Reports

Structure of the Constant and 3′ Untranslated Regions of the Murine γ2b Heavy Chain Messenger RNA

Abstract. *The complete coding sequence for the constant region of the mouse γ2b immunoglobulin heavy chain and the 3′ untranslated region has been determined. The coding portion of the sequence is 1008 nucleotides long (amino acid residues 114 to 449), and the 3′ noncoding region contains 102 nucleotides preceeding the polyadenylate. An extra carboxyl-terminal lysine residue which had not been observed in the γ2b or other γ subclass protein sequences occurs in the nucleotide sequence and is probably processed posttranslationally. A 17-nucleotide sequence occurs with slight variation twice in* C_H1 *and once in* C_H2 *domains in the same relative location but with different translational phase. This sequence may be the site of crossover in a γ2b · γ2a heavy chain variant, an indication of possible recombinational activity of some kind.*

The genes for the constant regions of immunoglobulin heavy chains (C_H) comprise a multigene family of great biological interest. Rearrangement of the genetic material at DNA and RNA levels probably leads to attachment of a variable region, chosen from an immense repertoire, to a constant region in a manner analogous to that proved for light chains (1, 2). But in contrast to light chains, heavy chains continue a process of rearrangement under developmental control leading to a succession of constant regions of differing class attached to the same variable region (3). These constant region switches apparently do not alter antibody specificity but govern other biological properties such as cellular localization, ability to fix complement, turnover time, susceptibility to proteolytic cleavage, and ability to cross the placental barrier (4).

In the mouse there are at least seven heavy chain types (μ, δ, γ2a, γ2b, γ3, γ1, and α). In each type the constant region protein can be subdivided into C_H domains, which contain about 110 amino acids each. They are relatively independent within a single chain but interact with homologous domains in adjacent light and heavy chains to determine the three-dimensional conformation of the antibody molecule (4). Direct determination of the protein sequence for heavy chains is a considerable challenge, and complete sequences have been reported for only μ (5), γ1 (6), and γ2a (7) chains. The C_H domains are considered to have evolved by duplication from a common ancestor.

The techniques of DNA cloning and nucleotide sequence analysis provide a means of determining protein sequence and to the analysis of gene structure and rearrangement. We now report the complete sequence of the constant region of the γ2b messenger RNA (mRNA) and the 3′ untranslated region (8). In (8) we present the structure of the germ line gene.

The nucleotide sequence we determined for the γ2b heavy chain is presented in Fig. 1, with C_H domains, hinge, and 3′ untranslated regions aligned to facilitate comparisons. This sequence was determined primarily from plasmid pγ2b-(11)^7 which was constructed by reverse transcription of MPC 11 mouse myeloma mRNA by Schibler *et al.* (9), who synthesized a double-stranded complementary DNA (cDNA) copy and inserted it into the *Eco* RI site of pMB9 (10) using the method of Maniatis *et al.* (11). The resulting clone was verified through hybrid-arrested translation experiments (12). A restriction map was constructed (Fig. 2), and nucleotide sequence of the inserted cDNA was analyzed (13). Since the cDNA clone did not extend far enough to include the 5′ terminal third of the C_H1 domain, the first 102 nucleotide of the sequence were obtained from the genomic clone, Ch4A 144.11.γ2b, which is discussed in (8).

The entire 1110 base pairs (bp) of DNA sequence was determined on both strands with the exception of the final 10 bp at the 3′ end. Most of the sequence was determined more than once by analysis of overlapping restriction fragments so that 70 percent of the bases were identified in at least three sets of data. To determine whether our sequence corresponds with the γ2b gene itself, we analyzed the corresponding regions of the genomic clone isolated from BALB/c mouse liver DNA (8).

To confirm the nucleic acid sequence, it was useful to compare it to the protein. Approximately 40 percent of the amino acid sequence of the γ2b C_H region has been determined by analysis of protein either from MPC 11 (14) or MOPC 141 myelomas (15). We find that the amino acid sequences from residues 116 to 135, 219 to 236, 236 to 281, 317 to 345, and 415 to 448 agree with that predicted from our DNA analysis with three exceptions: (i) we find Ser (16) instead of Leu at position 241, (ii) Ser instead of Trp at 269, and (iii) an extra Lys at the carboxyl terminal (see asterisks in Fig. 1). Since all three of our assignments are confirmed in the germ line clone, we believe that the two internal protein sequence differences may reflect experimental error in the protein sequencing. Spontaneous mutation occurring in cultured MPC 11 cells could be an alternative explanation, but these cases would require two non-

Table 1. Codon usage for the C region of γ2b heavy chain mRNA. Single letter abbreviations are given for line of comparison with Fig. 1.

F	(Phe)	UUU	1	S	(Ser)	UCU	6	U	(Tyr)	UAU	2	C	(Cys)	UGU	4
		UUC	8			UCC	10			UAC	8			UGC	7
L	(Leu)	UUA	0			UCA	6		(Ter)*	UAA	0		(Ter)*	UGA	1
		UUG	2			UCG	0			UAG	0	W	(Trp)	UGG	6
L	(Leu)	CUU	2	P	(Pro)	CCU	6	H	(His)	CAU	2	R	(Arg)	CGU	0
		CUC	8			CCC	8			CAC	6			CGC	0
		CUA	1			CCA	16	Q	(Gln)	CAA	2			CGA	0
		CUG	8			CCG	2			CAG	7			CGG	2
I	(Ile)	AUU	2	T	(Thr)	ACU	8	N	(Asn)	AAU	4	S	(Ser)	AGU	6
		AUC	11			ACC	12			AAC	11			AGC	14
		AUA	1			ACA	10	K	(Lys)	AAA	13	R	(Arg)	AGA	4
M	(Met)	AUG	4			ACG	2			AAG	11			AGG	1
V	(Val)	GUU	2	A	(Ala)	GCU	5	D	(Asp)	GAU	7	G	(Gly)	GGU	5
		GUC	11			GCC	3			GAC	9			GGC	2
		GUA	3			GCA	2	E	(Glu)	GAA	1			GGA	5
		GUG	17			GCG	0			GAG	14			GGG	6

*Terminator.

0036-8075/79/1214-1299$01.00/0 Copyright © 1979 AAAS

```
     GCCAAAACAACACCCCCATCAGTCTATCCACTGGCCCCTGGGTGTGGAGATACAACTGGTTCCTCCGTGACCTCTGGGTGCCTGGTCAAGGGGTACTTCCCTGAGCCAGTGACTGTGACTTGGAACTCTGGATCCCTGTCCAGCAGTGTGCACACCTTCCCA
1    ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--
     A  K  T  T  P  P  S  V  Y  P  L  A  P  G  C  G  D  T  T  G  S  S  V  T  S  G  C  L  V  K  G  Y  F  P  E  P  V  T  V  T  W  N  S  G  S  L  S  S  S  V  H  T  F  P
                       120               CH1   |     130            CH1            140                           150                           160         CH1
                                 LIGHT CHAIN ATTACHMENT?                           $

     GAGCCCAGCGGGCCCATTTCAACAATCAACCCCTGTCCTCCATGCAAGGAGTGTCACAAATGCCCA
292  --------+---------+---------+---------+---------+---------+-------
     E  P  S  G  P  I  S  T  I  N  P  C  P  P  C  K  E  C  H  K  C  P
        HINGE                   220            HINGE          230

     GCTCCTAACCTCGAGGGTGGACCATCCGTCTTCATCTTCCCTCCAAATATCAAGGATGTACTCATGATCTCCCTGACACCCAAGGTCACGTGTGTGGTGGTGGATGTGAGCGAGGATGACCCAGACGTCCAGATCAGCTGGTTTGTGAACAACGTGGAAGTA
357  --+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------
     A  P  N  L  E  G  G  P  S  V  F  I  F  P  P  N  I  K  D  V  L  M  I  S  L  T  P  K  V  T  C  V  V  V  D  V  S  E  D  D  P  D  V  Q  I  S  W  F  V  N  N  V  E  V
                 CH2      240*                          250                           260         CH2            *  270                           280
                                                                                               $

     GGGCTAGTCAGAGCTCCACAAGTATACACTTTGCCGCCACCAGCAGAGCAGTTGTCCAGGAAAGATGTCAGTCTCACTTGCCTGGTCGTGGGCTTCAACCCTGGAGACATCAGTGTGGAGTGGACCAGCAATGGGCATACAGAGGAGAACTACAAGGACACC
688  --+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------
     G  L  V  R  A  P  Q  V  Y  T  L  P  P  P  A  E  Q  L  S  R  K  D  V  S  L  T  C  L  V  V  G  F  N  P  G  D  I  S  V  E  W  T  S  N  G  H  T  E  E  N  Y  K  D  T
        CH3               350                           360         CH3               370                           380                           390
                                                                                   $

                                                                     3' COMMON SEQUENCE
                                                                            +----+
     TGAGCTCAGCACCCACAAAGCTCTCAGGTCCTAAGAGACACTGGCACCCATATCCATGCATCCCTTGTATAAATAAAGCATCCAGCAAAGCTGGTACCATGTAAAAAAAAAAAA
1009 -+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+-- 1122
     TER
                                    3'UT
```

Fig. 1. Nucleotide sequence of the C_H domains, hinge, and 3′ untranslated region of mouse γ2b mRNA. The symbol + appears under every tenth nucleotide, and the number to the left of each line of sequence denotes the first nucleotide of that domain. The left third of the C_H1 domain sequence to nucleotide 102 is from Tucker *et al*. (*8*). 5-Methylcytosine residues occur at positions 82, 222, 598, 744, and 769 and on the other strand at these positions less 2 as a result of methylation in *Escherichia coli* at the sequence CC(A/T)GG (*13*). The amino acids are displayed directly below the first base of the corresponding codon and are numbered to preserve the registration with the γ2b protein sequences at amino acid 236. The invariant cystine residues proposed to participate in intradomain disulfide linkages are indicated with $. Cystine 128 may be cross-

adjacent base changes in each codon (*17*). The additional Lys coding for the Pro-Gly at the carboxyl terminal probably can be explained in another way. Eight different γ class proteins have been sequenced so far, and all of them end with the dipeptide Pro-Gly with the exception of the human γ4 subclass, which has Leu-Gly (*18*). Considering the number of amino acid sequence determinations at this site, it seems probable that the extra Lys residue in the gene is proteolytically removed after translation, probably by a protease in the serum. It would be premature to speculate on whether a precursor with an extended carboxyl terminus of only one amino acid could have physiological significance, but there is some evidence for a carboxyl extension in the human μ chain (*19*). So far as we know these are the only reported cases of carboxyl terminal processing in eukaryotes. We have learned that Sakano and Tonegawa have also observed a terminal lysine in the mouse γ1 chain (*20*).

The 3′ untranslated sequence of 102 nucleotides for pγ2b(11)[7] is approximately the length expected from the mRNA measurement of 1800 nucleotides determined by Marcu *et al*. (*21*), assuming about 50 nucleotides of 5′ noncoding sequences (*22*) and 200 nucleotides of 3′ polyadenylate [poly(A)] extension. This length is shorter than that anticipated from the sequence of other eukaryotic mRNA's and shows no extensive homology with any of them. For example, more than 40 percent of the mouse κ light chain mRNA (*23*) and mouse β globin mRNA (*24*) is comprised of 3′ noncoding sequence.

It is conceivable that the method (*11*) employed to construct our double-stranded γ2b cDNA could lead to truncation at the 3′ end or, alternatively, that in plasmid replication of the hybrid DNA, deletions could occur near or in the homopolymeric tracts. However, four indirect points of evidence render these potential artifacts unlikely. First, the prototypic sequence AAUAAA, found about 20 nucleotides from the poly(A) attachment site at the 3′ end in all eukaryotic untranslated regions sequenced to date (*23*–*26*), occurs as expected in the γ2b sequence 24 nucleotides from the 3′ poly(A) end. Second, hybridization by the Southern method of reverse transcripts of purified γ2b mRNA to genomic clone 144.11.γ2b gave absolutely no hybridization to the region immediately downstream of the

Table 2. Comparison of mouse γ subclass C_H amino acid sequences. The amino acid sequences of corresponding domains of three heavy chains were aligned to maximize homology by introducing gaps. The lower right portion of the square gives the number of gaps required followed by the total number of amino acids included in the gaps. The upper portion of the square gives the percent homology (amino acid identities × 100/amino acids compared plus the number of gaps). Gaps in the alignment are counted as single disagreements. Numbers in parentheses within the squares are percent nucleotide agreement (upper portion) and number of nucleotide gaps required (lower square) for alignment, based on comparison of our sequence with the mRNA sequence of Rogers *et al*. (*29*), (C_H3) and Sakano *et al*. (*30*) (hinge).

Comparison	C_H1	Hinge	C_H2	C_H3	Entire C_H
γ2b vs. γ2a	81% 2,2	67% 2,6	82% 0,0	59% 1,1	74% 5,9
γ2b vs. γ1	76% 1,3	33% (56%) 2,9 (2,27)	71% 0,0	56% (68%) 0,0 (0,0)	65% 3,14
γ2a vs. γ1	80% 3,18	55% 2,6	62% 0,0	61% 1,1	66% 5,25

```
                                                   HOMOLOGY                                        HOMOLOGY
                                              +---------------+                                  +-----------+
GCTCTCCTGCAGTCTGGACTCTACACTATGAGCAGCTCAGTGACTGTCCCCTCCAGCACCTGGCCAAGTCAGACCGTCACCTGCAGCGTTGCTCACCCAGCCAGCAGCACCACGGTGGACAAAAAACTT
--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+----
A  L  L  Q  S  G  L  Y  T  M  S  S  S  V  T  V  P  S  S  T  W  P  S  Q  T  V  T  C  S  V  A  H  P  A  S  S  T  T  V  D  K  K  L
      170                           180                           190                  CH1      200                     210
                                                                                 $




                                                                      HOMOLOGY
                                                                 +-----------------+
CACACAGCTCAGACACAAACCCATAGAGAGGATTACAACAGTACTATCCGGGTGGTCAGCACCCTCCCCATCCAGCACCAGGACTGGATGAGTGGCAAGGAGTTCAAATGCAAGGTGAACAACAAAGACCTCCCATCA CCCATCGAGAGAACCATCTCAAAAATTAAA
+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+------- --+---------+---------+--------
H  T  A  Q  T  Q  T  H  R  E  D  Y  N  S  T  I  R  V  V  S  T  L  P  I  Q  H  Q  D  W  M  S  G  K  E  F  K  C  K  V  N  N  K  D  L  P  S  P  I  E  R  T  I  S  K  I  K
         290                  CH2      300                           310                           320               CH2         330            CH2            340
                                    +-----+                                                                 $
                            CARBOHYDRATE ATTACHMENT?

GCACCAGTTCTTGACTCTGACGGTTCTTACTTCATATATAGCAAGCTCAATATGAAAACAAGCAAGTGGGAGAAAACAGATTCCTTCTCATGCAACGTGAGACACGAGGGTCTGAAAAATTACTACCTGAAGAAGACC ATCTCCCGGTCTCCGGGTAAA
+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+------- --+---------+-------
A  P  V  L  D  S  D  G  S  Y  F  I  Y  S  K  L  N  M  K  T  S  K  W  E  K  T  D  S  F  S  C  N  V  R  H  E  G  L  K  N  Y  Y  L  K  K  T  I  S  R  S  P  G  K
   CH3   400                           410                           420               CH3         430            CH3            440                        *
                                                                                          $
```

linked to the light chain. A proposed carbohydrate recognition site is indicated (amino acid 300). Three homology regions discussed in the text are indicated above the corresponding DNA sequences. The common sequence beginning at nucleotide position 1080 appears in all 3′ untranslated regions so far sequenced. Asterisks are used to indicate deviations from published partial protein sequence data (see text). The single letter abbreviations are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr [see (*16*)].

Kpn I site at the 3′ end of our cDNA sequence. Third, our nucleotide sequence of about 180 bp to the right of this *Kpn* I site in the genomic clone revealed no sequence of the type AAUAAA. And fourth, C_H clones for α and μ classes constructed in our laboratory also have short 3′ untranslated regions. The report by Faust *et al.* (*27*) suggesting that mouse γ subclass mRNA's must contain 600 to 700 noncoding nucleotides is inconsistent with our data.

In analyzing the protein subunits of immunoglobulins, it is essential to know the domain boundaries. The exact amino acid positions of all but the left boundary of the C_H1 domain have been assigned (Fig. 1) from the location of intervening DNA sequences in the germ line gene (*8*). The left boundary of the C_H1 domain has been tentatively assigned at amino acid position 114 by homology with the $\gamma1$ and $\gamma2a$ sequences (*6, 7*). Since neither our cDNA nor genomic clones contain variable (V) region sequence, and the amino acid sequence for this region has not been determined, unequivocal identification of the joint must await further analysis.

Since the Cys residues in our sequence are relatively few in number, once domain boundaries have been established it is possible to predict plausible disulfide linkages. The invariant intradomain disulfide bridges (*4*), which link approximately 60 residues are: C_H1, 140 and 195; C_H2, 263 and 323; C_H3, 369 and 427. The Cys residue at position 128 may provide the disulfide bond to the light chain. These are indicated by $ in Fig. 1. Within the $\gamma2b$ hinge region (residues 211 to 232), four potential inter- or intrachain (or both) disulfide linkages and six Pro residues should afford an exceptionally rigid pivot point for the flexible portion of the hinge.

The usual recognition tripeptide sequence for carbohydrate attachment, Asn-X-(Thr or Ser) (*28*), where X may be any amino acid, occurs only once in the protein (position 300) in the C_H2 domain.

Codon usages for the C_H region of the $\gamma2b$ mRNA are given in Table 1. As observed in other eukaryotic mRNA's (*20-24*), codon selection is nonrandom. Our observations include the following. Certain codons are used repeatedly (UUC for Phe, AUC for Ile, GAG for Glu), and others are used very little or not at all (GGC for Gly, CGC for Arg, UUA for Leu, GUU for Val). The order of third base preference is C > G > U ≃ A, and A is chosen at a significantly higher frequency (~18 percent of $\gamma2b$ C_H codons) than in most other mRNA's [for example, ~ 4 percent of the codons in either mouse, rabbit, or human hemoglobin (*24, 26*)]. The frequency of the dinucleotide CG is characteristically low in the C_H portion of this mRNA, yet its occurrence in the second and third position (such as CCG for Pro and ACG for Thr) is no rarer than in the first and second positions (CGG for Arg). The need for Arg has apparently elevated the latter usage in the hemoglobin mRNA's.

It is generally assumed that all immunoglobulin constant region domains have evolved through duplication and subsequent mutation from a common ancestral sequence (*4*). Some clues to this evolutionary history can be found by align-

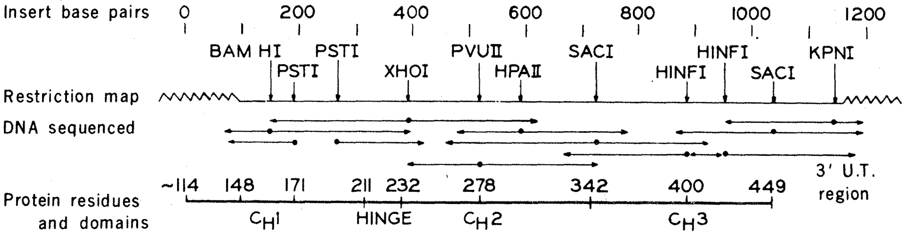


Fig. 2. Endonuclease restriction map for the cDNA insert in pγ2b(11)[7]. In agreement with Fig. 1, nucleotides are numbered from 5′ to 3′ on the coding strand, beginning with the residue corresponding to amino acid 114, the putative beginning of the C_H1 domain. The beginning of the cDNA insert of γ2b mRNA in the plasmid is nucleotide 103. The first amino acid residue encoded by the cDNA is number 148. The wavy lines on each side of the insert represent the poly(A) tails (~40 nucleotides on the 5′ side and 150 nucleotides on the 3′ side) and pMB9 vector DNA. Only the restriction sites from which sequence was derived are shown. The direction and length of sequence obtained from these sites (designated below by dots) are indicated by the arrows. The double-headed arrow between *Hinf* I sites indicates that the strands of this end labeled fragment were separated and both were sequenced.

ing the sequences of domains and counting the nucleotide or amino acid matches. The usual method (*4*) is to introduce gaps into the sequences so as to bring the invariant intradomain Cys and Trp residues into alignment and, where necessary, to insert additional gaps to bring other regions of homology into alignment. These gaps can be thought of as corresponding to single evolutionary deletion or insertion events.

When the DNA sequence of the C_H3 domain of $\gamma1$ determined by Rogers *et al.* (*29*) is compared with our sequence for $\gamma2b$–C_H3, no gaps are needed (Table 3). In this case, 68 percent of the bases and 56 percent of the amino acids match, indicating these genes are very closely related. Of 59 conserved amino acids, 42 were coded by conserved triplets, giving 29 percent rate of alteration at the third positions of these triplets, which is slightly less than the 32 percent overall rate of alteration between the genes. Thus the so-called silent positions of codons for invariant amino acids are at least as well conserved as the rest of the gene, an indication that selection pressure on these third positions must be at least as strong as that acting to conserve the amino acids. In the codons for the 47 amino acids that are not conserved, the number of silent alterations is 23 for a much higher rate of 49 percent mismatch. This correlation of silent mutations with nonsilent ones may indicate that, once an amino acid has changed in evolution, there may be selective pressure to alter the third position to use a more preferred codon.

The other constant region domain with which our DNA sequence can be compared is the hinge region of $\gamma1$ (*30*). In this case, the lengths of the DNA sequences differ (66 for $\gamma2b$, and 39 for $\gamma1$). When gaps were introduced in the $\gamma1$ strand of 24 nucleotides at position 10 and three nucleotides at position 29, the nucleic acid homology was 59 percent, and only 33 percent of the amino acids matched, evidence that the hinge has evolved at a more rapid rate than C_H3.

Additional protein sequence homology comparisons between corresponding C_H domains of our $\gamma2b$ sequence and the other two γ subclasses for which complete protein sequence is available are shown in Table 2. The C_H1 and C_H2 domains appear to be more similar to each other than to C_H3 or hinge, although several deletions must be assumed in the C_H1 comparisons whereas few or no gaps are needed when comparing the C_H2 or C_H3 domains.

The high variability of C_H3 domains in γ chains seems to be concentrated at the carboxyl terminal and contrasts to the relative evolutionary stability of carboxyl terminal domains in μ and α chains (*4, 5*). This may reflect a relative lack of selective pressure on that domain in γ chains since the circulating antibodies do not require membrane association.

Table 3. Internal comparison of mouse $\gamma2b$ domains. The protein sequences were aligned and matches were computed as described for Table 2.

Domains	Amino acid matches (%)	Nucleotide matches (%)	Number of gaps	Amino acids in gaps
C_H1-C_H2	16	36	5	9
C_H2-C_H3	22	42	3	5
C_H1-C_H3	27	40	2	3

When the domains of $\gamma2b$ were compared by alignment of invariant Cys and Trp residues, these more distantly related sequences showed from 36 to 42 percent base homology (Table 3).

One striking short region of base homology that was found between C_H1 (near amino acid 184) and C_H2 (near amino acid 309) domains deserves mention:

C_H1
5′ TCCCC - TCCAGCACCTGG 3′
. . . .
C_H2 TCCCCATCCAGCACCAGG

This stretch of 17 almost perfectly homologous nucleotides occurs at exactly corresponding positions in C_H1 and C_H2 when invariant Cys residues are aligned; but, owing to translational reading frame shift, the amino acid sequence is different. Removal of the underlined A would bring them into register over this stretch (although another base would be needed to restore the phase downstream). The same sequence of C_H2 (near amino acid 309) also matches C_H1 farther down the chain (near amino acid residue 200) although not so perfectly:

C_H1 CACCCAGCCAGCAGCACC
. . . .
C_H2 TCCCCATCCAGCACCAGG

Again the match is out of codon register.

The same sequence of C_H2 appears to be at or near the crossover point between $\gamma2a$ and $\gamma2b$ in a variant of the MPC 11 myeloma isolated by Birshtein *et al.* (*31*) that synthesizes a $\gamma2b \cdot \gamma2a$ hybrid H chain. Birshtein's amino acid sequencing experiments have established that the crossover point must occur in the C_H2 domain between the $\gamma2b$ residues Thr 307 (Thr in $\gamma2b$ and the hybrid) and Ser 332 (Ala in $\gamma2a$ and the hybrid). This localization is compatible with but does not prove an involvement of the region of C_H2 that is twice duplicated in C_H1. It will be very informative to know exactly where the crossover point occurred at the DNA (or RNA) level.

In considering the implications of out-of-phase matches, we find that by extending the homologies discussed above, the C_H1 and C_H2 domains can be matched surprisingly well over substantially their entire lengths by the use of out-of-phase registrations. One scheme with four gaps including a total of eight nucleotides yielded 37 percent match. Another utilizing two gaps totaling 11 nucleotides yielded 36 percent match. It was not possible to find out-of-phase homologies approaching this strength for the other C_H domain comparisons. Still, it raises the possibility that mammalian evolution like that of phage ϕX174 may make use of information encoded in several phases of the DNA code (*32*). It is not necessary to suppose that more than one phase is expressed at any one epoch to see the evolutionary advantage of such arrangements in DNA.

PHILIP W. TUCKER*
KENNETH B. MARCU†
JERRY L. SLIGHTOM
FREDERICK R. BLATTNER
Laboratory of Genetics,
University of Wisconsin, Madison 53206

References and Notes

1. C. Brack, M. Hirama, S. Lenhard-Schuller, S. Tonegawa, *Cell* **15**, 1 (1978).
2. J. G. Seidman, A. Leder, M. Nau, B. Norman, P. Leder, *Science* **202**, 11 (1978).
3. R. M. E. Parkhouse and M. D. Cooper, *Immunol. Rev.* **37**, 105 (1977).
4. F. W. Putnam, in *The Plasma Proteins*, F. W. Putnam, Ed. (Academic Press, New York, 1977), pp. 1-153.
5. M. Kehry *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 2932 (1979).
6. K. Adetugbo, *J. Biol. Chem.* **253**, 6068 (1978).
7. M. Fougereau, A. Bourgois, C. dePreval, J. Rocca-Serra, C. Schiff, *Ann. Immunol. (Paris)* **127c**, 607 (1976).
8. P. Tucker, K. B. Marcu, N. Newell, J. Richards, F. R. Blattner, *Science* **206**, 1303 (1979).
9. U. Schibler, K. B. Marcu, R. P. Perry, *Cell* **15**, 1495 (1978).
10. H. W. Boyer, M. Getlack, F. Bolivar, R. L. Rodriquez, H. L. Heyneker, J. Shine, H. M. Goodman, in *Recombinant Molecules: Impact on Science and Society*, R. F. Beers, Jr., and E. G. Bassett, Eds. (Raven, New York, 1977), pp. 9-20.
11. T. Maniatis, S. G. Kee, A. Efstratiadis, F. C. Kafatos, *Cell* **8**, 163 (1976).
12. B. M. Paterson, B. E. Roberts, E. L. Kuff, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 4370 (1977).
13. A. M. Maxam and W. Gilbert, *ibid.*, p. 560.
14. T. Francus and B. K. Birshtein, *Biochemistry* **17**, 4324 (1978).
15. C. dePreval, J. R. L. Pink, C. Milstein, *Nature (London)* **228**, 930 (1970).
16. Abbreviations: Ala, alanine; Cys, cysteine; Asp, aspartic acid; Glu, glutamic acid; Phe, phenylalanine; Gly, glycine; His, histidine; Ile, isoleucine; Lys, lysine; Leu, leucine; Met, methionine; Asn, asparagine; Pro, proline; Gln, glutamine; Arg, arginine; Ser, serine; Thr, threonine; Val, valine; Trp, tryptophan; Tyr, tyrosine.
17. P. Coffino and M. D. Scharff, *Proc. Natl. Acad. Sci. U.S.A.* **68**, 219 (1971).
18. *Atlas of Protein Sequence and Structure, 5 Suppl. 3*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, D.C., 1978), pp. 197-228.

19. P. B. Williams, R. T. Kubo, H. M. Grey, *J. Immunol.* **121**, 2435 (1978).
20. S. Tonegawa, personal communication.
21. K. B. Marcu, O. Valbuena, R. P. Perry, *Biochemistry* **17**, 1723 (1978).
22. R. L. Jilka and S. Peska, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5692 (1977).
23. P. H. Hamlyn, G. G. Brownlee, C. C. Cheng, M. J. Gait, C. Milstein, *Cell* **15**, 1067 (1978).
24. D. A. Konkel, S. M. Tilghman, P. Leder, *ibid.*. p. 1125.
25. P. H. Seeberg, J. Shine, J. A. Martial, J. D. Baxter, H. M. Goodman, *Nature (London)* **270**, 486 (1977).
26. A. Efstratiadis, F. C. Kaftos, T. Maniatis, *Cell* **10**, 571 (1977); C. A. Marotta, J. T. Wilson, B. G. Forget, S. M. Weissman, *J. Biol. Chem.* **252**, 5040 (1977).
27. C. H. Faust, I. Heim, J. Moore, *Biochemistry* **18**, 1106 (1977).
28. R. D. Marshall, *Annu. Rev. Biochem.* **41**, 673 (1972).
29. J. Rogers, P. Clarke, W. Salser, *Nucleic Acid Res.* **6**, 3305 (1979).
30. H. Sakano, J. H. Rogers, K. Hüppi, C. Brack, A. Traunecker, R. Maki, R. Wall, S. Tonegawa, *Nature (London)* **277**, 627 (1979).
31. B. K. Birshtein, R. Campbell, M. L. Greenberg, in preparation.
32. F. Sanger *et al.*, *Nature (London)* **265**, 687 (1977).
33. We thank Drs. B. Birshtein, J. Rodgers, P. Clarke, W. Salser, and S. Tonegawa for communicating results to us prior to publication; Dr. W. Fitch for his computer search analyses and insightful discussions; J. Schroeder for implementing additional computer search routines; and Dr. R. B. Perry for making pγ2b(11)^7 available. All experiments were carried out in accordance with the NIH Guidelines on Recombinant DNA Research. Supported by grant GM 21812 (F.R.B.), GM 06526 (P.W.T.), and GM 20069 (Dr. O. Smithies).

* Present address: Department of Biochemistry, University of Mississippi, School of Medicine, Jackson 39216.

† Visiting investigator. Permanent address: Department of Biochemistry, State University of New York, Stony Brook 11794.

13 August 1979; revised 10 October 1979

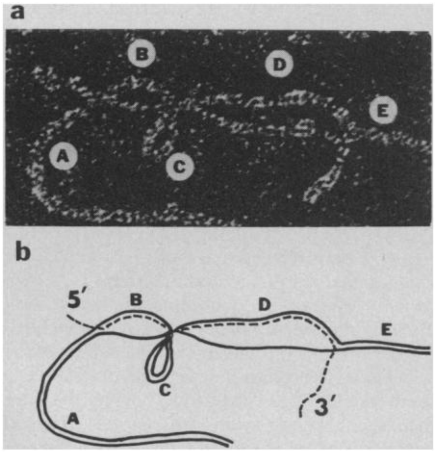
Sequence of the Cloned Gene for the Constant Region of Murine γ2b Immunoglobulin Heavy Chain

Abstract. *The complete nucleotide sequence of the γ2b constant region gene cloned from BALB/c liver DNA is reported. The sequence of approximately 1870 base pairs includes the 5′ flanking, 3′ untranslated, and 3′ flanking regions and three introns. The Cγ2b coding region is divided by these introns into four segments corresponding to the homology domains and hinge region of the protein. The introns separating the hinge from the C_H2 domain and the C_H2 from the C_H3 domain are small (106 and 119 base pairs). A larger intervening sequence of 314 base pairs separates the C_Hl and hinge regions. The stretch of DNA comprising this large intron plus the hinge shows a strong homology with the other C_H domains.*

The protein chains comprising immunoglobulins as discussed (*1*) are organized into domains, and this organization is reflected at the gene level. The variable (V) and constant (C) region domains of light chains are encoded at separate loci, and even in the committed lymphocyte they are separated by introns (*2, 3*). Two studies of DNA cloning of myeloma tumor DNA have appeared showing that the C_H domains of heavy chains are also separated by introns. Early *et al.* (*4*), using electron micrographic and restriction mapping techniques, identified two introns in the α constant region gene from the MOPC 603 myeloma. They concluded that within measurement errors, the introns were located at presumptive domain boundaries. Using mapping techniques plus partial DNA sequence analysis, Sakano *et al.* (*5*) studied a γ1 class constant region gene from the MOPC 21 myeloma in much greater detail. They reached the same conclusion and determined placement of introns between the domains of the Cγ1 gene. They also showed that the γ1 hinge region is encoded by a separate DNA segment.

We have now extended these findings to BALB/c mouse DNA from a tissue not committed to immunoglobulin production (liver). We assume that this DNA is identical in organization to that of the germ line. We have analyzed several constant region genes and report here the complete DNA sequence of one of them, γ2b. Our results show that, for this gene, introns occur precisely between domain and hinge encoding segments. Moreover, since the sequence is complete, we have also proved that no small introns within domains or the 3′ untranslated region have escaped notice. These results in conjunction with the messenger RNA (mRNA) sequence presented (*1*) allow determination of the splice sites at which mRNA presumably is processed to eliminate the introns.

A shotgun collection was constructed by inserting 10- to 20-kilobase-pair (kbp) *Eco* RI fragments from partially digested BALB/c liver DNA into the bacteriophage vector Charon 4A (*6*). The shotgun collection was screened on megaplates (*7*) with a mixture of ^{32}P-labeled plasmid C_H probes (*8*) which included pγ2b(11)^7. Several γ2b positive phages



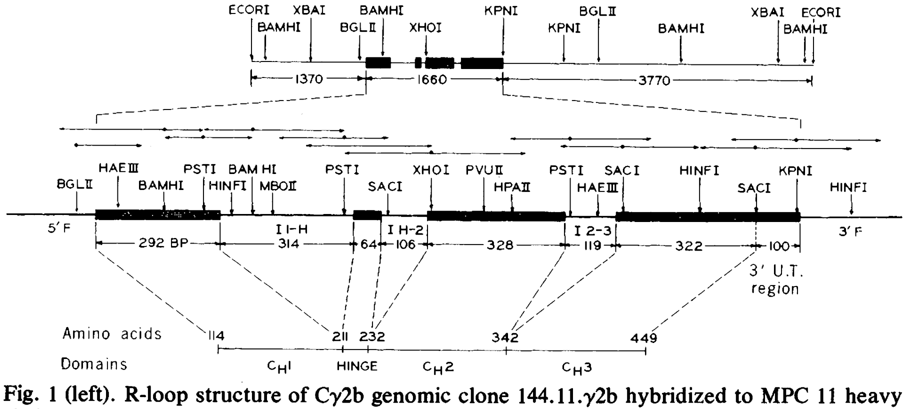


Fig. 1 (left). R-loop structure of Cγ2b genomic clone 144.11.γ2b hybridized to MPC 11 heavy chain mRNA. (a) Electron micrograph; *B* and *D*, correspond to the DNA-mRNA hybrid, *C* corresponds to double-stranded DNA. *A* leads to the right end of the Ch4A vector, *E* to the left end. Lengths given in text were measured in 23 molecules with double-stranded (pCo1E1) and single-stranded (G4) standards. (b) Interpretative drawing; MPC 11 heavy chain mRNA is represented by a dotted line, and single-stranded DNA by solid line. Fig. 2 (right). Endonuclease restriction map of the Cγ2b gene and flanking DNA regions. The cloned 6.8-kbp *Eco* RI fragment in the 5′ to 3′ orientation is shown in the upper panel. The position of the Cγ2b coding region (denoted by boxes) was initially determined by R-looping (legend to Fig. 1) and refined by restriction analysis and Southern hybridization techniques. We have no explanation other than chance for the unusual palindromic arrangement of restriction sites in this DNA. A fivefold magnification of the 1660 nucleotide transcribed Cγ2b gene and the adjoining flanking regions (5′F and 3′F) is shown in the same orientation below. The coding segments are interrupted by three introns: I1-H, IH-2, and I2-3. The dots and arrows above the restriction enzymes denote the sequencing strategy; fragments were cleaved and labeled at the dots, and the corresponding arrows give the direction and length of the sequence obtained. The C_H portion of the γ2b heavy chain protein is illustrated at the bottom of the figure. Dashed lines denote the corresponding positions in the protein and DNA. All nucleotide distances are in base pairs.