Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 1566-1570, March 1988
Genetics

Immunoglobulin heavy chain variable region gene evolution: Structure and family relationships of two genes and a pseudogene in a teleost fish

MELANIE R. WILSON, DARLENE MIDDLETON, AND GREGORY W. WARR*

Department of Biochemistry and Molecular Biology, Medical University of South Carolina, Charleston, SC 29425

Communicated by Susumu Ohno, October 2, 1987

ABSTRACT **Nucleotide sequences for two immunoglobulin heavy chain variable region (V_H) genes and one pseudogene in the goldfish (*Carassius auratus*) and the family relationships and distribution of these genes in individual fish are presented. Comparison of the nucleotide and inferred amino acid sequences of goldfish and other vertebrate V_H genes indicates that goldfish V_H genes show the major V_H gene regulatory and structural features (5′-putative promoter region, split hydrophobic leader, three framework and two complementarity-determining regions, and 3′-recombination signals for V_H to diversity region joining) and that goldfish V_H genes are not more closely related to one another than they are to V_H genes of evolutionarily distant vertebrates such as the mammals. Goldfish V_H genes appear to exist in distinct families, and individual goldfish can carry from none to apparently >15 genes of a given family. These results suggest that whereas the basic structure of V_H genes has been conserved in evolution, there may be substantial variation in the nature and population distribution of V_H gene families in the vertebrates.**

The great diversity of antibody specificities is derived from multiple germ-line copies of the gene segments encoding the antigen-binding site, the variety of possible recombinations of these elements, and somatic events that introduce mutations and junctional imprecision during recombination (1–3). Of the three elements [heavy chain variable region (V_H), diversity, and heavy chain joining region] encoding the binding-site-containing domain of the immunoglobulin heavy chain in mammals, the V_H genes are present in the greatest number in the germ line and have also been suggested to encode the region that is the primary determinant of binding-site specificity (4). In the course of vertebrate evolution, the immune system appears to have increased in complexity, for example, in terms of the appearance of multiple structurally diverse classes of antibody (5), and in its efficiency, e.g., the appearance of antibodies with higher-affinity binding sites, and in the phenomenon of rapid increase of antibody affinity during an antibody response. A complete understanding of the evolution of the immunoglobulin molecule will be dependent on our knowing the structure, diversity, organization, and expression of the immunoglobulin gene elements in diverse vertebrate groups. Whereas the observation of the unusual organization of V_H, diversity, heavy chain constant, and joining region genes in an elasmobranch (6) in multiple repeating clusters of ≈10 kilobases is clearly consistent with the restricted antibody response seen in this species (7), it is not known if this pattern of gene organization is restricted to the elasmobranchs or is representative of the ectothermic vertebrates in general. The teleost fish are one of the major vertebrate classes for which large gaps exist in our knowledge. No information on the primary structure of their immunoglobulin molecules or the genes encoding them is available. Here we describe the structure of V_H genes in a teleost fish,† the goldfish (*Carassius auratus*), and present evidence that the organization and expression of V_H gene families in this species is different from what has been observed in the mammals.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Cloned Genomic DNA of the Goldfish. The construction of a genomic DNA library in phage λ EMBL4, its screening under P1/EK1 containment, and the initial characterization of seven recombinant phage showing cross-hybridization with the S107 murine V_H probe (8) have been described elsewhere (9). DNA fragments of interest were subcloned and sequenced by the dideoxynucleoside triphosphate chain-termination method (10).

DNA Blotting. Genomic DNA (10 μg) or phage DNA (1.0 μg) was digested to completion with restriction enzymes, electrophoresed in one dimension on agarose gels, and blotted onto nitrocellulose sheets by standard techniques (11), as described (9). DNA probes were nick-translated to a specific activity of 2–5 × 10^8 cpm/μg, and hybridization was carried out exactly as described (9). Filters were washed four times for 30 min under low-stringency conditions (52°C; 0.15 M NaCl/0.015 M sodium citrate/0.1% $NaDodSO_4$) or high-stringency conditions (65°C; 0.015 M NaCl/0.0015 M sodium citrate/0.1% $NaDodSO_4$) before drying and exposing to x-ray film at −70°.

RESULTS AND DISCUSSION

Structure of V_H Genes in the Goldfish. The functional regions of and sequencing strategy for two V_H genes and one pseudogene cross-hybridizing with the murine S107 probe are shown in Fig. 1*A*. Gene *5A* and pseudogene *5B* were derived from the same recombinant λ phage, a partial restriction map of which is shown in Fig. 1*B*. Segments *5A* and *5B* were separated by ≈5.7 kilobases and were in the same transcriptional orientation. The nucleotide sequences (Fig. 2) indicated that the hybridizing regions of *3* and *5A* represented apparently functional V_H genes. They possessed 5′-flanking sequences typical of V_H promoters, including the highly conserved (12) octameric sequence (ATGCAAAT, position −145, Fig. 2) and possible "TATA boxes" (TTAAT, position −117 in gene *3*, and ATGAAAA, position

Abbreviation: V_H, immunoglobulin heavy chain variable region gene.

*To whom reprint requests should be addressed at: Department of Biochemistry and Molecular Biology, Medical University of South Carolina, 171 Ashley Avenue, Charleston, SC 29425.

†The sequences reported in this paper are being deposited in the EMBL/GenBank data base (Bolt, Beranek, and Newman Laboratories, Cambridge, MA, and Eur. Mol. Biol. Lab., Heidelberg) (accession nos. J03616, goldfish V_H gene *3*; J03617, goldfish V_H gene *5A*; and J03618, goldfish pseudogene *5B*).

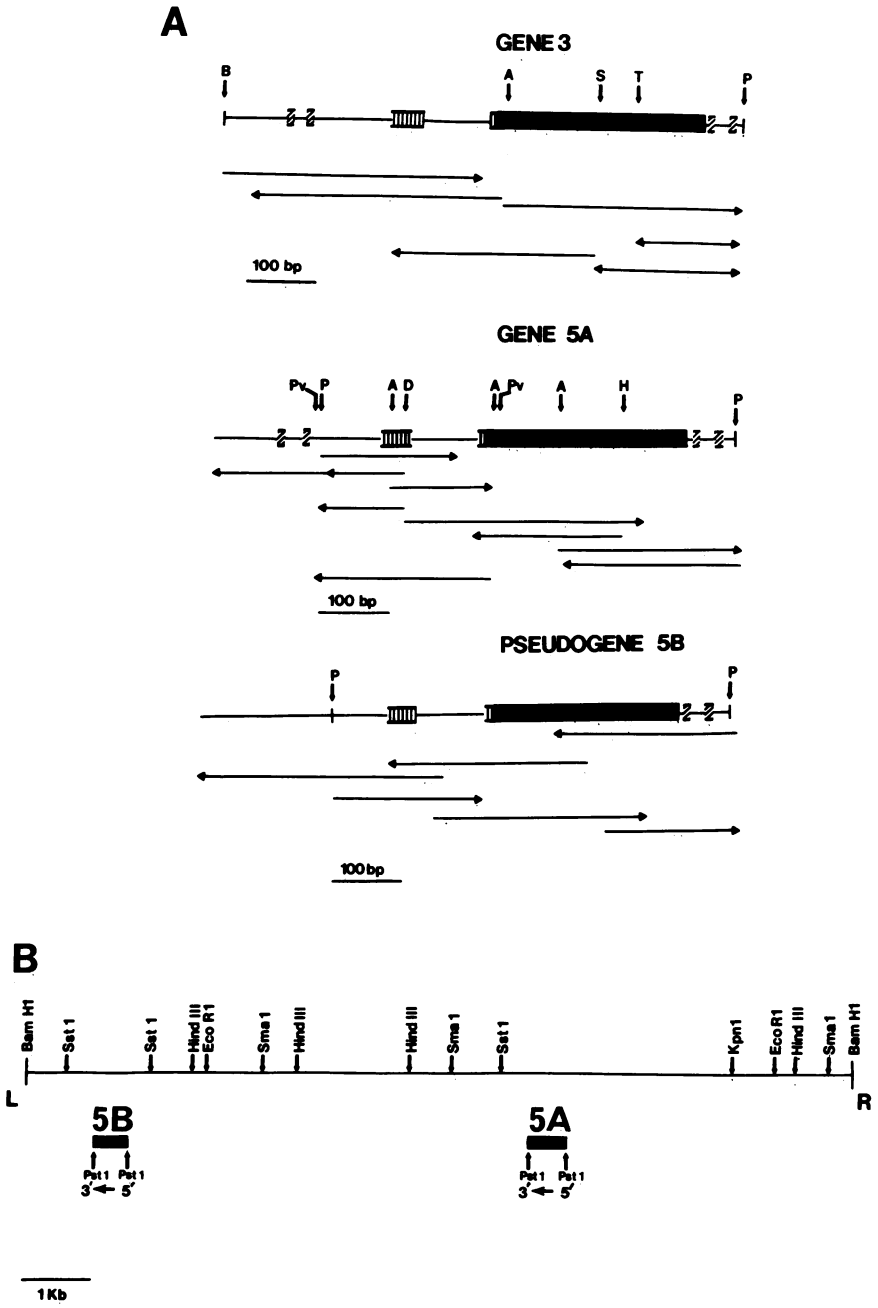

FIG. 1. Partial restriction maps of goldfish V_H-positive DNA showing the sequencing strategy for genes *3* and *5A* and pseudogene *5B* (*A*) and the localization of gene *5A* and pseudogene *5B* on recombinant λ phage 5 (*B*). (*A*) ▬, regions encoding mature polypeptide; ▥, regions encoding leader; hatched box, regulatory regions (putative promoter octamer sequence and TATA box and 3′-recombination signals). The arrows indicate sequences determined from cloning sites or by extension priming. Sequence upstream of the 5′ *Pvu* II site on gene *5* was determined on a *Pvu* II–*Pvu* II partial digest fragment. Sequence upstream of the 5′ *Pst* I site on pseudogene *5B* was determined on an *Sst* I–*Sst* I fragment (*B*). Restriction sites: A, *Alu* I; B, *Bam*HI; D, *Dra* I; H, *Hae* III; S, *Sma* I; T, *Taq* I; Pv, *Pvu* II. (*B*) Complete and partial digestion products of end-labeled DNA were used to derive the map. The regions containing the S107 cross-hybridizing DNA, which defined the majority of gene *5A* and pseudogene *5B* and which were bounded by *Pst* I sites, are indicated by the solid boxes. Gene orientation (noncoding strand) is indicated by the arrows underneath these boxes. L and R, left and right arms, respectively, of the EMBL4 vector.

− 119 in gene *5A*). Genes *3* and *5A*, encoded (Fig. 3) in open reading frame, complete V_H genes comprising three framework regions, two complementarity-determining regions, and a hydrophobic 19-amino acid NH_2-terminal signal peptide, the coding region for which was split by a single intron. The DNA on the 3′ side of the coding region contained sequences typical of recombination signals [conserved heptamer-23-base-pair (bp) spacer-conserved nonamer] for V_H to diversity region joining. The S107 cross-hybridizing region *5B* was an apparent pseudogene (Fig. 2). It contained a stop codon in the first framework-encoding region and lacked the putative promoter region (conserved octamer and TATA boxes), although in all other aspects it resembled a typical V_H gene. Thus, goldfish V_H genes resemble, in their major structural features, the apparently universal pattern of V_H gene structure seen not only in mammals but also in reptiles and elasmobranchs (13, 14).

Translation of the nucleotide sequences for genes *3* and *5A* into inferred amino acids (Fig. 3) showed, in an alignment with S107, a clustering of sequence identities in the framework regions (especially near the cysteines) and the much greater divergence of sequence in the complementarity-

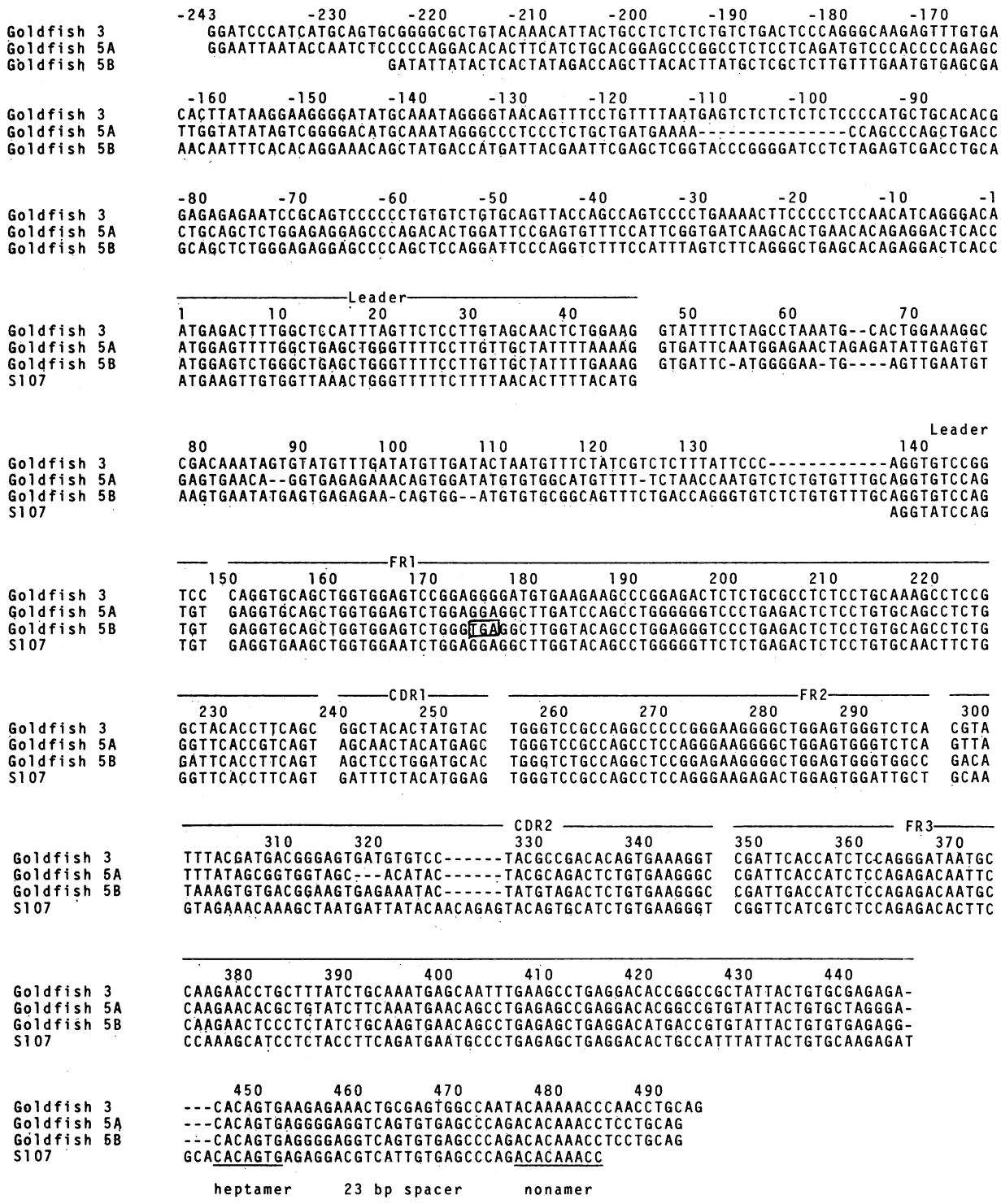

FIG. 2. Nucleotide sequences of goldfish V_H genes *3* and *5A* and pseudogene *5B* are aligned with that of mouse V_H gene S107 (8). Within the coding regions, gaps were introduced only within complementarity-determining region 2 (CDR2). Gaps outside the coding region were introduced to allow alignment of the conserved octameric sequence ATGCAAAT (position −145), the start codon, and the conserved variable–diversity recombination signals on the 3′ side of the coding region. In addition, some gaps were introduced to allow alignment of obviously similar regions of sequence in the intron splitting the leader region. The numbering refers to gene *3*. The stop codon of gene *5B* in framework region 1 (FR1) is boxed.

determining regions. Alignment of the sequences could be accomplished without the introduction of gaps in any part of the coding region except for the second complementarity-determining region. Variations in the length of the second complementarity-determining region are characteristic of distinct V_H gene families (15) and suggest that genes *3* and *5A* could be regarded as representatives of distinct families. They are 73% identical in their nucleotide sequences in the coding regions and 67% identical in their inferred amino acid sequence which is below the level (80% identity) suggested by Brodeur and Riblet (16) as characteristic of members of V_H families in the mammals. The close similarity in structure of gene *5A* and pseudogene *5B* (86% nucleotide identity in the coding region) and their close association in the DNA

```
                        ———Leader ————    ——————— FR1——————————————   CDR1  ———FR2———
                      -19                  1                     30
Goldfish 3 protein    MRLWLHLVLLVATLEGVRS  QVQLVESGGDVKKPGDSLRLSCAASGYTFS  GYTMY  WVRQAPGKGLEWVS
Goldfish 5A protein   MEFWLSWVFLVAILKGVQC  EVQLVESGGGLIQPGGSLRLSCAASGFTVS  SNYMS  WVRQPPGKGLEWVS
Goldfish 5B protein   MESGLSWVFLVAILKGVQC  EVQLVESG*GLVQPGGSLRLSCAASGFTFS  SSWMH  WVCQAPGKGLEWVA
S107 protein          MKLWLNWVFLLTLLHGIQC  EVKLVESGGGLVQPGGSLRLSCATSGFTFS  DFYMD  WVRQPPGKRLEWIA

                        ———CDR2———          ——————— FR3 ———————————
                                60                         90
Goldfish 3 protein    RIYDDGSDVS--YADTVKG  RFTISRDNAKNLLYLQMSNLKPEDTGRYYCAR
Goldfish 5A protein   VIYSGGS-TY--YADSVKG  RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
Goldfish 5B protein   DIKCDGSEKY--YVDSVKG  RLTISRDNAKNSLYLQVNSLRAEDMTVYYCVR
S107 protein          ASRNKANDYTTEYSASVKG  RFIVSRDTSQSILYLQMNALRAEDTAIYYCARD
```

FIG. 3. Inferred amino acid sequences of goldfish V_H genes *3* and *5A* and pseudogene *5B*. The alignment of the sequence with that for the mouse V_H S107 (8) is identical to that shown in Fig. 2. The one-letter amino acid code is used.

(which was not a result of rearrangement during cloning or propagation of the library, Fig. 4) strongly suggests that they are the products of a gene-duplication event. The identity values in the protein-coding regions, in terms of nucleotides (and in parentheses in terms of inferred amino acids) for the alignments shown in Figs. 2 and 3 are as follows: for gene *3* vs. gene *5A*, 73% (67%); for gene *3* vs. S107, 60% (53%); and for gene *5A* vs. S107, 76% (68%). A search of the sequence data base of the Protein Identification Resource‡ indicated that the closest relationships for mature protein sequences of the two goldfish genes were to be found among human V_H III regions. In the case of gene *3*, there was a 76% identity with KOL (17), and in the case of gene *5B* there was an 86% identity with BUT (18). However, there was an even closer similarity observed in a comparison of leader regions, involving gene *5A* and human gene *H11* (19) in which there was 96.5% identity at the nucleotide level and 89.5% identity at the inferred amino acid level. While the strong conservation of an 18- or 19-amino acid leader peptide in heavy chains is notable since apparently approximately one in five of randomly generated peptides can function as signal peptides for protein secretion (20), it is perhaps more surprising that there is conservation, in the V_H genes, of the intron that always splits the leader-encoding region after the first base of the codon for amino acid −4. This intron has no known function, and its presence suggests that, in the course of V_H gene evolution, mechanisms other than those acting to maintain on the one hand functionally essential structures and on the other hand a sufficient diversity of binding sites to meet antigenic challenge may have operated. One possible mechanism is molecular drive (21) that could have acted by processes of nonreciprocal exchange to maintain V_H gene structures independently of their selective value.

V_H Families Related to Genes *3* and *5A* in the Goldfish. To determine the number and distribution of V_H genes closely related to genes *3* and *5A*, Southern blot hybridization analysis under high-stringency conditions was undertaken. Of 35 independently derived, unique, S107 cross-hybridizing recombinant phage clones recovered from the library, probes for genes *3*, *5A*, and *5B* hybridized only with the phage in which they were cloned (data not shown). Hybridization of probe 5A back onto restriction digest fragments of the original DNA from which the library was constructed (Fig. 4) showed only two hybridizing components of identical size to those present in the digest of the phage, suggesting that no rearrangement of the DNA in this region had occurred during construction or propagation of the library and that phage clone 5 contained all the members of the V_H gene family present in the original genomic library.

When the hybridization of probes for genes *3* and *5A* on DNA from other individual goldfish was tested, it was observed that 14 out of 15 fish lacked sequences cross-hybridizing under high stringency conditions with these probes. The DNA from the 15th fish, however, showed multiple restriction fragments hybridizing with probes for gene *3* and gene *5A* (Fig. 5). The patterns of hybridizing fragments for the two probes appeared to overlap significantly but not completely (Fig. 5). The Southern blot hybridization patterns for this DNA were essentially identical under low- and high-stringency conditions (data not shown) and always rather blurred (in six repeated experiments), indicating the possibility of multiple hybridizing fragments with mobilities close enough to form broad overlapping bands. There appeared, by conservative estimate, to be a minimum of 15 bands hybridizing with the gene *5A* probe and 9 bands hybridizing with the gene *3* probe.

Thus it appears that the goldfish V_H genes differ significantly from those of the mammals in terms of their family relationships, both within and among individuals. In the mouse, for example, there appears to exist a much larger number of V_H genes than we were able to detect with heterologous or homologous probes in the goldfish. The number of mouse V_H genes, originally estimated in the hundreds (16), now appears likely to be >1000 (22). These V_H genes have been divided into nine families on the basis of

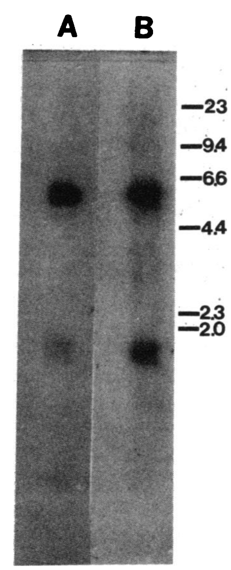

FIG. 4. Southern blot hybridization analysis of genomic DNA with goldfish V_H gene *5A*. DNA (10 μg) from which the library was constructed and DNA (1 μg) from λ phage clone 5 was digested to completion with *Kpn* I and *Sst* I, electrophoresed on 1% agarose gels, and transferred to nitrocellulose for hybridization with the ^{32}P-labeled nick-translated 584-bp *Pst* I–*Pst* I DNA fragment containing the coding sequence for gene *5A*. Lanes: A, λ phage 5 DNA (autoradiograph exposed for 2 hr); B, genomic DNA (autoradiograph exposed for 24 hr). Size markers are in kilobases.

‡Protein Identification Resource (1987) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), searched on-line April/May of 1987.

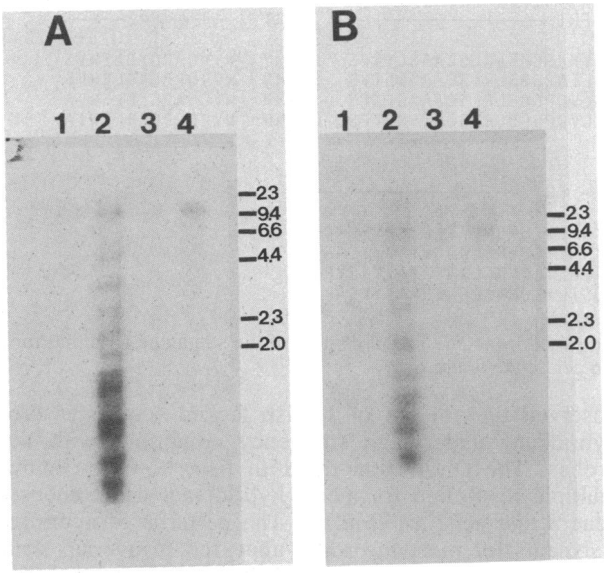

FIG. 5. Southern blot hybridization analysis of genomic DNA from four goldfish with V_H genes *3* and *5A*. DNA (10 μg) from four individual fish (lanes 1–4) was digested to completion with *Sst* I and analyzed on 1% agarose gels prior to transfer to nitrocellulose and hybridization with ^{32}P-labeled nick-translated purified DNA bearing either gene *3* (*B*) or gene *5A* (*A*). The nick-translated fragments were, for gene *3*, the 737-bp *Bam*HI–*Pst* I fragment and, for gene *5A*, the 584-bp *Pst* I–*Pst* I fragment. Size markers are shown in kilobases. Autoradiographs were exposed for 16 hr. Wash conditions were low-stringency.

sequence similarities (16, 23, 24), and these families can also be defined, operationally, by Southern blot hybridization analysis under high-stringency conditions, which detects degrees of nucleotide sequence similarity corresponding to that used to define a V_H family (16). Our studies clearly suggest that distinct families of V_H genes exist in the goldfish. However, in view of the relatively close similarity between gene *3* and gene *5A*, it may be premature to conclude that they are not members of the same family. In the mouse, all individuals appear to possess genes of all V_H families (24), and, although the size of the V_H gene families varies substantially, variations in the number of genes of any one family possessed by different mice is usually small, leading to the conclusion that in the mammals V_H gene families have evolved by the gradual gain and loss of small numbers of V_H genes (24). The situation in the goldfish appears quite different. Genes *3* and *5A* seem to belong to a V_H gene family (or families) with numbers that vary greatly within the population of goldfish, even being apparently absent in significant numbers of individuals.

Thus, while we can conclude that the goldfish possesses apparently typical vertebrate V_H genes, our initial observations on their organization into V_H families would lead us to suggest that the mode of evolution of these families, on a population basis, seems likely to have been quite different, in a teleost fish, from what has been seen for the mammals.

This work was supported by grants from the National Science Foundation (PCM 8408484 and 8709877).

1. Tonegawa, S. (1983) *Nature (London)* **302,** 571–581.
2. Davis, M. M., Calame, K., Early, P. W., Livant, D. L., Joho, R., Weissman, I. L. & Hood, L. (1980) *Nature (London)* **283,** 733–739.
3. Alt, F. W., Blackwell, T. K., DePinho, R. A., Reth, M. G. & Yancopoulos, G. D. (1986) *Immunol. Rev.* **89,** 5–30.
4. Ohno, S., Mori, N. & Matsunaga, T. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2945–2949.
5. Warr, G. W. & Marchalonis, J. J. (1982) in *The Reticuloendothelial System, A Comprehensive Treatise, Phylogeny and Ontogeny*, eds. Cohen, N. & Sigel, M. M. (Plenum, New York), Vol. 3, pp. 541–567.
6. Hinds, K. R. & Litman, G. W. (1986) *Nature (London)* **320,** 546–549.
7. Makela, O. & Litman, G. W. (1980) *Nature (London)* **287,** 639–640.
8. Early, P., Huang, H., Davis, M., Calame, K. & Hood, L. (1980) *Cell* **19,** 981–992.
9. Wilson, M. R., Middleton, D., Alford, C., Sullivan, J. T., Litman, G. W. & Warr, G. W. (1986) *Vet. Immunol. Immunopathol.* **12,** 21–28.
10. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
11. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
12. Parslow, T. G., Blair, D. L., Murphy, W. J. & Granner, D. K. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 2650–2654.
13. Litman, G. W., Murphy, K., Berger, L., Litman, R., Hinds, K. & Erickson, B. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 844–848.
14. Litman, G. W., Berger, L., Murphy, K., Litman, R., Hinds, K. & Erickson, B. W. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2082–2086.
15. Dildrop, R. (1984) *Immunol. Today* **5,** 85–86.
16. Brodeur, P. H. & Riblet, R. (1984) *Eur. J. Immunol.* **14,** 922–930.
17. Schmidt, W. E., Jung, H.-D., Palm, W. & Hilschmann, N. (1983) *Z. Phys. Chem.* **364,** 713–747.
18. Torano, A. & Putnam, F. W. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 966–969.
19. Ram, D., Benneriah, Y., Cohen, J. B., Zakut, R. & Givol, D. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 4405–4409.
20. Kaiser, C. A., Preuss, D., Grisafi, P. & Botstein, D. (1987) *Science* **235,** 312–317.
21. Dover, G. A. & Flavell, R. B. (1984) *Cell* **38,** 622–623.
22. Livant, D., Blatt, C. & Hood, L. (1986) *Cell* **47,** 461–470.
23. Winter, E., Radbruch, A. & Krawinkel, U. (1985) *EMBO J.* **4,** 2861–2867.
24. Riblet, R., Brodeur, P., Tutter, A. & Thompson, M. A. (1987) in *Evolution and Vertebrate Immunity: The Antigen-Receptor and MHC Gene Families*, eds. Kelsoe, G. & Schulze, D. (Univ. Texas Press, Austin, TX), pp. 53–62.