SEQUENCES OF PROTEINS OF IMMUNOLOGICAL INTEREST

FOURTH EDITION

Tabulation and Analysis of
Amino Acid and Nucleic Acid Sequences of
Precursors, V-Regions, C-Regions, J-Chain,
T-Cell Receptor for Antigen, T-Cell Surface Antigens,
β_2-Microglobulins, Major Histocompatibility Antigens,
Thy-1, Complement, C-Reactive Protein, Thymopoietin,
Post-gamma Globulin, and α_2-Macroglobulin

1987

Elvin A. Kabat, Tai Te Wu †, Margaret Reid-Miller ‡,*
Harold M. Perry ‡, and Kay S. Gottesman ‡

* Depts. of Microbiology, Genetics and Development, and Neurology, Cancer Center/ Institute of Cancer Research, College of Physicians and Surgeons, Columbia University, New York, NY 10032 and the National Institute of Allergy and Infectious Diseases, Bethesda, MD 20892.

† Depts. of Biochemistry, Molecular Biology, and Cell Biology, and Engineering Sciences and Applied Mathematics and Biomedical Engineering, Northwestern University, Evanston, IL 60201 and the Cancer Center, Northwestern University Medical School, Chicago, IL 60611

‡ Bolt Beranek and Newman Inc., Cambridge, MA 02238

The collection and maintenance of this data base is sponsored through Contract N01-RR-8-2158 by the following components of the National Institutes of Health, Bethesda, MD 20892:

Division of Research Resources
National Cancer Institute
National Institute of Allergy and Infectious Diseases
National Institute of Arthritis, Diabetes, Digestive and Kidney Diseases
National Institute of General Medical Sciences

U.S. DEPARTMENT OF HEALTH
AND HUMAN SERVICES
Public Health Service
National Institutes of Health
(1987)

Our listing of sequences will be kept up to date. Investigators are invited to send additional sequence data when accepted for publication. Send two copies of the manuscript together with a letter of acceptance from a journal to:

Dr. E.A. Kabat
National Institutes of Health
Building 8, Room 126
9000 Rockville Pike
Bethesda, Maryland 20892

It would be extremely helpful if you can send us your sequence data on magnetic tapes or floppy diskettes or a clean copy of the sequences. The file formats should be such that they can be read by a generic word processor.

When published, three reprints should be provided.

If any published sequences have been overlooked or if any errors are found, please bring them to our attention.

INTRODUCTION

Our earlier "Variable Regions of Immunoglobulin Chains" (1), the second edition "Sequences of Immunoglobulin Chains" (2) and the third edition "Sequences of Proteins of Immunological Interest" (3) have been further expanded herein to include amino acid and nucleotide sequences of precursors, variable regions, constant regions, J-chain, β_2-microglobulins, antigens of the major histocompatibility complex (HLA, H-2, Ia, DR) as well as of Thy-1, complement, T-lymphocyte receptors for antigens, other T-cell antigens of the immunoglobulin superfamily, interleukins and various other proteins related to immune functions. The identification and sequencing of clones obtained using recombinant DNA techniques has yielded nucleotide sequences of signal, variable, and constant regions of immunoglobulins (4,5), and these nucleotide sequences have been translated into amino acid sequences. Instances of the latter have been included in the tables of amino acid sequences and are indicated by an apostrophe followed by CL after the name of the clone. We have continued to use the PROPHET Computer System of the Division of Research Resources, National Institutes of Health (6,7) to tabulate the sequences.

In compiling the data we have tried to be as up-to-date as possible and have included only sequences which have been published or which have been accepted for publication. Residues which have not been definitely determined have been excluded. It should be remembered that sequences are often published in review articles without detailed documentary evidence. These have often been revised. We have listed such revisions in the notes in many instances; others can readily be found by comparison with sequences in previous editions.

Since the preparation of camera-ready copy for printing the pages is carried out in sequence from page 1 in batches, the amino acid sequences were set several months before the nucleotide sequences. We have continued to include new nucleotide sequences up to the point at which camera-ready copy for them had to be set, but translated amino acid sequences were not able to be included. Thus many nucleotide sequences appear without translation. When antibody activities were known, they have been listed at the end of the nucleotide sequences and are included in the index.

When doubts arise as to the validity of any residue in a sequence, the original reference should be examined to ascertain whether definitive evidence for the sequence has been provided. We have sent the amino acid and nucleotide sequences as stored in the computer to the original authors for verification. If so verified, this is denoted by "(checked by author)" at the end of each reference. Except for the earliest sequences, the date on which the checked sequence was returned to us is given. Whenever possible, nucleotide sequences from GenBank (8) have been used. Programs for converting a GenBank sequence to the codon format of our tables have been developed. The correctness of the table sequence has been verified by converting back into the linear form and comparing with GenBank. When this has been done the sequence is listed as "(from GenBank)". If the sequences were entered by us from the literature and then checked with GenBank, this is indicated by "(checked with GenBank)". We have entered many nucleotide sequences which were not available from GenBank. In general, we have not included stretches of sequence such as enhancers, switch regions and introns for which no codification of the nucleotide sequences is as yet agreed upon with respect to function, etc. Much information about such stretches may be found in references 9, 10.

It is also possible, by examining the numbers of sequences at the end of each table and the summary tables, to evaluate the probability that a given amino acid at a given position may not be correct. This is most readily done for the framework residues of the V-region and for the C-region; in the complementarity-determining regions this is more difficult because of the high variability.

Amino Acid Sequences

The first column in each table gives the residue number. Except for complement, T-cell surface antigens and miscellaneous proteins, the second column is a tabulation of invariant residues. Since exceptions to invariance are found, the frequency, if less than 1.0 and greater than or equal to 0.95, is indicated alongside the residue listed as invariant; when only a single sequence is available, this is not given. Each sequence is tabulated in each subsequent column. Dashes (---) indicate that no amino acid is present at that position and that the sequence continues. In all instances residues

considered uncertain by the authors have not been included in the table. In some instances the symbol # is used to indicate that several amino acid residues were found in one position, and these residues are listed in the notes. The four columns at the end of each table give:

1. the number of residues sequenced at that position,
2. the number of different amino acids found at that position,
3. the number of times the most common amino acid occurred and that amino acid in parentheses, and
4. the variability.

Variability is calculated (11) as:

$$\text{Variability} = \frac{\text{Number of different amino acids occurring at a given position}}{\text{Frequency of the most common amino acid at that position}}$$

An invariant position would have a variability of one; if 20 amino acids occurred with equal frequency, the variability would be 20 divided by 0.05 equals 400. If, for example, four different amino acids Ser, Asp, Pro, and Thr occurred at a given position, and of 100 sequences available at that position, Ser occurred 80 times, the variability would be 4/0.8 = 5. When any of the amino acid residues sequenced were not identified completely and are listed as Glx (or Asx), two values, separated by a colon, are given in the last three columns. The first value in each of these columns is calculated assuming that only one of the two possibilities, e.g., Glu or Gln (or Asp or Asn) occurred, while the second considers that both were present and maximizes variability. In the variability plots, the horizontal bars indicate the two values.

When two or more amino acids are most common and occur with equal frequency, they are tabulated as a note, and the symbol + is used in the next to last column. If no sequence data have been reported for any position, there are no entries in the last four columns. Variability is not calculated for insertions or if only a single sequence is known. When the translated sequence of a clone corresponds to a previously listed sequence of a plasmacytoma from which it was prepared, only one sequence is listed so that the variability computations are not affected, and a note is included.

If a given sequence is associated with any antibody activity, this is indicated by an asterisk alongside the protein heading, and the antibody specificities are given in a separate list with binding constants if available. The notes list the a-allotypes for the rabbit heavy chain V-region and the b-allotypes for the constant domain of the rabbit kappa light chain. A key reference to the sequence is given; generally the most recent reference since it is usually the most nearly complete, but often several references are included, especially when revisions of a sequence have been made. Notes are now of two types; general notes about a table indicated by the symbol #, and specific notes indicated by the sequence number.

Signal Sequences

The signal (precursor) amino acid sequences of immunoglobulin chains are listed in three tables: one for kappa light chains, one for lambda light chains, and one for heavy chains. They were obtained either by direct sequencing of signal proteins (12-14) or by translating nucleotide sequences from DNA clones. Signal segments range from 17-29 amino acid residues in length and are thus numbered from -29 to -1. Genomic DNA clones contain introns of varying length that interrupt the coding sequence of the precursor within the codon for position -4, and in rare cases for position -6. Thus, the L-gene encodes the leader peptide to position -4 and the 5′ end of the V-gene codes for positions -4 to -1.

The signal amino acid sequences of the T-cell receptors for antigens, β_2-microglobulins, major histocompatibility complex proteins, and complement components are listed in separate tables.

By conformational energy calculations, the core V_χ hydrophobic Leu-Leu-Leu-Trp-Val-Leu-Leu-Leu (MOPC321, MOPC63) exists in an alpha helical conformation, terminated by chain reversal conformations in the four C-terminal residues Trp-Val-Pro-Gly; the four amino terminal residues are compatible with the alpha helix (15).

Variable Region Sequences

The variable regions (16) of immunoglobulins have been shown to contain hypervariable segments in their light (11,17-23) and heavy (22,24-27) chains, of which certain residues have been affinity labeled (28-41). Three hypervariable segments of light chain were delineated from a statistical examination

of sequences of human V_κ, human V_λ, and mouse V_κ light chains aligned for maximum homology (11,22). These and the three corresponding segments of the heavy chains (22,26,27) were hypothesized (11,22) to be the complementarity-determining regions or segments (CDR) containing the residues which make contact with various antigenic determinants, and this has been verified by X-ray diffraction studies at high resolution (42-67). The rest of the V-region constitutes the framework (11,22,66-68). It is convenient to identify the framework segments (FR1, FR2, FR3, and FR4) and the complementarity-determining segments (CDR1, CDR2, and CDR3) with the three CDRs separating the four FRs. The residue numbers for these segments are as follows:

Segment	*Light Chain*	*Heavy Chain*
FR1	1-23 (with an occasional residue at 0, and a deletion at 10 in V_λ chains)	1-30 (with an occasional residue at 0)
CDR1	24-34 (with possible insertions numbered as 27A,B,C,D,E,F)	31-35 (with possible insertions numbered as 35A,B)
FR2	35-49	36-49
CDR2	50-56	50-65 (with possible insertions numbered as 52A,B,C)[a]
FR3	57-88	66-94 (with possible insertions numbered as 82A,B,C)
CDR3	89-97 (with possible insertions numbered as 95A,B,C,D,E,F)	95-102 (with possible insertions numbered as 100A,B,C,D,E,F,G,H,I,J,K)
FR4	98-107 (with a possible insertion numbered as 106A)	103-113

[a] In the rabbit, Mage *et al.* (69) consider position 65 in V_H to be in FR3, since it is allotype related.

In the tables of V-regions, the FR and CDR are separated by horizontal lines for convenience in reading. One mouse kappa light chain, MPC11, has an extra segment of 12 amino acid residues between position 1 and the signal sequence (70). Several chains have internal deletions.

In the tables, the V-genes for the light chains code to amino acid position 95, and the J-minigenes from position 97 to 107 for lambda and 108 for kappa light chains. Position 96 is usually the site of V-J joining by recombination and may be coded partly by the V-gene and partly by the J-minigene. Because the site of V-J recombination could occur at different positions within a codon, different amino acid residues may result at this position. We have changed the location of the inserted residues from 97A-F (2) to 95A-F, since it makes for better alignment by confining chains of different lengths to the V-gene region. In V_κ chains, J1 and J2 were used 5 to 10 times more frequently than J4 and J5 (71).

The V-genes for the heavy chains code up to amino acid position 94 and are followed by the D- and J-minigenes. Because of the extensive variation in the lengths of D-minigenes, the exact boundary between D and J is not always located at the same amino acid position. In addition, the lengths of the J encoded amino acid sequences vary by a few amino acid residues. Moreover, the process of D-J joining appears to involve insertions of extra nucleotides between V and D and between D and J, termed the N region (72-76) and correlates with the appearance of terminal deoxytransferase in B cells (75). The original numbering system for the heavy chains has therefore been retained. Wysocki *et al.* (76) have provided some evidence suggesting a non-random origin for the V_H-D_H junction, perhaps a minigene, rather than random addition of the N nucleotides.

It has become evident that a critical understanding of the architecture of antibody combining sites and the genetics of the generation of diversity and of antibody complementarity will depend to a great extent on the evaluation of a large number of sequences of the variable regions and especially of the complementarity-determining segments of light and heavy chains of immunoglobulins of different species. Ability to locate residues in the site making contact with antigenic determinants (77) and to predict (67,78-82) the structures of antibody combining sites will depend heavily upon such sequences.

Figures 1 and 2 are stereoviews of the α-carbon skeletons of the four Fv regions for which high resolution X-ray structures have been determined, NEWM (44), KOL (62), MCPC603 (47, 48, 63), and J539 (64). The residues in the CDRs are shown as solid circles. In Fig. 1 the combining site is at the

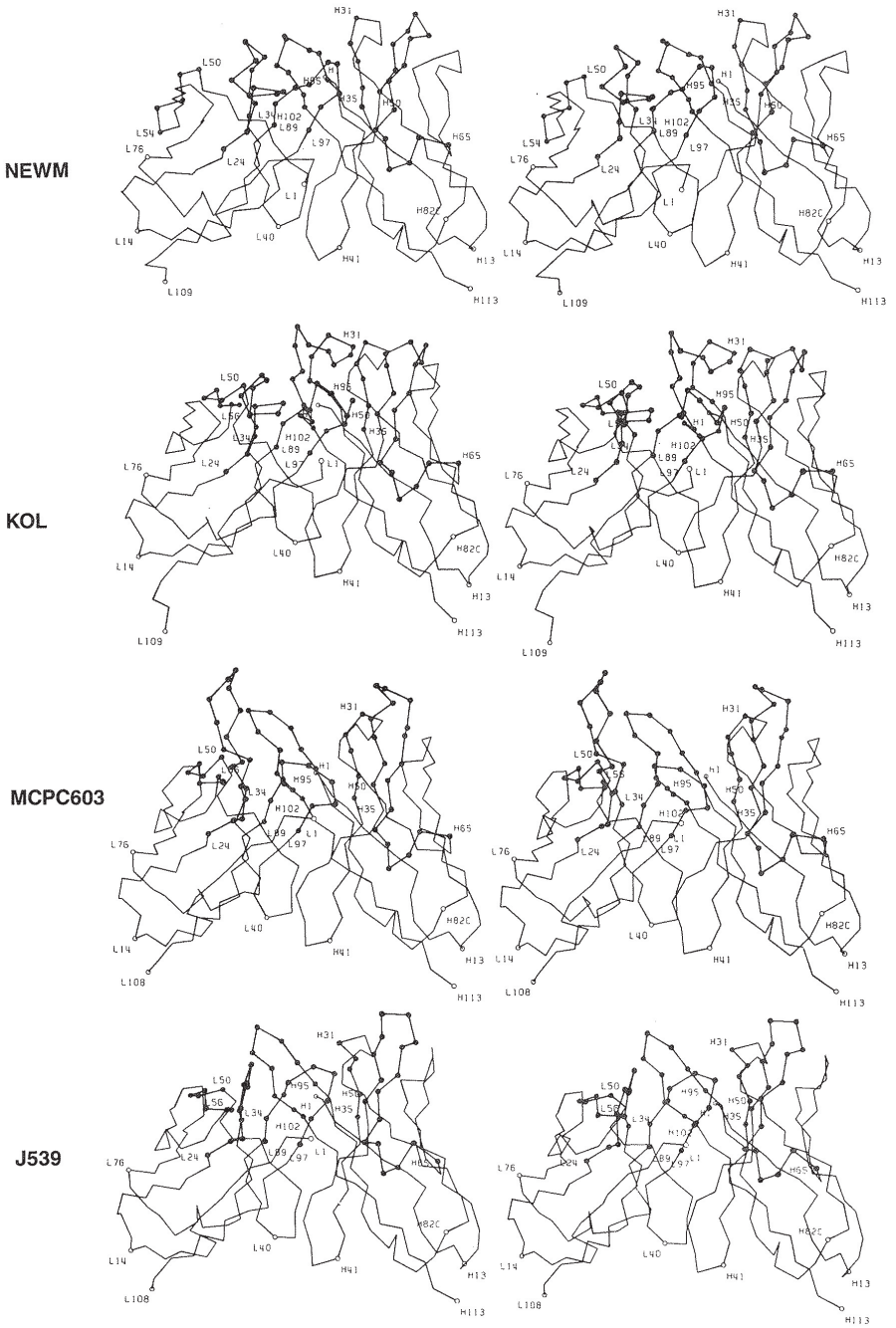

FIG. 1. Stereodrawings of the α-carbon skeletons of four Fv regions studied crystallographically. Top to bottom: NEWM(43,44,49,59), KOL (62), MCPC603(47, 48, 53, 55, 63), J539(64). Coordinates for NEWM, KOL from the Protein Data Bank (Bernstein *et al.* 1977, J. Mol. Biol. 112:532-544); for MCPC603 and J539 courtesy of David R. Davies. The stereodrawings of Figures 1,2,3 and 4 were prepared by Dr. Eduardo Padlan.

V_L is on the left, V_H on the right. The first and last residues of each chain as well as several other residues are shown as open circles for reference; residues of the CDRs are shown as solid circles. The Fv's are aligned by least-squares superposition using the program ALIGN (G.H. Cohen, NIH); the stereoplots were prepared using the program PLUTO (S. Motherwell, Cambridge, England). The view shown is with the combining site at the top. With a stereo viewer it is possible to see two adjacent models at the same time, so that a comparison may be made in three dimensions.

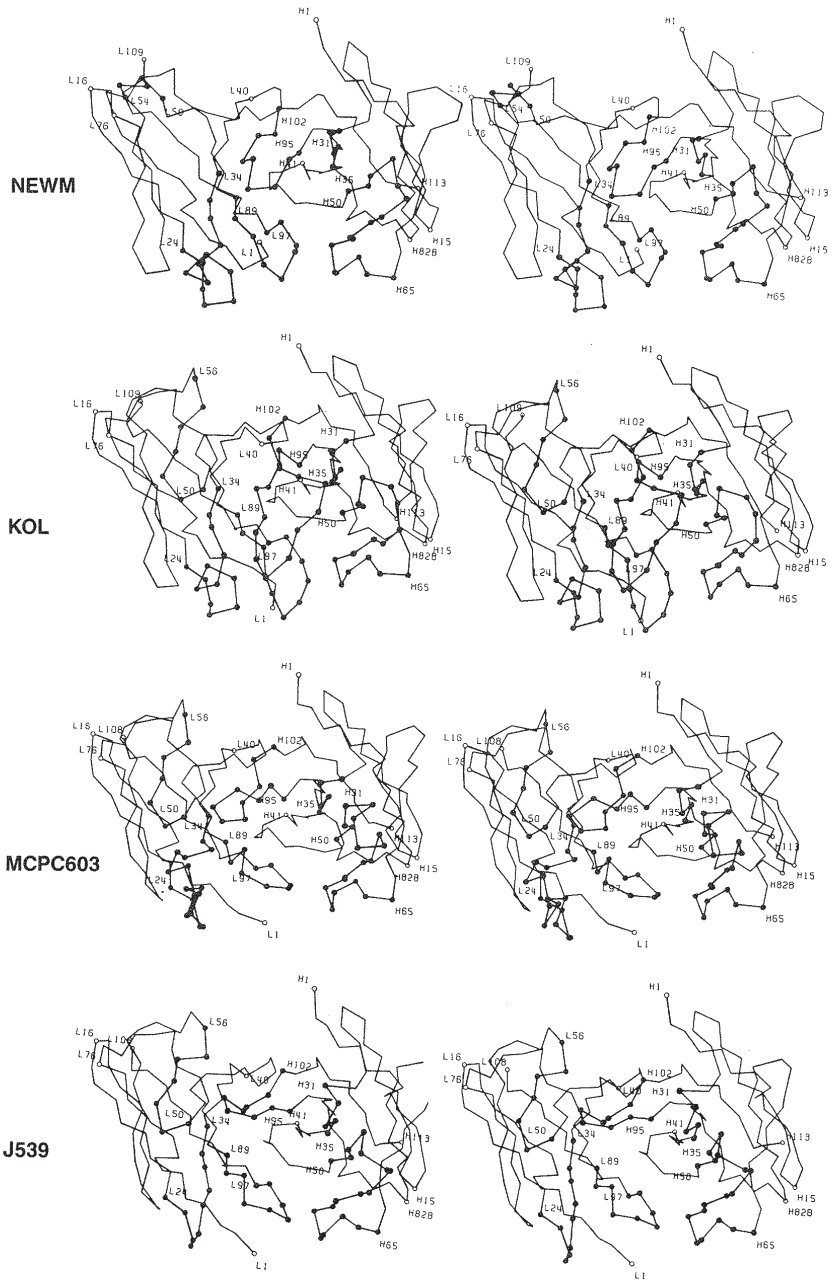

FIG. 2. Same as Fig. 1, but looking into the combining site; models perpendicular to those of Fig. 1.

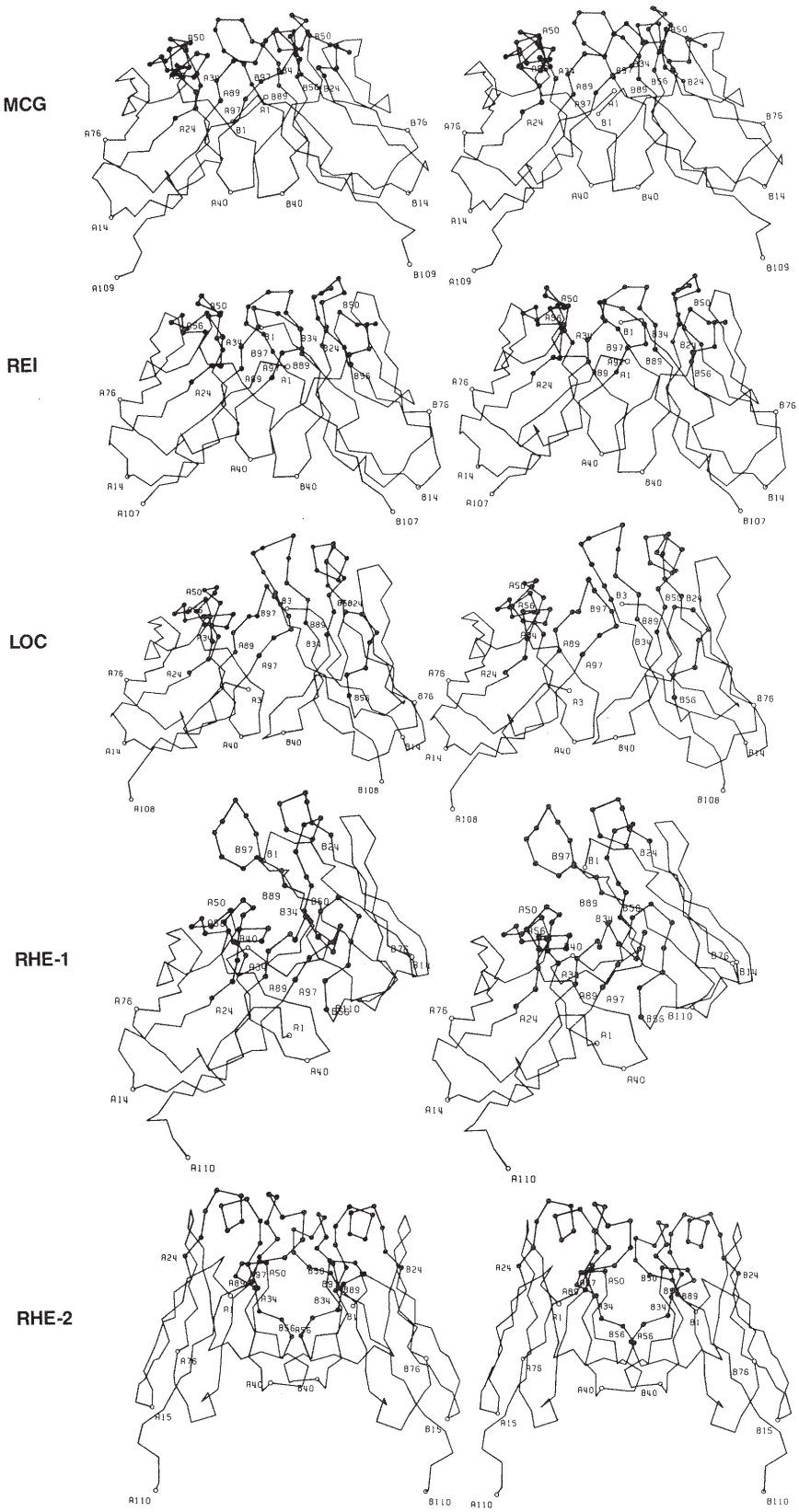

FIG. 3. Stereodrawings of the α-carbon skeletons of four Bence Jones V_L dimers studied crystallographically. Top to bottom: MCG(42,46,50), REI (45, 51, 57), LOC(61), RHE(60). The bottom view is of RHE with its twofold axis toward the top of the page. The left V_L is oriented like the V_L of the Fv's in Fig. 1. The view shown is like that of Fig. 1. Coordinates for MCG, REI, RHE from the Protein Data Bank; for LOC, courtesy of Dr. Marianne Schiffer.

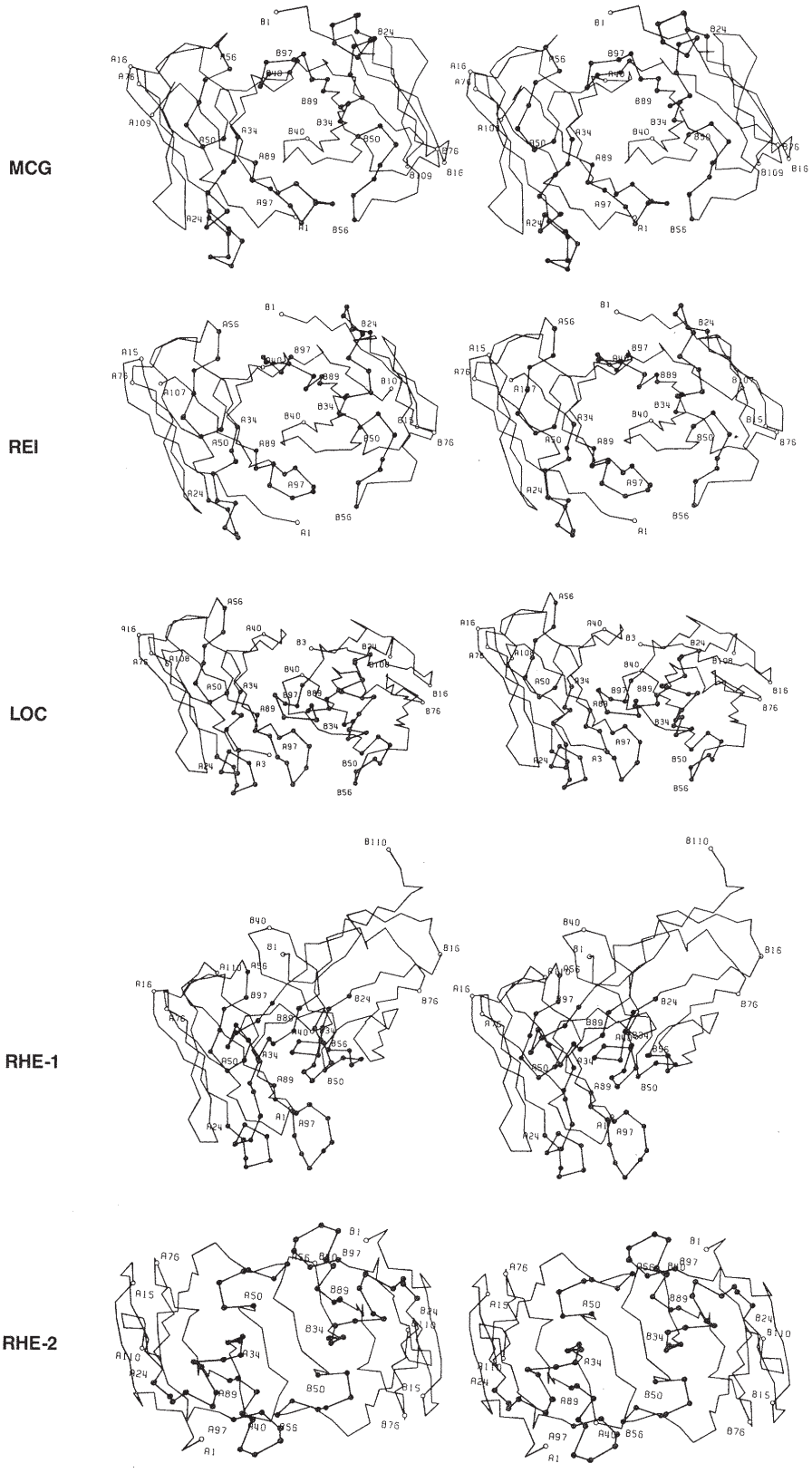

FIG. 4. Same as Fig. 3. but looking into the combining site. Models perpendicular to those of Fig. 3. The bottom view of RHE is looking into its twofold axis.

top; the view of Fig. 2 is perpendicular to Fig. 1 and is looking into the combining site. The different orientations of the loops containing the complementarity-determining regions provide some insight into how specificity of various sites might differ (67,82). If the amino acid side chains were included, the differences would become much more detailed.

Figures 3 and 4 are stereoviews of the four V_L dimers, Bence Jones proteins, for which high resolution X-ray structures are available, MCG (42,46,50), REI (45, 51, 57), LOC (61) and RHE (60). The V_L chains each contribute four β-strands to the V_L-V_L or V_H-V_L interaction whereas the V_H chains each provide five (65). Thus, although in a Bence Jones V_L dimer one V_L assumes the position of V_H (50), nevertheless the absence of one β-strand in each V_L may make the sites of V_L dimers less specific than those of the Fab fragments. This is supported by the finding that the V_L dimer of MCG binds a wide variety of ligands whereas no ligand which binds has yet been found for the MCG protein (83).

A recent high resolution x-ray crystallographic study (84) of a crystalline complex of lysozyme with a monoclonal anti-lysozyme shows that contact between lysozyme and antibody occur on a rather flat surface with the interactions largely due to protruberances and depressions formed by the amino acid side chains producing a tightly packed region of interaction. The lysozyme determinant involves two noncontiguous stretches, residues 18 to 27 and 116 to 119 of its polypeptide chain. All six CDRs of the antibody and two residues outside the CDRs but adjacent to the CDRs, Tyr 49 in V_L and Thr 30 in V_H, make contact with the lysozyme. Ten of the 17 contacting residues are in V_H. Four of the 10 contacting V_H residues and three of the seven contacting residues in V_L are in the corresponding CDR3s. Table 1 lists the residues on the anti-lysozyme and on lysozyme which are in contact. These findings, if and to the extent applicable to anti-carbohydrate sites, with respect to interactions of side chains on essentially flat surfaces could have substantial implications for our understanding of these antigen-antibody interactions. However, the J539 site would appear to be some type of groove complementary to a tetrasaccharide. Unfortunately thus far the crystal form has not allowed the ligand to enter the site (64). Figure 5 is a stereoview of the lysozyme-antilysozyme carbon skeleton showing the region of interaction.

Bence Jones dimer LOC (61) has a convex (male) binding site quite different from the usual Bence Jones dimers MCG and REI. If such a male type combining site were to be found for an Fab, the possibility would have to be considered that a reciprocal type of antigen-antibody interaction might occur in which the side chains of the CDRs would fit into a groove or depression on the surface of an antigen. The possibility has been noted that interactions might occur with the CDRs of the two faces of the projecting convex site (61). RHE also has a quite distinct type of binding site based on a unique V_L-V_L interaction. The basis for such differences is not understood but could contribute a new parameter to site complementarity and diversity.

The sequence data may be used to make rough screens of a new sequence for homology with the V-region. If the sequence to be compared is aligned with the large V-region summary tables, one can ascertain whether any homology exists. If homology involves the less frequently occurring residues, they can be found in the individual tables and homology evaluated.

The variable region a-group allotypes and allotype a-negative rabbit V_H chains have been correlated with certain amino acids in FR1 and FR3 as follows (69):

Allotype	Amino Acid Position																	
	FR1							FR3										
	5	8	10	12	13	16	17	65	67	70	71	74	75	76	84	85	87
V_H a1	glu	gly	ARG	val	THR	thr pro gly	pro gly ser	gly	phe	ser	lys	thr	[-][a]	[-]	THR	GLU	thr
V_H a2	LYS	GLU	gly	PHE	lys	ASP	THR	SER	SER	THR	ARG ser	ASN	GLU	asn	ala GLY ala	GLN ala ala	thr
V_H a3[b]	glu	gly	ASP	val	lys	ala	ser	gly	phe	ser	lys	thr	[-]	[-]	ala	ala	thr
V_H a100	glu	gly	gly	val	gln	ala	ser		thr	ser	lys	ser	[-]	[-]	?	?	MET
V_H a-	glu val	gly	gly	val	gln	gly glu thr	gly ser	gly	phe	ser	ser	ala	gln	asn	ala	ala	thr

[a] Square brackets indicate gaps to maximize homology. Allotype related residues are in capitals.

[b] In some a3-like genes, codons for amino acids 75 and 76 were found (85).

In Figs. 1 and 2, the location of the allotypic regions may clearly be seen to be on the outside of V_H away from the combining site. Residues 13 and 65 of V_H are numbered and will facilitate location of the V_Ha allotypes. The few cDNA sequences (69) available have provided no evidence as yet that germ line sequences encoding latent allotypes may exist in some rabbits. Antisera to rabbit V_Ha allotypes crossreact with human IgG (86), various other species of IgM and IgG, and with the Galapagos shark 7S immunoglobulin and correlate with the N-terminal amino acid sequence (87).

There are substantial species differences between the human, rat and rabbit C_κ allotypes. The amino acid sequences of rabbit C_κ allotypic determinants K-1, b4, b5 and b9 differ at 47 of 106 positions, the differences occuring in clusters; the K-2 bas isotype differs at three additional positions (88) whereas the human C_κ allotypes differ by two positions (89) and the rat R1-1a and R1-1b differ at 11 positions (90).

TABLE 1

Antibody Residues Involved in Contact with Lysozyme

Antibody residues		Lysozyme residues in contact
Light Chain		
CDR1	His 30	Leu 129
	Tyr 32	Leu 25, Gln 121, Ile 124
FR2	Tyr 49	Gly 22
CDR2	Tyr 50	Asp 18, Asn 19, Leu 25
CDR3	Phe 91	Gln 121
	Trp 92	Gln 121, Ile 124
	Ser 93	Gln 121
Heavy Chain		
FR1	Thr 30	Lys 116, Gly 117
CDR1	Gly 31	Lys 116, Gly 117
	Tyr 32	Lys 116, Gly 117
CDR2	Trp 52	Gly 117, Thr 118, Asp 119
	Gly 53	Gly 117
	Asp 54	Gly 117
CDR3	Arg 99 (96)	Arg 21, Gly 22, Tyr 23
	Asp 100 (97)	Gly 22, Tyr 23, Ser 24, Asn 27
	Tyr 101 (98)	Thr 118, Asp 119, Val 120, Gln 121
	Arg 102 (99)	Asn 19, Gly 22

Sequence positions are numbered as in this book except for V_H CDR3, where the numbers are given in parentheses; the others are sequential. (From (84). Amit, Mariuzza, Phillips, and Poljak (1986) *Science* **233**:747-753; courtesy of Dr. Roberto Poljak and *Science*. Copyright 1986 by the AAAS.)

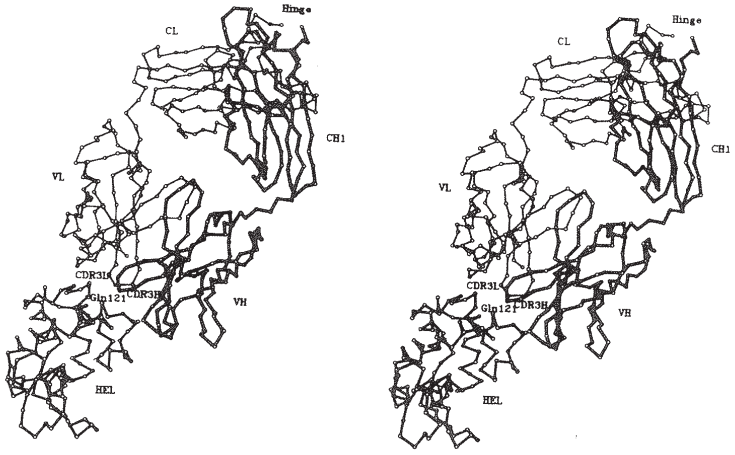

FIG. 5. Stereo diagram of the Cα skeleton of the complex. Fab is shown (upper right) with the heavy and light chains with thick and thin bonds, respectively. The lysozyme active site is the cleft containing the label HEL. Antibody-antigen interactions are most numerous between lysozyme and the heavy chain CDR loops. (From (84). Amit, Mariuzza, Phillips, and Poljak (1986) *Science* **233**:747-753; courtesy of Dr. Roberto Poljak and *Science*. Copyright 1986 by the AAAS.)

It has proven extremely useful, except for mouse V_κ chains, to order the V_L and V_H sequences into sets (68) such that all chains with identical FR1 are listed together, the set with the most members being listed first. Chains differing in sequence from this set by a single residue are then listed in order of substitution, beginning at residue 22 for light chains and residue 30 for heavy chains and proceeding in decreasing position number to residue 1. These are then followed by chains with two amino acid differences, again listing in the same decreasing order, and followed by chains with three amino acid substitutions, etc. Amino acid residues differing from the major FR1 sequence are given in lower case letters, so that one can readily see the pattern of substitution. In this ordering procedure, missing residues are treated as potentially different from the main sequence. If residues are missing at position 23 in the light chain and position 22 in the heavy chain, they are assumed to be Cys to preserve the essential V-domain structure. Finally, sequences which are incomplete in FR1 are given. Within a given FR1 set, identical FR2 sets are also listed together.

The human V_κ rearranged and germ-line genes of all four subgroups have been sequenced (91-95). Human V_κIV has but a single germ-line gene (92, 93); thus somatic mutation must play a dominant role in the utilization of this gene. V_κII genes are characterized by a much longer intron between the signal and V-region (94) than the other subgroups. Unlike the mouse the human V_κI, V_κII, and V_κIII genes are not separated in the genome, but a large section of the V_κ locus has been duplicated; both sections existing as two non-allelic clusters, containing eight and six genes of all three subgroups (91,95). All genes are in the 5′→3′ orientation. These findings have necessitated reclassification of some incomplete sequences (RPMI-6410′CL).

The tables of mouse V_κ light chains have been rearranged (96,97). In previous editions (1-3), mouse V_κ light chains were listed in one table with the length from residue 1 to Trp 35 specified. They have now been separated into eight tables. The first six tables vary from 41 to 34 residues by the different lengths of CDR1 (residues 27A-F); the sixth also lacks residue 28 and the seventh is also missing residue 22 in FR1. If residues 1 through 35 have not been determined completely, the sequences are listed in the eighth table, unless they show good homology to more complete sequences in one of the other tables. In each table, the group number is given below the name of the chain. When residues 1 through 35 have not been determined, the earlier group designation based on residues 1-23 (96) is given in square brackets [ ]. However, for each table the same principles have been used in ordering the chains. In all instances, residues 1-35 for the largest group are given in capitals. Variations from this sequence are in lower case, beginning at residue 34 and proceeding toward residue 1 as in the other tables.

The mouse V_H sequences have been revised to take into account their division into families by Dildrop (98) based on amino acid sequences and of Brodeur and Riblet (100) based on nucleotide sequences of completely sequenced V-regions. We have however retained the earlier classification of V_H chains into subgroups but have subdivided the subgroups to list the families as follows:

		Antibody specificities found	Family: Dildrop (98) Dildrop et al (99)	Family: Brodeur and Riblet (100, 101) Winter et al (102)
Subgroup I	A	Ars, DNP, HEL, DIG, poly-GA	3	V_H36-60
	B	2-PHEOX, HEL	2	V_HQ52
			8	V_H3609, MV31
Subgroup II	A	α(1→3)α(1→6)DEX, RNA, Ars, α(1→6)DEX, DIG, HEL, IdAc38, GAT		V_HJ558
			9	VMU-1, VGAM3-8
	B	NP, GAT, DIG, PC, IdAc38, 2-PHEOX	1	
	C	GAT, GA, H-2K-k		V_HJ606
Subgroup III	A	PC, DNP, HEL, DNA	7	V_HS107
	B	β(1→6)GAL, α(1→6)DEX, NAcMAN, β(2→6)FRU, GAT, STR-A(β-DGlcNAc)	4	V_HJ558, V_HX24
	C	β(2→1)FRU, STR-A, SRBC, GAT	6	V_HJ606
	D	DIG, H-2K-k, SRBC	1	J558
Subgroup V	A	ARS CRI+, ARS CRI-		
	B	ARS, ARS CRI+	1	J558
Miscellaneous		ARS, DNA, HEL, H-2K-k, GAT, DIG, 2-PHEOX, CEA	5	V_H7183

For newly sequenced V_H regions, the nucleotide or amino acid sequence homologies have been compared with the previously classified sequences.

It is evident that antibody specificities for a given antigen fall into several of these subgroups and that many framework residues which are characteristic of the subgroups are sufficiently different to indicate that different germ-line genes may give rise to antibodies of a given specificity (103).

The classification given is in better accord with the amino acids (98) than with the nucleotides (100), since the latter used probes whereas the former was based on complete sequences. Two rabbit germ-line a-negative V_H clones showed the greatest nucleotide sequence homology, 70.4-79.4% to V_H X24, V_H 7183, V_H J606, and V_H S107 (100) and considerably less homology to the other families.

Each table of mouse V_κ and V_H chains is followed by a list of strains other than BALB/c.

The members of identical FR and identical CDR sets are given in the notes. Members of individual FR1 sets may be associated with different FR2 sets, etc. (68), suggesting independent assortment of FR sets.

A sequence identical to the FR2 sequence of the light chain of McPC603 has been found in two human V_κI, one human V_κ IV, 31 mouse V_κ (14 NZB and 17 BALB/c), one each mouse V_κI and V_κ VI, and 15 rabbit V_κ sequences, and thus has been preserved for about 80 million years (104). It and the corresponding loop of the heavy chain are seen at the bottom center of each of the four stereo figures (Fig. 1); L40 and H41 are numbered to facilitate its location. Despite its preservation, there are 12 other FR2 sets in the mouse and 8 in the rabbit with sequence variation which may involve 13 of the 15 positions, only Trp 35 and Gln 39 being invariant. The loop is in a relatively open position so that substitutions are readily permitted (104,63). The evidence of assortment of FR segments (68) suggested the hypothesis that the V-region was coded for by sets of minigenes for the FR and CDR segments and that these minigenes were assembled somatically during embryogenesis. FR2 of V_H also shows substantial preservation, one set having six mouse and eight rabbit chains of identical amino acid sequence. It is extraordinary that one human V_HIII genomic clone, V_H26 (105), and a rabbit cDNA clone (106), were identical in nucleotide sequence of the codons for amino acids 36 through 47, differences being seen in codons for amino acids 48 and 49.

Since only FR sets were used in demonstrating the assortment (68), it would be independent of and would be seen whether or not any CDR residues assorted with any FR. The early cloning studies (4,5) showed that the genes coding for residues comprising FR1 almost through CDR3 of mouse V_λ and V_κ light chains were assembled in twelve-day-old mouse embryo DNA, and each was followed by an intervening sequence; in one V_λI clone, two residues of CDR3 were included with FR4 (107). In an adult V_λ myeloma, the genes coding for the entire V-region were assembled (108). Thus, the minigene coding for FR4 plus the last two residues of CDR3 (107), termed the J segment (108), had been joined to the rest of the V-region between the twelfth day of embryonic life and the adult, and thus was added somatically by recombination. Milstein (17), in his original description of human V_κ groups, had pointed out that subgroup associated residues extended only through residue 94 and that it seemed as if frequent crossing over occurred beyond residue 94. Weigert *et al.* (109), in studying the V_κ 21 group in NZB myelomas, assorted the last two or three residues of CDR3 together with FR4 and suggested that this would contribute to the generation of diversity. With rabbit light chains, it was possible to assort the individual FR and CDR segments considering FR4, plus the last two residues of CDR3, as a J-minigene (110).

The subsequent demonstration and sequencing of five J nucleotide segments from BALB/c, each separated by an intervening sequence in the mouse genome, supported the assortment data (68) and established J as a minigene. Four of the J segments encoded amino acid residues 96-108, but J3 appeared to be a pseudogene (111,112). In mouse V_H four J-(113-115), in human V_H six J-(116), in rat V_κ six J-(117,118), in human V_κ five J-(119), and in mouse V_λ four J-minigenes (120,121) were recognized; J_λ4 may not be functional. In rabbits of b4 and b5 allotypes, there are five V_κ J-minigenes (122, 123); only J2 being functional. There appear to be two or more J_κ minigenes that can be expressed with the b9 allotype and three potentially functional J_κs are found associated with the K2 isotype (124, 125).

Unlike the findings for light chains, the mouse genomic V_H and J_H segments did not encode complete V-regions (113,114). Segments coding for five and for 14 amino acids, most of CDR3, were missing. Subsequently, genomic DNAs capable of encoding three, four, and six amino acids were located, using as a probe an incompletely rearranged D-J segment including the flanking signal sequences (126-128). Some of these mouse D-minigenes coded for as many as six amino acids of CDR3, matches of 9 to 18 nucleotides having been found.

Four human D-minigenes have been sequenced (129), but their presumed coding regions did not correspond to any CDR3. However, one V_HIII human genomic clone (105) contained fourteen nucleotides in CDR2 which completely matched the D2 minigene coding segment (130). Another human V_HII genomic clone (131) matched at 13 of these 14 nucleotides, the first of the fourteen nucleotides being *c* instead of *g* (131). Other instances were a match of 12 nucleotides in mouse (132) and a rabbit cDNA (133). The amino acids coded for by these segments (130) were present as residues 53-56

in two fractions of type III rabbit antipneumococcal antibody from one rabbit (134,135); additional matched sequences *tatt* and *tact* include residues 51 and 58 in the human V_HIII gene (130). The V_H encoding sequence of a rabbit cDNA clone (106) also matched for eight residues of the D1 minigene in CDR2 (129) and also had homologous sequences *tatt* on the 5′ and *tat* on the 3′ side. Thus the question arises as to how and why these nucleotide sequences appear in CDR2. Moreover, Takahashi *et al.* (136) have described a rearranged immunoglobulin pseudogene with all of CDR2 deleted. D and J minigenes are also being found in the T-cell receptor for antigen. The D minigenes in both B (137) and T (138) cells were found to have a promotor 5′ to D and D_H-J_H-C_H and D_T-J_T-C_T rearrangements and both made mRNA. A D_H-J_H-C_H miniprotein was actually expressed (137); its role has not been established, but since much of the idiotypic specificity is associated with D_H, the functions of such miniproteins should be studied especially since anti-idiotypic sera are being produced by immunization with synthetic peptides of CDR3 coupled to keyhole limpet hemocyanin (see 139, 140).

The sequence data on mouse V_λ chains, in which 12 sequences were identical over the entire V-region and seven others showed amino acid replacements essentially confined to the CDR that resulted from single base substitutions, were the first indication that somatic mutation could be contributing to antibody diversity (141,142). There are only two germ line genes, one V_λI and one V_λII, so that the other V_λ sequences arose by somatic diversification (120,121). Subsequent data on the mouse V_κ 21 group (143), amino acid (144,145) and nucleotide (146) data on phosphorylcholine specific myelomas and hybridomas supported this interpretation. Nucleotide sequences in several other systems (147-150) also showed single base changes in the CDR as well as silent third base changes.

Recently somatic mutation has been shown to play a major role in generating diversity in mouse V_λ chains, the changes being predominantly in the CDRs (151). It was estimated that the number of V region variants in V_λI alone is close to the total number of B-cell clones in the individual mouse (151). This finding makes it important to sequence the V_λ chains of the $\alpha(1\rightarrow3)\alpha(1\rightarrow6)$ anti-dextrans, the anti-NP and other systems with V_λ chains.

The chicken also rearranges a single V_λ gene to a single V_λ-J_λ-C_λ sequence (152). However, the chicken has a large number of pseudogenes 5' to the functional V_L sequence. These pseudogenes are composed of fragments of V regions. Weill (153) has proposed that the chicken generates diversity of antibody combining sites in the bursa of Fabricius by a series of gene conversions which create independent assortment of CDRs resembling the proposed minigene assortment (68).

V-J and V-D-J joining provide another important mechanism for generating diversity in that recombinations may occur at different nucleotides within a codon generating diversity at the boundaries (junctional diversity) (109,111,112,150). But not all of the substitutions at these junctions can be accounted for by this mechanism (154). It should be borne in mind that diversity generated by recombination or addition of N sequences at the V-J and V-D-J boundaries also makes it difficult to distinguish between it and somatic mutation at these junctions. Somatic mutation thus is most clearly analyzable in CDR1 and CDR2 of each chain, subject to the uncertainties created by finding human D2 (130,131) and D1 (106) minigene sequences in a portion of CDR2 of human, mouse, and rabbit V_H genes and by the finding that the complementary strand of the D2 and D4 minigenes has the nucleotide sequence of a portion of CDR1 of V_κ (155).

The findings, by Southern blotting and sequencing, that nucleotides coding for V_L, excluding J, and for V_H, excluding D and J, occur as contiguous coding sequences in the germ line has been considered to be in conflict (113,156) with the assortment data (68,104,110). This has been largely resolved by the suggestion (157,158) that gene conversion could be creating such assortment in the V-region. Moreover, assortment has been shown at the nucleotide level (159) using probes each containing FR1, FR2, or FR3, as well as an intact germ line V_κ probe (not containing J), using a dot blotting technique much more sensitive than Southern blotting. Twenty-eight clones were isolated by hybridization to a cDNA plasmid containing the entire V-region of MOPC21, which did not hybridize to J or to C region probes and thus contained unrearranged V_κ embryonic genes. Of these, six hybridized to all three individual FR probes, three hybridized only to the FR1, 10 only to the FR2, and nine only to the FR3 containing probes. Thus Southern blotting may not be sensitive enough to reveal the assortment, and the intact V probes select only intact V-region genes. Additional evidence of assortment of two idiotypic determinants by recombination involving residues in FR3 has been presented (160). Assortment by somatic recombination could rescue functional segments of pseudogenes (158) and thus contribute to the generation of diversity as is found for the chicken (153).

The apparent conflict between the assortment data and the occurrence of contiguous sequences of germ-line genes through FR3 in V_H and up to J in V_L has been clarified by a study in which nucleotide sequences of various human V_κI chains were related to the amino acid sequences (161). Assortment of amino acids of the FRs and CDRs was seen. However, the corresponding nucleotide sequences showed that the sites of recombination did not occur at the FR-CDR boundaries but varied from one sequence to another, being generally at different codons in the FR segments in which the higher degree of homology favored recombination. Thus the amino acid data appeared to indicate the existence of

minigenes, but from the nucleotide data it was evident that other mechanisms such as gene conversion could have led to increased diversity of germ-line genes in evolution (161). These data however support the original idea that assortment of CDR segments could contribute importantly to increasing the diversity of V_H and V_L chains; it is not clear whether such recombinations occur somatically. Substitutions involving short sequences of amino acids in the phosphorylcholine system were also ascribed to gene conversion (162). High frequency of inverted repeats was seen in the FRs as compared with the CDRs (163) which could be related to gene conversions.

It is becoming of great importance, with all of the different mechanisms which are clearly generating diversity, to evaluate the extent to which each type of diversity, other than those resulting in pseudogenes, contributes noise rather than functional differences in complementarity of antibody combining sites (164, 165).

Ohno *et al.* (166, 167) have proposed that the genes coding for variable domains of the light and heavy chains arose from tandem repeats of a primordial nucleotide sequence about 48 base pairs in length which subsequently diverged by mutations and deletions producing a resemblance to FR1, FR2, and FR3 (167). The complementary strand of the primordial 48 base pair repeat of V_L became the primordial V_H. The finding (155) that the complementary strands of the human D2 and D4 minigenes coded for a portion of CDR1 of V_κ tends to support this hypothesis. A 45 base pair primordial building block has been proposed for the gene for the class I major histocompatibility complex (168).

Constant Region Sequences

The constant region sequences were aligned in such a manner as to permit various comparisons of the light chain (C_L) and the individual domains of the heavy chain (C_H1, C_H2, C_H3, and C_H4). This was accomplished by sequential numbering on the left with gaps inserted for alignment. The following numbering system is used:

108 to 215 of C_L;
114 to 223 of C_H1, plus the first part of hinge (224 to 241),
the end of hinge (242 and 243),
and the first two residues of C_H2 (244 and 245);
246 to 360 of C_H2;
361 to 496 of C_H3;
497 to 628 of C_H4.

The gene quadruplication in the human IgG3 hinge region (169) is numbered differently using letters 241A to 241Z, and 241AA to 241SS, and these residues should not be used in aligning domains for homology. The next two columns in the heavy chain tables indicate the EU (170) and OU (171) residue numbers, respectively. The succeeding columns which are numbered give the sequence data. The C_H and hinge domains conform to the findings of Sakano *et al.* (172), who defined each domain precisely by sequencing the coding and intervening nucleotide sequences bordering each domain.

The extensive nucleotide sequence data on exons for the constant regions of heavy chains have provided exact boundaries for C_H1, hinge, C_H2, C_H3, and C_H4. Usually the introns separating these domains fall within the codon for a single amino acid. We have included that amino acid residue with the domain, the exon of which contains two of the three coding nucleotides. The constant regions of heavy chains thus contain four domains: C_H1, hinge, C_H2, and C_H3, or in IgM and IgE C_H1, C_H2, C_H3, and C_H4. Several C_H3 domains are extra long. These have been listed in a separate table after C_H3. An additional table is included to list several C_H4 domains that are very long and the C-terminal portions of membrane bound segments. Such extra long sequences in C_H3, C_H4, and membrane bound segments are listed retaining their individual protein numbers.

Figure 6 (173, 67) shows the domain structure for IgG1 protein EU. Numbering on the left half indicates the CDR for the light and heavy chains (67, 164), while that on the right half gives the EU numbering (170).

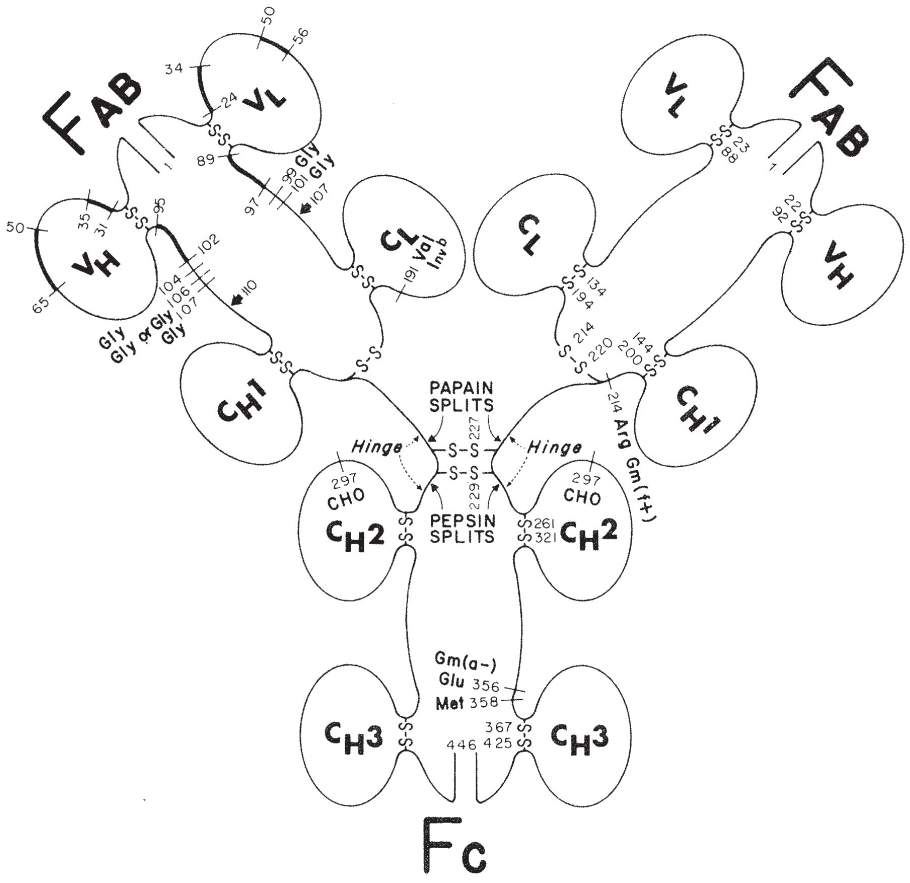

FIG. 6. Schematic view of four-chain structure of human IgG1$_\kappa$ molecule. *Numbers on right side:* actual residue numbers in protein EU [Edelman *et al.* (170)]; *Numbers* of Fab fragment on *left side* aligned for maximum homology; light chains numbered as in Wu and Kabat (11) and heavy chains as in Kabat and Wu (22). Heavy chains of EU have residue 52A, three residues 82A,B,C, and lack residues termed 100A,B,C,D,E,F,G,H,I,J,K, and 35A,B. Thus residue 110 (end of variable region) is 114 in actual sequence. Hypervariable regions, complementarity-determining segments or regions (CDR): *heavier lines.* V_L and V_H: light and heavy chain variable region; C_H1, C_H2, C_H3: domains of constant region of heavy chain; C_L: constant region of light chain. *Hinge region* in which two heavy chains are linked by disulfide bonds is indicated approximately. Attachment of carbohydrate is at residue 297. *Arrows* at residues 107 and 110 denote transition from variable to constant regions. Sites of action of papain and pepsin and locations of a number of genetic factors are given. Modified from 67.

The notes for the constant region list those residues associated with the Gm determinants in human and allotypic determinants in rabbit IgG (174-176). The proteins are also classified as to subgroups in the notes, Table 2 summarizes those residues associated with various Gm Groups; for the rabbit C_H allotypes, the codons are also given (176).

TABLE 2

Heavy Chain Allotype Markers Associated with Defined Sequence Differences

Heavy Chain	Homology Region	Allotype	Sequence			
Human			355			358
$\gamma 1$	CH3	Gm(1)(a)	Arg	Asp	Glu	Leu
		Gm(non-1)(non-a)	Arg	Glu	Glu	Met
			214			
$\gamma 1$	CH1	Gm(17)(z)	Lys			
		Gm(3)(f,b^w,b^2)	Arg			
			436			
$\gamma 2$	CH3	Gm(11)(b^0)	Phe			
		Gm(non-11)(non-b^0)	Tyr			
			296			
$\gamma 2$	CH2	Gm(5,non-21)(b,non-g)	Phe			
		Gm(non-5,21)(non-b,g)	Tyr			
			355			358
$\gamma 4$	CH3	Gm(non-1)[a]	Gln	Glu	Glu	Met
			309			311
$\gamma 4$	CH2	Gm(4a)	Val	Leu	His	
		Gm(4b)	Val	---	His	
Rabbit			309[b]			
γ	CH2	e14	acg	Thr		
		e15	gcg	Ala		
			225[b]			
γ	hinge	d11	atg	Met		
		d12	acg	Thr		
			408[b]			
γ	CH_3	e14[c]	aac	Asn		
		e15	agc	Ser		

[a] Only in IgG4
[b] EU numbering
[c] There is no evidence that these amino acid substitutions contribute to an allotypic determinant.
From (174, 175)

For comprehensive reviews of the constant regions including their biological functions, domain structures, disulfide bridges and three-dimensional structures, see references 175, 177-182.

X-ray studies on whole IgG, except for two proteins Dob (183, 184) and Mcg (185) in which the hinge is deleted, show a lack of electron density in the Fc region which has been related to flexibility of the hinge (186, 187). The deletions in the hinge region appear to cause impairment of biological activities (181, 188, 189). This is considered a serious gap in our knowledge of IgG structure (181).

A variety of immunological activities have been reported for peptide sequences that are identical to portions of various immunoglobulin domains (190-192). The tetrapeptide tuftsin—Thr-Lys-Pro-Arg residues 289-292 (EU numbering)—occurs in the C_H2 domain of IgG1, IgG3, and IgG4. It has been found to increase the capacity of antigen-pulsed macrophages to educate T-lymphocytes. When coupled to bovine serum albumin via its C-terminus, it increases the antibody response to bovine serum albumin (190). A 24 amino acid containing peptide—corresponding to the first 24 residues of C_H3 of IgG1, 335-358 (EU numbering)—has been shown to be as active as intact Fc in polyclonal stimulation of

mouse B cells to secrete antibody (191). Tuftsin plus the four amino acids of IgG at the N-terminus, the four amino acids of IgG at the C-terminus to give octa- and decapeptides had comparable activity in binding to and stimulating phagocytosis of opsonized sheep cells by peritoneal macrophages (192, 193); thus the receptors for tuftsin and Fc on macrophages may be related or even identical.

From a study of the constant region sequences of various immunoglobulins and correlation of sequence with degree of binding of the Fc receptors on monocytes, a potential monocyte binding site has been localized on human IgG (194).

J-Chain Sequences

Amino acid sequences from a human J-chain and from a clone of mouse cDNA (195) and a human genomic clone (196) derived from the DNA sequence are given. Nucleotide sequences for both are also tabulated. Secondary structure was characterized by computer-assisted modelling and circular dichroism studies on isolated J-chain. It resembled V_L domains and superoxide dismutase and was considered to be an 8-stranded anti-parallel β-barrel with about 37% β-sheet and the remainder reverse turns or random coil (197).

T-Cell Receptors for Antigen

Since the third edition (3), there has been enormous progress in the cloning, characterization, and sequencing of the T-cell receptors for antigen (Ti) (198-202). They are expressed on the T-cell surface having signal, V, J, C, transmembrane and cytoplasmic regions. The derived amino acid sequences show them to belong to the immunoglobulin superfamily, being built of V- and C-domains each with an S-S bond with about as many amino acids in the loop as in immunoglobulin light and heavy chains, but with a relatively low degree of homology to immunoglobulin chains. The T-cell receptors for antigen have two chains, α and β; the V-genes of each can join to J-minigene segments and the β chain has D-minigene segments as well; whether the α chain has a D is uncertain.

Variability plots (203-209) are as yet based on small numbers of chains so that they are clearly not definitive. Tentative interpretations suggested considerably more hypervariable regions than were present in immunoglobulin chains. The occurrence of two sets of identical V-regions in but 21 T_β sequences may indicate a relatively small number of V-regions. One set of identical V-regions uses different Js but both chains of the second pair use the same J. Human, mouse and rabbit V-regions of the β-chains of T-cell receptors for antigen have been divided into two and possibly three subgroups (210) based on their ability or inability to form a salt bridge and the occurrence of a different invariant amino acid residue adjacent to one of the residues involved in the salt bridge. (Repeats of identical sequences have been excluded.) Subgroup I has an invariant Phe at position 65 and a salt bridge can be formed between the Asp at position 86 and the 11 Arg, 2 Lys and 1 His residues at position 64 in 5 human, 1 rabbit, and 9 mouse β-chains. A tenth mouse β-chain has Phe 65 and Lys 64 but has a Tyr at position 86 and could not form a salt bridge. Subgroup II has an invariant Tyr at position 65, 8 chains have Gly at position 64 and three have Ala at position 63 and Asp at position 64. Those with Gly 63 cannot form a salt bridge in this region of the chain. The three with Asp 64 have Arg at position 86, could form a reverse salt bridge and might constitute a third subgroup. The β-chains of the T-cell receptors for antigen thus do not constitute a homogeneous population and variability plots for which a homogeneous population of chains is assumed may be misleading. Examination of the total number of sequences of V_L and V_H chains in the present volume show that each consists of a homogenous population of sites with respect to their ability to form a salt bridge between Arg 61 and Asp 82 with an essentially invariant Phe 62 in V_L and between Arg and Lys 66 and Asp 86 with Phe, Ala or Leu at position 67 in V_H. Thus although homogeneity of the populations of V_L and V_H chains was an assumption at the time variability plots were introduced (11, 22), subsequent studies have amply justified it and X-ray crystallographic studies (cf. 210) have shown salt bridges to be present in V_L and V_H chains. The role of these subgroups in the functions of the T-cell receptors for antigen remains to be determined. A gamma chain for the T-cell receptor for antigen has also been found which appears early in development. For additional sequences to those reported see (211).

The T-cell receptor for antigen is associated on the cell surface with T3, a T-cell surface antigen composed of three chains $\delta, \varepsilon, \gamma$ (212-217). The expressed V-regions of the T-cell receptor for antigen differ from the expressed immunoglobulin V_L and V_H chains by the larger number of J-minigene segments, the β chains having two sets of about six Js each associated in the germ-line via an intron with a C_β region termed $J_\beta 1$ 1-6 and $C_\beta 1$ and the second termed $J_\beta 2$ 1-6 and $C_\beta 2$. Each of the J_β sets has a D-minigene about 500-600 nucleotides 5′ of the first J segment termed $D_\beta 1$ and $D_\beta 2$ respectively (198). It is not established whether there are more D_β minigenes in the germ-line. The alpha chains of the murine and human T-cell receptors for antigen have many more Js, perhaps 20 in each species (218). These have not yet been localized in the genome. The T_γ chain has three V-genes and three different J-C genes separated by undetermined lengths of DNA; only one of these appears to be transcribed (219). Thus a greater proportion of the diversity of T-cell receptors for antigen may be created by V_T-J_T and V_T-D_T-J_T joining (204, 210).

A number of T-cell surface antigens have been cloned and sequenced in addition to T3. T4 (220) and T8 (221-223) are, with some exceptions, associated with T-helper and T-cytotoxic cells, T4 with class II and T8 with class I antigens of the major histocompatibility complex (MHC). These are included in the table on T-cell surface antigens. Their relation to suppressor T-cells is unclear.

A major question to be resolved at the structural level is how the T-cell receptors for antigen function at the cell surface in recognizing antigen only in association with class I and class II proteins of the major histocompatibility complex. Both the α and β chains are essential; loss variants of each chain were non-functional but fusion of an α loss variant and a β loss variant restored recognition of antigen plus MHC (213). MHC restriction has also been found in antibody recognition of influenza viral antigens on infected cells (214). Mutant clones of the T-cell receptor for antigen showed structural differences and had changed their MHC restriction from one haplotype to another (215). Thus far, the sequences of both the α and β and the class I and class II chains have given no insight into the nature of these interactions (218). Somatic mutation seems to be relatively infrequent in the α and β chains (218). For reviews of the differentiation markers of human T-lymphocytes, see references 224-226. V_T genes have been classified into subfamilies based on similarity in nucleotide sequence of 75% or more (224, 226) as was done in Brodeur and Riblet (100) for V_H genes. For subfamilies of the T-cell receptor for antigen, the first digit denotes the subfamily followed by a period, and the second digit indicates the individual member of that family; there are 14 $V_T\beta$ and 11 $V_T\alpha$ subfamilies (205, 206).

The $C_T\alpha$ and $C_T\beta$ regions are each encoded in the germ-line by four exons, the first corresponding to the C-regions of immunoglobulins; $C_T\alpha$ has a considerably smaller disulfide loop, 49 residues, than do all the other C-regions (224).

For mechanisms of V_T-J_T and V_T-D_T-J_T joining, see reference 224; evidence for rearrangement by deletion, homologous but unequal chromatid exchange, inversion, and reintegration of deleted sequences have been proposed. Rearrangement of T_β genes frequently occurs on both chromosomes but only one has the correct reading frame and is functional. $V_T\beta$ rearrangements occurred uniformly in mouse T-helper, T-cytotoxic, and T-cell tumors but apparently not in many T-suppressor cells (224).

N regions and junctional diversity are also found in assembled V_T-D_T-J_T chains. $D_T\beta 1$ and $D_T\beta 2$ minigenes may be used in all three reading frames. These are apparently used with equal frequency, unlike D_H-minigenes which are usually joined in the same reading frame. From the different degrees of utilization of the various mechanisms, estimates of the potential numbers of $V_T\beta$ chains may not be less than those of assembled V_H and V_κ chains (224). The variation would occur predominantly in CDR3 rather than more uniformly distributed among the three CDRs as in immunoglobulin chains.

Induction of antiidiotypic suppressor T cells has been produced by immunization with the light chain of MOPC315 ($\lambda 2$) or its V-region. Comparison of three V_λII chains showed four differences in MOPC315, three in CDR3 (Phe-Arg-Asn residues 92-94) and one in FR2 (Ile 36) which were not present in two other V_λI chains which did not induce this effect (227). Disruption of the disulfide bond of the domain destroys the induction of suppressor activity.

Sequences of related proteins

Among the related proteins, we have listed β_2-microglobulins, major histocompatibility complex antigens, I-region antigens, Thy-1, human complement components, C-reactive protein, thymopoietin, and post-gamma globulins. The major histocompatibility antigens are divided into five domains: N, C1, C2, membrane, and cytoplasmic, based on the location of the splice sites. Similarly, the I-region antigens consist of A-chains and B-chains and are divided into three domains each: A1, A2, and A-membrane (228) and B1, B2, and B-membrane. The complement and miscellaneous related proteins have at least one sequence that is too long for a page, and since no domains have been distinguished, they run sequentially 130 amino acid positions per page, retaining their individual protein numbers.

Table 3 lists the chromosomes on which the genes coding for immunoglobulins and various proteins have been found.

TABLE 3

Location of Immunoglobulin, MHC, T-cell receptor for antigen, β_2-Microglobulin, Complement and Interleukin Genes on various Chromosomes[a].

	Chromosome number		
	Human	Mouse	Rat
Light Chain			
κ	2p11(229)	6(230,231)[a]	
λ	22q11(234, 235)	16(236)	
Heavy Chain	14q32(237, 244)[b]	12(230)[c]	6(232)
J Chain	4q21(245)		
T-Cell Receptor for Antigen			
α chain	14q11(239, 241-243)[d]	14cd(233)	
β chain	7q35-36(240, 248, 251, 252)[e]	6b(246, 247)	
γ chain	7p15(249,250)	13a2-3(250)	
T-Cell Surface Antigens			
T3 δ chain	11q23-11qter(243, 252, 255)	9(252)	
T4	12pter(253, 254)		
T8/Leu-2	2(256)	(Lyt2,3) 6(229, 257-260)	
CTLA-1[f]		14(261)	
MHC	6(264)[g]	17(267, 268)	
Ia invariant chain		18(269)	
Natural Killer Cell Susceptibility	6(265)		
β_2-microglobulin	15(270)	2(271, 272)	
Thy-1	11q23-24(273)	9(274)	
Complement			
C1 inhibitor	11(275)		
C3	19(276)		
C6		15(277)	
Interleukin 2 Receptor	10p14-15(278)		
Immune Interferon Receptor	6q(279)		

[a] Frequent translocations to portions of chromosomes carrying immunoglobulin genes have been found (280) including chromosome 15 to 12, and a reciprocal translocation of chromosome 6 and 15 in mouse. Burkitt's lymphoma cells show exchanges between chromosomes 8 and 14, between 8 and 22, and 8 and 2. The cellular myc gene has been mapped to chromosome 8q24 in Burkitt's lymphoma cells and the myc gene may also occur as an oncogene (281). In mouse plasmacytoma c-myc is translocated to chromosome 12 from chromosome 15 (282). For the organization of human V_κ genes, see (283).

[b] A processed human epsilon heavy chain has been found on chromosome 9 (284).

[c] The order of the mouse heavy chain genes on chromosome 12 has been found to be 5′-J_H-6.5kb-C_μ-4.5kb-C_δ-55kb-C_γ3-34kb-C_γ1-21kb-C_γ2b-15kb-C_γ2a-14kb-C_ε-12kb-C_α-3′ (285).

[d] The breakpoint occurs between V_α and C_α leaving V_α on chromosome 14; C_α is translocated to chromosome 11 (11 P+). (238).

[e] Chromosomal rearrangements in T cells map close to this region in normals and patients with ataxia telangectasia (240, 262); some related sequences were also found on the short arm of chromosome 7p15-21 (262). An 8.8 Kb deletion in NZB mice of a T_β chain involved the loss of $C_T\beta1$, $D_T\beta2$ and $J_T\beta2$; the strain had functional T-cells (263).

[f] Cytotoxic T-lymphocyte associated.

[g] For a map of the short arm of chromosome 6, see (266).

Nucleotide Sequences

The codons for signal, variable, and constant regions of immunoglobulins and related proteins are listed in the same format as that for amino acid sequences, so that a side-by-side comparison between amino acid sequences and nucleotide sequences is feasible.

D-minigene sequences

The tables have been designed to facilitate rapid visual comparisions. Since the D-minigenes are located by recognition sequences surrounding the presumed coding regions (126, 128, 129), the table is arranged as follows: the 5′ nonanucleotide signal sequence is given on one line, each nucleotide of the twelve nucleotide spacer is listed on a separate line, and the heptanucleotide signal sequence is given on one line. Each nucleotide of the presumed coding sequence is given on a different line. Since the coding frame is not known, all three possible reading frames are translated into amino acid sequences which are listed next to the presumed coding region. This is followed by the 3′ heptanucleotide signal sequence on one line, the 11 or 12 nucleotide spacer with each nucleotide on a separate line, and the nonanucleotide signal sequence on one line. The complementary strands are also listed in the same format and are denoted by "-C". There are horizontal lines above and below the signal nona-and heptanucleotides.

J-minigene sequences

Unlike the D-minigenes, the reading frame for the J-minigenes is known. Thus, the coding regions are listed as codons together with the translated amino acid sequences. The non-coding regions 3′ to the coding regions are listed as segments of ten nucleotides. At the 5′-end, the signal regions are given as a row of nine nucleotides, followed by 21 to 24 single nucleotides, then by a row of seven nucleotides; 5′ to that region, the non-coding segments are listed as rows of ten nucleotides. There are horizontal lines above and below the signal nona-and heptanucleotides.

The intervening regions between the kappa light chain J-minigenes have been completely sequenced. Therefore, the first column can be joined to the second column, etc., giving the entire J-region nucleotide sequences. The lambda light chain J-minigenes are separated by constant region genes, so they are not contiguous (120,121). The heavy chain J-minigenes are similar to those of the kappa light chains.

Pseudogenes

Every pseudogene's possible coding region is listed in four different columns. The first column contains the codons as read off with zero frame shift. The other three columns are the translated amino acid sequences with zero, one, and two nucleotide frame shifts, respectively. Possible useful amino acid sequence segments are listed in upper case. All other amino acid residues are in lower case. Termination codons are indicated as trm. A note gives the reason for classifying it as a pseudogene.

Variability Plots

A set of variability plots is given for those tables for which sufficient data are available; some graphs are combinations of several tables or are portions of a given table as indicated. Horizontal bars signify variabilities when values due to Glx (or Asx) were computed as only one of the two possible choices. The variability plots have ordinates with the ranges largely determined by the numbers of available sequences but with the same ordinate scale so that comparisons can be made among different graphs.

Although the variability plots are based on the sequences in the current tables, the three complementarity-determining segments (CDR) can usually be seen quite clearly despite the deviations of the collection of sequences from the probably more random data in the earlier editions. The current data include many sequences on selected populations, e.g., those with a particular antibody specificity, those in a given group or subgroup, those with unblocked N-termini, etc. The division of mouse V_κ chains into subgroups based on length to the first Trp and the division of the V_H regions based on the germ-line families of Dildrop (98) and Brodeur and Riblet (100) have made it possible to evaluate variability in each of these categories and variability plots are given for those groups for which there are reasonable numbers of sequences. The three CDRs as originally defined can be seen in variability plots of light

and heavy chains, kappa light chains, human light chains, human heavy chains, human kappa light chains, human kappa light chains subgroup I and human heavy chains subgroup III, since the collection of human light and heavy chains remains largely random. Human lambda light chains subgroups I, II and III all show three hypervariable regions although the values of variability are smaller. Rabbit kappa light chains are essentially limited to anticarbohydrate antibodies of a few specificities. They show no hypervariability in second CDR except perhaps at position 50, and show high variabilities at position 2, and some at 1, 22 and 70. The CDR2 peak is also not significant in human kappa light chains subgroups II and III.

Variability plots for the subgroups of mouse kappa light chains based on length to invariant Trp 35 show less well defined CDR peaks, perhaps due to the non-random collections of sequence data. Similar features are observed for the mouse heavy chains classified both by subgroups and the Dildrop (98) and Brodeur and Riblet (100) germ-line families. These mouse subgroups generally have a large number of antibodies of a limited number of specificities and thus variability within each subgroup in the CDRs may predominantly be related to differences in specificity. Certain mouse heavy chain subgroups may not exhibit variability in all three CDRs. For example, mouse heavy chains subgroup IIB show considerable variability in CDRs 2 and 3, and subgroups IIC and IIIB show hypervariability only in CDR3. As more data are accumulated it may be possible to ascribe fine specificity features of different antibodies as being primarily influenced by one or another CDR in each chain.

Several of the graphs show moderate variability in other portions of the V-regions or occasionally at individual residues. While the lengths and locations of the segments showing hypervariability differ for the various graphs, this may be largely due to the selection of data and various ways of combining them. The best values for the lengths of CDR are those based on light chains and heavy chains as originally computed (11,22).

The uniqueness of variability plots of immunoglobulin chains may be seen by comparison with a similar plot for 67 cytochromes c from sequences provided by Dr. E. Margoliash of Northwestern University that does not resemble the variability plots for light or for heavy chains.

In the light chains, the first and third CDR begin after Cys 23 and Cys 88 and are terminated by invariant Trp and Phe at residues 35 and 98, respectively. The second CDR is followed at position 57 by 417 Gly of 427 chains sequenced.

In the heavy chains, the first and third CDR are located several residues from Cys 22 and Cys 92 toward the C-terminus. Position 25 shows low variability with 488 Ser, 39 Thr, 2 Ala, 2 Phe, 1 Pro, 1 Lys, 1 Gly, and 1 Asn in 535 sequences; Gly 26 is essentially invariant, occurring 527 times in 535 sequences with 1 Val, 1 Ala, 1 Leu, 1 Asp, and 4 Glu, and the first complementarity-determining segment is followed by an invariant Trp 36, as in the light chain. The second CDR is not immediately preceded by invariant residues; position 66 has a variability of 6.9. CDR3 is not preceded by an invariant residue, but is followed by Trp 103 which occurred in 302 of 310 sequences.

Positions 99 and 101 of the light chain are nearly invariant glycines in all species examined; positions 104 and 106 in the heavy chains are also invariant Gly, (286) except for a human myeloma protein EU in which two adjacent Gly residues occur at positions 106 and 107; an incompletely sequenced cryoglobulin PAV (287) in which Ser and Ala have been reported at 104 and 106. It was proposed that these glycines (19,11,288) serve as a pivot permitting adjustment of the CDR of the light and heavy chains to make better contact with the antigenic determinant. (For discussions see references 59 and 67).

In two of the myeloma proteins studied by X-ray crystallography, McPC603 (47) and Newm (44,49,59), the second complementarity-determining segment of the light chain does not participate in the site. In the former it is shielded from the site by an insertion of six residues in the first complementarity-determining segment, while in the latter a deletion of seven residues which follow the second CDR removes it from the site. These additional parameters increase further the way in which diversity of sites is generated (175). In the anti-lysozyme site, all six CDRs make contact with the lysozyme (Fig. 5, Table 1) (84). In mouse V_λ chains position 48 is not invariant and since there is an insertion of three residues in CDR1 of all mouse V_λ chains, it has been suggested (67) that CDR2 of V_λ might extend from 48-56.

Variability plots of the T-cell antigen receptor α chains and β chains as well as β chain subgroups I and II (210) are given. The variability plot has been used to locate hypervariable regions of the envelope proteins of various strains of AIDS virus (289).

Composite tables of codon usage for the variable regions of the light and heavy chains are also given.

The major histocompatibility complex in the mouse and in humans is characterized by extensive allelic polymorphisms (290-295). The mouse H-2K^b class I mutants have been localized to the N and C1 domains of the sequence involving amino acid substitutions at residues 77, 89, 116, 121, 155, 156 and 165 (290). In class II the A_α, A_β and E_β chains show allelic hypervariability of the amino acids in the first domain involving residues 10-15, 44-49, 53-59 and 69-77 for A_α; 9-17, 63-68 and 84-89 for A_β; and 1-13, 27-39, 68-75 and 87-93 for E_β derived (291, 292, 294) using the variability equation (11,22). Similar allelic variability is seen in a plot of the human HLA-DRβ locus (294). All of the hypervariability occurs in the first domain with the second domain showing minimal variability; for a different type of plot (*cf.* 295). Based on the sequences accumulated, separate allelic variability plots are given for human DR_β and DC_β and mouse IA_α and IE_β Class II MHC antigens.

Germline genes have been reported which might prove to be non-functional. One V_H from a hybridoma making anti-NP (4-hydroxy-3-nitrophenylacetyl) antibodies had Ser replacing Cys at position 22, so that the disulfide loop would be absent; a second genomic clone had one base deleted from the codon for Cys 22 and thus might also be expected to be non-functional (147). Similarly, a cDNA clone ABE48 from antilevan myeloma, ABPC48, was found to have a codon for Tyr at position 92 instead of Cys (296). Rudikoff and Pumphrey (297) demonstrated that the ABPC48 plasmacytoma made a protein which bound and precipitated bacterial and rye grass levans. Amino acid sequencing showed that the ABPC48 protein actually contained Tyr in place of the second half cysteine in V_H. Thus the presence of a disulfide bridge in the V_H-domain is not obligatory for a functional antibody and indicates that the V_H-domain may be held in a conformation suitable for reactivity with antigen despite its inability to form a disulfide bond. Another cDNA clone from antilevan myeloma protein, UPC10, had a codon coding for Val at position 106 instead of Gly (296). Somatic mutants arising in culture of phosphorylcholine binding myeloma S107 which showed evidence of altered binding have been examined (298, 299). One, U4, had a replacement in CDR1 of Glu 35, which has been shown to be a contacting residue, by Ala; this change would eliminate the hydrogen bond between Glu 35 in V_H and Tyr 94 in V_L and would materially change the structure of the site (63); this mutant no longer bound phosphorylcholine but had acquired the capacity to bind double-stranded DNA, phosphorylated protamine and cardiolipin (300). A second had a reduced capacity to hemagglutinate phosphorylcholine coupled to sheep erythrocytes and to bind phosphorylcholine coupled to KLH (keyhole limpet hemocyanin), but its Ka with phosphorylcholine does not differ from the parent S107; the only change in the V_H sequence was the replacement of Asp 101 by Ala. A more detailed analysis of the effects of various amino acid substitutions on specificity is given in reference 165.

Acknowledgements

The authors are indebted to Drs. C. Croce, T. Kindt, E. Long, R. Mage, C. Milstein, M. Potter, W.F. Raub and L. Steiner for helpful suggestions, and to Dr. Eduardo A. Padlan for making the stereomodels of the V_L dimers and the Fabs, and to Dr. Roberto Poljak for making Table 1 and Figure 5 available before publication.