The EMBO Journal vol.11 no.2 pp.603-609, 1992

Immunoglobulin V_H clan and family identity predicts variable domain structure and may influence antigen binding

Perry M.Kirkham[1], Frank Mortari[2], J.Allen Newton[3] and Harry W.Schroeder, Jr[1,2]

Division of Developmental and Clinical Immunology, the Comprehensive Cancer Center, and the Departments of Microbiology[1], Medicine[2] and Pediatrics[3], University of Alabama at Birmingham, Birmingham, AL 35294, USA

Communicated by C.H.Chothia

Mammalian immunoglobulin V_H families can be grouped into three distinct clans based upon sequence conservation in two of the three framework (FR) intervals. Through replacement/silent site substitution analysis, molecular modeling and mathematical evaluation of known immunoglobulin crystal structures, we demonstrate that this conservation reflects preservation of protein sequence and structure. Each clan contains a characteristic FR 1 interval that is solvent-exposed and structurally separated from the antigen binding site. Families within a clan contain their own unique FR 3 interval that is capable of either influencing the conformation of the antigen binding site or interacting directly with antigen. Our results provide a structural context for theories that address differential use of V_H families in the immune response.

Key words: antibodies/molecular evolution/molecular modeling

Introduction

Immunoglobulin heavy chain variable domains are generated by the combinatorial joining of discontinuous germline V_H, D_H and J_H gene segments (Tonegawa, 1983). Peptide sequence variations among variable domains are typically clustered in three distinct linear intervals, termed hypervariable or complementarity determining regions (CDRs). These CDR domains are juxtaposed in the tertiary structure of the protein to form the classic antigen binding site (Kabat *et al.*, 1987). The V_H gene segment encodes CDRs 1 and 2, as well as the first three framework (FR) sections that separate them (Kabat *et al.*, 1987).

Mouse and human V_H elements were likely derived from three distinct progenitor V_H elements whose descendants populate three known clans of V_H gene segments (Tutter and Riblet, 1989; Schroeder *et al.*, 1990). These clans are defined based upon nucleotide sequence homology between families (both within and across species) in the 6–24 codon interval in FR 1 and the 67–85 interval in FR 3. Clan I includes the human V_H1, V_H5 and V_H7 (described in this report) families, and the murine J558 and $V_{gam}3.8$ (V_H9) families. Primordial clan II split into two distinct sets of families before the divergence of mouse and man: the V_HII subclan (human V_H2, mouse 3609), and the V_HIV subclan (human V_H4, V_H6 and mouse Q52, 36-60, V_H12). Clan III consists of the human V_H3 family and the murine 7183, T15, J606, X24, V_H10, V_H11 and V_H13 families (Schroeder *et al.*, 1990; and this paper). Nucleotide conservation of the FR 1 and FR 3 intervals is evident in all clans, with the greatest conservation seen in the clan III members. Interestingly, clan III members are preferentially expressed during fetal life (human V_H3 and mouse 7183) (Schroeder *et al.*, 1987; Yancopoulos *et al.*, 1984; Perlmutter *et al.*, 1985a; Schroeder and Wang, 1990). To test the hypothesis that clan and family identity have structural and functional meaning, we undertook a systematic analysis of the translation products of FR 1 and FR 3 nucleotide intervals of all known human and mouse V_H families. Our analysis reveals that the peptide residues in these regions of conserved nucleotide sequence define immunoglobulin clan- and family-specific subdomains whose structures have been maintained across species and evolutionary barriers. The observed patterns of conservation support the hypothesis that these subdomains play an important role in antibody function.

Results

Analysis of the effects of evolution on immunoglobulin variable genes requires cross-species sequence comparisons. However, many germline-encoded V_H elements are demonstrably non-functional because they contain stop codons, deletions or frameshift mutations (Kodaira *et al.*, 1986). The sequences of other germline gene segments may appear functional, yet these gene segments are rarely used by adult B cells. The T15 family V13 gene segment, for example, is 96% homologous to the related V1 and V11 gene segments, yet this gene segment is rarely found in functional rearrangements (Feeney, 1990). Similarly, the V_H81X gene segment of the 7183 family, although preferentially rearranged in fetal pre-B cells and able to make a functional protein, is a rare contributor to the adult antibody repertoire (Decker *et al.*, 1991). Thus, because of the extensive diversity of the germline repertoire of V_H gene segments, it can be difficult to interpret the significance of the similarity or differences between two individual gene segments drawn at random from the germline pool.

Because V_H gene segments undergo gene conversion within a family (Perlmutter *et al.*, 1985b) and thus appear to evolve as a family unit, we reasoned that the confounding influence of variably functional V_H gene segments could be minimized by comparing each gene segment with a group norm and then using that group norm for evolutionary comparisons. We searched the literature for immunoglobulin heavy chain sequences and entered all the human and mouse germline sequences we could find into a database. Known pseudogenes were excluded and cDNA sequences were included only when less than two germline sequences were available. We focused our attention on the FR 1 and FR 3 intervals that we had previously demonstrated to be

most predictive of family and clan identity (Schroeder *et al.*, 1990). Family-specific group norms, or consensus sequences, were generated by identifying the most commonly utilized nucleotide at each base pair position.

If the peptide sequence of a protein domain is essential for the function of that protein, then its sequence will be preserved during evolution (Bowie *et al.*, 1990). The effects of this pressure to maintain peptide sequence can be quantified by taking advantage of the redundancy of the genetic code. For example, nucleic acid changes in the third codon position often do not change the amino acid of the translated product. Such mutations are termed silent (S) in contrast to the replacement (R) mutations which alter the peptide sequence. On average, codons undergoing random mutation will yield a replacement to silent amino acid substitution ratio (R/S ratio) of 2.9 (Jukes and King, 1979; Shlomchik *et al.*, 1987). Ratios <2.9 indicate preservation of peptide sequence. Ratios significantly >2.9 can only be achieved through selection for diversity. For example, the catalytic site of a soluble enzyme will exhibit the lowest R/S ratio, followed by the hydrophobic interior that stabilizes the structure, while the hydrophilic exterior often demonstrates an R/S ratio typical of random mutation.

We first examined the extent of sequence variation that existed within each family. For each set of germline or near-germline gene segments belonging to a given family, we determined the number of replacement and silent codon changes from the family consensus sequence (Jukes and King, 1979; Shlomchik *et al.*, 1987; Schroeder *et al.*, 1990). Within families, R/S ratios ranged from 0 to 14.0 for FR 1 and 0 to 5.0 for FR 3 (Table I). Ratios >3.0 in FR 1 were typically due to single outlier gene segments; notably, the 7183 V_H81X and the T15 V13 gene segments were the single outlier elements for their respective families. The average R/S ratio of FR 3 within a family was 1.9, indicating that pressure had been exerted to maintain this peptide sequence within a family. Although the average R/S ratio for FR 1 was 2.7, or near random, the significance of this value was unclear due to the contribution of similar outlier gene segments.

We then examined the extent of sequence variation between families, both within and across species barriers. We chose to use the group norms of the major human families as our standards for comparisons between families. For each set of family consensus sequences belonging to a given clan, we determined the number of replacement and

Table I. Replacement/silent site substitution analysis of the V_H repertoire of mouse and man

	V_H family	Number of sequences	Family		Clan					
			FR 1 R/S	FR 3 R/S	Base homology (%)	Peptide homology (%)	FR 1 R/S	Base homology (%)	Peptide homology (%)	FR 3 R/S
Human										
Clan I	V_H1	6	0.5	5.0	–	–	–	–	–	–
	V_H5	2	1.0	0.0	79	79	0.7	70	50	11.0
	V_H7	2	1.0	2.0	91	84	1.5	64	55	1.7
Clan II	V_H2	3	3.0	4.0	72(IV)	74	0.6(IV)	61	45	2.4(IV)
	V_H4	10	6.0	2.0	–	–	–	–	–	–
	V_H6	1	0	0	81	79	0.7	79	63	2.3
Clan III	V_H3	16	0.5	1.0	–	–	–	–	–	–
Mouse										
Clan I	J558	21	1.3	2.0	90	84	3.0	79	68	2.3
	V_{gam}3.8	14	*3.0	3.4	81	74	1.3	61	36	2.8
Clan II	3609	3	1.3	2.5	72(II)	58	2.7(II)	76	59	1.8(II)
					65(IV)	63	0.8(IV)	61	50	1.6(IV)
	V_H12	3	1.0	0	72	74	0.6	73	55	2.5
	36-60	3	1.0	2.2	79	84	0.4	73	55	2.5
	Q52	2	0.7	1.5	81	79	0.6	52	36	2.8
Clan III	7183	5	*8.0	3.3	91	89	0.7	86	82	1.3
	T15	3	*3.0	1.0	91	75	0.3	85	82	1.0
	X24	2	0	0	88	75	0.2	82	73	1.5
	J606	2	0	0.5	84	85	0.6	76	89	0.8
	V_H10	4	1.5	0	79	79	0.7	80	68	1.8
	V_H11	7	0.9	1.6	91	95	0.3	82	82	0.8
	V_H13	6	1.0	0	79	68	1.5	77	68	1.4
Weighted average	(within a family)	115	2.7	1.9						
Average	(between families)	20			82	81	0.9	73	62	2.4

The number of sequences examined within a given family is listed in column 3. An asterisk (*) marks families wherein a single outlier sequence has raised the R/S ratio in FR 1 above 2.9. Columns 4 and 5 present the R/S ratios calculated within a family in the FR 1 and FR 3 intervals, respectively. Columns 6 and 9 report the nucleotide sequence homology of the family consensus sequences in these same intervals compared with the clan I, II (subclan IV) and III consensus sequences. Columns 7 and 10 similarly report the peptide sequence homology. Columns 8 and 11 display the family R/S ratios as compared with the clan consensus sequences. Sequences and clan R/S ratios of the murine 3609 and human V_H2 families were calculated against both subclan II and IV consensus sequences. Weighted averages were obtained by multiplying the R/S ratio for a given family times the number of elements, adding all the weighted R/S ratios for a given comparison, and then dividing by the total number of elements compared. The R/S ratio for the FR 1 interval between the consensus sequences of clan I and clan III is 3.7, the ratio for the FR 3 interval is 6.5. The family consensus sequences have been reported to EMBL under the identifiers X59907-23 and X59972-74. A complete listing of the gene segments examined is available from the authors.

silent codon changes from the human V_H1 for clan I sequences, V_H2 or V_H4 for sub-clan II or IV sequences, respectively, and V_H3 for clan III sequences (Table I). Unexpectedly, these inter-family comparisons revealed strong evidence of preservation of protein sequence in the FR 1 interval. Family consensus FR 1 interval sequences exhibited R/S ratios ranging from 0 to 3.0 (average of 0.9) when compared with the prototype human clan consensus sequences. Indeed, only one family consensus sequence, J558, exhibited an R/S ratio >1.5 in comparison with the human consensus.

The strong conservation of FR 1 across families within a clan was not reflected in the FR 3 interval. The clan I and II FR 3 intervals exhibited R/S ratios as high as 11 and the average FR 3 R/S ratios for all families was 2.4, or near random. The clan III FR 3s were an exception, in that they were highly conserved (R/S ratio range of 0.8−1.8). Upon closer inspection, it became clear that the primary discriminator between families which belong to the same clan was the sequence of their respective FR 3 intervals (Table I). For example, the human V_H5 family is derived from the same progenitor as V_H1, yet although the FR 1 R/S ratio between V_H5 and V_H1 is 0.7, the FR 3 R/S ratio is 11. Thus, the codons within FR 1 were most predictive of clan identity, whereas FR 3 codons appeared to differentiate families within a clan.

Analysis of clan II members provided further support for the hypothesis that clan identity reflects evolutionary selection

```
                                     <--    FR 1     -->        <--     FR 3      -->
                                     6                24        67     75 77       85
                     Family  Gene    ^  ^  ^^  ^  ^   ^                ^  ^
Clan I                               * * * * *** * * *          * * * **** * * ** ***
                     Consensus       QSGAEVKKPGASVKVSCKA        VTMTRDTSTSTAYMELSSLRSE
            Human    VH1     HG3     ...................        ...........V..........
                             20P3    ...................        ........I........R...D
                             51P1    ..........S........        ..I.A.E...............
                     VH5     H251    ...........E.L.I....       ..ISA.K.I....LQW...KAS
                     VH7     N10P1   ..AS.L.............        FVFSL...V....LQI....A.

            Mouse    J558    B4      .P...LV.......L....        A.L.V.KPS.....Q....T..
                             V186-2  .P...LV.......L....        A.L.V..PS.....Q....T..
                             V108A   ...P.LV.......I....        A.L.V.N.S..........T..
                             HyHEL5  .....LM.......I....        A.F.A...S.....Q.N..T..  <--
                             R19.9   E....LVAASS...M....        T.L.V.R.S.....Q.R..T..  <--

Clan II
            VH4      Consensus       ESGPGLVKPSETLSLTCTV        VTISVDTSKNQFSLKLSSVTAA
            Human    VH4     58P2    ...................        ......................
                             V2-1    ...................        ..........H...........
                             NEW     .......R..Q........        ..ML..........R.......  <--
                     VH6     C17P3   Q.........Q......AI        I..NP.........Q.N...PE

                     VH2     M60     ....A....TQ..T....L        L...K......VV.TMTNMDPV

            Mouse    3660    VH3660  ....S.....Q......S.        IS.TR......YY.Q.N...TE
                             HyHEL10 ....S.....Q......S.        IS.TR......YY.D.N...TE  <--
                     Q52     VH101   Q......Q..QS..I....        LS..K.N..S.VFF.MN.LQSN
                             PJ14    .......A..QS..I....        LS..K.N..S.VF..MN.LQTD

                     3609    VH23P9  .C...ILQ..Q......SF        L...K..YN..VF..IAN.DT.

Clan III
                     Consensus       ESGGGLVQPGGSLRLSCAA        FTISRDNSKNTLYLQMNSLRAE
            Human    VH3     30P1    ...................        ......................
                             38P1    ...................        .....E.A..S..........G
                             56P1    .....V....R........        ......................
                             KOL     .....V....R......SS        .............F...D...P.  <--

            Mouse    7183    VHE415  .......K.....K.....        .......A..N.....S...S.
                             VH39P1  .......K.....K....T        .......A..N.....S...S.
                             VH50P1  .............K....T        .......A........SR.KS.
                             VH81X   ..........RE..K...ES       .I.....T.K......S...S.
                     X24     VH441   .............K.....        .I.....A........SKV.S.
                             J539    .............K.....        .I.....A..S.....SKV.S.  <--
                     T15     V1      ..................T        .IV...T.QSI......A....
                             MCPC603 ..................T        .IV...T.QSI......A....  <--
                             V11     ..................T        ........QSI...........
                             V13     ..........A......E.        ........Q.I......T....
                     J606    VH22P1  .............MK...V.       ......D..SSV.....N....
```

Fig. 1. Peptide sequences with codons 6−24 (FR 1) and 67−85 (FR 3) (Schroeder *et al.*, 1990) of representative germline gene segments are compared with their respective clan consensus sequences. The peptide sequences of the crystal structures used for modeling studies are listed in the figure according to their family and clan identity and are marked by arrows. Residues 6, 9, 12, 13, 16, 19 and 23 from FR 1, and 75 and 77 from FR 3 serve as reference residues and are indicated by ∧ at the top of the figure. Those residues which protrude outward (using the KOL Fab structure as a reference) are indicated by asterisks.

```
        |<--                                 Framework 1                                       -->|<--  CDR 1  -->|<--                 Framework 2                  -->|
 N10P1: CAGGTGCAGCTGGTGCAATCTGCGTCTGAGTTGAAGAAGCCTGGGGCCTCAGTGAAGGTTTCCTGCAAGGCTTCTGGATATACCTTCACA AGCTATGCTATGAAT TGGGTGCGACAGGCCCCTGGACAAGGGCTTGAGTGGATGGGA
RF-TS3: ......................G...........................................................C........T ............... ..........................................
  51p1: .................G....G.G.....G...............T....G........C..................GGC.......GC ...........C.GC ..........................................

        |<--                 CDR 2                    -->|<--                                  Framework 3                                                    -->|
 N10P1: TGGATCAACACCAACACTGGGAACCCAACGTTTGCCCAGGGCTTCACAGGA CGGTTTGTCTTCTCCTTGGACACCTCTGTCAGCACGGCATATCTGCAGATCAGCAGCCTAAGGGCTGAGGACACTGCCGTGTATTACTGTGCGAGA
RF-TS3: ...............................A................... ..............................................................A......C.......................
  51p1: G......T.C.T.T.TT...T.CAG...AC.AC..A...AAG...CAG..C A.AG.CACGA.TA..GC....GAA..CACG.....A..C..CA..G..C.G........G..AT..........G......................
```

Fig. 2. Nucleotide sequences of two V_H7 sequences, N10p1 and RF-TS3 (Pascual *et al.*, 1990) compared with a V_H1 family member, 51p1 (Schroeder *et al.*, 1987). A dot denotes sequence identity with V_H7. The N10p1 sequence has been submitted to EMBL under the identifier X59906.

pressure on the protein sequence of FR 1. By nucleotide sequence comparison, the members of the clan II-subclan II group (human V_H2, mouse 3609) are more similar to each other than to subclan IV elements (human V_H4, V_H6; mouse 36-60, Q52, V_H12). Hence, the progenitor of these families likely diverged from ancestral clan II elements prior to the separation of man and mouse. In spite of this nucleotide similarity, by peptide sequence (Figure 1) and R/S ratio analysis (Table I) mouse 3609 is more similar to the human subclan IV consensus than to human subclan II consensus. Thus, analysis of the known 19 V_H families in man and mouse suggested that all V_H gene segments belong to one of three clans, and that each clan is under pressure to maintain a FR 1 interval protein sequence which is unique to each clan.

Based upon these observations, we postulated that if novel human families exist they would follow this same pattern: i.e. nucleotide and peptide sequence similarity to one of the three clan consensus sequences in FR 1 with divergence in FR 3. We generated an unrestricted cDNA library from mononuclear cells derived from the cord blood of a female term infant and screened the library for Cμ-containing clones.

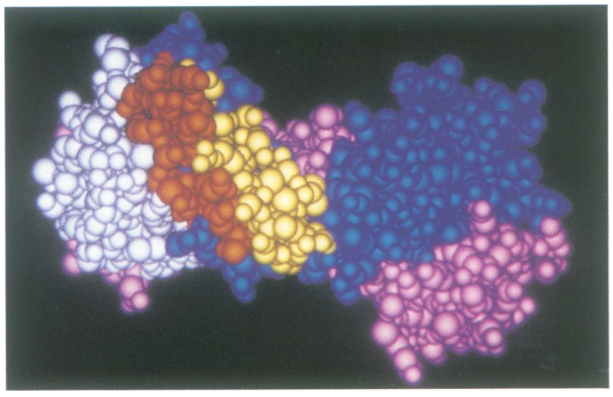

Fig. 3. The Fab region of clan III KOL is displayed with the regions of interest differentially colored: the CDRs of the heavy and light chain are in white; the heavy chain is blue, the light chain is violet; FR 1 residues (6–24) are in yellow and FR 3 (67–85) are in red. Note that the FR 1 and FR 3 residues are solvent-exposed and form a bridge between the CDR and CH1 domain of the heavy chain (to the right).

One of the recombinants isolated, N10p1, exhibited 79% homology to members of the V_H1 family (Figure 2) with 91% nucleotide identity in the FR 1 interval region and 64% nucleotide identity in the FR 3 interval region. Because these changes could theoretically result from somatic mutation, we searched the literature for additional examples and found a similar sequence (98% identity to Np1P1) which had been isolated from an EBV cell line secreting an antibody with rheumatoid factor activity (Pascual *et al.*, 1990). The V_H7 consensus sequence exhibited 84% amino acid identity to the clan I consensus sequence in FR 1 with an R/S ratio of 1.5, whereas in FR 3 there was only 55% amino acid identity with an R/S ratio of 1.7 (Table I). Analysis of the immunoglobulin compilation of Kabat *et al.* (1987) revealed sets of sequences with homology in FR 1 and variation in FR 3, suggesting that additional human families likely exist.

By inspection, we had previously demonstrated that nucleotide motifs within the FR 1 and 3 family-specific intervals were conserved across species barriers (Schroeder *et al.*, 1990). Examination of the translated products of these nucleotide motifs revealed characteristic peptide residues which contributed to the low R/S ratios in the FR 1 and FR 3 interval (see codons 9, 12 and 23 of FR 1, and codons 75 and 77 of FR 3 in Figure 1). Although these residues or like-polarity substitutions were not found in every sequence, their frequent presence in regions exhibiting low R/S ratios indicated that these residues might play an important role in the structure and/or function of the antibody.

Through molecular modeling we had previously determined that the FR 1 and 3 intervals are highly solvent-exposed (Figure 3) (Schroeder *et al.*, 1990). These studies were extended to include all immunoglobulins whose crystal coordinates had been deposited in the Brookhaven database. We were fortunate that structures from all three clans were represented. We found that the major non-CDR structural differences between heavy chains of different clans were in the β loop regions that contain the majority of the FR 1 and FR 3 conserved peptide motifs.

Using molecular modeling, the morphology of the FR 1 and FR 3 intervals in these immunoglobulins was examined and compared. Each clan was found to express a distinctive FR 1 loop structure unique to that clan (Figures 4A and 5A). Indeed, immunoglobulins within the same clan exhibit

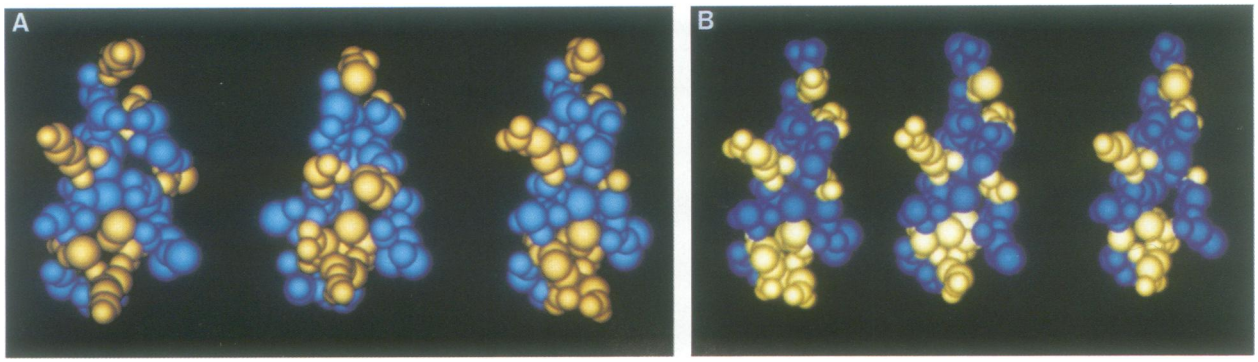

Fig. 4. (**A**) Three-dimensional generation of FR 1 interval residues of molecules representing the three clans. The atoms of the FR 1 intervals (codons 6–24) of representative members of clan I HyHEL-5 (right), clan II NEW (middle) and clan III KOL (right) are displayed in three-dimensional (raster) form. Reference residues 6, 9, 12, 13, 16, 19 and 23 are colored yellow to aid in visualizing the morphology of the FR 1 loop. In this perspective, the CDR would be at the top and the CH1 domain at the bottom. The width and shape of this β loop clearly differ between clans. (**B**) Three-dimensional generation of FR 1 interval residues representing three clan III antibodies. Human V_H3 KOL (left), mouse T15 MCPC603 (middle) and mouse X24 J539 (right) are shown. Coloring of residues is as in (A). Note the similarity in the shape of the FR 1 loop as compared with that displayed in (A).

virtually identical FR 1 loop morphology (Figures 4B and 5B) both within and across species. For example, the three clan III heavy chains, human V_H3 KOL, mouse T15 MCPC603 and mouse X24, J539, demonstrate extensive sequence and structural similarity (Table I and Figures 4B and 5B), as do the clan I (mouse R19.9 and HyHEL-5) heavy chains (Figure 5C), and clan II (human V_H4 New and mouse 36-60 HyHEL-10) (molecular models not shown).

A pair-wise comparison of the root mean square deviation in the distance between the α carbons of the FR 1 intervals of each of the seven Fab crystal structures was performed using the algorithm HOMOLOGY (Rao and Rossmann, 1973) (Table II). The similarities which were so clearly apparent upon visual inspection were confirmed by this mathematical analysis. The strict preservation of this FR 1 loop structure within clan normals (Figure 5C and Table II) is more remarkable when one notes the deviation of the peptide sequence of R19.9 from the clan I consensus and from HyHEL-5 (Figure 1).

In contrast, the variation in nucleotide and peptide sequence of FR 3, which differentiated familities within a clan, was mirrored at the structural level. For example, clan II human V_H4 New and mouse 36-60 HyHEL-10, and clan I mouse R19.9 and HyHEL-5 were quite different from each other (molecular models not shown). Correlation of R/S ratio and sequence with structure was predictive but not absolute. For instance, the FR 3 of KOL was more similar structurally to MCPC603 than to J539, even though by peptide sequence J539 was more homologous to KOL (Figure 1). Thus, clan identity, as originally defined by nucleotide homology, predicts similarity in the FR 1 β loop, whereas family identity predicts similarity in the FR 3 β loop, with each family exhibiting its own characteristic FR 3 sequence and structure.

Discussion

The residues of exterior protein domains that are neither engaged in specific intermolecular interactions (e.g. receptor or catalytic function) nor contributors to essential structure are in general subject only to polarity constraints (e.g. solubility) (Bowie *et al.*, 1990). The β loops which define these FR 1 and FR 3 intervals are both solvent-exposed and free to diverge without affecting the basic structure of the immunoglobulin fold. Therefore, the sequence preservation of these two solvent-exposed β loop subdomains suggests that they likely play an important role in the function of the antibody.

The FR 3 β loop is adjacent to the CDR 1 and CDR 2 domains of the heavy chain and the top of the loop can be considered an extension of the antigen binding site. Recent studies have suggested that residues within this loop can participate directly in the binding of antigen to the antibody (Radic *et al.*, 1989). Molecular modeling studies have suggested that a limited set of conformations exists for CDR 1 and CDR 2 and that residue 71, which is a part of the FR 3 interval, can influence the conformation and position of CDR 2 (Chothia *et al.*, 1989). Thus, conservation of the residues of the FR 3 interval within a family and across species may provide initial family-associated constraints to antibody affinity for specific antigen epitopes. This conservation may explain why specific classes of auto-

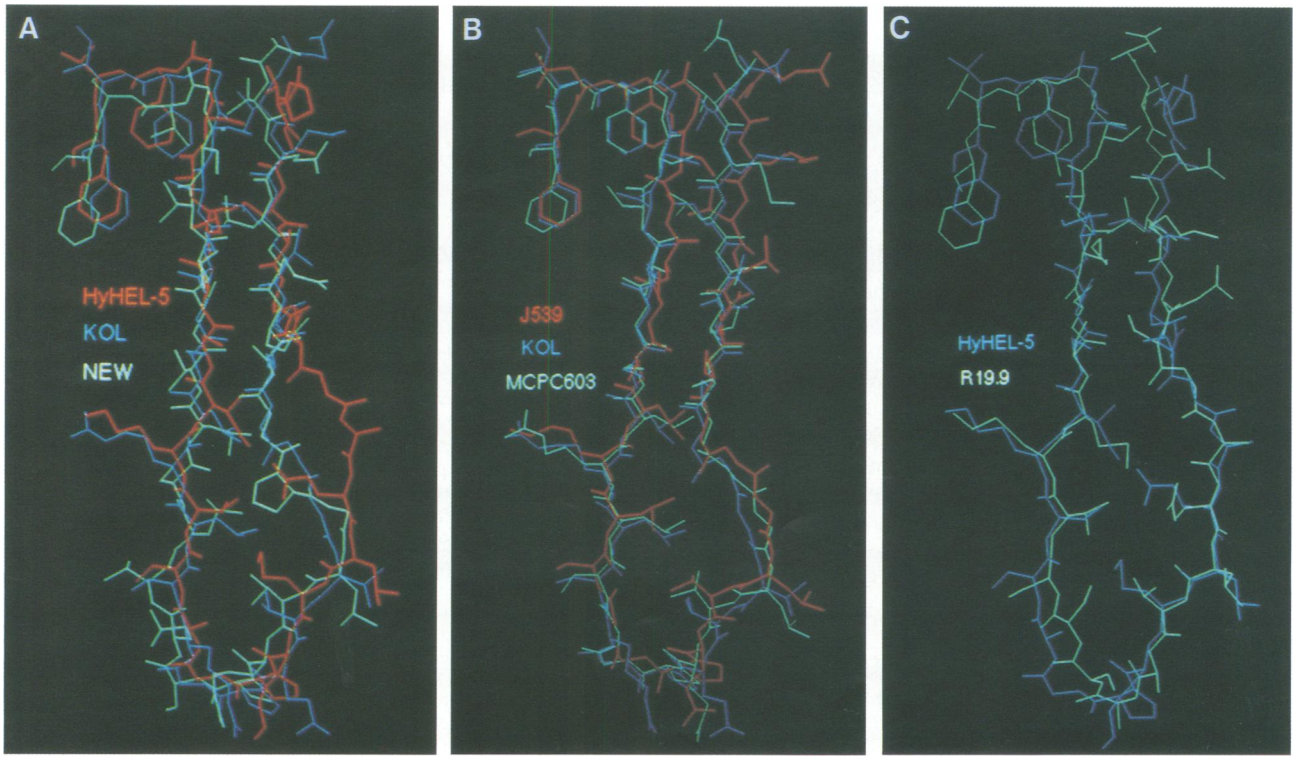

Fig. 5. (**A**) Superposition of the FR 1 intervals from three molecules representing the three clans. HyHEL-5 and NEW were superpositioned on KOL. Residue side chains are included and further illustrate the differences between the three structures. (**B**) Superposition of the FR 1 intervals of MCPC603, J539 and KOL, all of which are from the same clan of origin (clan III). MCPC603 and J539 structures were compared with KOL. Note that residue side chains are extremely similar in orientation and size. (**C**) Superposition of the FR 1 interval of HyHEL-5 and R19.9, both from clan I. R19.9 was superpositioned on HyHEL-5.

Table II. Root mean square (r.m.s.) deviation between the FR 1 intervals of seven Fab structures

	KOL	J539	HyHEL-10	NEW	HyHEL-5	R19.9
MCPC603	0.587	0.696	0.887	1.062	1.121	1.315
KOL	–	0.767	1.112	1.327	1.232	1.402
J539	–	–	1.066	1.272	1.323	1.485
HyHEL-10	–	–	–	0.606	1.342	1.535
NEW	–	–	–	–	1.421	1.565
HyHEL-5	–	–	–	–	–	1.106

The lowest value obtained for each Fab structure is enlarged. Three distinct groups of structural homology are seen: MCPC603, KOL and J539 fall into one group, HyHEL-10 and NEW compose a second group, and a third group contains the R19.9 and HyHEL-5 Fab structures. All deviation distances are given in Å.

antibodies (e.g. anti-DNA) are often associated with families of the same clan in both mouse and man (e.g. V_H3 and T15) (Dersimonian *et al.*, 1987; Behar *et al.*, 1991). The enhanced affinity for antigen demonstrated by antibodies of the secondary response which have undergone somatic hypermutation is associated with an R/S ratio significantly greater than 2.9, indicating positive selection for high affinity antigen binding sites (Shlomchik *et al.*, 1987). We can speculate that the highly divergent FR 3 sequences seen in variant families such as V_H5 (FR 3 R/S ratio of 11 versus the V_H1 consensus) may be the result of pressure to generate unusual classes of antigen binding sites.

A more direct FR 3–ligand interaction is also possible. Superantigens are bacterial or self-antigens which recognize antigen receptors in a family-specific fashion. The region in TCR Vβ gene segments homologous to the immunoglobulin FR 3 loop region has been identified as the site of superantigen binding (Pullen *et al.*, 1990). It is provocative, therefore, that Staph protein A (Sasso *et al.*, 1989) recognizes a preponderance of human V_H3 elements (10 of 11 V_H3 IgM, 0 of 7 V_H1, 0 of 6 V_H2).

Our studies suggest that all functional antibodies in man and mouse contain one of the three FR 1 β loops which define the clans. Although this sequence is not absolutely conserved, it is remarkable that in at least two cases (V13 and V_H81X) sequences divergent from the family consensus are rarely used in mature antibodies. The fetal heavy chain repertoire is enriched for use of a small subset of V_H gene segments. In both mouse and man, this repertoire is enriched for members of clan III. We had previously shown that the greatest homology between human and mouse fetal antibodies was exhibited by V_H30p1 and V_HE415. Remarkably, the 30p1 fetal sequence, which is also found in germline or near-germline form as a component of autoantibodies, is identical to the clan III consensus. This pattern of fetal expression of consensus or near-consensus sequence is also found in clan I and clan II: the V_H1 20p3 and V_H4 58p2 are identical to the clan I and clan II FR 1 consensus, respectively. These findings would suggest that the evolutionary pressure to maintain the FR 1 sequence may be first exerted at the earliest stages of lymphocyte development.

Pre-B cells exhibit an unusual surface immunoglobulin receptor which consists of fully rearranged heavy chains coupled to a light chain equivalent formed from the λ5 and V-preB molecules (Karasuyama *et al.*, 1990; Tsubata and Reth, 1990). These light chain-like polypeptides differ from normal light chain by the addition of two peptide subdomains at the site normally occupied by the light chain CDR 3. Although no such structures have yet been crystallized, molecular modeling suggests that these additions would have the effect of blocking the antigen binding pocket. If so, then the major source of variation in the pre-B surface immunoglobulin would lie in the FR 1 and FR 3 solvent-exposed subdomain which we have described in this paper. If a monitoring system exists to prevent expansion of B-cells bearing variant framework structures, we propose that normal framework boundaries may be determined by the antigen receptors borne on pre-B cells which are expressed during the earliest stages of development.

Recent advances in genetic engineering technology have made it possible to create chimeric mouse–human antibodies by uniting mouse variable domains with human constant regions or 'humanizing' mouse antibodies by splicing mouse CDRs onto human variable domains. These heroic changes are necessary because of the variable immunogenicity of the unmodified mouse immunoglobulins. If our hypothesis that the FR 3 interval plays a major role in the conformation and structure of the antigen binding site is correct, then the likelihood of reproducing the appropriate antigen specificity of a 'humanized' antibody may be increased if mouse CDRs are spliced onto human frameworks from the same clan. It is also possible that FR 1 and FR 3 intervals from mouse antibodies which diverge significantly from the human consensus could be immunogenic. If so, changes in the residues of these chimeric antibodies that would bring them closer to the human consensus might reduce their antigenicity.

Our studies have demonstrated striking preservation of two framework intervals within immunoglobulin V_H genes. The function of these framework intervals, and in particular their relationship to both normal and pathologic autoreactive specificities, will be important areas for future research.

Materials and methods

Cloning

Mononuclear cells from 20 cc of cord blood were isolated by Ficoll–Hypaque gradient centrifugation. An oligo(dT) primed cDNA library of 1.2×10^6 recombinants was generated from 1 μg of poly(A)$^+$ RNA and screened for Cμ-containing cDNA clones (Schroeder *et al.*, 1987). The sequence of clone N10p1 was obtained by subcloning into pUC19 and then double strand sequencing as previously described (Schroeder *et al.*, 1987).

Structural analysis

The structures of the FR 1 and FR 3 intervals for all solved immunoglobulin structures available from the Brookhaven database (Bernstein *et al.*, 1977; Abola *et al.*, 1987) were modeled on an IRIS 4D/220GTX computer system (Silicon Graphics) using the QUANTA (Polygen, Waltham, MA) software package. The crystallographic structures of the mouse HyHEL-5 (J558; file 2HLF; 2.54 Å resolution) (Sheriff *et al.*, 1987) and mouse R19.9 (J558; file 1F19; 2.8 Å resolution) (Lascombe *et al.*, 1989) represented clan I. Human NEW (V_H4; file 3FAB; 2.0 Å resolution) (Saul *et al.*, 1978), and mouse HyHEL-10 (36-60; file 3HFM; 3.0 Å resolution) (Padlan *et al.*, 1989) represented clan II. Finally, human KOL (V_H3; file 2FB4; 1.9 Å resolution) (Marquardt *et al.*, 1980), mouse J539 (amily X24; file 2FBJ; 2.6 Å resolution) (Suh *et al.*, 1986) and mouse MCPC603 antibodies (family T15; file 1MPC; 2.7 Å resolution) (Satow *et al.*, 1986) represented clan III.

Fab structures were compared and superpositioned by applying a least squares comparison on all atoms of the FR 2 residues. After the superposition was completed, the coordinate files were rewritten to retain the superposition coordinates and all atoms outside of the FR 1 interval (residues 6–24, inclusive) were removed.

Root mean square (r.m.s.) deviations were performed on the seven crystal structures mentioned above using the program HOMOLOGY (Rao and

Rossmann, 1973). The α carbon atoms of residues 6–24 of the FR 1 interval in each Fab structure were compared with each of the other structures. The results of the comparisons are given in Å of deviation between structures.

Acknowledgements

We wish to thank Drs Gillian Air, Max Cooper, Stephen Ealick, Stephen Harvey and Ming Luo for invaluable discussions during the preparation of this manuscript, and Jin Yi Wang for excellent technical assistance. This research was supported in part by National Institutes of Health grants AI26394, AI30879, AR03555 and GM08111, and by a grant from the National Science Foundation to S.H. (DIR-89-08155). H.W.S. is an RJR Nabisco Research Scholar in Immunology. F.M. is a fellow of the Medical Research Council of Canada. J.A.N. is a St Judes Research Fellow and holds a Pediatric Scientist Research Award.

References

Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Crystallographic Database—Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Bonn, pp. 107–132.

Behar,S.M., Lustgarten,D.L., Corbet,S. and Scharff,M.D. (1991) *J. Exp. Med.*, **173**, 731–741.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimaouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Bowie,J.Y., Reidhaar-Olson,J.F., Lim,W.A. and Sauer,R.T. (1990) *Science*, **247**, 1306–1310.

Chothia,C., Lesk,A.M., Tramontano,S., Levitt,M., Smith-Gill,S.J., Air,G., Sheriff,S., Padlan,E.A., Davies,D., Tulip,W.R., Colman,P.M., Spinelli,S., Alzari,P.M. and Poljak,R.J. (1989) *Nature*, **342**, 877–883.

Decker,D.J., Boyle,N.E., Koziol,J.A. and Klinman,N.R. (1991) *J. Immunol.*, **146**, 350–361.

Dersimonian,H., Schwartz,R.S., Barrett,K.J. and Stollar,B.D. (1987) *J. Immunol.*, **139**, 2496–2501.

Feeney,A.J. (1990) *J. Exp. Med.*, **172**, 1377–1390.

Jukes,T.H. and King,J.L. (1979) *Nature*, **281**, 605–606.

Kabat,E.A., Wu,T.T., Reid-Miller,M., Perry,H.M. and Gottesman,K.S. (1987) *Sequences of Proteins of Immunological Interest.* 4th edn. US Department of Health and Human Services, Bethesda, pp. vii–804.

Karasuyma,H., Kudo,A. and Melchers,F. (1990) *J. Exp. Med.*, **172**, 969–972.

Kodaira,M., Kinashi,T., Umemura,I., Matsuda,F., Noma,T., Ono,Y. and Honjo,T. (1986) *J. Mol. Biol.*, **190**, 529–541.

Lascombe,M.-B., Alzari,P.M., Boulot,G., Saludjian,P., Tougard,P., Berek,C., Haba,S., Rosen,E.M., Nisonoff,A. and Poljak,R.J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 607–611.

Marquardt,M., Deisenhofer,J,. Huber,R. and Palm,W. (1980) *J. Mol. Biol.*, **141**, 369–391.

Padlan,E.A., Silverton,E.W., Sheriff,S., Cohen,G.H., Smith-Gill,S.J. and Davies,D.R. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 5938–5942.

Pascual,V., Randen,I., Thompson,K., Sioud,M., Forre,O., Natvig,J. and Capra,J.D. (1990) *J. Clin. Invest.*, **86**, 1320–1328.

Perlmutter,R.M., Kearney,J.F., Chang,S.P. and Hood,L.E. (1985a) *Science*, **227**, 1597.

Perlmutter,R.M., Berson,B., Griffin,J.A. and Hood,L.E. (1985b) *J. Exp. Med.*, **162**, 1988.

Pullen,A.M., Wade,T., Marrack,P. and Kappler,J.W. (1990) *Cell*, **61**, 1365–1374.

Radic,M.Z., Mascelli,M.A., Erikson,J., Shan,H., Shlomchik,M. and Weigert,M. (1989) *Cold Spring Harbor Symp. Quant. Biol.*, **54**, 933–946.

Rao,S.T. and Rossmann.,M.G. (1973) *J. Mol. Biol.*, **76**, 241–256.

Sasso,E.H., Silverman,G.J. and Mannik,M. (1989) *J. Immunol.*, **142**, 2778–2783.

Satow,Y., Cohen,G.H., Padlan,E.A. and Davies,D.R. (1986) *J. Mol. Biol.*, **190**, 593–604.

Saul,F.A., Amzel,L.M. and Poljak,R.J. (1978) *J. Biol. Chem.*, **253**, 585–595.

Schroeder,H.W., Jr and Wang,J.Y. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 6146–6150.

Schroeder,H.W., Jr, Hillson,J.L. and Perlmutter,R.M. (1987) *Science*, **238**, 791–793.

Schroeder,H.W., Jr, Hillson,J.L. and Perlmutter,R.M. (1990) *Int. Immunol.*, **20**, 41–50.

Sheriff,S., Silverton,E.W., Padlan,E.A., Cohen,G.H., Smith-Gill,S.J., Finzel,B.C. and Davies,D.R. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 8075–8079.

Schlomchik,M.J., Marshak-Rothstein,A., Wolfowicz,C.B., Rothstein,T.L. and Weigert,M.G. (1987) *Nature*, **328**, 805–811.

Suh,S.W., Bhat,T.N., Navia,M.A., Cohen,G.H., Rao,D.N., Rudikoff,S. and Davies,D.R. (1986) *Proteins*, **1**, 74–80.

Tonegawa,S. (1983) *Nature*, **302**, 575–581.

Tsubata,T. and Reth,M. (1990) *J. Exp. Med.*, **172**, 973–976.

Tutter,A. and Riblet,R. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 7460–7464.

Yancopoulos,G.D., Desiderio,S.V., Paskind,M., Kearney,J.F., Baltimore,D. and Alt,F.W. (1984) *Nature*, **311**, 727–733.

Received on October 2, 1991