Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 10884–10888, November 1995
Immunology

Analysis of the relation between the sequence and secondary and three-dimensional structures of immunoglobulin molecules

(immunoglobulin sequence comparison/pattern recognition/statistical significance)

ISRAEL M. GELFAND AND ALEXANDER E. KISTER

Department of Mathematics, Rutgers University, New Brunswick, NJ 08903

Contributed by Israel M. Gelfand, August 3, 1995

ABSTRACT **Methods of structural and statistical analysis of the relation between the sequence and secondary and three-dimensional structures are developed. About 5000 secondary structures of immunoglobulin molecules from the Kabat data base were predicted. Two statistical analyses of amino acids reveal 47 universal positions in strands and loops. Eight universally conservative positions out of the 47 are singled out because they contain the same amino acid in >90% of all chains. The remaining 39 positions, which we term universally alternative positions, were divided into five groups: hydrophobic, charged and polar, aromatic, hydrophilic, and Gly-Ala, corresponding to the residues that occupied them in almost all chains. The analysis of residue–residue contacts shows that the 47 universal positions can be distinguished by the number and types of contacts. The calculations of contact maps in the 29 antibody structures revealed that residues in 24 of these 47 positions have contacts only with residues of antiparallel β-strands in the same β-sheet and residues in the remaining 23 positions always have far-away contacts with residues from other β-sheets as well. In addition, residues in 6 of the 47 universal positions are also involved in interactions with residues of the other variable or constant domains.**

The large amount of sequence and structural data for immunoglobulins makes this superfamily one of the best objects for the analysis of the relation between the sequence and secondary and three-dimensional (3D) structures. Based on the unique collection of immunoglobulin sequences (Kabat data base), Kabat suggested that conservative residues define conservative conformations of the variable domains (1, 2). Padlan (3, 4) and Chothia and colleagues (5–8) examined the immunoglobulin sequences and structures and found that conservative residues have the same structural role in the packing of the β-sheet framework.

In this work, we predict the secondary structures for about 5000 sequences from GenBank (Kabat data base) and assign each residue in a sequence to the position of the strand or loop. This allowed us to perform a statistical analysis and to detect the universal positions whose residues are identical or share a common feature (for example, hydrophobic) in almost all of the chains. To determine the structural characteristics of these positions, 29 structures of antibody molecules were compared. To compare the molecules, we introduced an invariant system of coordinates and calculated contacts between the residues in the universal positions.

The secondary structure prediction of immunoglobulin chains is based on our analysis of the 3D structures of light chain variable region and heavy chain variable region (V_l and V_h, respectively) domains. Beginning with the 29 immunoglobulin x-ray structures, from Brookhaven data base, we calculated a secondary structure for each of them. We divided each of these sequences into 21 fragments, which we term words. Each word corresponds to a strand or a loop. The multialignment of these words permits us to construct patterns for each word. By using these patterns, we then predicted the secondary structures for the approximately 5000 immunoglobulins.

Statistical analysis, performed by two independent methods, allows us to determine 47 positions. Eight of the 47 positions are singled out because, in more than 90% of all immunoglobulin chains, they contain the same amino acids. In these positions, which we term universally conservative positions, there are Cys residues in the B and the F strands, Trp residues in the C strands, Asp residues in the EF loop, Tyr residues in the F strand, and two Gly residues and one Thr residue in the G strand. The remaining 39 positions are divided into five groups: hydrophobic, charged and polar, aromatic, hydrophilic, and Gly-Ala, corresponding to the residues that occupy them. In more than 90% of all chains, the residues in these universally alternative positions share common characteristics.

The 47 universal positions play an essential role in our secondary structure prediction algorithm. An additional benefit is that they also allow us to determine whether a given sequence belongs to the immunoglobulin family. Analysis of 1500 mouse κ chains shows that no fewer than 44 of the 47 universal positions are always occupied by the amino acid predicted for these positions. This leads us to believe that the classification and selection of positions that we propose is a reliable guide for secondary structure prediction of immunoglobulins.

For different molecules, conservative positions were selected separately. For example, in the mouse κ chains, aside from the 8 universal conservative positions, 26 conservative positions were selected. Among the unexpected results, a Pro-occupied conservative position is found in the antigen-binding region in almost all chains. Another observation shows that placement of a particular residue in certain positions can be critical for determining a sequence. For example, the first residue of a chain can, in certain cases, determine membership in a family in the sense of the Kabat data base.

The 47 universal positions are of considerable interest with regard to secondary structure definition and understanding the spatial structure. To determine the structural role of the amino acids in these positions, we examined 29 x-ray structures and calculated their residue–residue contacts (number of residues with which a given residue shares contacts). The results show that residues in 28–30 positions have more than the average number of contacts, which we calculated to be between 6 and 7. It should be noted that 22 of these positions are among the 47 universal positions. Further, all these positions, which we term high-contact positions, were found to be

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

Abbreviations: V_l, light chain variable region; V_h, heavy chain variable region; 3D, three dimensional.

*The structures are identified by Protein Data Bank names: 1BAF, 1BBD, 1BBJ, 1BJL, 1CBV, 1DBA, 1F19, 1FDL, 1GGC, 1GGI, 1HIL, 1IGF, 1IGI, 1MAM, 1MCO, 1MCP, 1NBV, 1TET, 2DBL, 2FB4, 2FBJ, 2HFL, 2IG2, 3FAB, 4FAB, 6FAB, 7FAB, 8FAB, and 3HFM.

the same in all of the analyzed molecules. Furthermore, the analysis of residue–residue contacts demonstrates that the 47 universal positions can be divided into two groups: (*i*) the 24 positions whose residues have contacts only with residues in antiparallel β-strands (these contacts define mutual orientation of the strands in a β-sheet) and (*ii*) the 23 positions in which residues have far-away contacts with residues of another β-sheet as well as contacts with the antiparallel β-strands. We suggest that residues in these positions are largely responsible for the interactions between the two β-sheets. In addition, residues in the 4 out of the 47 positions are involved in the V_l and V_h domain interactions and residues from 2 other positions form contacts with residues in a constant domain.

The calculation of the contact map for the residues from the universal positions resulted in almost identical maps for all molecules examined. Residues in certain positions were uniformly found to have contacts with each other. It appears that these "conservative" contacts are, to a large extent, responsible for the β-sheet framework.

Determination of Strands and Loops in 3D Structures of Immunoglobulins

Materials and Methods. Our calculations were made by using the 29 x-ray structures of Fab molecules from the Brookhaven data base.* Usually, the secondary structure of proteins with known 3D structure is assigned by using the program DSSP (9). However, these calculations may not be sufficient to distinguish a strand and a loop exactly. To make more accurate calculations, we applied visual analysis of the structures on the screen, calculation of backbone dihedral angles, calculation of torsion angles around virtual C_α–C_α bonds, and analysis of H-bonds between strands in β-sheets.

Results. The large number of calculations we performed leads to a reliable definition for secondary structures. Ten strands were determined in every molecule: A, A′, B, C, C′, C″, D, E, F, and G. The first 3 residues of a chain do not belong to the A strand; these are designated the 0A fragment. The identification of loops is derived from two strands that are connected by that loop: thus, loops are termed AA′, A′B, CC′, C′C″, C″D, DE, EF, and FG. The connection between B and C strands, which has a unique M-like conformation with 1 residue deeply inserted into the structure (5), is subdivided into two loops, BC and CB.

To compare secondary structures of different molecules, we define the term word to denote a fragment of a sequence that characterizes a strand or a loop. For example, in 1CBV, a C word describes the C strand consisting of Leu, His, Trp, Tyr, Leu, and Gln. Every residue in a word is designated by an index. For example, Trp, which is the third residue in the C strand, has index C3. These indices define the positions of a residue in a secondary structure (6, 10).

In the 29 examined x-ray structures, the sequences of the V_l and V_h domains divide into 21 words: 0A, A, AA′, A′, A′B, B, BC, CB, C, CC′, C′, C′C″, C″, C″D, D, DE, E, EF, F, FG, and G words, in accordance with our secondary structure definition.

Multialignment of Words: Patterns of Immunoglobulin Secondary Structure

Materials and Methods. The procedure we used incorporates the following steps:

The sequence multialignment of words. Each set of identical words that describe the same strands or loops of 29 molecules (set of 29 0A words and so on) was considered separately. We accepted that identical words can be of different length; i.e., they can have various numbers of positions. No gaps inside a word are allowed.

Structural superposition multialignment. Strands and loops from different molecules were superimposed on a screen and compared visually to determine residues in the same positions.

Residue–residue contact multialignment. All atom–atom distances between residues were calculated for every molecule. (We assumed that 2 residues are in contact when any two heavy atoms are closer than 5.0 A.) On the assumption that residues in the same positions have approximately the same lists of contacts, we verified the multialignment of words by comparing the lists of contacts.

C_α coordinate multialignment. For comparison of the spatial location of residues from different molecules, we instituted an invariant system of coordinates. The examination of residue

Table 1. The secondary structural multialignment of 0A, A, AA′, A′, A′B, B, BC, CB, C, CC′, and C′ words of V_l domains from 29 x-ray structures.

0A	A	AA'	A'	A'B	B	BC	CB	C	CC'	C'	
QIV	LTQ	SPAI	MSAS	PGE	KVTMTCSA	SSS	VYY	MYWYQQ	KPGSSP	RLLIY	1BAF
DIV	MTQ	SPSS	LTVT	TGE	KVTMTCKS	SQS	LLNSRTQKNY	LTWYQQ	KPGQSP	KLLIY	1BBD
DIQ	MTQ	SPAS	LSVS	VGE	TVTITCRA	SEN	IYSN	LAWYQQ	KQGKSP	QLLVY	1BBJ
PSV	LTQ	PPS	ASGT	PGQ	RVTISCSG	SSSN	IGENS	VSWYQH	LPGTAP	KLLIY	1BJL
DVV	MTQ	TPLS	LPVS	LGD	QASISCRS	SQS	LVHSNGNTY	LHWYLQ	KPGQSP	KLLIY	1CBV
DVV	MTQ	IPLS	LPVN	LGD	QASISCRS	SQS	LIHSNGNTY	LHWYLQ	KPGQSP	KLLMY	1DBA
DIQ	MTQ	TTSS	LSAS	LGD	RVTISCRA	SQD	ISNY	LNWYQQ	KPDGTV	KLLVY	1F19
DIQ	MTQ	SPAS	LSAS	VGE	TVTITCRA	SGN	IHNY	LAWYQQ	KQGKSP	QLLVY	1FDL
DIV	LTQ	SPGS	LAVS	LGQ	RATISCRA	SES	VDDDGNSF	LHWYQQ	KPGQPP	KLLIY	1GGC
DIV	LTQ	SPGS	LAVS	LGQ	RATISCRA	SES	VDDDGNSF	LHWYQQ	KPGQPP	KLLIY	1GGI
DIV	MTQ	SPSS	LTVT	AGE	KVTMSCTS	SQS	LFNSGKQKNY	LTWYQQ	KPGQPP	KVLIY	1HIL
DVL	MTQ	TPLS	LPVS	LGD	QASISCRS	NQT	ILLSDGDTY	LEWYLQ	KPGQSP	KLLIY	1IGF
DVV	MTQ	TPLS	LPVS	LGD	QASISCRS	SQS	LVHSNGNTY	LNWYLQ	KAGQSP	KLLIY	1IGI
DIQ	MTQ	TTSS	LSAS	LGD	RVTISCRA	SQD	IYNY	LNWYQQ	KPDGTV	KLLIY	1MAM
PSA	LTQ	PPS	ASGS	LGQ	SVTISCTG	TSSD	VGGYNY	VSWYQQ	HAGKAP	KVIIY	1MCO
DIV	MTQ	SPSS	LSVS	AGE	RVTMSCKS	SQS	LLNSGNQKNFL	LAWYQQ	KPGQPP	KLLIY	1MCP
DVV	MTQ	TPLS	LPVS	LGD	QASISCRS	SQS	LVHSNGNTY	LHWYLQ	KPGQSP	KLLIY	1NBV
DVL	MTQ	TPLS	LPVS	LGD	QASISCKS	SQS	IVHSSGNTY	FEWYLQ	KPGQSP	KLLIY	1TET
DVV	MTQ	IPLS	LPVN	LGD	QASISCRS	SQS	LIHSNGNTY	LHWYLQ	KPGQSP	KLLMY	2DBL
QSV	LTQ	PPS	ASGT	PGQ	RVTISCSS	TSSN	IGSST	VNWYQQ	LPGMAP	KLLIY	2FB4
EIV	LTQ	SPAI	TAAS	LGQ	KVTITCSA	SSS	VSS	LHWYQQ	KSGTSP	KPWIY	2FBJ
DIV	LTQ	SPAI	MSAS	PGE	KVTMTCSA	SSS	VNY	MYWYQQ	KSGTSP	KRWIY	2HFL
QSV	LTQ	PPS	ASGT	PGQ	RVTISCSG	TSSN	IGSST	VNWYQQ	LPGMAP	KLLIY	2IG2
?SV	LTQ	PPS	VSGA	PGQ	RVTISCTG	SSSN	IGAGNH	VKWYQQ	LPGTAP	KLLIF	3FAB
DVV	MTQ	TPLS	LPVS	LGD	QASISCRS	SQS	LVHSQGNTY	LRWYLQ	KPGQSP	KVLIY	4FAB
DIQ	MTQ	IPSS	LSAS	LGD	RVSISCRA	SQD	INNF	LNWYQQ	KPDGTI	KLLIY	6FAB
ASV	LTQ	PPS	VSGA	PGQ	RVTISCTG	SSSN	IGAGHN	VKWYQQ	LPGTAP	KLLIF	7FAB
E	LTQ	PPS	VSVS	PGQ	TARITCSA	NA	LPNQY	AYWYQQ	KPGRAP	VMVIY	8FAB
DIV	LTQ	SPAT	LSVT	PGN	SVSLSCRA	SQS	IGNN	LHWYQQ	KSHESP	RLLIK	3HFM

coordinates in the invariant system is an analytical analog of the visual superposition of different molecules. Since the residues in the same positions can be different, we compared only the C_α coordinates of the residues. (A detailed description of the invariant system of coordinates will be published elsewhere.)

H-bond multialignment. The residues (in different molecules) whose main chain atoms are involved in H-bonds between the strands were compared. It is suggested that H-bonds are formed between residues from the same positions.

Values of accessibility multialignment. For every residue, the extent of residue exposure in a protein was calculated by applying the method for assessing the relative degree of solvent exposure of residues (11). By assuming that residues in the same position in different molecules have approximately equal values of accessibility, we verified a multialignment of the words.

Results. The result of secondary structural multialignment for 0A, A, AA′, A′, A′B, B, BC, CB, C, CC′, and C′ words are presented in Table 1. After performing all of the steps of the multialignment procedure, we were able to construct patterns for every word in the light chains and in the heavy chains. For example, the pattern for the C word of light chains is C1 [Met, Leu, Val, Phe, or Ala], C2 [Tyr, Thr, Ala, Ser, His, Asn, Thr, Glu, Lys, or Arg], C3 [Trp], C4 [Tyr], C5 [Gln or Leu], and C6 [Gln or His]. In the C1 position, only hydrophobic residues were found. The residues in the C2 positions are very different, but in the C3, C4, and C6 (with one exception), positions with a single residue were found. Only Gln or Leu can occupy the C5 position. An analysis of the CB words shows that the residues in the CB1 position are always hydrophobic, but the residues in the remaining positions are very different.

For V_l sequences, we obtained a secondary structure consensus by assembling 21 patterns of words. The same procedure, with (a different set of) 21 patterns of words for the heavy chains, gave us a secondary structure consensus for V_h sequences. We use these to define strands, loops, or their fragments in antibody molecules.

The Secondary Structure Prediction for V_l and V_h Sequences

Materials and Methods. Various approaches have been used for secondary structure prediction (12–18). One of the most promising is the method of prediction by analogy with homologous proteins of known 3D structures (19, 20). In this paper, we predict immunoglobulin secondary structures based on our analysis of the 3D structures of V_l and V_h domains from the Brookhaven Protein Data Bank (21). About 5000 sequences of the V_l and V_h domains were compiled from the Kabat data base. To divide a sequence into words, the secondary structure consensus was aligned with every sequence in question. Starting with the 0A word, all words in the sequence are defined in consecutive order. Each residue determines a definite position in a word.

Results. Secondary structures were predicted for the following cases: for human, mouse, and rat, κ, λ, and heavy chains; for rabbit, κ and λ chains; for chicken, λ chains; and for shark, heavy chains.

Statistical Analysis of Positions in Amino Acid Sequences

Materials and Methods. Knowledge of the secondary structure of about 5000 immunoglobulin sequences makes it possible to carry out statistical analysis of residues in various positions. At first, different sequences (differing in at least one residue) were selected. The amino acid distributions in each position for human κ, human λ, and the other chains were calculated. Our calculations show that strands and loops can be conventionally divided into two groups—those with many and those with relatively few amino acid substitutions. For example, there are a vast number of sequences that differ only by substitutions in the antigen binding regions. These substitutions can exert a substantial influence on the results of a statistical analysis of residues in "nonvariable" strands and loops. Therefore, for further statistical analysis, we chose sequences that differ in at least one of the nonvariable A, B, C, C′, D, E, EF, F, and G words. This approach has the advantage that the statistical analysis for the positions of these words does not depend on the presence of a large number of sequences with amino acid substitutions outside of the strands and the loops under consideration. For example, from 1426 mouse κ chains, 515 sequences with different A, B, C, C′, D, E, EF, F, and G words were selected.

Results. A comparison of residues occupying the same positions in the 5000 immunoglobulin sequences can furnish a large number of coincidences. The term coincidence can be thought of in many different ways: residues are identical or have similar properties (for example, all residues are aromatic or have approximately the same coordinates for main chain atoms, or the list of residue–residue contacts are similar, etc.). The number of coincidences is a criterion of how accurately the positions were assigned. In this paper, we point out two types of coincidences: the comparison of the residues and the residue–residue contacts.

Comparison of the residues: Conservative and alternative positions. Statistical analysis of the residues reveals that certain positions are occupied by the same residues in more than 90% of human κ, human λ, or the other chains. Conservative positions and residues in these positions may vary in different molecules. For example, 0A2 is the conservative position for human κ chains but not for mouse κ chains. However, there are eight positions—B6, C3, EF6, F3, F5, G3, G5, and G6—that are found as conservative positions in almost all molecules (Table 2).

The analysis of sequences allows us to single out 39 positions also because in more than 90% of different chains, residues with similar properties are observed, for example, hydrophobicity. These universally alternative positions were classified and divided into five groups: (*i*) 13 hydrophobic positions (Ile, Val, Leu, Phe, Ala, Met, and Trp); (*ii*) four charged and polar positions (Glu, Asp, and Gln); (*iii*) two aromatic positions (Phe

Table 2. Conservative and alternative positions of immunoglobulin chains

Group	Position(s)
Conservative	B6 (Cys^t, +), C3 (Trp^t, +), EF6 (Asp^t, +), F3 (Tyr^t, +), F5 (Cys^t, +), G3 (Gly^t, −), G5 (Gly^t, −), G6 (Thr^t, +)
Hydrophobic	A1 (+), B2 (+), B4 (+), CB1 (+), C1 (+), C′3 (+), C″D3 (−), D2(F^l, +), E4(L^l, +), E6(I^l, +), EF2 (+), G8 (+), G10 (+)
Aromatic	F4 (+), G2 (Phe^l, −)
Gly-Ala	F1 (Ala^h, +)
Hydrophilic	AA′1 (−), A′B2 (Gly^t, −), B1 (−), B3 (−), B5 (−), BC1 (Ser^k, −), BC2 (−), CC′3 (−), CC′4 (−), C′C″3 (Ser^m, −), C″D2 (Gly^k, −), C″D4 (Gly^m, −), D5 (Ser^k, −), DE1 (Gly^k, −), E7 (−), EF1 (−), EF3 (−), G9 (−)
Charged and polar	0A1 (−), A3(Gln^l, +), C6(Gln^k, +) D1(Arg^l, +)

The universally conservative and alternative positions for all immunoglobulin chains are shown. Superscripts: t, residues in the universally conservative positions; l, residues in the conservative positions for the light chains; k, residues in the conservative positions for the κ chains; m, residues in the conservative positions for the mouse κ chains; h, residues in the conservative positions for the heavy chains. Residues in the position indicated have more than the average number of contacts (+) or less than the average number of contacts (−).

Table 3. Conservative positions of mouse κ chains

	MFAA		Words, no.			MFAA		Words, no.			MFAA		Words, no.	
P	Residue	%	a	b	P	Residue	%	a	b	P	Residue	%	a	b
A2	Thr	93	25	10	D2	Phe	99	49	44	F5	Cys	99	122	119
A3	Gln	98	25	16	D4	Gly	98	49	42	FG5	Pro	94	493	429
A'B2	Gly	98	46	35	D5	Ser	97	49	35	G1	Thr	98	49	39
B6	Cys	99	151	147	D6	Gly	98	49	37	G2	Phe	99	49	44
BC1	Ser	97	65	45	D7	Ser	98	49	34	G3	Gly	99	49	46
C3	Trp	99	116	108	DE1	Gly	94	22	11	G5	Gly	99	49	44
C6	Gln	94	116	95	DE2	Thr	90	22	7	G6	Thr	99	49	43
C5	Ile	93	88	68	E4	Leu	95	107	91	G7	Lys	97	49	37
C'C"3	Ser	92	79	47	E6	Ile	98	107	97	G8	Leu	98	49	44
C"D2	Gly	99	61	56	EF5	Glu	93	145	119	G9	Glu	99	49	41
C"D4	Pro	97	61	48	EF6	Asp	98	145	134					
D1	Arg	99	49	41	F3	Tyr	99	122	114					

P, position. The percentage of sequences in which the most frequently occurring amino acid (MFAA) is encountered in a given position are shown. For words, the number of different words that describe the strand, loop, or a part of a loop with given position (columns a) and the number of words with this amino acid in the given position (columns b). For example, in the A2 position, Thr is encountered in 93% of all sequences; furthermore, Thr is found in 10 of the 25 words that describe the A strand.

and Trp); (*iv*) 19 hydrophilic positions with different nonpolar, polar, and charged residues, but excluding the residues from the hydrophobic group; (*v*) one Ala-Gly position (Table 2). In total, these 47 universally conservative and alternative positions that were found in all chains are a good test for distinguishing an immunoglobulin sequence. Note that analysis of 1500 mouse κ chains shows that there are no more than two or three exceptions per sequence to the 47 universal positions.

Comparison of residue–residue contacts in mouse κ molecules. The analysis of the 3D structures of 29 molecules shows that while the number of residue–residue contacts (from 0 to 16 contacts) varies greatly, on average a residue has contact with 6 or 7 other residues. We established that for all examined molecules, residues in 28–30 positions have more than the average number of residue–residue contacts. Our observations show that 22 of these high-contact positions are always the same in all molecules. Moreover, these 22 positions are in the universal conservative group, and in the hydrophobic, aromatic, Gly-Ala, and charged and polar groups of universal alternative positions but not in the alternative hydrophilic group (see Table 2).

Further, the calculations of the contacts for residues from the universal positions gave almost identical maps for different molecules (unpublished results). Evidently, residues in certain positions always have contacts with each other. These contacts are termed conservative contacts. For example, Tyr in F3, which has the greatest number of these contacts, interacts with residues in 11 universal positions, A3, B4, C3, C6, C'3, D2, E4, E6, EF6, G6, and G8.

The analysis of residue–residue contacts shows that the 47 universal positions can be distinguished by the types of contacts. It was found that residues in the 23 positions (A1, A3, B4, B5, B6, CB1, C1, C3, C'C"3, D1, D2, D5, DE1, E4, E6, EF2, EF6, F3, F4, F5, G2, G6, and G8) have contacts not only with the residues of the antiparallel strands but also with far-away residues of another β-sheet. The ability to form such interactions is characteristic for these positions in the light and the heavy chains. These conservative contacts permeate the entire structure and, possibly, determine the immunoglobulin-like folding to a large extent. Residues of the remaining 24 positions have contacts only with residues of antiparallel β-strands. Apparently, these contacts are responsible for the conformations of β-sheets.

Except for intradomain conservative contacts, involvement of residues from universal positions in V_l–V_h interactions was also observed (17). The residues from the C6, CC'4, F4, and G2 positions in both V_l and V_h domains form the conservative contacts between the light and heavy chains. Furthermore, residues from the two last positions of the G strand—G9 and G10—were observed to share contacts with residues of a constant domain.

Statistical Analysis of Positions in Words

Method. We present here another approach to the statistical analysis of positions. As an example, we conducted an analysis

Table 4. Analysis of conservative positions of mouse κ chains: The frequency of amino acids in the frequent and rare words

	FW			RW			FW			RW			FW			RW	
P	a	b	c	a	c	P	a	b	c	a	c	P	a	b	c	a	c
A2	5	Thr	4	12	4	D2	4	Phe	4	25	22	F5	25	Cys	25	52	51
A3	5	Gln	5	12	6	D4	25	Gly	22	49	42	FG5	11	Pro	10	309	256
A'B2	14	Gly	13	20	12	D5	4	Ser	4	25	15	G1	5	Thr	5	19	12
B6	20	Cys	20	82	80	D6	4	Gly	3	25	18	G2	5	Phe	5	19	15
BC1	8	Ser	8	25	15	D7	4	Ser	4	25	16	G3	5	Gly	5	19	18
C3	22	Trp	22	59	52	DE1	5	Gly	4	6	2	G5	5	Gly	5	19	16
C6	22	Gln	20	59	46	DE2	5	Thr	2	6	2	G6	5	Thr	5	19	14
C'5	13	Ile	12	36	125	E4	13	Leu	11	53	43	G7	5	Lys	5	19	12
C'C"3	22	Ser	19	79	47	E6	13	Ile	6	53	47	G8	5	Leu	5	19	16
C"D2	12	Gly	12	28	23	EF5	22	Glu	20	80	64	G9	5	Glu	5	19	14
C"D4	12	Pro	11	28	22	EF6	22	Asp	22	80	70						
D1	4	Arg	4	25	19	F3	25	Tyr	25	52	48						

P, position; FW, frequent words encountered in ≥1% of all sequences. Columns: a, the number of frequent words that describe a strand, loop, or part of a loop containing the specified position; b, most frequent residue at this location; c, number of frequent words with this amino acid in the given position. RW, the same information (columns a and c) for the words that are found only once (rare words).

of E words of mouse κ chains. For 1426 sequences, 107 E words were constructed. The frequency with which these words were encountered in the sequences varies; for instance, one E word occurred in 349 sequences whereas another was found only once. However, in this approach the number of times a given word is encountered is not taken into consideration. We concentrated on amino acid distributions that we calculated for 107 E words.

Comparison of Results of the Two Approaches for Mouse κ Chains. Calculation of amino acid distribution revealed 34 positions in which the same residues were found in ≥90% of 515 mouse κ chains (Table 3). The statistical analysis of words showed that for most of the conservative positions, results are in good agreement. However, in three conservative positions (A2, DE, and DE), the residues that dominate in almost all chains were found in less than half of A and of DE words. To explain the divergence of the results for these positions, we calculated the number of words that were encountered in more than 1% of all sequences (Table 4). The analysis of 25 A words showed that there are only 5 frequent words, 4 of which contain Thr in the A2 position. However, in 12 rare A words (words encountered only once in the sequences), Thr occurs only in 4 A words. The analysis of frequent and rare words was performed for all strands and loops (Table 4). Of 34 conservative positions of the mouse κ chains, 11 positions were selected (B6, C3, C″D2, E6, EF6, F3, F5, G2, G3, G6, and G8). These stable conservative positions are not only occupied by the same residue in ≥90% of all chains but also contain the same residues in 90% of all words. The detailed characteristics and classification of words will be published elsewhere.

We thank Drs. D. Leschiner, V. Nauchitell, I. Arnold, and G. Feldman for help in calculation and Drs. I. Muchnik and J. Vasiliev for many helpful discussions. We also thank Drs. K. Kulikowsky, B. Sofer, C.-H. Sah, and Mrs. M. Goldman for valuable comments on the manuscript. A.E.K. was supported by the Gabriella and Paul Rosenbaum Foundation.

1. Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991) *Sequences of Proteins of Immunological Interest*, National Institutes of Health Publ. No. 91-3242. (DHHS, PHS, Natl. Inst. Health, Bethesda, MD), 5th Ed.
2. Kabat, E. A. (1978) *Adv. Protein Chem.* **32,** 1–75.
3. Padlan, E. A. (1994) *Mol. Immunol.* **31,** 169–217.
4. Padlan, E. A. (1979) *Mol. Immunol.* **16,** 287–296.
5. Tramantano, A., Chothia, C. & Lesk, A. M. (1989) *Proteins Struct. Funct. Genet.* **6,** 382–394.
6. Harpaz, Y. & Chothia, C. (1994) *J. Mol. Biol.* **238,** 528–539.
7. Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196,** 901–907.
8. Chothia, C., Lesk, A. M., Tramantano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W., Colman, P. M., Spinelli, S., Alzari, P. M. & Poljak, R. J. (1989) *Nature (London)* **342,** 877–883.
9. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.
10. Hazes, B. & Hol, W. G. J. (1992) *Proteins Struct. Funct. Genet.* **12,** 278–298.
11. Nauchitel, V. V. & Somorjai, R. L. (1993) *Proteins Struct. Funct. Genet.* **15,** 50–61.
12. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13,** 211–215.
13. Lim, V. I. (1974) *J. Mol. Biol.* **88,** 857–872.
14. Cohen, F. F., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1986) *Biochemistry* **25,** 266–275.
15. Taylor, W. R. & Orengo, C. A. (1989) *Protein Eng.* **2,** 505–519.
16. Rooman, M. J. & Wodak, S. (1991) *Proteins Struct. Funct. Genet.* **9,** 69–78.
17. Zhong, X., Mesirov, J. P. & Waltz, D. L. (1992) *J. Mol. Biol.* **225,** 1049–1063.
18. Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232,** 584–599.
19. Chothia, C., Lesk, A. M., Levit, M., Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1986) *Science* **233,** 755–758.
20. Havel, T. F. & Snow, M. E. (1991) *J. Mol. Biol.* **217,** 1–7.
21. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.