Mouse Immunoglobulin Heavy Chains Are Coded by Multiple Germ Line Variable Region Genes

(amino-acid sequence/sequenator/BALB/c mouse/hapten binding/selection)

P. BARSTAD*, V. FARNOWORTH*, M. WEIGERT†, M. COHN‡, AND L. HOOD*

* Division of Biology, California Institute of Technology, Pasadena, Calif. 91109; † Institute for Cancer Research, Philadelphia, Pennsylvania 19111; and ‡ Salk Institute, San Diego, California 92121

Communicated by Ray D. Owen, June 20, 1974

ABSTRACT The N-terminal 20 residues of 13 heavy immunoglobulin chains from myeloma proteins of the BALB/c mouse are compared with the same residues of 15 other heavy chains described in the literature. Sixteen of 28 sequences are different from one another. These proteins fall into four major sets, with 18 of the proteins in the largest set being further divisible into at least five subsets. This pattern of diversity suggests there are at least eight germ line genes coding for the variable regions of mouse heavy chain. Many of the immunoglobulins from which these heavy chains are derived exhibit binding activity for various haptens. The differing hapten specificities are closely correlated with distinct primary amino-acid sequences.

An understanding of the genetic mechanism responsible for antibody diversity is one of the most intriguing problems in modern immunology, for its resolution may provide insights into the problems of information storage and expression in other complex systems (1, 2). Two alternative hypotheses for antibody diversity have been proposed—the germ line theory suggests that diversity arises in antibody genes during the evolution of the species, whereas the somatic theories propose that antibody diversity is created anew during the development of each individual by somatic mutation (1–3). Insights into the nature of this problem have come, in part, through the patterns observed from the detailed amino-acid sequence analysis of homogeneous myeloma immunoglobulins. The immunoglobulin molecule is comprised of two different polypeptides, light and heavy chains (4). All immunoglobulin chains can be divided into two segments, the variable (V) and the constant (C) regions, which code, respectively, for the antigen-binding and the various effector functions of the antibody molecule (5). There are three families of immunoglobulin genes, which are unlinked in the mammalian genome. Two code for light chains [lambda (λ) and kappa (κ)] and the third codes for heavy chains (1). The V and C genes of a given family appear to be coded by separate germ line genes (6). Myeloma proteins from the inbred BALB/c mouse have proven to be a fruitful system for the analysis of amino acid sequence patterns because tumors can be induced at will and are readily transplanted from one individual to a second (7). In addition, these inbred mice are presumably genetically homogeneous, in contradistinction to the genetic polymorphism present in most outbred populations such as humans and rabbits.

Two important patterns relevant to theories of antibody diversity have emerged from the sequence studies on myeloma proteins. First, the V regions from most immunoglobulin families can differ so extensively that all immunologists agree that they are coded by multiple germ line V genes. Thus, all theories of antibody diversity are multigenic, the question at issue is how multigenic. Second, when the V regions from large numbers of myeloma light or heavy chains are compared, three to four sections of extreme sequence variability are noted, which are designated "hypervariable regions" (8, 9). X-ray crystallographic studies have established that antibody polypeptides fold so that hypervariable regions constitute the walls of the antigen-binding crevice (10, 11, 32). The presence of hypervariable regions that make up the antigen-binding site has divided somatic theories into two groups. Both agree that the total antibody diversity is coded in part by germ line V genes and in part by V genes somatically derived from the germ line by mutation and stepwise selection for superior antigen binding sites. One class of somatic theories argues that the selection for somatic variants occurs predominantly, if not exclusively, in the hypervariable regions (2), whereas, a second group of somatic theories argues that somatic variation can occur throughout the entire V region (3).

Previous studies from our laboratories have assessed the diversity present in myeloma proteins from the two light chain families of the BALB/c mouse, λ and κ (12–16). In this paper we extend these studies to include an analysis of the N-terminal 20 amino-acid residues of 28 myeloma heavy chains derived from BALB/c mice. These sequences lie outside the hypervariable regions of the heavy chain and will, accordingly, place different constraints on the various theories of antibody diversity outlined above.

METHODS AND RESULTS

Mouse myeloma proteins were isolated from serum of tumor-bearing BALB/c mice by immunoadsorption for those examples with known specificities; J558 and MOPC 104E on dextran B-1355–polyacrylamide gel (17), W3129 and W3434 on Sephadex G-75, and S23 on di-nitrophenyl (Dnp) Sepharose (18). J606 and Y5606 were precipitated from serum by dialysis against 0.01 M NH4HCO3 (19). Y5476, Y2020, Y5444, W3082, and S176 were isolated by the methods described by Potter (20). The purified proteins were partially reduced and alkylated and heavy and light chains were separated as described by Bridges and Little (21). The heavy chains were desalted, lyophilized, and approximately 10 mg per run was loaded on the sequencer.
Each chain was subjected to 20 steps of automated Edman degradation on an updated Beckman 890A protein sequenator employing a standard Quadrol or DMBA program. Samples were dried under N₂ and converted to the phenylthiohydantoin amino acids (PTH) by treatment with 1 M hydrolysis in 6N HCl at 110°C for 18 hr at 100°C. The amino acids were then regenerated by hydrolysis in hydrochloric acid for 18 hr at 120°C (22) and analyzed on a Durum D-500 amino-acid analyzer. Throughout these experiments phenylthiohydantoin norleucine was added to the sequencer tubes prior to conversion, to normalize for any losses due to handling. The methodologies used have been described in detail in a previous publication (16). Yields of approximately 50% of the theoretical values were noted at step 1 and repetitive yields averaged 90% for all chains.

**DISCUSSION**

**Selection of Proteins.** The heavy chains used in this analysis were generally selected because the immunoglobulins from which they were derived exhibited specific binding for one or more haptenes (Fig. 1).

**Patterns of Diversity.** The amino terminal 20 residues of 13 heavy chains are compared with 15 sequences taken from the literature (Fig. 1). Sixteen of the 28 sequences differ by one residue or more. By sequence homology, these sequences can be divided into four major sets, which contain, respectively, 22, 3, 2, and 1 proteins. Prototype sequences can be derived for each of these sets by selecting the major residue alternative expressed at each position (Fig. 1) and these are respectively designated V̂ NH, V̂ NH, V̂ NH, and V̂ NH. The four sequence sets differ from one another by 25-55% of their residues (Table 1). Prototype set V̂ NH can be broken down into at least four subsets that are defined by two or more linked residues. For example, the heavy chains from W3082 and J606 have a glutamic acid at position 5 and a methionine at position 18 that distinguish these proteins from the remaining proteins of this set (Fig. 1). In this fashion, the heavy chains from H8, T15, S107, M603, and M167 as well as Y5606 and M21A can also be divided into two additional subsets by the presence of linked amino acids (Fig. 1).

The variable region of the heavy chain (V̂̂ region) is about 120 residues in length and hence the N-terminal 20 residues represent about 1/6 of the V̂̂ region. Accordingly, a rough
estimate of the number of amino-acid differences between two 
\( V_H \) regions can be determined by multiplying the difference in 
diversity shown over the N-terminal 20 residues by six (this is 
an underestimate because no hypervariable regions are 
included in this region). Thus, proteins in different sets will 
probably differ by 48 residues or more, whereas proteins within 
a set can, from this estimate, differ by as many as 24 or more 
residues. The only two nearly complete mouse 
\( V_H \) sequences from different sets, M315 and T173, are consistent with the 
above analysis in that they differ by 10 residues in the first 20 
and by 62/104 residues and two sequence gaps.

**Theories of V Gene Diversity.** A more sophisticated examination 
of these proteins can be carried out by a genealogic analysis (1). The sequences within any set can be related to one 
other by an evolutionary tree which depicts the minimal number of genetic events (i.e., single base substitutions) 
required to generate this set of sequences from a single 
ancestral sequence (gene). This procedure permits a step by 
step analysis of the single base substitutions required to generate 
diversity in a given set of sequences. One can then ask 
what events are compatible only with germ line evolution and 
what events might be explained by somatic evolution. The genealogic analysis in Fig. 2 shows that the relationships of these sequences to one another are somewhat more complex at 
the nucleotide than at the protein level because certain single 
amino-acid differences are coded by two base substitutions. In 
spite of this, four distinct branches are noted on the genealogic 
tree, which correspond to the four sets of sequences described 
previously, \( V_H \) through \( V_H^{1\!1\!1} \). The \( V_H^{1\!1\!1} \) branch is divided into 
five or more sub-branches by the presence of linked nucleotide 
bases in certain of the sequences (e.g., T15-167-5444; 3082- 
606; 5476-3129-J1; 2020-5606-M21A-406; and 175—see Fig. 2). 
Obviously additional data will be required (more and longer sequences) to more sharply delineate this genealogic tree, 
particularly with regard to the sub-branch structure of \( V_H^{1\!1\!1} \).

The somatic theory which argues for somatic variation only 
in hypervariable regions (2) and the germ line theory would 
agree on the conclusions to be drawn from this genealogic tree; 
namely, each of the \( V \) regions that differ from one another is 
coded by distinct germ line \( V \) genes that arose by gene duplication, sequence divergence (mutation, recombination, etc.) and 
selection during the evolution of the species. Thus both of 
these theories would agree that the diversity noted in these \( V_H \) 
regions must be coded by at least 16 germ line \( V_H \) genes. If, 
however, it is postulated that replacements in the N-terminal 
20 residues result from somatic mutation and selection, two 
questions must be asked. First, how many single base substitutions 
are reasonable to postulate during the somatic development 
of a given lymphocyte line? Second, how many parallel 
(or identical) mutations may occur in lymphocyte lines in 
separate individuals? If somatic diversification occurs by 
ordinary random mutation followed by selection, then it 
would appear likely that relatively few (i.e., 1–5) mutations 
occurred in examining primarily \( V_H \) regions, one for each 
mutant clone of lymphocytes (2). In addition, 
selective forces adequate to explain significant parallel 
mutation have been postulated to occur somatically, thus each 
set of sequences with multiple linked residues is probably 
derived from a different germ line gene (2). Accordingly, from 
the viewpoint of random somatic mutation that can occur 
throughout the entire \( V \) gene, at least eight germ line \( V_H \) genes 
must be postulated for this group of \( V_H \) regions, one for each 
distinct branch and sub-branch of the genealogic tree. A discussion of other somatic theories is beyond the scope of this 
report.

The mouse heavy chain sequences examined here appear to 
exclude less diversity than the mouse \( \kappa \) family and far more 
diversity than the mouse \( \lambda \) family. This generalization should 
be qualified, however, because most of these \( V_H \) sequences 
were derived from immunoglobulins selected because they 
exhibited binding activity for various haptens. Thus, these 
heavy chains certainly constitute a selected subset of BALB/c 
heavy chain myeloma pool. The degree of selection that 
ocurred in examining primarily \( V_H \) regions derived from 
myeloma proteins with binding activity is entirely unknown.
It will be very important to analyze the diversity exhibited by unselected $V_H$ regions to obtain a more reliable estimate of the diversity present in the BALB/c pool of myeloma heavy chains. In addition, mouse heavy chains blocked at their N-termini comprise about 78% of the normal serum immunoglobulin (23). Thus it will be particularly valuable to analyze the extent of diversity present in these blocked sequences.

**Normal Pooled Mouse Heavy Chains Appear to be More Restricted in Sequence Than the Pool of Myeloma Heavy Chains.** A recent study reports that normal pooled mouse heavy chains differ in two respects from the sequences reported here. First, as mentioned earlier, the normal mouse $H$ chains have a blocked $\alpha$-amino group in 78% of the proteins (23). Only one of 20 myeloma $H$ chains we examined was blocked (HOPC-1). Second, the pool of unblocked mouse $H$ chains obtained from the serum appeared to have a single residue at the 95% level at most of the N-terminal 28 positions. This is in striking contrast to the myeloma data we have reported here. For example, in the pool of myeloma heavy chains, aspartic acid is seen in 14% of the sequences at position 1; glutamine is seen in 32% of the sequences at position 3; the alternatives at position 5 are evenly spread among a number of residues. Thus, the myeloma pool of heavy chains reveals a heterogeneity not detectable by sequencing the normal pool, because the myeloma population, not surprisingly, expands individual sequences that are normally represented infrequently or not at all in the serum pool. Hence one must be cautious about drawing conclusions concerning the nature and extent of $V$ region diversity from studies on normal pooled sequences.

It appears that the myeloma immunoglobulins are themselves highly selected in that 6% or more of them bind to a few determinants (7). This is unexpected in view of the enormous functional heterogeneity of the normal immune response. Accordingly, both the normal and myeloma pools could be interpreted not to represent the true $V$ region diversity that is coded in the BALB/c genome.

One obvious difference between the myeloma pool and the normal serum pool is the fact that the artificial induction of myeloma proteins in BALB/c mice selects primarily IgA proteins, whereas the normal serum pool is mostly of the IgG type (7). Obviously, an analysis has yet to be carried out of the extent and nature of amino-acid sequence diversity in the $V_H$ regions associated with BALB/c $\gamma$ heavy chains. In any case, a critical question arises. Are the $V_H$ regions associated with $\gamma$ heavy chains distinct from those associated with $\alpha$? If the $V_H$ regions from $\gamma$ and $\alpha$ myeloma proteins are equally diverse, then the normal and myeloma heavy chain pools express the same family of germ-line $V_H$ genes. If the $V_H$ regions from $\gamma$ myeloma proteins are less diverse than those from their $\alpha$ counterparts, then selection must occur in the association of certain germ-line $V_H$ genes with given $C_H$ genes.

**Some Mouse $V_H$ Regions Are Similar to Their Human Counterparts, Whereas Others Are Quite Distinct.** A rough comparison of the similarity of mouse and human heavy chains at their N-termini can be made by comparing the prototype sequences of the mouse and human $V_H$ sets (Table 1). The mouse and human prototypes are quite distinct from one another (30-60%) apart from the mouse and human $V_H$ sequences (about 15% difference). The comparison of the single nearly complete human $V_H$ sequence (Nie) with two mouse $V_H$ regions (M173-39% and M315-61%) suggests that the group of mouse sequences to which M173 belongs is more closely related to the human $V_{HIII}$ sequences than other mouse $V_H$ regions are, although they are not as closely related as suggested by the N-terminal data. The fact that many mouse $V_H$ prototypes are as closely related to one another as to their human counterparts suggests that multiple $V_H$ genes existed in the mouse-human ancestor (Table 1). Thus some of the $V_H$ genes in human and mouse may have been derived from a common $V_H$ gene in the human–mouse ancestor, whereas others appear to have diverged from distinct genes in this ancestor. The heterogeneity of the myeloma heavy chains at the N-terminal compared with their normal counterparts also suggests that from the analysis of normal pooled sequences it is not possible to conclude that species- or phylogenetically associated residues are coded by a major portion of the germ-line $V$ genes.

**The Hapten Binding Properties of Certain Myeloma Proteins Appear to Correlate with Their Heavy Chain Sequences.** Six groups of myeloma proteins with binding activity for different haptens are shown in Fig. 1. The binding activity for a specific hapten appears to correlate with the sequence of the N-terminal 20 residues (Figs. 1 and 2). Indeed, each of the major branches on the genealogic tree with just a few sequences correlates with a single hapten binding activity (e.g., dinitrophenyl–$V_{HI}$, $\alpha$-l-3 dextran–$V_{HII}$, and 5-acetyluracil–$V_{HIV}$). The binding specificities for heavy chains in the $V_{HIII}$ branch generally correspond to the linked sets of proteins or subbranches that occur within this branch (e.g., the phosphorylcholine and levan binding heavy chains). Other studies suggest that the correlation of specificity with distinct amino-acid sequences is much greater with the heavy than with the light chain sequences (24).

**The Correlation of Particular Antigen Binding Specificities with Particular N-Terminal Sequences Suggests That Selection for Function Occurs Outside As Well As Within Hypervariable Segments of $V$ Regions.** This observation renders unlikely theories of antibody diversity that argue that special hyper-mutational mechanisms operate only in the hypervariable regions or that the hypervariable regions are coded by separate episome-like genes which are integrated into a framework $V$ gene. Clearly, mutation and selection occur through the entire $V$ gene. The relative contributions of germ line versus somatic diversification to this process remain uncertain.

**Heavy Chains May Have a Precursor with Additional Residues at the N-Terminus.** The amino-acid sequence analysis of S176 revealed a mixture of two sequences, probably identical to one another but for a single extra residue at the N-terminus (Fig. 1). A recent report has demonstrated that a myeloma light chain synthesized in vitro has an additional 10–15 residues on the N-terminus (25) which are, presumably, cleaved off by a post-translational event in vivo. Accordingly, we suggest that heavy chains also have a precursor form that is modified post-translationally. The presence of the normal heavy chain in this tumor product suggests that the extra N-terminal residue cannot be accounted for by a mutational event at the DNA level.

---

*Species- or phylogenetically associated residues are those found at a particular position in most of the serum immunoglobulin chains of one species and which are distinct from the homologous residues of a second species (23, 27).*
Are There Any Differences in the Set of $V_H$ Regions Associated with $\lambda$ As Compared to $k$ Light Chains? None of the heavy chains associated with $\lambda$ chains has a lysine at position 3, a residue found in the heavy chains associated with 18 of 21 $k$ type molecules (Fig. 1). Eight of the nine heavy chains derived from $\lambda$ type molecules have glutamine at position 3. On the other hand, the $V_H$ regions from M315 and M460 as well as Y5606 and M21A are quite similar, yet in each case one molecule is of the $\lambda$ and the other of the $k$ type. Obviously a great deal more sequence data must be available before conclusions can be drawn regarding $V_H$ associations with $V_\lambda$ and $V_k$ regions.

This investigation was supported by NIH Grant AI 10781 and NSF Grant GB 27605 to L.H.; NIH AI 105875 and CA 05213 to M.C.; and NIH Grant GM 20664 to M.W. L.H. has an NIH Career Development Award. M.W. has an American Cancer Society Faculty Research Award no. PRA-59. We thank M. Potter for tumor MOPC 104E.