(Reprinted from Nature, Vol. 280, No. 5720, pp. 288-294, July 26 1979)

Sequences at the somatic recombination sites of immunoglobulin light-chain genes

Hitoshi Sakano, Konrad Hüppi, Günther Heinrich & Susumu Tonegawa

Basel Institute for Immunology, 487 Grenzacherstrasse, Postfach, CH-4005 Basel, Switzerland

The entire nucleotide sequence of a 1.7-kilobase embryonic DNA fragment containing five joining (J) DNA segments for mouse immunoglobulin κ chain gene has been determined. Each J DNA segment can encode amino acid residues 96–108. Comparison of one of the five J DNA sequences with those of an embryonic variable (V) gene and a complete κ chain gene permitted localisation of a precise recombination site. The 5'-flanking regions of J DNA segments could form an inverted stem structure with the 3'-non-coding region of embryonic V genes. This hypothetical structure and gel-blotting analysis of total embryo and myeloma DNA suggest that the somatic recombination may be accompanied by excision of an entire DNA segment between a V gene and a J DNA segment. Antibody diversity may in part be generated by modulation of the precise recombination sites.

IMMUNOGLOBULIN polypeptide chains consist of two regions, the amino-terminal half, called the variable (V) region, and the carboxyl-terminal half, called the constant (C) region. In embryonic cells the two regions are encoded in widely separated DNA segments. During differentiation of bone marrow-derived (B) lymphocytes the two DNA segments are brought into proximity by site-specific recombination, producing a single transcription unit[1,2].

The details of the sequence organisation before and after recombination have been worked out in the mouse λI light-chain system by analysis of cloned DNA fragments [3,4]. These studies revealed that in embryo cells the conventionally defined V region is itself encoded in two separate DNA segments, one (V DNA) coding for the major part of the V region (residues 1–97) and the other (J DNA) coding for the rest of the V region (residues 98–109). Although the distance between the V and the C DNA segments is unknown, the J DNA segment has been mapped at 1.2 kilobases towards the 5'-side, relative to the direction of transcription, of the C DNA segment.

Somatic recombination takes place at the 3'-end of the V DNA segment and the 5'-end of the J DNA segment. Consequently, in a λI chain synthesising B lymphocytes or plasma cells, although the DNA segment coding for the entire V region is much closer to the C DNA segment, the two DNA segments are still separated by 1.2 kilobases. This untranslated sequence (intron)[5] separating the J and C DNA segments is thought to be removed at the RNA level by a process called splicing[6,7]. Hence, the J DNA segment and its flanking sequences seem to be involved in two key processes essential for the expression of an immunoglobulin gene: site-specific V–J recombination and J–C splicing[3].

More recently, we have analysed the organisation of mouse κ light-chain genes[8,9]. A comparison of the sequence organisation of the two light-chain gene systems revealed both common and differing features. As in the λ gene system, a secreted κ chain is encoded in three separate DNA segments in embryonic cells, Vκ, Jκ and Cκ. In addition, somatic recombination takes place between embryonic Vκ and Jκ DNA segments. However, unlike the λI gene system, the mouse genome contains multiple Vκ and Jκ DNA segments, the latter being clustered in the vicinity of a single copy of the Cκ DNA segment.

For a better understanding of the structural features of the recombination sites we have determined the nucleotide sequences of the Jκ cluster as well as the 3'-end region of an embryonic Vκ DNA. This information, in addition to Southern gel-blotting experiments, was used to formulate a model for somatic recombination. We also discuss V–J joining with respect to generation of antibody diversity and evolution of immunoglobulin genes.

Sequencing strategy for the J DNA cluster

We have previously reported cloning of a 15-kilobase embryonic DNA fragment (clone Ig146κ) containing a single Cκ DNA copy[8]. Analyses of R-loops formed between this DNA fragment

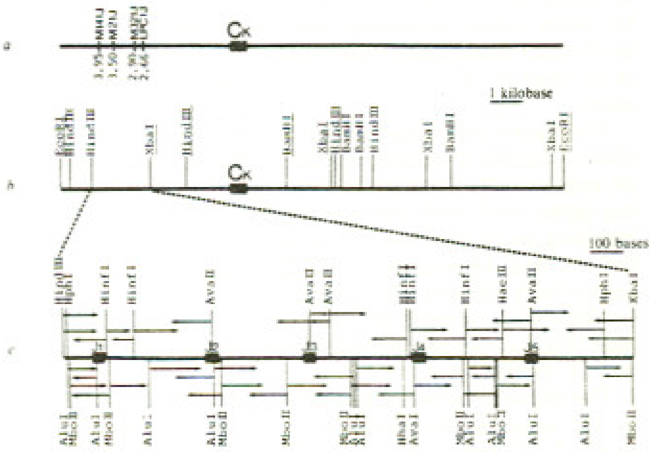


Fig. 1 *a*, R-loop map of the 15-kilobase insert of Ig146. κ mRNAs from four different myelomas, MOPC141, MOPC21, MOPC321 and LPC1, were individually annealed with the Ig146 *Eco*RI DNA fragment and the position of the corresponding J DNA segment was determined by electron microscopy (see ref. 9 for details). The bars represent the J DNA segments. The numbers indicate their distance from the Cκ DNA segment. *b*, Restriction enzyme cleavage map of the 15-kilobase insert of Ig146. The whole Ig146 phage DNA and the purified vector phage (λWES)[30] arm DNA (1 μg each) were digested either with one of the indicated enzymes or with that enzyme plus *Eco*RI, and separated by electrophoresis in a 1% agarose gel. Comparison of the single and double digestion patterns allowed localisation of the outermost cleavage sites of each enzyme on the 15-kilobase insert. Other cleavage sites were determined by comparing the pattern obtained by a single enzyme digestion (enzyme A or B) with that obtained by a double enzyme digestion (enzyme A plus B). The position of the Cκ DNA segment relative to the underlined cleavage sites was determined by examination of R-loop structures formed between MOPC321 κ mRNA and the Ig146 DNA predigested with the corresponding enzyme. *c*, Sequencing strategy and a restriction map of the 1.7-kilobase *Hin*dIII–*Xba*I fragment. The rightmost cleavage site of each enzyme was identified by comparing the cleavage pattern of the 1.7-kilobase *Hin*dIII–*Xba*I fragment with that of the 2.7-kilobase *Hin*dIII–*Hin*dIII fragment. For determination of additional cleavage sites the *Hin*dIII–*Xba*I fragment was end-labelled with [γ-^{32}P]ATP and separated into two fragments by *Hha*I digestion followed by electrophoresis. The purified *Hin*dIII–*Hha*I and *Hha*I–*Xba*I fragments were digested with one of the indicated enzymes in limited conditions and the cleavage sites were deduced according to the method of Smith and Birnstiel[31]. Horizontal arrows indicate the direction and the extent of sequence determination. The boxes indicate positions of J DNA segments as determined by the nucleotide sequence analysis.

and five different κ mRNAs isolated from myelomas MOPC141, MOPC21, MOPC321, LPC1 and HOPC8 identified at least four different J DNA segments (HOPC8 J and LPC1 J may be identical) in the region 2.5–4.0 kilobases to the left of the Cκ DNA segment (Fig. 1*a*)[9]. To determine the nucleotide sequence of this region, we first constructed a restriction map of the 15-kilobase Ig146κ DNA using restriction enzymes *Hin*dIII, *Bam*HI, and *Xba*I (Fig. 1*b*). The position of the Cκ DNA segment in the restriction map was determined by examining R-loops formed between MOPC321κ mRNA and the Ig146κ DNA predigested with the enzymes listed above. From a comparison of this restriction map with the R-loop map previously constructed for J segments we concluded that the J DNA cluster is on the 1.7-kilobase *Hin*dIII–*Xba*I fragment (Fig. 1*a*,*b*). We then determined the entire nucleotide sequence of this fragment using a strategy shown in Fig. 1*c*.

Complete nucleotide sequence of the J-rich 1.7-kilobase *Hin*dIII–*Xba*I fragment

Figure 2 shows the complete nucleotide sequence of the *Hin*dIII–*Xba*I fragment. To date, about 30 mouse κ chains of BALB/c and NZB strains have been sequenced around the J region[10,11]. If one assumes that the Jκ region begins with residue 96 and ends with residue 108 (see below) based on the analogy to the Jλ region[4], the 30 Jκ regions fall into 10 different classes (Table 1). We therefore scanned the nucleotide sequence determined for each of the 10 predicted J sequences, and identified four (J1, J2, J4 and J5). Taking advantage of the fact that all known J regions are extensively homologous, we looked for additional J or J-like sequences and found one (J3) for which no sequenced κ chain is known. A systematic computer search for other J-like sequences that are related to the known or predicted J sequences by more than 50% homology revealed no more than the five Js in the 1.7-kilobase *Hin*dIII–*Xba*I fragment.

The amino acid sequences of the MOPC21 and MOPC321 κ chain J regions correspond to the predicted sequences of J2 and J4, respectively. Positions of these Js as determined by sequencing agree well with the positions of Js identified by R-loop mapping using the respective κ mRNA. We therefore conclude that J2 and J4 encode MOPC21 and MOPC321 J regions, respectively. Amino acid sequences of the MOPC141 and LPC1 (or HOPC8) J regions are unknown, but based on the R-loop map (Fig. 1*a*) we predict that these J regions are encoded in J1 and J5, respectively.

The distance between adjacent Js is quite constant. The five Js are arranged as follows: J1-310 base pairs-J2-246 base pairs-J3-268 base pairs-J4-299 base pairs-J5. We have recently completed sequencing a stretch of about 1 kilobase on each side of the *Hin*dIII–*Xba*I fragment and have found no J or J-like sequences. If all Js are clustered and regularly arranged every 300 or so base pairs, there should be no J sequences other than those identified here. On the other hand, if all J peptides (see Table 1) are germ-line encoded, there should be additional J DNA segments elsewhere. Alternatively, some J peptides may be somatically generated (see below).

Precise boundaries of J DNA segments

Our previous nucleotide sequencing studies of embryonic Vλ, embryonic Cλ and myeloma Vλ + Cλ DNA clones identified the site of V–J recombination within one of two alternative pairs of phosphodiester bonds[4]. To obtain similar information for a κ chain gene, it is necessary to identify a germ-line V DNA of a specific κ chain for which the J DNA sequence is known, and determine the nucleotide sequences of both germ-line V DNA and myeloma (V + C) DNA. To this end the germ-line V gene of MOPC321 κ chain, Vκ21-C, was cloned and sequenced (details to be published elsewhere). We also determined the partial nucleotide sequence of the cDNA clone of MOPC321 κ mRNA, 5D10 (ref. 3). Nucleotide sequences around the putative recombination sites on Vκ21-C and J4 (which encodes the Jκ peptide of MOPC321 κ chain, see above) as well as the sequence of the corresponding region of 5D10 are:

```
                    9 5
            AsnGlnAspProPheThrPheGly
5D10        AATGAGGATCCATTCACGTTCGGC

Vκ21-C      AATGAGGATCCTCCCACAGTGGCT

J4          GAATCACTGTGATTCACGTTCGGC
```

Examination of these sequences enabled us to conclude that the recombination is at the phosphodiester bond linking the second and third nucleotides of the proline codon at residue 95. This corresponds to residue 97 in the λ chain numbering system. In the case of one λ gene we have shown that recombination occurs either between the His(97) and Phe(98) codons or between the first and second bases of the latter codon[4]. Thus, the exact sites of recombination in the two cases seem to differ by one or two bases. It remains to be seen whether this difference is due to characteristics of the two recombination systems, one for κ chains and the other for λ chains. An alternative possibility is that the recombinases allow a limited flexibility in the sites of breakage and reunion regardless of the chain type (see below). In this respect it is relevant that all mouse κ chains sequenced to date carry proline at residue 95. As none of the Js reported here can code for this proline (Fig. 2) it is very likely that the first residue fully encoded in J is always at position 96. However, precise recombination sites can only be defined by comparative sequencing analysis of the three DNA segments involved.

Where does coding by J end? The sequences of J4 and 5D10 in the relevant regions are as follows:

```
                        1 0 8
            LeuGluIleLysArgAlaAspAla
5D10        TTGGAAATAAAACGGGCTGATGCT

J4          TTGGAAATAAAACGTAAGTAGACT
```

These sequences indicate that J4 can code for MOPC321 κ chain up to the second base of the Arg(108) codon, CGG. However, as J is followed by an intron preceding the C DNA segment and as most introns start with GT (ref. 12), we suggest that coding by J actually ends with the first base of the triplet CGT, as indicated by the vertical line. As shown in Fig. 3, sequences around the splice site are highly conserved among the five Js. In all cases except for J3, for which no κ chains are known, the doublet GT is present in the last J triplet (CGT) that can partially encode κ chains. It is likely, therefore, that coding by J always ends with the first base of this triplet. In the equivalent position of J3 a Pro codon, CCT, replaces the Arg codon, CGT. It is not clear whether J3 is a genuine J, and if so, where the coding precisely ends.

Inverted repeat-stem structure could be formed between embryonic V and J sequences

Figure 3 compares the sequences of five Js in the coding and flanking regions. Sequences are highly conserved in the coding regions and around the RNA splice sites. In contrast, sequences in the rest of the flanking regions are rather diverse, although the non-coding sequences between J3 and J4, and J4 and J5 are somewhat similar. Nevertheless, there are two short conserved sequences in the 5′-non-coding regions. One is a decamer, GGTTTTTGTA, located about 30 base pairs away from all Js, and the other is a palindromic hexamer interrupted by an AT base pair at the centre of symmetry, CACTGTG, immediately preceding the Js. The same or closely related sequences are also present at the equivalent position in the λI chain gene clone[4].

The significance of these conserved sequences with respect to V–J joining became apparent when the sequences of an embryonic Vκ gene clone, IgVκ21-C, and an embryonic VκI clone,

```
1 Hind III AAGCTTTCGCAGCTACCCACTGCTCTGTTCCTCTTCAGTGAGGAGGGTTTTTGTACAGCCAGACAGTGGAGTACTACCAC 80
           TTCGAAAGCGTCGATGGGTGACGAGACAAGGAGAAGTCACTCCTCCCAAAAACATGTCGGTCTGTCACCTCATGATGGTG

J1 TrpThrPheGlyGlyGlyThrLysLeuGluIleLysArg
TGTGGTGGACGTTCGGTGGAGGCACCAAGCTGGAAATCAAACGTAAGTAGAATCCAAAGTCTCTTTCTTCCGTTGTCTATGTCTGTGGCTTCTATGTCTAAAAATGATGTATAAAATCTT 200
ACACCACCTGCAAGCCACCTCCGTGGTTCGACCTTTAGTTTGCATTCATCTTAGGTTTCAGAGAAAGAAGGCAACAGATACAGACACCGAAGATACAGATTTTTACTACATATTTTAGAA

ACTCTGAAACCAAGATTCTGGCACTCTCCAAGGCAAAGATACAGAGTAACTCCGTTAAGCAAAGCTGGGAATAGGCTAGACATGTTCTCTGGAGAATGAATGCCAGTGTAATAATTAACA 320
TGAGACTTTGGTTCTAAGACCGTGAGAGGTTCCGTTTCTATGTCTCATTGAGGCAATTCGTTTCGACCCTTATCCGATCTGTACAAGAGACCTCTTACTTACGGTCACATTATTAATTGT

CAAGTGATAGTTTCAGAAATGCTCAAAGAAGCAGGGTAGCCTGCCCTAGACAAACCTTTACTCGGTGCTCAGACCATGCTCAGTTTTTGTATGGGGGTTGAGTGAAGGGACACCAGTGTG 440
GTTCACTATCAAAGTCTTTACGAGTTTCTTCGTCCCATCGGACGGGATCTGTTTGGAAATGAGCCACGAGTCTGGTACGAGTCAAAAACATACCCCCAACTCACTTCCCTGTGGTCACAC

J2 TyrThrPheGlyGlyGlyThrLysLeuGluIleLysArg
TGTACACGTTCGGAGGGGGGACCAAGCTGGAAATAAAACGTAAGTAGTCTTCTCAACTCTTGTTCACTAAGTCTAACCTTGTTAAGTTGTTCTTTGTTGTGTGTTTTTCTTAAGGAGATT 560
ACATGTGCAAGCCTCCCCCCTGGTTCGACCTTTATTTTGCATTCATCAGAAGAGTTGAGAACAAGTGATTCAGATTGGAACAATTCAACAAGAAACAACACACAAAAAGAATTCCTCTAA

TCAGGGATTTAGCAAATCCATCTCAGATCAAGTGTTAAGGAGGGAAAACTGCCCACAAGAGGTTGGAATGATTTTCAGGCTAAATTTTAGGCTTTCTAAACCAAAGTAACTAAACTAGGG 680
AGTCCCTAAATCGTTTAGGTAGAGTCTAGTTCACAATTCCTCCCTTTTGACGGGTGTTCTCCAACCTTACTAAAAGTCCGATTTAAAATCCGAAAGATTTGGTTTCATTGATTTGATCCC

                                                                  J3 IleThrPheSerAspGlyThrArgLeuGluIleLysPro
GAAGAGGGATAATTGTCTACCTAGGGAGGGTTTTGTGGAGGTAAAGTTAAAATAAATCACTGTAAATCACATTCAGTGATGGGACCAGACTGGAAATAAAACCTAAGTACATTTTTGCTC 800
CTTCTCCCTATTAACAGATGGATCCCTCCCAAAACACCTCCATTTCAATTTTATTTAGTGACATTTAGTGTAAGTCACTACCCTGGTCTGACCTTTATTTTGGATTCATGTAAAAACGAG

AACTGCTTGTGAAGTTTTGGTCCCATTGTGTCCTTTGTATGAGTTTGTGGTGTACATTAGATAAATGAACTATTCCTTGTAACCCAAAACTTAAATAGAAGAGAACCAAAAATCTAGCTA 920
TTGACGAACACTTCAAAACCAGGGTAACACAGGAAACATACTCAAACACCACATGTAATCTATTTACTTGATAAGGAACATTGGGTTTTGAATTTATCTTCTCTTGGTTTTTAGATCGAT

CTGTACAAGCTGAGCAAACAGACTGACCTCATGTCAGATTTGTGGGAGAAATGAGAAAGGAACAGTTTTTCTCTGAACTTAGCCTATCTAACTGGATCAGCCTCAGGCAGGTTTTTGTAA 1040
GACATGTTCGACTCGTTTGTCTGACTGGAGTACAGTCTAAACACCCTCTTTACTCTTTCCTTGTCAAAAAGAGACTTGAATCGGATAGATTGACCTAGTCGGAGTCCGTCCAAAAACATT

                             J4 PheThrPheGlySerGlyThrLysLeuGluIleLysArg
AGGGGGGCGCAGTGATATGAATCACTGTGATTCACGTTCGGCTCGGGGACAAAGTTGGAAATAAAACGTAAGTAGACTTTTGCTCATTTACTTGTGACGTTTTGGTTCTGTTTGGGTAAC 1160
TCCCCCCGCGTCACTATACTTAGTGACACTAAGTGCAAGCCGAGCCCCTGTTTCAACCTTTATTTTGCATTCATCTGAAAACGAGTAAATGAACACTGCAAAACCAAGACAAACCCATTG

TTGTGTGAATTTGTGACATTTTGGCTAAATGAGCCATTCCTAGCAACCTGTGCATCAATAGAAGATCCCCCAGAAAAGAGTCAGTGTGAAAGCTGAGCGAAAAACTCGTCTTAGGCTTCT 1280
AACACACTTAAACACTGTAAAACCGATTTACTCGGTAAGGATCGTTGGACACGTAGTTATCTTCTAGGGGGTCTTTTCTCAGTCACACTTTCGACTCGCTTTTTGAGCAGAATCCGAAGA

GAGACCAGTTTTGTAAGGGGAATGTAGAAGAAAGAGCTGGGCTTTTCCTCTGAATTTGGCCCATCTAGTTGGACTGGCTTCACAGGCAGGTTTTTGTAGAGAGGGGCATGTCATAGTCCT 1400
CTCTGGTCAAAACATTCCCCTTACATCTTCTTTCTCGACCCGAAAAGGAGACTTAAACCGGGTAGATCAACCTGACCGAAGTGTCCGTCCAAAAACATCTCTCCCCGTACAGTATCAGGA

    J5 LeuThrPheGlyAlaGlyThrLysLeuGluLeuLysArg
CACTGTGGCTCACGTTCGGTGCTGGGACCAAGCTGGAGCTGAAACGTAAGTACACTTTTCTCATCTTTTTTTATCTGTAAGACACAGGTTTTCATGTTGGAGTTAAAGTCAGTTCAGAAA 1520
GTGACACCGAGTGCAAGCCACGACCCTGGTTCGACCTCGACTTTGCATTCATGTGAAAAGAGTAGAAAAAAATAGACATTCTGTGTCCAAAAGTACAACCTCAATTTCAGTCAAGTCTTT

ATCTTGAGAAAATGGAGAGGGCTCATTATCAGTTGACGTGGCATACAGTGTCAGATTTTCTGTTTATATCAGCTAGTGAGATTAGGGGCAAAAAGAGGCTTTAGTTGAGAGGAAAGTAAT 1640
TAGAACTCTTTTACCTCTCCCGAGTAATAGTCAACTGCACCGTATGTCACAGTCTAAAAGACAAATATAGTCGATCACTCTAATCCCCGTTTTTCTCCGAAATCAACTCTCCTTTCATTA

TAATACTATGGTCACCATCCAAGAGATTGGATCGGAGAATAAGCATGAGTAGTTATTGAGATCTGGGTCTGACTGCAGGTAGCGTGGTCTTCTAGA
ATTATGATACCAGTGGTAGGTTCTCTAACCTAGCCTCTTATTCGTACTCATCAATAACTCTAGACCCAGACTGACGTCCATCGCACCAGAAGATCT Xba I 1736
```

Fig. 2 Complete nucleotide sequence of the 1.7-kilobase *Hin*dIII–*Xba*I fragment containing the J cluster. For sequencing, a 2.7-kilobase *Hin*dIII–*Hin*dIII fragment carrying the J cluster (Fig. 1*b*) was subcloned into a plasmid vector pBR322 (ref. 32). A 300-μg portion of the chimaeric plasmid DNA was digested with a mixture of *Hin*dIII and *Xba*I and separated in a 6% polyacrylamide gel (20 × 40 × 0.5 cm); the 1.7-kilobase *Hin*dIII–*Xba*I fragment (Fig. 1*b*) was then purified. A 5-μg aliquot of this fragment was digested with *Hin*fI, *Mbo*II, *Alu*I, *Ava*I, *Hph*I, *Hae*III or *Ava*II, end-labelled and sequenced by the method of Maxam and Gilbert[33] as described previously[26]. In some experiments, 3′-ends of the DNA fragments were labelled with [α-^{32}P]nucleoside triphosphates (NTPs) using the Klenow fragment of *Escherichia coli* DNA polymerase I (ref. 34). For this, 1 μg of DNA was incubated with 20 μCi each of four [α-^{32}P]-NTPs (3,000 Ci mmol^{-1}, NEN) and 1 unit of the Klenow enzyme (Boehringer, Mannheim) at 25 °C for 20 min in 50 μl of 6 mM Tris-HCl, *p*H 7.5, 50 mM NaCl, 6 mM $MgCl_2$ and 6 mM β-mercaptoethanol. Aliquots (50 mM) of each of the four NTPs and 0.5 unit of the enzyme were then added to the reaction mix and the incubation was continued for another 10 min. The reaction was terminated by heating the mixture at 70 °C for 10 min and DNA was precipitated by ethanol. For strand separation the labelled DNA samples were dissolved in 20 μl of 50% (v/v) dimethyl sulphoxide, 10 mM Tris-borate, *p*H 8.3, 1 mM EDTA, 0.5% (v/v) xylene cyanol and 0.05% (v/v) bromphenol blue, heated at 90 °C for 3 min, chilled on ice and fractionated in a 6% or 8% strand-separation gel (bisacrylamide/acrylamide = 1/62.5) (A. Maxam, personal communication). Amino acid sequences predicted from the identified J DNA segments are shown in italics.

Ig99λ (ref. 4), were examined. Both V DNA clones contain sequences closely related to the decamer and heptamer described above on the anti-sense strands in the 3′-non-coding regions. The presence of these sequences could permit the formation of an inverted repeat stem structure between the 3′-end of an embryonic V sequence and the 5′-end of a J sequence. Examples of such stem structures are shown in Fig. 4*a–d*. Vκ21-C can form a stem structure of similar stability with any one of the five Js. Another Vκ gene, VK-2 (ref. 13), which codes for a V region remotely related to the MOPC321 V region, can also form such a stem structure with any one of the five Jκs (Fig. 4*c*). We note that the majority of bases conserved in the 3′-non-coding regions of the two Vκ genes participate in base pairing with Jκs. A stem structure can also be constructed between embryonic Vλ and embryonic Jλ, although the homology is lower than that of the Vκ–Jκ stems (Fig. 4*d*). We propose that the hybrid stem structures are part of the recognition signals for the putative site-specific recombinase.

In all cases known, a few base pairs occur between either one or both cut points of recombination and the stem bases. Thus, for the V–J joining leading to the MOPC321 κ chain gene, the only case in which the exact cut points are known, three base pairs intervene between the cuts on V21κ-C and the stem bases, whereas the cuts on J4 are actually at the bases (Fig. 4*a*). We do not know whether this relationship between the cut points and the stem bases applies to other recombination events between various pairs of Vκs and Jκs, nor whether recombination between Vκ21-C and J4 always occurs at the same sites as shown in Fig. 4*a*. As residue 96 is a hot spot[10,11], one intriguing possibility is that the recombinase can cut and rejoin the DNA strands of a given pair of V and J at slightly different positions so that different triplet codons are reconstituted at the recombination site. For example, the hypothetical recombination events between V21κ-C and J1 (Fig. 4*b*) may occur in any one of

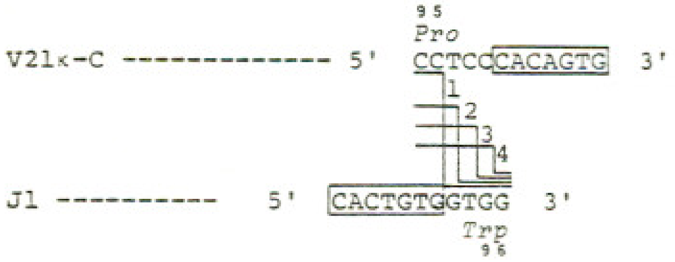
Table 1 Classification of BALB/c and NZB myeloma κ chains

Group	Myeloma	Jκ Peptide 96												108	Jκ DNA Segment
I	M141* M63 B32 2880 7210 M70 M41 A17 1229 6308 C101 T111 X-44 7769	Trp	Thr	Phe	Gly	Gly	Gly	Thr	Lys	Leu	Glu	Ile	Lys	Arg	J1
II	M21 A22 9245 3741	Tyr	Thr	Phe	Gly	Gly	Gly	Thr	Lys	Leu	Glu	Ile	Lys	Arg	J2
III	M173 6684	Arg	Thr	Phe	Gly	Gly	Gly	Thr	Lys	Leu	Glu	Ile	Lys	Arg	
IV	M11 7940	Pro	Thr	Phe	Gly	Gly	Gly	Thr	Lys	Leu	Glu	Ile	Lys	Arg	
V	T124	Trp	Thr	Phe	Gly	Ser	Gly	Thr	Lys	Leu	Glu	Ile	Lys	Arg	
VI	M321 7043	Phe	Thr	Phe	Gly	Ser	Gly	Thr	Lys	Leu	Glu	Ile	Lys	Arg	J4
VII	X-24	Ile	Thr	Phe	Gly	Ser	Gly	Thr	Lys	Leu	Glu	Ile	Lys	Arg	
VIII	LPC1* M511 4045 3154 HOPC8* 7175 2485 7183	Leu	Thr	Phe	Gly	Ala	Gly	Thr	Lys	Leu	Glu	Leu	Lys	Arg	J5
IX	T601	Ile	Thr	Phe	Gly	Ala	Gly	Thr	Lys	Leu	Glu	Leu	Lys	Arg	
X	2413	Trp	Thr	Phe	Gly	Gly	Gly	Thr	Asp	Leu	Glu	Ile	Glu	Arg	
XI		Ile	Thr	Phe	Ser	Asp	Gly	Thr	Arg	Leu	Glu	Ile	Lys	Pro	J3

BALB/c and NZB myeloma κ chains are classified into 10 groups according to the amino acid sequence of the Jκ peptide (residues 96–108)[10,11]. The DNA segments coding for four Jκ peptides, J1, J2, J4 and J5, have been identified on the embryonic Cκ clone, Ig146κ, in the present study. No myeloma κ chain is known for the fifth J DNA segment, J3, also identified in the present study. Those amino acid residues different from the corresponding residues of the J1 peptide are underlined.

* Myelomas classified by R-loop mapping rather than by amino acid sequencing (see text).

the following four ways:



where recombinations 1 and 2 give the same dipeptide Pro(95)–Trp(96), whereas recombinations 3 and 4 generate Pro(95)–Arg(96) and Pro(95)–Pro(96), respectively. The latter two recombination events can account for the group III and group IV J peptides, respectively (see Table 1). It follows that the total number of Jκ DNA segments might be considerably less than the number of different Jκ peptides.

The inverted repeat structures described above have some resemblance to those found at the ends of prokaryotic insertion elements. Ohtsubo and Ohtsubo[14], and Grindley[15] determined sequences of IS1 and found that some 30 bases at one end are invertedly repeated on the other end. Inverted repeat sequences of similar length have also been found at the ends of most insertion elements. Integration of insertion elements into host DNA generates 5- or 9-base pair direct repeats at the outer ends of the inverted repeat sequences[15–22,36]. The conserved palindromic sequences, CACAGTG near V genes and CACTGTG near Js, can be considered to be equivalent to these direct repeats observed at the margins of integrated insertion elements. Although these structural similarities may be coincidental, they could reflect a common evolutionary origin or a common enzymatic mechanism for the two recombination systems, or both.

V–J joining probably results in deletion

Several models have previously been proposed for recombination of immunoglobulin genes[2]. In the 'copy–insertion' model, a specific V DNA segment is duplicated and the copy is inserted at a site adjacent to a J DNA segment. The 'excision–insertion' model suggests that a specific V DNA segment is excised into an episome-like structure, and this in turn is integrated adjacent to a J DNA segment. According to the 'deletion model' the entire DNA in the interval between a particular pair of V and J DNA segments loops out, is excised, and is diluted out on subsequent cell multiplication. In the 'inversion' model the V and the C DNA (also J DNA) segments are arranged in opposite directions, and a segment of chromosome between a particular V DNA (inclusive) and a J DNA (exclusive) is inverted.

In the context of these models, let us consider the fate of the DNA sequences immediately adjacent to the 3′-end of V, or immediately adjacent to the 5′-end of J. In all the models except for the deletion model, either one of the two sequences or both should recombine with DNA sequences elsewhere in the genome on V–J joining. In contrast, the deletion model predicts that both sequences should be eliminated. To test these two alternatives, we investigated the λI chain gene in which both V and J are unique. We digested embryo and myeloma (H2020) DNA with various restriction enzymes, fractionated the digests in a 0.8% agarose gel, and analysed the DNA by Southern gel-blotting procedures. As the hybridisation probes we used a DNA fragment of about 300 base pairs from an embryonic VλI clone, Ig99λ, or from an embryonic CλI clone, Ig25λ, that includes the respective recombination site as shown in Fig. 5*E*. Both DNA fragments contain not only the sequence to be examined (*b* or *c*) but also an adjacent sequence (*a* or *d*) known to be incorporated into the rearranged λ chain gene. We deliberately chose such DNA fragments as hybridisation probes to let the *a* and *d* part serve as internal hybridisation controls.

Figure 5*A* compares hybridisation patterns of embryo and myeloma DNA that were digested with *Eco*RI and analysed with the *ab* probe. Embryo DNA gave two bands of 4.8 and 3.5 kilobases that had previously been identified as VλII- and VλI-containing fragments, respectively[3–5]. In the myeloma DNA an additional band of 7.4 kilobases known to contain the rearranged, complete λI gene[4] was observed, but no previously unidentified band was present, which would have been expected if the DNA sequence corresponding to the *b* part of the probe had recombined with another DNA sequence. Results of analogous experiments obtained with the *cd* probe are shown in Fig. 5*C*. Embryo DNA gave a band of 8.6 kilobases known to contain the CλI sequence, and another band, X. The latter band contained a sequence homologous to JλI and its flanking regions (unpublished observations). Myeloma DNA gave these two bands and an additional band of 7.4 kilobases containing the rearranged λI chain gene[3]. Again, we observed no myeloma-specific band that could be attributed to the hypothetically recombined DNA containing the *c* part of the probe. Similar results obtained with *Hin*dIII are shown in Fig. 5*B* and *D*. We did not observe myeloma-specific DNA fragments that could not be accounted for by the rearranged, complete λI chain gene with either probe. We have carried out similar experiments using *Kpn*I, *Bam*HI, *Hin*dII and *Hae*III, and obtained results that lead us to the same conclusion (data not shown). Although the results do not unequivocally prove the deletion model (more complicated models in which elements of the other models are combined cannot be ruled out), they suggest strongly that it is the correct one.

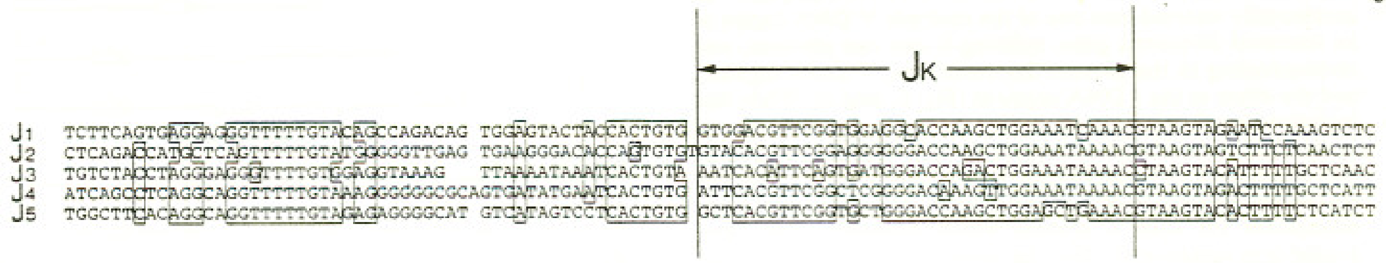

Jκ
J1 TCTTCAGTGAGGAGGGTTTTTGTACAGCCAGACAG TGGAGTACTACCACTGTG GTGGACGTTCGGTGGAGGCACCAAGCTGGAAATCAAACGTAAGTAGAATCCAAAGTCTC
J2 CTCAGACCATGCTCAGTTTTTGTATGGGGGTTGAG TGAAGGGACACCAGTGTGTGTACACGTTCGGAGGGGGGACCAAGCTGGAAATAAAACGTAAGTAGTCTTCTCAACTCT
J3 TGTCTACCTAGGGAGGGTTTTGTGGAGGTAAAG TTAAAATAAATCACTGTA AATCACATTCAGTGATGGGACCAGACTGGAAATAAAACCTAAGTACATTTTTGCTCAAC
J4 ATCAGCCTCAGGCAGGTTTTTGTAAAGGGGGGCGCAGTGATATGAATCACTGTG ATTCACGTTCGGCTCGGGGACAAAGTTGGAAATAAAACGTAAGTAGACTTTTTGCTCATT
J5 TGGCTTCACAGGCAGGTTTTTGTAGAGAGGGGCAT GTCATAGTCCTCACTGTG GCTCACGTTCGGTGCTGGGACCAAGCTGGAGCTGAAACGTAAGTACACTTTTTCTCATCT

Fig. 3 Comparison of nucleotide sequences of the five J-coding and the flanking regions. Nucleotides common in at least four Js are outlined. Vertical lines indicate determined or predicted ends of the J-coding DNA segments.

Evolution of light-chain genes

We thus suggest that V–J joining is a recombination event accompanied by excision of the DNA sequence between the V and J DNA segments. This DNA sequence, which we call *excison*, carries at its ends characteristic inverted repeat sequences consisting of about 30 bases that could form a stem-loop structure somewhat similar to those of prokaryotic insertion elements. In the light of these findings and the intron–exon structure of immunoglobulin genes we will consider some aspects of the evolution of light-chain genes.

The V and C regions of light chains show a weak sequence homology and are similar in their size, in arrangement of intra-chain disulphide bonds and in three-dimensional structure[23–25]. Based on these observations it has been generally thought that the DNA segments coding for the two regions arose by duplication of a primordial gene coding for a single homology unit (Fig. 6, step *a*). As we discussed elsewhere this duplication step probably generated a 'dimeric gene' in which the ancestral V- and C-coding DNA segments are separated by a spacer[26]. When this gene is expressed, the entire DNA is transcribed into a single RNA molecule and the spacer sequence is presumed to have been removed by splicing of the V- and C-coding RNA sequences. We assume that the J DNA segment was an integral part of the ancient V DNA because the J region is homologous to the carboxyl end of the C region[35].

The ancestral V DNA, together with some flanking sequences, duplicated, triplicated, and so on under pressure to increase V-region diversity (Fig. 6, step *b*). As splicing is most probably an intramolecular reaction, one requirement of such a 'polymeric gene' would be that the V and C DNA sequences are co-transcribed. In the long RNA transcript every V-coding sequence would have one half-pair of RNA splice sites at its 3′-end (the other half-pair is at the 5′-end of the C-coding sequence), so that a series of mature mRNAs containing different V-coding sequence and the same C-coding sequence could be generated. However, the V–C co-transcription requirement would impose a limit on the size of such a polymeric gene because the transcription unit would become intolerably long as the gene grows. To accommodate 200 V-coding DNA segments, a number estimated for mouse κ chains[2,8,13,27,28], the transcript would have to be 100 kilobases long even if no spacer between two adjacent V DNA segments is postulated. The hypothetical transcript may be much larger because at least one such spacer seems to be 10 times longer than a V DNA segment[9]. One obvious solution is to duplicate the entire polymeric gene of a still allowable size. Indeed, all mammals studied so far seem to carry several unlinked sets of light-chain V DNA clusters each associated with a C-coding DNA segment.

An alternative solution is to rearrange DNA sequences somatically so that the V DNA to be expressed is brought in the vicinity of the C DNA in the lymphocytes. We propose that such

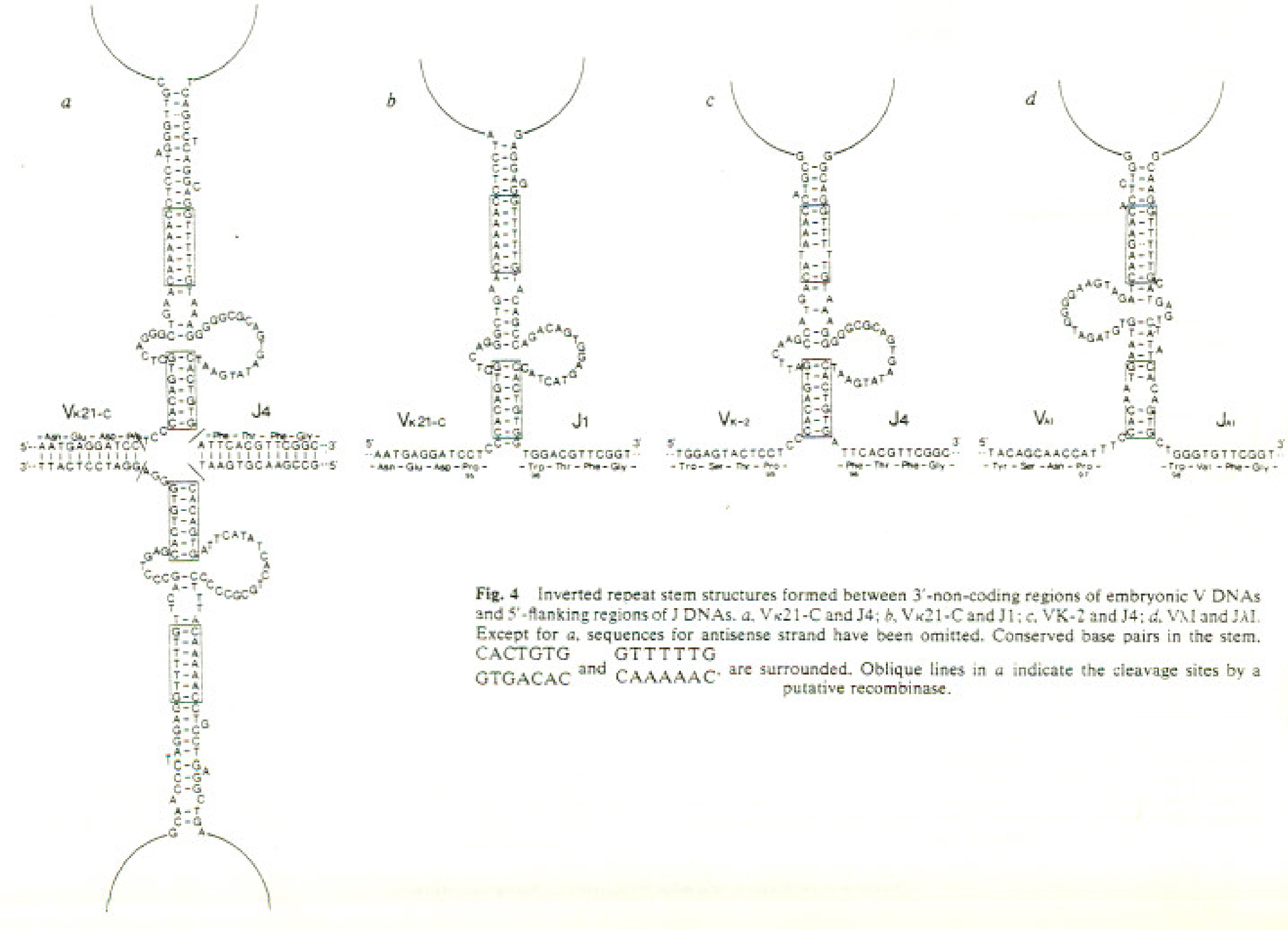


Fig. 4 Inverted repeat stem structures formed between 3′-non-coding regions of embryonic V DNAs and 5′-flanking regions of J DNAs. *a*, Vκ21-C and J4; *b*, Vκ21-C and J1; *c*, VK-2 and J4; *d*, Vλl and Jλl. Except for *a*, sequences for antisense strand have been omitted. Conserved base pairs in the stem, $\begin{matrix}\text{CACTGTG}\\ \text{GTGACAC}\end{matrix}$ and $\begin{matrix}\text{GTTTTTG}\\ \text{CAAAAAC}\end{matrix}$, are surrounded. Oblique lines in *a* indicate the cleavage sites by a putative recombinase.

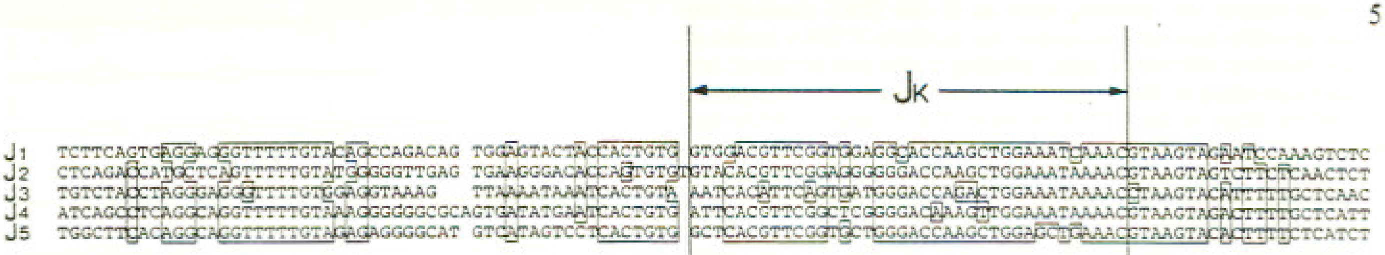

Jκ
J1 TCTTCAGTGAGGAGGGTTTTTGTACAGCCAGACAG TGGAGTACTACCACTGTG GTGGACGTTCGGTGGAGGCACCAAGCTGGAAATCAAACGTAAGTAGAATCCAAAGTCTC
J2 CTCAGACCATGCTCAGTTTTTGTATGGGGGTTGAG TGAAGGGACACCAGTGTGTGTACACGTTCGGAGGGGGGACCAAGCTGGAAATAAAACGTAAGTAGTCTTCTCAACTCT
J3 TGTCTACCTAGGGAGGGTTTTGTGGAGGTAAAG TTAAAATAAATCACTGTA AATCACATTCAGTGATGGGACCAGACTGGAAATAAAACTTAAGTACATTTTTGCTCAAC
J4 ATCAGCCTCAGGCAGGTTTTTGTAAAGGGGGGCGCAGTGATATGAATCACTGTG ATTCACGTTCGGCTCGGGGACAAAGTTGGAAATAAAACGTAAGTAGACTTTTGCTCATT
J5 TGGCTTGACAGGCAGGTTTTTGTAGAGAGGGGCAT GTCATAGTCCTCACTGTG GCTCACGTTCGGTGCTGGGACCAAGCTGGAGCTGAAACGTAAGTACACTTTTCTCATCT

Fig. 3 Comparison of nucleotide sequences of the five J-coding and the flanking regions. Nucleotides common in at least four Js are outlined. Vertical lines indicate determined or predicted ends of the J-coding DNA segments.

Evolution of light-chain genes

We thus suggest that V–J joining is a recombination event accompanied by excision of the DNA sequence between the V and J DNA segments. This DNA sequence, which we call *excison*, carries at its ends characteristic inverted repeat sequences consisting of about 30 bases that could form a stem-loop structure somewhat similar to those of prokaryotic insertion elements. In the light of these findings and the intron–exon structure of immunoglobulin genes we will consider some aspects of the evolution of light-chain genes.

The V and C regions of light chains show a weak sequence homology and are similar in their size, in arrangement of intra-chain disulphide bonds and in three-dimensional structure[23–25]. Based on these observations it has been generally thought that the DNA segments coding for the two regions arose by duplication of a primordial gene coding for a single homology unit (Fig. 6, step *a*). As we discussed elsewhere this duplication step probably generated a 'dimeric gene' in which the ancestral V- and C-coding DNA segments are separated by a spacer[26]. When this gene is expressed, the entire DNA is transcribed into a single RNA molecule and the spacer sequence is presumed to have been removed by splicing of the V- and C-coding RNA sequences. We assume that the J DNA segment was an integral part of the ancient V DNA because the J region is homologous to the carboxyl end of the C region[35].

The ancestral V DNA, together with some flanking sequences, duplicated, triplicated, and so on under pressure to increase V-region diversity (Fig. 6, step *b*). As splicing is most probably an intramolecular reaction, one requirement of such a 'polymeric gene' would be that the V and C DNA sequences are co-transcribed. In the long RNA transcript every V-coding sequence would have one half-pair of RNA splice sites at its 3′-end (the other half-pair is at the 5′-end of the C-coding sequence), so that a series of mature mRNAs containing different V-coding sequence and the same C-coding sequence could be generated. However, the V–C co-transcription requirement would impose a limit on the size of such a polymeric gene because the transcription unit would become intolerably long as the gene grows. To accommodate 200 V-coding DNA segments, a number estimated for mouse κ chains[2,8,13,27,28], the transcript would have to be 100 kilobases long even if no spacer between two adjacent V DNA segments is postulated. The hypothetical transcript may be much larger because at least one such spacer seems to be 10 times longer than a V DNA segment[9]. One obvious solution is to duplicate the entire polymeric gene of a still allowable size. Indeed, all mammals studied so far seem to carry several unlinked sets of light-chain V DNA clusters each associated with a C-coding DNA segment.

An alternative solution is to rearrange DNA sequences somatically so that the V DNA to be expressed is brought in the vicinity of the C DNA in the lymphocytes. We propose that such

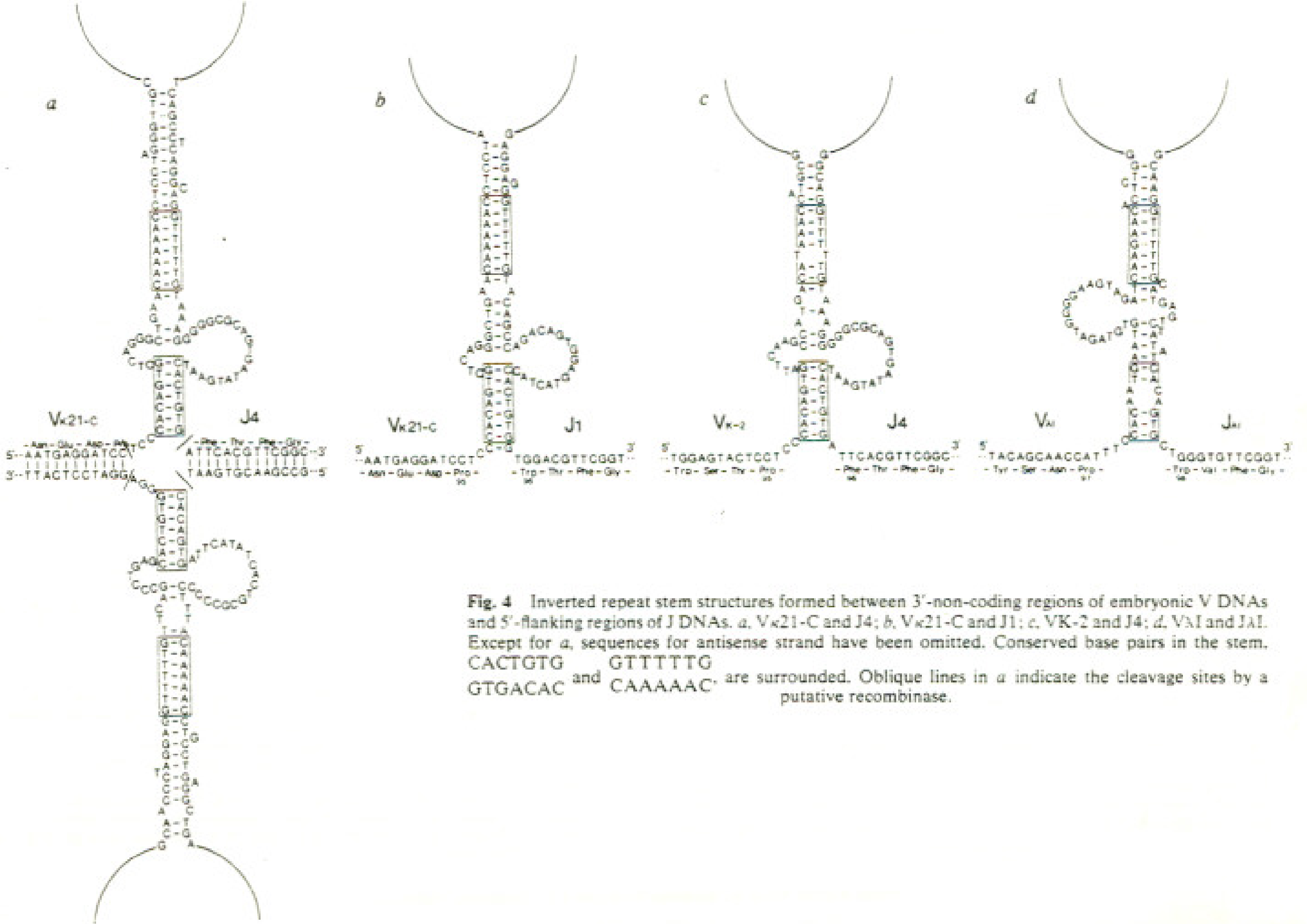


Fig. 4 Inverted repeat stem structures formed between 3′-non-coding regions of embryonic V DNAs and 5′-flanking regions of J DNAs. *a*, Vκ21-C and J4; *b*, Vκ21-C and J1; *c*, VK-2 and J4; *d*, VλI and JλI. Except for *a*, sequences for antisense strand have been omitted. Conserved base pairs in the stem, $\begin{matrix}\text{CACTGTG}\\\text{GTGACAC}\end{matrix}$ and $\begin{matrix}\text{GTTTTTG}\\\text{CAAAAAC}\end{matrix}$, are surrounded. Oblique lines in *a* indicate the cleavage sites by a putative recombinase.