresponding genealogical trees (Smith *et al.*, 1971; Hood, 1973). However, our approach allows a quantification of this difference in diversity which can, as previously indicated, be employed to determine the nature of the relationship between these two populations. Line 3 in Table I shows a distance of 10.45 between the human and mouse V_κ regions, whereas line 4 gives the average distance between the BALB/c population and a subpool of size 40 (the same size as the human V_κ population) as only 2.0. Use of even the most crude estimate (Chebyshev's inequality) indicates that there is less than 0.002 probability that the human population could be a subpool of the BALB/c population. This again confirms what we know intuitively, that the V_κ genes in man and mouse have changed significantly over the evolutionary time period since the divergence of the two species (75 million years).

The second example compares the V_κ regions associated with 11 IgG immunoglobulins (population G) with those associated with 14 IgA immuno-

Table II. Diversity Indices for a Population of V_L Regions from 11 IgG (G) and 14 IgA (A) Immunoglobulins of the BALB/c Mouse[a]

1.	$W(B,B) = 18.62$
2.	$W(B,G) = 17.84$
3.	$W(B,A) = 19.73$
4.	$\sigma(B,1) = 2.83$
5.	$D(B,G) = 5.18$
6.	$\mu(B,11) = 5.94$
7.	$\sigma(B,11) = 0.97$
8.	$D(B,A) = 6.55$
9.	$\mu(B,14) = 5.55$
10.	$\sigma(B,14) = 0.92$

[a] The B population is the pool of V_κ regions derived from myeloma proteins of the BALB/c mouse and given in Fig. 1.

globulins (population A). The pool of all available V_κ regions derived from myeloma tumors in the BALB/c mouse is designated population B (Fig. 1). The corresponding diversity indices are summarized in Table II. A comparison of lines 1, 2, and 3 in Table II shows that there is no significant difference in the degree of variation in the three pools, B, A, G. The standard deviation for the variation, $\sigma(B,1)$, in the entire BALB/c population is 2.83; thus both $V(G,G)$ and $V(A,A)$ are less than 1 SD from $V(B,B)$. Moreover, not only are the variations of these populations similar but also lines 5 through 10 in Table II demonstrate that the distances between B and G and between B and A are what would be expected if all were derived from a common population. Thus we tentatively conclude by our analysis that subsets of V_κ regions from the κ family of the BALB/c mouse are *not* selectively associated with certain C_H regions. This implies that individual κ-chains are not restricted in their associations with the various classes of immunoglobulins. Obviously, this analysis should be extended to the entire V region and should include more V_H regions before any firm conclusions are drawn.

These mathematical approaches can be extended to many other aspects of molecular immunology. A search for hypervariable residues and polypeptide segments could be conducted by examining the variation index at individual residue positions throughout these polypeptide chains. It would be interesting to compare this approach to the analysis of hypervariable regions with that already published (Wu and Kabat, 1970). A second possibility is that V-region subgroups could be defined quantitatively by searching for subpools within the population whose individual members have very low distance indices. The concept of V-region subgroups has had important implications for theories of antibody diversity (Hood, 1973; Cohn, 1974), yet

subgroups have been defined only in qualitative terms in the past. Third, comparative measures of diversity in the antibody gene families of mouse and man will have important implications for the evolution of "information" in this complex and multigenic system (Hood, 1973). Fourth, interesting questions about the differentiation of antibody-producing cells can be approached. For example, are the V_H regions associated with various C_H regions drawn from a common or separate pools? The former would be consistent with contemporary models of the V_H gene switching from C_μ to C_γ to C_α during the differentiation of the antibody-producing cell (Sledge *et al.*, 1975). Finally, the distance between individual residue positions in homologous populations from two species (e.g., mouse and human V_κ regions) could be examined to search for species-associated residues. This concept has been important in considering mechanisms for the evolution of antibody gene families and once again this approach will permit a quantitative definition of species-associated residues. The overall importance of this approach lies in the fact that the diversity existing within populations of proteins can be compared quantitatively.

The mode of analysis described in this article does have several limitations. First, the quantification of populations of proteins using amino acid vs. position probability matrices loses the linkage relationships of the amino acid residues contained in individual proteins. However, it is precisely this loss of information when large populations are compressed into matrices that permits us to analyze the diversity indices for large populations of proteins. Second, we cannot determine the precise number of proteins that should be randomly selected to give the general properties of a pool because we do not know the pool size or the shape of the distribution curve. In spite of these limitations, this approach appears to be very useful in comparing the diversity indices of immunoglobulin gene families.

The antibody gene families are the only well-documented examples of a complex informational multigene family (Hood *et al.*, 1975). A multigene family is defined as closely linked clusters of genes whose individual members exhibit sequence homology and carry out related or overlapping functions. An informational family has gene members whose sequences (and functions), although homologous, differ from one another. It appears likely that in the future many additional informational multigene families will be described and characterized (e.g., *Ir* genes, cell surface differentiation molecules, immunoglobulin T-cell receptors). The methods that we have developed in this article can also be extended to analyze these other multigene families. It appears likely that some of these multigene families will be related to one another. Our methodologies could even be employed to begin to trace back the ancient relationships of interrelated multigene families.

As the amount of amino acid sequence information increases for the antibody gene families and other multigene families, new mathematical

approaches will have to be developed to extract the information contained in these gene products regarding the regulation, evolution, differentiation, and genetics of these gene systems. The two indices that we have defined, based on the amino acid vs. position probability matrices, have provided a method for quantifying both the distance between two protein pools and the internal variation of a single pool. Preliminary applications of these indices have proved useful in providing information about the relative differences of various sequences of antibody families.

APPENDIX

In this section, we give formulas for the mean and expectation of the distribution of $D(R,X)$, where X runs through all samples of size m taken from the population P. R contains n proteins and m is less than n. We first give the expression for the jth residue position. Let $\mu(R,m)$, $\mu_j(R,m)$, $\sigma(R,m)$, and $\sigma_j(R,m)$ denote the means and standard deviations of the $D(R,X)$ and $D_j(R,X)$ distributions, respectively.

Let n_{ij} represent the number of proteins of the population R that have the ith amino acid at the jth residue position. Recall that p_{ij} is the probability of finding the ith amino acid in the jth position. Finally, let $E[Y]$ denote the mathematical expectation of the random variable Y. Then

$$\mu_j(R,m) = E[D_j(R,X)] = \sum_{i=0}^{20} \sum_{r=0}^{K_j} \left(\frac{K_j}{r}\right) p_{ij}^r (1-p_{ij})^{(K_j-r)} \left| p_{ij} - \frac{r}{K_j} \right| \quad (1)$$

where K_j = minimum $\{m, n_{ij}\}$.

$$\sigma_j(R,m) = \{E[D_j(R,X)^2] - E[D_j(R,X)]^2\}^{\frac{1}{2}} \quad (2)$$

$$\mu(R,m) = \sum_{j=1}^{T} \mu_j(R,m) \quad (3)$$

$$\sigma(R,m) = \left\{ \sum_{j=1}^{T} (\sigma_j(R,m)^2) \right\}^{\frac{1}{2}} \quad (4)$$

ACKNOWLEDGMENTS

We thank the Occidental College Computing Center for their generous gift of computing time.

REFERENCES

Amzel, L., Poljak, R., Saul, F., Varga, J., and Richards, F. (1974). The three dimensional structure of a combining-site ligand complex of immunoglobulin NEW at 3.5-Å resolution. *Proc. Natl. Acad. Sci.* 71:1427.

Barstad, P., Rudikoff, S., Potter, M., Cohn, M., Konigsberg, W., and Hood, L. (1974). Immunoglobulin structure: Amino terminal sequences of mouse myeloma proteins with phosphorylcholine-binding activity. *Science* **183**:962.

Barstad, P., Hubert, J., Black, B., Eaton, B., Weigert, M., and Hood, L. (1975). Immunoglobulins with hapten-binding activity: Structure–function correlations and genetic implications. *Proc. Natl. Acad. Sci.* (submitted).

Cohn, M. (1974). A rationale for ordering the data on antibody diversification. In Brent, L., and Holborow, J. (eds.), *Progress in Immunology*, Vol. II, American Elsevier, New York, pp. 261–284.

Doolittle, R. (1966). The amino-terminal acid sequences of rabbit immunoglobulin light chains. *Proc. Natl. Acad. Sci.* **55**:1195–1201.

Edelman, G. M., Cunningham, B. A., Gall, W., Gottlieb, P., Rutishauser, U., and Waxdal, M. (1963). The covalent structure of an entire γG immunoglobulin molecule. *Proc. Natl. Acad. Sci.* **63**:78.

Gally, J. (1973). Structure of immunoglobulins. In *The Antigens*, Vol. I, Academic Press, New York, pp. 162–299.

Gally, J. A., and Edelman, G. M. (1970). Somatic translocation of antibody genes. *Nature* **227**:341.

Gally, J. A., and Edelman, G. M. (1972). The genetic control of immunoglobulin synthesis. *Ann. Rev. Genet.* **6**:1.

Gutman, G., Loh, E., and Hood, L. (1975). Structure and regulation of immunoglobulins: Kappa allotypes in the rat have multiple amino acid differences in the constant region. *Proc. Natl. Acad. Sci.* **72**:5046.

Hilschmann, N., and Craig, L. C. (1965). Amino acid sequence studies with Bence-Jones proteins. *Proc. Natl. Acad. Sci.* **53**:1403.

Hood, L. (1972) Two genes: One polypeptide factor fiction. *Fed. Proc.* **31**:177.

Hood, L. (1973). The genetics, evolution and expression of antibody molecules. *Stadler Symp.* **5**:73.

Hood, L., Gray, W. R., Sanders, B. G., and Dreyer, W. J. (1967). Light chain evolution. *Cold Spring Harbor Symp. Quant. Biol.* **32**:133.

Hood, L., Eichman, K., Lackland, H., Krauss, R., and Ohms, J. (1970). Rabbit antibody light chains and gene evolution. *Nature* **228**:1040.

Hood, L., Barstad, P., Loh, E., and Nottenburg, C. (1974). Antibody diversity: An assessment. In *The Immune System: Genes, Receptors, Signals*, Academic Press, New York, pp. 119–139.

Hood, L., Campbell, J. H., and Elgin, S. C. R. (1975). The organization, expression and evolution of antibody gene families and other multigene families. *Ann. Rev. Genet.* **9**:305.

Mage, R., Lieberman, R., Potter, M., and Terry, W. D. (1973). Immunoglobulin allotypes. In *The Antigens*, Vol. I, Academic Press, New York, pp. 300–377.

Milstein, C. (1967). Linked groups of residues in immunoglobulin κ chains. *Nature* **216**:330.

Milstein, C., Brown, M. G. C., Cartwright, E. M., Jarvis, J. M., and Proudfoot, N. J. (1974). Sequence analysis of immunoglobulin light chain messenger RNA. *Nature* **252**:354.

Putnam, F. (1969). Immunoglobulin structure: Variability and homology. *Science* **163**:633.

Sledge, C., Fair, D. S., Black, B., Kruger, R. G., and Hood, L. (1976). Antibody differentiation: Apparent sequence identity between variable regions shared by IgA and IgG immunoglobulins. *Proc. Natl. Acad. Sci.* **73**:923.

Smith, G. (1973). Unequal crossover and evolution of multigene families. *Cold Spring Harbor Symp. Quant. Biol.* **38**:507.

Smith, G., Hood, L., and Fitch, W. (1971). Antibody diversity. *Ann. Rev. Biochem.* **40**:969.

Wu, T., and Kabat, E. (1970). An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma chains and their implications for antibody complementarity. *J. Exp. Med.* **132**:211.