Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 12562–12567, November 1997
Immunology

A very limited number of keywords (main patterns) describes all sequences of the human variable heavy (V_H) and κ (V_κ) domains

(Ig sequence/sequence comparison/pattern recognition)

ISRAEL M. GELFAND* AND ALEXANDER E. KISTER

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

Contributed by Israel M. Gelfand, August 19, 1997

ABSTRACT **Sequences of the variable heavy (V_H) and κ (V_κ) domains of Ig structures were divided into 21 fragments that correspond to strands, loops, or parts of these structural units of the variable domains. Amino acid sequences of fragments (termed "words") were collected from the 1,172 human heavy and 668 human κ chains available in the Kabat database. Statistical analysis of words of 17 fragments was performed (fragments that comprise the complementary determining regions′ fragments will not be discussed in this paper). The number of different words (those with different residues in at least one position) ranged, for various fragments, from 11 to 75 in the κ chains, and from 23 to 189 in the heavy chains. The main result of this study is that very few keywords, or main patterns of words, were necessary to describe over 90% of the sequences (no more than two keywords per fragment in the κ and no more than five per fragment in the heavy chains). No identical keywords were found for different fragments of the variable domains. Keywords of aligned fragments of the V_H and V_κ domains were different in all but two instances. Thus, knowing the keywords, one can determine whether any given small part of a sequence belongs to a heavy or κ chain and predict its precise localization in the sequence. In addition, by using all of the keywords obtained through analysis of the Kabat database, it was possible to describe completely the sequences of the human V_H and V_κ germ-line segments.**

Analysis of residues that maintain a common Ig structure has been the subject of many investigations (1–13). At present, the numbers of Ig sequences in the Kabat database and Ig structures in the Protein Data Bank are large enough to carry out a statistical analysis of sequence repertoire of the variable domains and to find the relationship between the sequence of a protein and its three-dimensional structure. Analysis of approximately 5,300 sequences, 111 gene segments, and about 100 three-dimensional structures of the variable domains allowed us to determine how frequently a residue was encountered at each position and to give a description of sequence determinants of Ig fold (refs. 12 and 13; C. Chothia, I.M.G., and A.E.K., unpublished work).

In this paper we focus on the correlation among residues within small fragments of the human heavy and κ chains in the Kabat database. We term amino acid sequences of the fragments "words" (12). Comparison of aligned fragments in the chains revealed a large number of variations among words of each fragment. In spite of this diversity, statistical analysis of words allowed us to select a very limited number of keywords, or main patterns, for each fragment, which characterize more than 90% of all Ig chains. The possibility of defining the keywords presupposes the presence of a large number of correlation among residues. Our approach, which involved dividing the sequences into words, permits such correlation to be revealed.

Methods of Analysis and Results

Fragmentation of Sequences. Prediction of secondary structures for all sequences of the Kabat database (12) allowed us to divide the human variable heavy (V_H) and κ (V_κ) sequences into 21 fragments that correspond, approximately, to strands, loops, or parts of these structural units. A, A′, B, C, C′, C″, D, E, F, and G fragments correspond to strands, and A′B, CC′, C′C″, C″D, DE, EF, and FG fragments correspond to loops. Because of its unique two-arch conformation, the loop between B and C strands was divided into two fragments: BC and CB. The first three residues of the sequences were defined as the OA fragment.

A position in a fragment was assigned to each residue in a sequence. In this paper, a residue, in addition to its Kabat numbering, is referred to by an index that contains the name of the fragment and its position therein. Referencing amino acids this way permits us to compare amino acids that are located in the same positions of various sequences.

Collections of Words from the Kabat Database Chains. In this work 1,172 human heavy chains and 668 human κ chains in the Kabat database were analyzed. The numbers of sequences containing particular fragments differs because not all chains in the database were complete. The numbers of sequences with each of the 17 fragments is presented in Table 1 (results for the remaining fragments, CB, C′C″, C″, and FG will be considered elsewhere).

Analysis of the words revealed that some words of a given fragment were encountered many times, whereas others were seen very rarely, or just once. In the F fragment, for example, the word AVYYCAR was found in 394 human heavy chains, but the word AVYYCTR was in only five chains. (Data on frequency of occurrence of individual words are not presented in this paper.) For each fragment we selected different words (those with differing residues in at least one position). The numbers of different words varied considerably for different fragments, ranging in the heavy chains, from 23 for the AA′ fragment to 189 for the E fragment (Table 1).

Classification of Amino Acids in Ig Sequences. For the purpose of analysis of words, we divided amino acids into several groups. This classification of amino acids is based on our previous analysis of residues' frequencies in 5,300 Ig sequences (C. Chothia, I.M.G., and A.E.K., unpublished work; ref. 14). Inspection of these data allowed us to group residues that usually occupy the same positions and are of similar chemical character. All residues were divided into nine groups:

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_H, variable domain of the heavy chain; V_κ, variable domain of the κ (kappa) chain.

*To whom reprint requests should be addressed. e-mail: igelfand@math.rutgers.edu.

Table 1. Numbers of sequences with fragments of the human heavy and κ chains and numbers of different words in each fragment

Fragment	Heavy chains		κ chains	
	Numbers of chains	Numbers of different words	Numbers of chains	Numbers of different words
OA	743 (82)	43 (6)	444 (32)	29 (9)
A	799 (82)	35 (8)	483 (32)	12 (2)
AA′	803 (82)	23 (7)	472 (32)	29 (13)
A′	840 (82)	61 (13)	506 (32)	32 (12)
A′B	841 (82)	38 (15)	485 (32)	26 (8)
B	811 (82)	119 (15)	465 (32)	74 (12)
BC	828 (82)	66 (10)	439 (32)	34 (5)
C	820 (83)	95 (23)	425 (32)	43 (12)
CC′	828 (82)	84 (17)	419 (32)	47 (13)
C′	807 (83)	135 (33)	427 (32)	43 (11)
C″D	828 (82)	124 (18)	430 (32)	50 (6)
D	825 (82)	129 (22)	421 (32)	32 (4)
DE	830 (82)	112 (21)	425 (32)	11 (1)
E	838 (82)	189 (18)	415 (32)	46 (6)
EF	827 (82)	140 (20)	411 (32)	58 (12)
F	844 (78)	99 (15)	411 (32)	51 (7)
G	478	98	335	75

Numbers of sequences containing a given fragment (numbers of chains) and numbers of different words are shown for 17 of 21 fragments of the human heavy and κ chains; analogous data for 16 available V_H and V_κ domain fragments of germ-line sequences are presented in parentheses.

1) L, M, V, I, F, and A (at positions 0A2, A1, B2, B4, C′3, CB1, C1, E4, E6, EF2, G8, and G10); 2) S and T (at positions A4, B1, B5, BC1, D3, D5, and G6); 3) G, A, and S (at positions AA′3, A′B2, B8, C″D5, DE2, F1, G3, and G5); 4) F, Y, and W (at positions C3, F3, F4, and G2); 5) D and E (at positions OA1, A3, A′1, and EF6); 6) K, R, and H (at positions A′3, A′4, B3, CC′4,C5, C6, and D1); 7) Q and N (at positions OA3, A3, and A′4); 8) C (at positions B6 and F5); and 9) P (at position CC′2). Three residues (A, F, and S) each are found in two groups. This is because our classification takes into account not only chemical structure, but also structural role of residues. For instance, in loops, residues S and A are found mostly in same positions and so, together with G, form one group. In strands, however, S and A are classified into two distinct groups: S "shares" positions with T, while A belongs to one group with hydrophophic residues.

Keywords, or Main Patterns of Words. Statistical analysis of words in human heavy and κ chains (which involved determining the number of chains containing particular words, frequency of occurrence of each word in the chains, number of different words, residues' frequencies at the positions of words, and so on) allowed us to define keywords, or main patterns of words, for the 17 fragments. Keywords demonstrate correlation among residues that most frequently are encountered in positions of words. We require that residues at any one position of a keyword belong to the same amino acid group (as per amino acid classification outlined above). The keywords are listed in Table 2. Let us consider, by way of an example, the keywords of the DE fragment of the heavy chains. The three keywords differ mostly in residues at the DE3 and DE4 positions. The first keyword represents the correlation between residues S or A (both of the same group) at DE2, K at DE3, and N at DE4. The DE1 position is variable (marked with an X) because no correlation was found between residues at this position and the residues at the other positions. This pattern (X at DE1, S or A at DE2, K at DE3, and N at DE4) describes 12 different words that were found in 461 human heavy chains. Similar analysis was performed for all keywords of the 17 fragments in the human heavy and κ chains.

Clusters of Words. The keywords defined above serve as a basis for dividing collections of words into clusters. Each cluster corresponds to a particular keyword and has three levels. The first level (*a*) includes those words that are exactly identical to the keyword. (The *a* level of the first cluster of DE words already has been discussed.)

In the second, *b* level, all words were included that had the following property: the residue at any position of these words is of the same amino acid group as the residue in the respective position of the keyword. For instance, the *b* level of the first cluster of the DE fragment of the heavy chain will include the words that have either R or H at DE3 position, as these two residues belong to the same group as K (which occupies DE3 in the keyword). The same considerations apply to all other positions. Our analysis showed that in the *a* and *b* levels of this cluster 16 different words were found in 467 human heavy chains (Table 2).

Inspection of words showed that many words cannot be included in the *a* or *b* level of a cluster because a residue in one position of these words is not of the same group as the residue(s) in the respective position of the keyword. It can be said that these words belong to a cluster with a mutation in one position. To incorporate these words into a cluster, we defined an additional *c* level. Our analysis revealed 53 different words in 539 human heavy chains that belong to all three levels (*a*, *b*, and *c*) of the first cluster of the DE fragment (Table 2).

In this work we defined the keywords for 17 fragments of the human heavy and κ chains and used them to form clusters (Table 2). For each level of every cluster we present two characteristics: the number of the chains containing all words of this level and the number of different words. Our analysis showed that more than 93% of all words can be assigned to 50 clusters of the heavy chains and 26 clusters of the κ chain.

Comments to Table 2

AA′ Words. *The heavy chains.* Residues in all positions are from the GAS or P amino acid groups.

The κ chains. Words in the κ chains are longer by one position than words in the heavy chains.

B Words. *The heavy chains.* The two keywords differ mainly at the B3 and B8 positions. It is interesting to note that residues S and T, which usually share one position and belong to the same group, are separated in these patterns. In the first pattern, S residues at the B1 and B5 positions are correlated with R or K at B3, and G or A at B8, whereas in the second pattern, T residues at B1 and B5 are correlated with S or T at B3, and I or V at B8. All patterns are characterized by hydrophobic residues at the even positions, hydrophilic residues at the odd positions, and C at the B6 position. [1]At the hydrophobic positions B2 and B4, the following pairs of residues are usually found: L-L, V-V, and L-I. [2]The B7 position is occupied by residues from different amino acids group: residues A, K, and T are found in about 90% of sequences and about 60% of different words.

The κ chains. The main motif of both keywords is the same as in the heavy chains. In contrast to the patterns of the heavy chains, the B7 position is not variable in the κ chains. R and K residues are found in 90% of the sequences. The main difference between B words of the V_H and V_κ domains is at the B1 position, which is occupied by S and T in the heavy chains and by P, or positively charged R and K, in the κ chains. [1]Combinations of residues A-L-S, or V-I-T at B2, B4, and B5, positions, respectively often are observed in the first pattern.

BC Words. *The heavy chains.* Three of four positions are conservative positions. [1]Fifteen residues are found at BC4 of which S and T, which are found in 50% of different words and about 90% of sequences, are most common. The three keywords are distinguished by residues at only the BC3 position.

Table 2. Keywords (main patterns) in the human heavy and κ chains

Heavy chains	#	1	2	3	*a*		*b*		*c*		TOTAL	
		0A1	0A2	0A3	dwords	chains	dwords	chains	dwords	chains	dwords	chains
	1	**Q**	**V**	**Q**	1 (1)	298 (42)	4 (3)	321 (48)	16 (5)	360 (53)	86%	92%
	2	**E**	**V**	**Q**	1 (1)	288 (29)	8 (1)	314 (29)	21 (1)	326 (29)		
Kappa	#	1	2	3								
chains		0A1	0A2	0A3								
	1	**DE**	**I**	**V**	2 (2)	231 (13)	13 (3)	258 (16)	19 (4)	278 (17)	90%	99%
	2	**D**	**I**	**Q**	1 (1)	157 (9)	3 (1)	160 (9)	3 (3)	160 (12)		

Heavy chains	#	4	5	6	*a*		*b*		*c*		TOTAL	
		A1	A2	A3	dwords	chains	dwords	chains	dwords	chains	dwords	chains
	1	**L**	**VL**	**E**	2 (2)	270 (26)	7 (2)	276 (26)	12 (2)	283 (26)	89%	98%
	2	**L**	**V**	**Q**	1 (1)	198 (30)	6 (1)	210 (30)	9 (2)	213 (31)		
	3	**L**	**Q**	**E**	1 (1)	105 (16)	1 (1)	105 (16)	6 (3)	156 (21)		
	4	**L**	**Q**	**Q**	1 (1)	90 (4)	1 (1)	90 (4)	4 (1)	131 (4)		
Kappa		4	5	6								
chains	1	**ML**	**T**	**Q**	2 (2)	468 (32)	5 (2)	476 (32)	10 (2)	481 (32)	83%	99%

Heavy chains	#	7	8	9		*a*		*b*		*c*	TOTAL	
		AA'1	AA'2	AA'3	AA'4	dwords	chains	dwords	chains	dwords	chains	dwords	chains
	1	**S**	**G**	**GA**	**-**	2 (2)	536 (52)	9 (5)	557 (57)	16 (6)	587 (60)	87%	97%
	2	**S**	**G**	**P**	**-**	1 (1)	187 (22)	2 (1)	187 (22)	4 (1)	189 (22)		
Kappa		7	8	9	10								
chains	1	**S**	**P**	**GAS**	**ST**	6 (4)	347 (17)	6 (4)	347 (17)	19 (9)	407 (23)	86%	99%
	2	**ST**	**P**	**L**	**S**	2 (2)	47 (7)	2 (3)	47 (8)	5 (4)	50 (9)		

Heavy chains	#	10	11	12	13	*a*		*b*		*c*		TOTAL	
		A'1	A'2	A'3	A'4	dwords	chains	dwords	chains	dwords	chains	dwords	chains
	1	**E**	**V**	**K**	**K**	1 (1)	229 (25)	9 (3)	258 (28)	18 (3)	276 (28)	97%	99%
	2	**G**	**L**	**V**	**K**	1(1)	220 (20)	14 (4)	291 (26)	26 (6)	323 (29)		
	3	**G**	**LV**	**V**	**Q**	2 (2)	191 (23)	9 (3)	216 (24)	15 (4)	240 (25)		
kappa		11	12	13	14								
chains	1	**L**	**S**	**LVA**	**S**	3 (3)	379 (18)	15 (9)	451 (27)	26 (10)	468 (28)	94%	99%
	2	**L**	**P**	**V**	**T**	1 (1)	32 (4)	4 (1)	36 (4)	4 (2)	36 (5)		

Heavy chains	#	14	15	16	*a*		*b*		*c*		TOTAL	
		A'B1	A'B2	A'B3	dwords	chains	dwords	chains	dwords	chains	dwords	chains
	1	**P**	**G**	**GAS**	3 (3)	380 (40)	4 (4)	383 (41)	13 (7)	401 (45)	84%	99%
	2	**P**	**GS**	**E**	2 (2)	203 (19)	4 (3)	212 (21)	7 (5)	219 (23)		
	3	**P**	**S**	**Q**	1 (1)	113 (4)	2 (1)	116 (4)	5 (2)	137 (8)		
	4	**P**	**G**	**R**	1 (1)	66 (6)	3 (1)	68 (6)	7 (1)	78 (6)		
kappa		15	16	17								
chains	1	**P**	**G**	**DE**	2 (2)	233 (10)	3 (2)	241 (10)	5 (3)	244 (11)	85%	99%
	2	**LV**	**G**	**ED**	4 (2)	207 (14)	10 (3)	225 (15)	17 (6)	235 (18)		

Heavy chains	#	17	18	19	20	21	22	23	24	*a*		*b*		*c*		TOTAL	
		B1	B2	B3	B4	B5	B6	B7	B8	dwords	chains	dwords	chains	dwords	chains	dwords	chains
	1	**S**	**LV**[1)]	**RK**	**LVI**[1)]	**S**	**C**	**X**[2)]	**GA**	35 (6)	502 (53)	46 (7)	515 (54)	84 (7)	568 (7)	90%	98%
	2	**T**	**LV**[1)]	**ST**	**LVI**[1)]	**T**	**C**	**X**[2)]	**IV**	10 (5)	198 (22)	14 (6)	200 (23)	23 (6)	228 (23)		
Kappa		18	19	20	21	22	23	24	25								
chains	1	**R**	**VA**[1)]	**T**	**LI**[1)]	**ST**[1)]	**C**	**R**	**A**	6 (2)	283 (17)	19 (4)	305 (19)	62 (10)	414 (25)	89%	97%
	2	**P**	**A**	**S**	**I**	**S**	**C**	**R**	**S**	1 (1)	32 (5)	3 (2)	35 (7)	4 (2)	36 (7)		

Heavy chains	#	25	26	27	28	*a*		*b*		*c*		TOTAL	
		BC1	BC2	BC3	BC4	dwords	chains	dwords	chains	dwords	chains	dwords	chains
	1	**S**	**G**	**FY**	**X**[1)]	19 (6)	533 (64)	23 (6)	537 (64)	34 (6)	551 (64)	80%	97%
	2	**S**	**G**	**G**	**ST**	2 (2)	123 (14)	5 (2)	130 (14)	13 (3)	167 (17)		
	3	**S**	**G**	**D**	**S**	1 (1)	80 (1)	4 (1)	85 (1)	6 (1)	87 (1)		
Kappa		26	27	27A-28	-								
chains	1	**S**	**Q**	**X**[1)]	**X**[1)]	9 (3)	389 (30)	15 (4)	398 (31)	28 (~~5~~)	427 (32)	90%	95%

Heavy chains	#	34	35	36	37	38	39	*a*		*b*		*c*		TOTAL	
		C1	C2	C3	C4	C5	C6	dwords	chains	dwords	chains	dwords	chains	dwords	chains
	1	**MI**	**X**[1)]	**W**	**V**	**R**	**Q**	21 (4)	491 (39)	47 (13)	564 (56)	65 (16)	594 (59)	95%	98%
	2	**W**	**X**[1)]	**W**	**I**	**R**	**Q**	7 (3)	179 (17)	18 (7)	203 (24)	25 (7)	212 (24)		
Kappa		33	34	35	36	37	38								
chains	1	**L**	**X**	**W**	**Y**	**Q**	**Q**	10 (6)	342 (22)	21 (9)	371 (27)	42 (12)	424 (32)	99%	99%

The results of the statistical analysis of words in *a*, *b*, and *c* levels of clusters are presented for 17 fragments of the human heavy and κ chains in the Kabat database; analogous data for germ-line database sequences is given in parentheses. We illustrate the use of Table 2 on the example of BC fragment of the heavy chains. The Kabat numbers (25–28, in the first row) correspond to indices BC1, BC2, BC3, and BC4. Three keywords: SG(FY)X; SGG(ST), and SGDS serve as a basis for dividing all words of BC fragment into three clusters. The words of these clusters are found in 97% of all chains and cover 80% of different words (last two columns—Total). Six middle columns show the statistics for three *a*, *b*, and *c* levels of each cluster. The *a* columns list the numbers of different words (dwords) and numbers of chains (chains) containing all words of the *a* level of each of the three clusters. *b* columns contain the data for *b* levels; and the number of different words and the total number of chains in each of clusters are presented in *c* columns. Thus, in *a* level of the first cluster, there are 19 different words (six in germ-line sequences), which are found in 533 chains (64 in germ-line sequences). Total number of chains in which were found all words of the first cluster (551) and total number of different words in *a*, *b*, and *c* levels (34) are given in *c* columns (respective data for germ-line sequences are 64 and 6). The superscripts refer to *Comment to Table 2* section of the paper.

Table 2. (continued)

Heavy chains	#	40 CC'1	41 CC'2	42 CC'3	43 CC'4	44 CC'5	45 CC'6	a dwords	a chains	b dwords	b chains	c dwords	c chains	TOTAL dwords	TOTAL chains
	1	X[1]	P	GS[2]	RK[2]	G	L	12 (5)	544 (52)	27 (7)	596 (58)	52 (11)	654 (63)	81%	97%
	2	A	P	G	Q	G	L	1 (1)	111 (11)	5 (2)	115 (13)	16 (5)	147 (16)		
Kappa chains		39	40	41	42	43	44								
	1	K	P	G	Q	X[1]	P	4 (4)	206 (11)	8 (5)	242 (13)	25 (8)	267 (16)	94%	99%
	2	K	P	G	K	A	P	1 (1)	107 (11)	4 (3)	112 (14)	19 (4)	149 (15)		

Heavy c chains	#	46 C'1	47 C'2	48 C'3	49 C'4	50 C'5	51 C'6	a dwords	a chains	b dwords	b chains	c dwords	c chains	TOTAL dwords	TOTAL chains
	1	E	W	VMI[1]	GAS	X[2]	I	52 (21)	594 (68)	93 (25)	663 (74)	118 (28)	696 (78	96%	99%
	2	E	W	L	G	X[2]	T	3 (2)	84 (2)	13 (2)	102 (2)	13 (2)	102 (2)		
Kappa chains		45	46	47	48	49	-								
	1	RK	L	L	I	Y	-	2 (2)	303 (17)	19 (3)	369 (18)	39 (10)	423 (31)	91%	99%

Heavy chains	#	61 C"D1	62 C"D2	63 C"D3	64 C"D4	65 C"D5	a dwords	a chains	b dwords	b chains	c dwords	c chains	TOTAL dwords	TOTAL chains
	1	X[1]	S	VL[2]	K	GS[2]	18 (6)	427 (50)	30 (7)	457 (52)	45 (8)	484 (53)	83%	98%
	2	Q	X[3]	FL	Q	G	7(3)	111(15)	10 (4)	114 (16	44 (8)	163 (22)		
	3	P	S	F	Q	G	1 (1)	100 (6)	4 (1)	104 (6)	8 (1)	113 (6)		
	4	X[4]	P	V	K	G	3 (1)	37 (1)	4 (1)	38 (1)	6 (1)	48 (1)		
Kappa chains		56	57	58	59	60								
	1	ST	G	VI[1]	P	SA[1]	3 (3)	187 (21)	9 (3)	196 (21)	32 (3)	238 (21)	96%	99%
	2	ST	G	VI	P	D	2 (2)	154 (10)	3 (2)	161 (8)	16 (2)	190 (8)		

Heavy chains	#	66 D1	67 D2	68 D3	69 D4	70 D5	71 D6	72 D7	a dwords	a chains	b dwords	b chains	c dwords	c chains	TOTAL dwords	TOTAL chains
	1	R	FV[1]	T	IM[1]	ST[1]	R	D	5 (3)	250 (35)	20 (7)	304 (43)	45 (13)	375 (50)	88%	97%
	2	R	V	T	I	ST	VA	D	2 (2)	123 (18)	20 (5)	183 (23)	50 (5)	256 (23)		
	3	Q	V	T	I	S	A	D	1 (1)	73 (3)	6 (1)	90 (3)	11 (2)	99 (6)		
	4	R	I	T	I	N	P	D	1 (1)	63 (1)	5 (1)	70 (1)	7 (1)	73 (1)		
Kappa chains		61	62	63	64	65	66	67								
	1	R	F	S	G	S	G	S	1 (1)	354 (29)	8 (2)	374 (30)	27 (4)	414 (32)	84%	98%

Heavy chains	#	73 DE1	74 DE2	75 DE3	76 DE4	a dwords	a chains	b dwords	b chains	c dwords	c chains	TOTAL dwords	TOTAL chains
	1	X[1]	SA	K	N	12 (6)	461 (50)	16 (7)	467 (51)	53 (8)	539 (52)	90%	99%
	2	X[1]	S	I	S	5 (2)	112 (10)	11 (7)	143 (18)	30 (8)	188 (19)		
	3	X[1]	S	T	S	5 (4)	74 (7)	10 (4)	77 (7)	18 (7)	95 (11)		
Kappa chains		68	69	-	-								
	1	G	T	-	-	1 (1)	398 (32)	11 (1)	407 (32)	11 (1)	425 (32)	100%	100%

Heavy chains	#	77 E1	78 E2	79 E3	80 E4	81 E5	82 E6	82A E7	a dwords	a chains	b dwords	b chains	c dwords	c chains	TOTAL dwords	TOTAL chains
	1	TS	L	Y	L	Q	M	N	2 (2)	189 (23)	14 (2)	213 (23)	51 (6)	272 (28)	75%	90%
	2	T	A	Y	L	Q	W	S	1 (1)	97 (6)	2 (1)	99 (6)	17 (2)	121 (7)		
	3	Q	F	S	L	K	L	S	1 (1)	91 (18)	6 (1)	108 (18)	26 (3)	130 (20)		
	4	T	A	Y	M	E	L	X[1]	4 (2)	103 (17)	11 (3)	113 (19)	33 (4)	141 (20)		
	5	Q	F	S	L	Q	L	N	1 (1)	69 (1)	4 (1)	72 (1)	13 (1)	86 (1)		
Kappa chains		70	71	72	73	74	75	76								
	1	DE	F	T	L	T	I	S	2 (2)	281 (20)	12 (3)	330 (21)	39 (6)	405 (32)	85%	98%

Heavy chains	#	82B EF1	82C EF2	83 EF3	84 EF4	85 EF5	86 EF6	87 EF7	a dwords	a chains	b dwords	b chains	c dwords	c chains	TOTAL dwords	TOTAL chains
	1	S	L[1]	R[1]	X	E	D	T	13 (3)	277 (33)	24 (5)	313 (36)	64 (12)	415 (49)	79%	94%
	2	S	V[2]	T[2]	A	A	D	T	1 (1)	111 (18)	4 (2)	116 (20)	16 (2)	141 (20)		
	3	S	L[1]	K[1]	A	S	D	T	1 (1)	92 (5)	5 (1)	97 (5)	22 (2)	125 (6)		
	4	S	V[2]	T[2]	P	E	D	T	1 (1)	77 (1)	3 (1)	79 (1)	8 (1)	100 (1)		
Kappa chains		77	78	79	80	81	82	83								
	1	S	L	Q	X	E	D	FV	9 (6)	160 (16)	15 (6)	188 (16)	40 (10)	264 (23)	99%	99%
	2	R	LV	E	X	E	D	FV	6 (2)	126 (9)	8 (2)	132 (9)	17 (2)	146 (9)		

Heavy chains	#	88 F1	89 F2	90 F3	91 F4	92 F5	93 F6	94 F7	a dwords	a chains	b dwords	b chains	c dwords	c chains	TOTAL dwords	TOTAL chains
	1	A	VM	Y	Y	C	A	RK	3 (3)	556 (61)	46 (8)	709 (69)	79 (14)	773 (77)	95%	98%
	2	A	VL	Y	YF	C	T	X	7 (1)	49 (1)	11 (1)	53 (1)	15 (1)	57 (1)		
Kappa chains		84	85	86	87	88	89	90								
	1	A	V	Y	Y	C	Q	Q	1 (1)	178 (7)	6 (1)	191 (7)	35 (3)	268 (15)	90%	98%
	2	A	T	Y	Y	C	Q	Q	1 (1)	99 (12)	3 (1)	108 (12)	11 (4)	133 (17)		

Heavy chains	#	102 G1	103 G2	104 G3	105 G4	106 G5	107 G6	108 G7	109 G8	110 G9	111 G10	a dwords	a chains	b dwords	b chains	c dwords	c chains	TOTAL dwords	TOTAL chains
98	1	X[1]	W	G	Q	G	T	LM	V	T	V	14	269	31	307	58	345	75%	92%
	2	X[1]	W	G	Q	G	T	T	V	T	V	3	56	5	59	15	93		
Kappa chain 75		97	98	99	100	101	102	103	104	105	106								
	1	T	F	G	X	G	T	KR	VL	ED	VI	23	260	42	291	71	320	95%	96%

The κ chains. Usually three positions are in each word; however, a few four-position words are observed. [1]BC3 or BC4 positions are variable, but most frequently occupied by S.

C Words. *The heavy chain.* [1]Position C2 is occupied by 16 different amino acids (S, H, N, and G are the most common).

The κ chains. The main differences between the patterns of the heavy and κ chains are at the C4 and C5 positions.

CC′ Words. *The heavy chains.* [1]At CC′1 position, A, P, M, and S residues are found most frequently. [2]The combination of G and K residues usually occupies CC′3 and CC′4 positions, as do S and R residues.

The κ chains. [1]A, P, and S residues are found at CC′5 position in more than 90% of the chains.

C′ Words. *The heavy chains.* The two patterns differ only at the C′6 position. [1]Three hydrophobic residues, V, M, and I, are found in an approximately equal number of words at C′3 position. [2]C′5 is a very variable position, which is occupied by any residue except D and P.

The κ chains. Peculiar words are found. Three hydrophobic residues are in a row, at positions C′2, C′3, and C′4. By contrast to the patterns of the heavy chains, in which the first position is occupied by negatively charged residues, the κ chains have positively charged residues at the first position. Also, κ chains words are shorter, by one position, than heavy chain words.

C″D Words. *The heavy chains.* [1]Different residues occupy the C″D1 position (mainly P and D), but Q never does. [2]When V occupies C″D3, then either G and S is at C″D5, but if C″D3 is taken by L, the residue at C″D5 is usually S. [3]When Q is at the C″D1 position, C″D2 can be occupied by many different residues (mostly K and N, but never S). [4]X equals A, T, D, P, or G residues at C″D1 position.

The κ chains. [1]In most cases, positions C″D3 and C″D5 are occupied by either V and S, or I and A, respectively.

D Words. *The heavy chains.* Hydrophobic residues at even positions are a common feature of the strands of the D, E, and B strands of the β-sheet. Positively charged residues, which take part in the salt bridge, are found at D1 in 85% of the heavy chains. Negatively charged residues occupy the last position (D7) in 95% of the sequences. [1]F at the D2 position always occurs with I at D4 and S at D5, and V at D2 correlates with T at D5.

The κ chains. One pattern is common to 98% of the sequences. The first three positions are similar in the heavy and κ chains, whereas the last four differ greatly in both chains. All positions from D3 on contain only G and S residues.

DE words. *The heavy chains.* [1]T, N, K, and D are the most common residues at the DE1 position (found in 322, 208, 135, and 82 sequences, respectively).

The κ chains. The keyword is two positions long.

E Words. *The heavy chains.* Hydrophobic residues were at the even positions, whereas hydrophilic residues were at the odd ones, as were B and D words. An exception was found in only 20 sequences, in which the E3 position was occupied by hydrophobic V residue. The E5 position almost always is occupied by charged and polar residues (most frequent are Q, E, and K residues, which are found in 56%, 19%, and 13% of sequences, respectively). We discovered two regularities involving residues at the E1 and E3 positions: (*i*) S and T at E1 never are found together with S at the E3 position; and (*ii*) S at the E3 position usually correlates with Q at E1. [1]The most frequently encountered residues at E7 were S and N.

The κ chains. Main motifs are similar to those of the heavy chains. However, in hydrophilic positions many differences between the two classes of chains were observed. D and E residues usually were found at E1 positions of the κ chains, but never at this position of the heavy chains. Also, in the κ chain, at the E3 position, T was found in 91% of the chains and S was seen very rarely, whereas the reverse situation was found in the heavy chains (S occurs in 27% of the sequences and T is never found).

EF words. *The heavy chains.* Four conservative positions were found: EF1 (contains S residue), EF2 (hydrophobic), EF6 (contains D), and EF7 (contains T). In the variable EF4 position, A, S, and P are the most frequent residues (88% of the chains); generally, these residues are located at those sites where a chain is "broken". [1]L at EF2 correlates, in almost all cases, with R or K at EF3 position. [2]Another correlation was observed between V at EF2 and T at EF3.

The κ chains. Just as in the heavy chain, EF4 is the variable position; A, S and P are most frequent residues at this position (88% of the chains).

F words. *The heavy chains.* Fifteen different amino acids can be found at position F6 and 17 at F7.

The κ chains. F1, F3, F4, and F5 are conservative positions. The same residues are found in the heavy and κ chains.

G words. *The heavy chains.* [1]Of the 16 residues found in the G1 position, most common were Y, V, I, and P (84% of the chains).

Keywords in the Germ Line of the Human V_H Segments

Although the number of sequences in the Kabat database is large (2), a question remains about whether the keywords obtained through analysis of these sequences will suffice for description of all V_H and V_κ domains. We, therefore, checked our results by using another database that contained the sequences of the complete human repertoire of the germ-line V_H and V_k segments (14, 15). The results of our analysis can be summarized as follows: (*i*) it was shown that the ratio of the number of different words to the total number of words is much higher for the fragments of the germ-line database sequences than for the fragments of the Kabat sequences (Table 1); (*ii*) all of the Kabat database keywords were found among the germ-line words (Table 2); (*iii*) the set of keywords was adequate for complete description of the entire germ-line repertoire, with the exception of five words of E and EF fragments; and (*iv*) comparison of the words' distribution in the three levels of clusters revealed that the fraction of words in the *c* level is significantly larger for the Kabat sequences than for the germ-line sequences. Considering the fact that the germ-line database (unlike the Kabat database) contains no sequences with somatic mutations, this last observation can be interpreted as supporting the hypothesis that most words in the *c* level arise through somatic mutations.

Conclusions

Division of sequences into fragments allowed us to perform alignment of all human heavy and human κ chains in the Kabat database. To describe a residue's location in a chain we used an index that included the name of the fragment and residue's position number therein. This permitted us to analyze residues at identical positions in different chains.

Statistical analysis of words of 17 fragments was performed. For each fragment we calculated the number of the chains containing this fragment and number of different words. The number of different words varied for each fragment, with the largest being 189 (E fragment) in the human heavy chains and 75 (G fragment) in the κ chains.

Very few keywords, or main patterns, for each fragment were found: 2–5 for the human heavy and 1–2 for the human κ chains. Each keyword served as the basis for constructing a cluster. It was shown that words of more than 90% of all sequences belong to clusters. This result demonstrates that all sequences of the variable domain can be described almost completely (with the exception of complementary determining regions) by a very limited number of patterns.

For the human heavy and human κ chains we suggested 50 and 26 keywords, respectively. The variable X positions, which are occupied by residues from different amino acid groups,

were found in 19 keywords of the heavy chains and six keywords of the κ chains. Only one variable position was in all of these patterns (with the exception of the BC keyword of the κ chains). The small number of variable X positions demonstrates that strong correlation is among residues for each fragment.

No identical keywords were found in more than one fragment of either human heavy or human κ chains. This observation demonstrates that different fragments of the variable domains have their own individual patterns (no repeat fragments are found).

In keywords of all fragments certain positions were found that are always, in both heavy and κ chains, occupied by residues from the same amino acid group. Residues at these positions constitute the common characteristic of the fragments. The F fragment, for instance, in both chains is characterized by Ala at F1, aromatic residues at F3 and F4, and Cys at F5 position (see Table 2).

However, keywords of aligned fragments of the human heavy and κ chains differed in all but two cases (0A and A′B fragments). Thus, human heavy and κ chains can be distinguished by the patterns of (almost) any one of their fragments. Moreover, because keywords of a fragment are unique, given a small segment of residues it is possible, knowing all the keywords, to determine the exact location of that segment in the κ or heavy chains.

Keywords obtained through an analysis of all chains in the Kabat database were used in the study of human repertoire of V_H and V_κ germ-line segments. It was shown that all available words of the germ-line sequences (with the exception of several words of the E and EF fragments) belong to the existing clusters. Furthermore, analysis of the Kabat database did not disclose any superfluous keywords, i.e., at least one word from a germ-line sequence was assigned to each cluster.

We are grateful to Drs. C. Chothia, C. Kulikowski, I. Muchnik, and O. Ptitsyn for very helpful discussions. We wish to acknowledge with deep gratitude the support of the Gabriella and Paul Rosenbaum Foundation and also to thank Mrs. M. Goldman for continuous encouragement. A.E.K. is supported by the Gabriella and Paul Rosenbaum Foundation.

1. Kabat, E. A. (1978) *Adv. Protein Chem.* **32,** 1–75.
2. Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991) *Sequences of Proteins of Immunological Interest* (Natl. Inst. Health, Bethesda), NIH Publ. No. 91–3442, 5th Ed.
3. Harpaz, Y. & Chothia, C. (1994) *J. Mol. Biol.* **238,** 528–539.
4. Tramantano, A., Chothia, C. & Lesk, A. (1989) *J. Mol. Biol.* **215,** 175–182.
5. Chothia, C., Boswell, D. R. & Lesk, A. M. (1988) *EMBO J.* **7,** 3745–3755.
6. Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196,** 901–917.
7. Bork, P., Holm, L. & Sander, C. (1994) *J. Mol. Biol.* **242,** 309–320.
8. Beale, D. & Coadwell, J. (1989) *Int. J. Biochem* **21,** 227–232.
9. Padlan, E. A. (1994) *Mol. Immunol.* **31,** 169–217.
10. Williams, A. F. & Barclay, A. N. (1988) *Annu. Rev. Immunol.* **6,** 381–405.
11. Gerstein, M. & Altman, R. B. (1995) *J. Mol. Biol.* **251,** 161–175.
12. Gelfand, I. M. & Kister, A. E. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10885–10889.
13. Gelfand, I. M., Kister, A. E. & Leshchiner, D. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 3675–3678.
14. Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992) *J. Mol. Biol* **227,** 799–817.
15. Tomlinson, I. M., Cox, J. P. L., Gherardi, E., Lesk, A. M. & Chothia, C. (1995) *EMBO J.* **14,** 4628–4638.