Proc. Natl. Acad. Sci. USA
Vol. 74, No. 6, pp. 2551–2555, June 1977
Immunology

Structural implications of sequence variability in immunoglobulins

(dissimilarities between amino acids/conservative and nonconservative replacements/specificity-determining residues)

EDUARDO A. PADLAN

Laboratory of Molecular Biology, National Institute of Arthritis, Metabolism, and Digestive Diseases, National Institutes of Health, Bethesda, Maryland 20014

Communicated by Elvin A. Kabat, April 7, 1977

ABSTRACT **Immunoglobulin sequences were compared by using a technique that takes into account the dissimilarity in physicochemical properties of amino acids. Exterior residues showed greater structural variability than interior residues. High structural variability was found at positions known from crystallographic studies to be involved in hapten binding.**

Two of the most interesting aspects of the binding of antibody to antigen are the high specificity of the reaction and the apparent existence of many different antibody-combining-site structures of diverse binding specificities. It is generally accepted that the specificity is due to the complementarity between the combining site and the antigenic determinant. The existence of many different specificities probably is mainly due to the presence of hypervariable (HV) regions (1–3), some of which are now known to form the combining site (see ref. 4 for a review). High-resolution crystallographic studies could pinpoint the protein residues crucial in providing specificity; yet these studies demand so large an investment in time and effort that the elucidation of only a few combining site structures might ever be attempted. Other methods are being used in attempts to discover which residues are specificity-determining, including (*i*) model-building studies from sequence data (5, 6), (*ii*) comparison of HV region sequences of proteins with the same specificity (e.g., ref. 7), (*iii*) correlation of HV region sequences with known specificities (8), and (*iv*) analysis of the frequency of occurrence of particular amino acid pairs in HV regions (9). Sequence comparisons are hampered by the limited number of HV region sequences from proteins of known specificity. Also, because HV loop structures appear to depend on their length (10), sequence comparisons are valid only when HV regions of the same length are considered. Immunoglobulin sequences are customarily compared by the method of Wu and Kabat (1), in which the variability at each position is computed as the number of different amino acids occurring at this position divided by the frequency of the most commonly occurring residue. This method does not distinguish between conservative substitutions, such as isoleucine for leucine, and structurally more drastic replacements, such as glycine for tryptophan.

This paper describes a comparison of immunoglobulin sequences in which the differences in the physicochemical properties of the amino acids are taken into account. The structural variability at each sequence position is correlated with whether it is in the domain interior, exposed to solvent, or in contact with the homologous domain. Furthermore, structural dissimilarities in HV regions of the same length and from molecules of different specificities are analyzed in an attempt to predict which residues may be specificity-determining. Ligand binding studies in the crystal have been performed on mouse M603 Fab (11), human myeloma protein NEWM Fab (12), and human Bence-Jones MCG dimer (13). Where pertinent, the crystallographic findings will be correlated with the results of the sequence comparisons.

Abbreviations: HV, hypervariable; V_H, heavy chain variable domain; V_L, light chain variable domain; ASD, average structural dissimilarity; C_L, light chain constant domain.

MATERIALS AND METHODS

The variable region sequences compared are from the tabulation by Kabat *et al.* (14), and their numbering scheme and HV region designations are used here. Thus, the first, second, and third light chain HV regions (L1, L2, and L3) include positions 24–34, 50–56, and 89–97, respectively; the first, second, and third heavy chain HV regions (H1, H2, and H3) include positions 31–35, 50–65, and 95–102, respectively. The additional HV region found in human heavy chains (3), which includes positions 81–85, is designated He. Each position in the sequences is classified (5, 15) according to whether it is completely or mainly exposed to solvent, partly exposed and partly buried, completely or mainly buried in the domain interior, in contact with the homologous domain, or in the interdomain interface but not in contact with the homologous domain.

Comparison of Nonhypervariable or Framework Regions. One molecule was chosen from each subgroup in different species. The sequences included in the analysis are as completely known as possible. The heavy chain variable domain (V_H) sequences compared are of: human proteins EU, DAW, and NIE; mouse M315, M21, and M603; rabbit BS-5; and guinea pig anti-DNP antibody (sequenced only to position 82). The light chain variable domain (V_L) sequences compared are of: human proteins HAU, MIL, TI, LEN, HA, VIL, X, SH, and BO; mouse M21, M70, M173, M41, M511, M11, and M104E; rat S211; and rabbit 4135, 2717, and 3315. In the V_H and V_L comparisons, only nonhypervariable or framework residues are considered.

Comparison of HV Regions. Only sequences that are completely known, from proteins with different specificities, are considered. Only one sequence per specificity was chosen. The sequence of one protein of unknown specificity from each subgroup in different species is included in the analyses. Only HV regions of the same length are compared, and the sequences of at least 10 different proteins are required per HV region comparison. The sequences are of the following proteins:

(*i*) L1 regions, 11 residues long—human AG, DAV, TIL, X, and SH; mouse M41, M46, M21, M11, U10, and E109; rat IR158; and rabbit 4135, BS-5, 3368, 3322C, 3547, and K4820;

(*ii*) L2 regions, 7 residues long—human AG, MIL, TI, LEN, HA, VIL, X, SH, BO, and MCG; mouse M41, M173, M70, M21, M511, M11, M104E, R20, T5606, and M315; rat S211; and rabbit 4135, BS-5, 2717, 3315, 3368, 3547, and K4820;

(*iii*) L3 regions, 9 residues long—human AG, TEW, TI, LEN, NEWM, and X; mouse M41, M173, M70, M21, M511, M104E, R20, Y5606, and M315; rat S211; and rabbit BS-5, 2717, and K4820;

(*iv*) H1 regions, 5 residues long—human EU, NIE, LAY, POM, and TUR; mouse M460, E109, M21, M173, M603, and anti-AR AB; rabbit BS-5 and 2690; and guinea pig anti-DNP, -ARS, and -TMA antibodies;

(*v*) H2 regions, 17 residues long—human EU, HE, NIE, LAY, POM, and TUR; mouse M21 and M173; and guinea pig anti-DNP, -ARS, and -TMA antibodies;

(*vi*) He regions, 8 residues long—human EU, DAW, NEWM, LAY, POM, and TUR; and mouse M21, M173, and M603.

Computation of Average Structural Dissimilarity (ASD) at Each Sequence Position. The structural dissimilarities between the 20 naturally occurring L-amino acids taken in pairs have been quantified by Sneath (16) and Grantham (17). Sneath compared the various residues by the methods of "numerical taxonomy" on the basis of 134 properties—e.g., charge, aromaticity, bifurcation of the sidegroup, etc. Grantham, on the other hand, compared the amino acids on the basis of only three attributes—namely, molecular volume, atomic composition, and polarity—the differences in each attribute being correlated with the relative substitution frequency of the particular amino acid pair. *A priori*, there is no basis for choosing Sneath's or Grantham's dissimilarity array over the other, and both are used. The dissimilarity tables of Sneath and of Grantham are shown in Table 1. Included in Table 1 is the mean dissimilarity, $\langle D\rangle j$, of the 20 naturally occurring residues relative to each other. Each $\langle D\rangle j$ represents the mean structural dissimilarity for random substitution at a given position with j being the residue at that position in the presumed ancestral sequence.

In the sequence comparisons, the structural variability at each position was defined as the ASD relative to the most commonly occurring residue, computed according to the formula:

$$\text{ASD} = 100 \times \frac{1}{N} \sum_{i=1}^{N} \frac{Dij}{\langle D\rangle j}$$

in which Dij is the dissimilarity (from Table 1) between residue i and the most commonly occurring residue j and the summation is over the N amino acids occurring at the specific position. The ASD represents structural variability expressed as percentage of random. Two measures of structural variability were computed: ASD-S calculated by using Sneath's dissimilarity array, and ASD-G calculated by using Grantham's dissimilarity array. If more than one residue was most commonly occurring (i.e., at the same frequency), then ASD values were computed relative to each and the value considered was that which was smallest. Where gaps existed, no sequence or structural variabilities were computed. If a residue was unidentified, it was ignored and the variabilities were computed only with the known residues. For an undetermined Asx (or Glx), the dissimilarity used was the average of those for Asp and Asn (or Glu and Gln). In the tally of residues, an Asx (or Glx) was counted as half an Asp and half an Asn (or half a Glu and half a Gln).

To obtain standards for structural conservation and hypervariability, ASD values were computed between the residues within and among the various amino acid families grouped by Dayhoff *et al.* (18) on the basis of their relative frequency of substitution in the known protein sequences. The measure of conservative replacements was based on the structural variation among the members of the groups Met-Leu-Ile-Val, Phe-Tyr-Trp, and Lys-Arg-His. These are three of the five distinct groupings of Dayhoff *et al.* who found that members of the same group frequently replace each other at the same sequence position but rarely replace a residue belonging to another group. The two other groups distinguished by Dayhoff *et al.* consist of Cys by itself in one group and Ala, Gly, Pro, Ser, Thr, Asn, Asp, Gln, and Glu in the other. These two groups were excluded here in the analysis of conservative and nonconservative replacements because of the heterogeneity of the latter group (divided into four subgroups) and because Cys plays a special structural role in immunoglobulin domain folding.

The ASD-S and ASD-G between the amino acids of the same group are 37.4, 35.8, and 46.9 (mean ± SD, 40.0 ± 4.9) and 12.5, 21.6, and 21.5 (18.5 ± 4.3) for the Met-Leu-Ile-Val, Phe-Tyr-Trp, and Lys-Arg-His groups, respectively. When an individual residue from each of the three groups is compared, the mean (± SD) ASD-S and ASD-G are 66.9 ± 6.1 and 47.3 ± 4.7, respectively, quite different from those obtained in the comparison of residues within a group. A similar comparison of the various members of the heterogeneous Ala-Gly-Pro-Ser-Thr-Asn-Asp-Gln-Glu group (one from each of the four subgroups) gives a mean (± SD) ASD-S of 67.6 ± 7.0 and a mean (± SD) ASD-G of 51.3 ± 4.7. Accordingly, in the comparisons that follow, the substitutions at a given position are considered conservative if ASD-S <44.9 or ASD-G <22.8. These values represent the mean ASD-S and ASD-G, respectively, between the residues in each of the three groups considered above plus 1 SD. Furthermore, the substitutions are expected to lead to structural hypervariability if simultaneously ASD-S >66.9 and ASD-G >47.3.

Table 1. Dissimilarities between amino acids

C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	⟨D⟩j	
195	126	107	113	60	86	94	106	96	84	111	27	91	112	99	58	64	148	112	94.4	A
	154	170	205	159	174	198	202	198	196	139	169	154	180	112	149	192	215	194	167.7	C
		45	177	94	81	168	101	172	160	23	108	61	96	65	85	152	181	160	110.4	D
			140	98	40	134	56	138	126	42	93	29	54	80	65	121	152	122	90.6	E
				153	100	21	102	22	28	158	114	116	97	155	103	50	40	22	95.8	F
					98	135	127	138	127	80	42	87	125	56	59	109	184	147	103.9	G
						94	32	99	87	68	77	24	29	89	47	84	115	83	75.3	H
							102	5	10	149	95	109	97	142	89	29	61	33	88.3	I
								107	95	94	103	53	26	121	78	97	110	85	89.8	K
									15	153	98	113	102	145	92	32	61	36	91.1	L
										142	87	101	91	135	81	21	67	36	84.4	M
											91	46	86	46	65	133	174	143	97.1	N
												76	103	74	38	68	147	110	86.0	P
													43	68	42	96	130	99	76.9	Q
														110	71	96	101	77	84.8	R
															58	124	177	144	100.0	S
																69	128	92	73.4	T
																	88	55	84.0	V
																		37	115.8	W
																			89.3	Y
13	30	34	26	9	29	17	26	15	25	25	16	26	37	16	20	12	36	34	22.3	A
	28	33	29	21	31	26	32	24	17	19	25	22	36	13	19	21	37	34	24.0	C
		7	35	33	35	28	34	25	31	14	40	22	39	25	29	28	45	34	28.1	D
			35	37	27	31	26	30	26	19	43	14	31	29	34	31	43	34	28.2	E
				29	18	22	28	19	24	24	27	24	34	25	28	26	13	13	23.9	F
					34	25	31	24	34	26	17	32	43	19	20	19	39	36	26.4	G
						28	27	25	30	24	36	27	31	28	34	31	25	23	27.1	H
							24	5	22	23	24	24	34	25	21	7	34	34	22.7	I
								23	24	27	31	21	14	31	34	26	34	34	26.3	K
									20	20	23	22	33	23	23	9	30	30	21.1	L
										21	31	13	28	22	25	23	31	32	23.9	M
											31	10	31	15	19	23	32	28	21.5	N
												33	43	24	25	20	37	37	28.1	P
													23	21	24	25	31	29	22.1	Q
														37	38	36	36	36	32.0	R
															12	20	35	29	22.4	S
																17	38	32	24.6	T
																	37	36	22.3	V
																		21	31.7	W
																			29.3	Y

Dissimilarity arrays from Grantham (17) (*Upper*) and Sneath (16) (*Lower*). $\langle D\rangle j$ is the average dissimilarity relative to residue j for random substitution.

RESULTS

The Wu–Kabat and structural variabilities are listed in Table 2 for each position in the sequences compared. Structural classifications and representative sequences are included in the table. The results are summarized in Table 3 where the average sequence and structural variabilities are listed for all the residues and for only the interior, exterior, or contact residues. Preliminary results of a similar analysis of light chain constant domain (C_L) sequences of human κ and λ, mouse κ and λ (15), rabbit κ (19), and pig λ (20) chains are included in Table 3 for purposes of comparison. The Wu–Kabat variability values listed in these tables were computed by using only the sequences considered in this paper and they differ from those calculated by Kabat *et al.* (14) on the basis of more (or different) sequences.

DISCUSSION

The analysis of sequence variability on the basis of the structural dissimilarities between amino acids has several advantages over the Wu–Kabat method (1). First, it permits a more correct weighting of sequence variations. Intuitively, larger differences

Table 2. Sequence and structural variability at each residue position for various regions

Seq. no.		s/c	WK	AS	AG
Light chain variable domain (M21)					
	N	0			
	I	3	10.0	41.6	48.6
	V	0	8.3	40.9	46.7
	M	4	8.9	45.9	9.4
	T	0	1.0	0.0	0.0
	Q	4	1.0	0.0	0.0
	S	1	11.1	47.9	36.3
	P	0			
	K	0			
10	S	0	8.6	27.2	31.4
	M	4	11.1	33.3	25.1
	S	0	5.7	24.1	25.1
	M	4	23.3	47.8	51.2
	S	0	6.2	24.1	27.3
	V	0	11.4	50.1	62.2
	G	0	1.0	0.0	0.0
	E	0	12.5	39.4	33.2
	R	0	8.9	65.5	57.1
	V	4	4.3	15.0	20.8
20	T	0	4.6	22.4	28.5
	L	4	2.5	4.4	1.1
	T	0	8.9	34.1	45.7
	C	4	1.0	0.0	0.0
				[L1]	
	W	4	1.0	0.0	0.0
	Y	C	5.7	20.1	10.0
	Q	3	3.5	10.2	10.6
	Q	C	3.0	12.9	11.9
	K	0	6.8	24.4	19.0
40	P	0	1.0	0.0	0.0
	E	0	5.1	27.3	19.1
	Q	1	12.7	45.9	26.4
	S	C	13.6	46.0	43.7
	P	C	6.3	18.1	21.4
	K	0	10.8	29.6	24.5
	L	C	7.1	32.6	33.4
	L	4	3.8	7.6	5.5
	I	4	1.0	0.0	0.0
	Y	C	5.0	18.3	25.6
				[L2]	
	G	0	1.0	0.0	0.0
	V	4	3.8	8.5	9.3
Heavy chain variable domain (M603)					
	E	0			
	V	2	4.0	33.0	34.1
	K	0	6.4	37.2	26.5
	L	4	1.0	0.0	0.0
	V	0	4.2	39.0	32.7
	E	2	2.7	12.4	8.0
	S	0	1.0	0.0	0.0
	G	4	1.0	0.0	0.0
	G	0	4.8	20.4	17.3
10	G	0	6.4	42.1	34.0
	L	0	2.7	10.6	8.8
	V	4	2.3	14.5	14.4
	Q	0	8.0	50.2	31.0
	P	0	1.0	0.0	0.0
	G	0	4.8	27.9	20.9
	G	0	24.0	72.0	55.0
	S	0	4.0	17.3	14.3
	L	4	4.0	24.8	18.4
	R	0	10.7	55.1	44.8
20	L	4	2.3	5.3	4.4
	S	0	3.2	20.0	21.8
	C	4	1.0	0.0	0.0
	A	0	13.3	52.7	51.0
	T	4	8.0	39.2	39.6
	S	0	2.3	6.7	7.3
	G	1	1.0	0.0	0.0
	F	4	4.0	21.9	22.8
	T	0	3.2	18.3	29.6
	F	4	4.8	31.3	8.5
30	S	0	2.3	6.7	7.3
				[H1]	
	W	4	1.0	0.0	0.0
	V	C	3.2	11.7	12.9
	R	3	1.0	0.0	0.0
	Q	C	1.0	0.0	0.0
40	P	2	4.8	32.5	22.1
	P	0	1.0	0.0	0.0
	G	0	2.3	17.5	11.8
	K	0	6.4	35.1	28.2
	R	0	6.4	39.3	37.5
	L	C	1.0	0.0	0.0

Various light chain hypervariable regions (AG)

L1 (11 residues long)						L2 (7 residues long)				
24	Q	0	10.0	50.2	32.5	50	D	1	60.7	86.2	78.9
25	A	4	3.9	11.7	14.0	51	A	1	15.1	47.4	41.9
26	S	0	2.3	12.4	7.2	52	S	0	5.6	24.2	16.1
27	Q	0	7.5	24.6	17.0	53	N	1	29.6	61.8	45.9
28	D	0	21.6	57.1	46.4	54	L	0	5.3	43.1	49.1
29	I	2	7.5	24.7	25.7	55	E	0	23.3	59.3	45.3
30	N	0	25.2	67.6	53.9	56	T	0	4.2	22.9	18.9
31	H	0	36.0	65.1	57.0						
32	Y	0	132.4	70.7	93.6						
33	L	4	5.5	17.1	20.6						
34	N	0	20.6	56.1	70.2						

Various heavy chain hypervariable regions (EU)

H1 (5 residues long)						H2 (17 residues long)				
31	R	0	11.4	65.1	46.4	50	G		16.5	67.7	65.2
32	S	4	8.9	39.5	47.4	51	I		9.2	43.7	32.6
33	A	2	42.7	81.4	82.6	52	V		25.7	74.1	79.0
34	I	4	7.3	28.4	14.1	52a	P		18.3	78.6	62.5
35	I	2	28.8	70.4	61.9	53	M		28.3	67.1	61.1
						54	F		8.8	47.5	41.2
						55	G		11.0	45.8	29.5
						56	P		22.0	65.7	49.6
						57	P		11.0	52.1	52.8

Seq. no.		s/c	WK	AS	AB
	P	1	2.4	12.8	12.9
60	D	0	18.7	50.9	38.3
	R	1	1.0	0.0	0.0
	F	4	1.0	0.0	0.0
	T	0	4.0	18.9	17.1
	G	4	1.0	0.0	0.0
	S	0	1.0	0.0	0.0
	G	4	9.2	38.8	37.5
	S	0	2.1	5.6	7.1
	A	0	3.8	15.9	14.8
	T	2	5.7	24.8	24.8
70	D	0	12.7	40.0	29.2
	F	4	5.0	38.0	37.7
	T	0	5.0	21.1	31.6
	L	4	2.1	4.5	1.2
	T	0	8.0	22.2	20.5
	I	4	3.3	2.6	1.9
	S	0	5.0	19.8	15.8
	S	0	17.1	65.5	48.0
	V	2	10.0	30.0	26.9
	E	2	1.0	0.0	0.0
	W	C	1.0	0.0	0.0
	I	4	10.7	30.8	21.3
	A	4	6.0	26.3	31.5
				[H2]	
	R	1	1.0	0.0	0.0
	F	4	4.8	37.1	15.9
	I	0	6.4	26.9	34.9
	V	4	2.7	7.7	8.2
70	S	0	2.7	13.4	14.5
	R	2	4.8	25.4	24.2
	D	0	4.8	25.4	14.8
	T	0	10.7	52.9	41.3
	S	0	6.4	30.6	23.5
	Q	0	8.0	54.6	36.9
	S	3	4.0	22.0	23.0
	I	0	8.0	49.8	43.9
	L	4	6.4	25.4	20.6
	Y	1	8.0	41.0	25.7
80	L	4	4.0	25.4	14.7

Seq. no.		s/c	WK	AS	AG
	Q	2	9.5	35.7	21.1
80	A	0	15.6	43.7	60.0
	E	0	11.2	21.6	24.2
	D	1	2.2	5.0	2.2
	L	2	23.3	81.6	94.9
	A	4	3.5	6.7	9.7
	D	c	20.0	70.2	79.0
	Y	4	1.0	0.0	0.0
	H	C	3.8	12.9	12.4
	C	4	1.0	0.0	0.0
				[L3]	
	F	C	1.0	0.0	0.0
	G	4	1.0	0.0	0.0
100	G	c	4.0	21.6	18.3
	G	0	1.0	0.0	0.0
	T	4	1.0	0.0	0.0
	K	0	10.4	35.6	27.1
	L	4	2.9	13.4	11.1
	E	0	5.5	52.7	41.6
	I	3	6.7	18.6	19.5
				[He]	
	D	2	1.0	0.0	0.0
	T	3	1.0	0.0	0.0
	A	4	1.0	0.0	0.0
	I	c	11.7	52.8	66.5
90	Y	4	1.0	0.0	0.0
	Y	C	2.8	12.7	7.0
	C	4	1.0	0.0	0.0
	A	c	1.0	0.0	0.0
	R	3	2.8	38.4	42.1
				[H3]	
	W	C	2.3	19.4	18.8
	G	4	2.3	19.5	20.2
	A	0	7.0	44.5	39.6
	G	4	1.0	0.0	0.0
	T	2	4.2	23.8	28.8
	T	0	5.3	46.6	51.6
	V	4	2.3	5.8	5.4
110	T	0	1.0	0.0	0.0
	V	4	1.0	0.0	0.0

L3 (9 residues long)

89	Q	C	11.4	51.6	53.4
90	Q	4	15.6	55.6	58.6
91	Y	c	16.3	59.8	60.5
92	D	2	41.8	78.1	98.9
93	T	0	19.0	71.7	60.0
94	L	0	59.7	80.8	80.9
95	P	0	19.0	59.8	53.8
96	R	C	34.2	71.4	52.6
97	T	3	6.3	34.2	47.9

He (8 residues long)

58	N		19.3	66.3	55.4	81	E	0	12.0	40.6	38.9
59	Y		2.4	8.1	4.5	82	L	4	2.5	16.7	3.6
60	A		3.7	14.7	15.1	82a	S	0	8.0	39.0	36.6
61	Q		14.7	60.2	44.9	82b	S	0	8.3	34.7	33.4
62	K		3.7	17.4	16.3	82c	L	4	2.9	12.8	10.5
63	F		6.3	22.8	19.3	83	R	0	8.0	46.9	38.3
64	Q		5.5	42.4	39.3	84	S	0	6.0	35.9	37.2
65	G		3.7	17.9	13.1	85	E	0	3.8	25.2	22.6

The structural classifications (*s/c*) are those assigned for the residues of M603 Fab (6): 0 = completely exposed to solvent; 1 = mainly exposed to solvent; 2 = partly exposed and partly buried in the domain interior; 3 = mainly buried; 4 = completely buried in the domain interior; *C* = in contact with the homologous domain; *c* = in the interdomain interface but not in contact with the homologous domain. WK, AS, and AG are short notations for the Wu–Kabat variability, ASD-S, and ASD-G, respectively. A representative sequence (of the protein identified in parentheses) is in the column between the sequence number and the structural classifications. No variability values are given where gaps occur in any of the sequences compared.

Table 3. Average sequence and structural variability for various framework and hypervariable regions

Region	Seq. no.	Residues	No. of positions	⟨WK⟩	⟨AS⟩	⟨AG⟩
C_L	7	All	101	6.1	29.4	26.4
		Interior	25	3.1	12.7	11.1
		Exterior	51	7.5	38.6	34.1
		Contact	17	7.0	29.3	28.3
V_L	20	Framework	76	6.4	22.3	21.0
		Interior	26	4.4	13.0	11.5
		Exterior	36	7.0	25.9	24.2
		Contact	8	5.8	20.1	19.8
V_H	8	Framework	76	4.2	20.4	17.6
		Interior	27	3.1	13.9	11.5
		Exterior	34	5.5	28.1	23.7
		Contact	6	1.9	7.3	6.4
L1	18	All	11	15.7	41.6	39.8
L2	28	All	7	20.5	49.3	42.3
L3	19	All	9	24.8	62.6	63.0
H1	16	All	5	19.8	57.0	50.5
H2	11	All	17	12.4	46.6	40.1
He	10	All	8	6.4	31.5	27.6

The interior, exterior, and contact residues are those assigned the structural classifications 3 or 4, 0 or 1, and *C*, respectively. ⟨WK⟩, ⟨AS⟩, and ⟨AG⟩ are the average Wu-Kabat variability, ASD-S, and ASD-G, respectively.

in the chemical natures of the residues involved in ligand binding would be expected to lead to greater differences in binding affinity or specificity. Second, it has permitted the evaluation, in structural terms, of sequence conservation and hypervariability. Third, the structural dissimilarity values are not as sensitive to the number of sequences as is the Wu-Kabat variability. In a comparison of N different sequences, the maximal attainable Wu-Kabat variability is $N \times N$ for $N < 20$ and 400 (20×20) for $N \geq 20$. In contrast, the maximal attainable ASD does not depend on N and can be achieved in a comparison of only a few sequences. The results, therefore, of two sequence analyses, one involving a certain number of sequences and the other involving a different number, can be compared to each other.

The results listed in Table 2 show that the Wu-Kabat and structural variabilities by and large are correlated. With few exceptions, ASD-S and ASD-G show a high degree of parallelism although they frequently disagree in the designation of whether the replacements at a given position are conservative, hypervariable or not. There are instances when the Wu-Kabat variability does not follow structural dissimilarity. For example, in V_L, positions 60 and 105 have Wu-Kabat variabilities of 18.7 and 5.5, respectively; yet their ASD-S and ASD-G values are 50.9 and 38.3, and 52.7 and 41.6, respectively. The Wu-Kabat variability of 18.7 is the fourth highest in the V_L framework comparison (Table 2) whereas 5.5 is close to the average. Furthermore, position 13, which has a Wu-Kabat variability of 23.3 (the highest in the V_L framework comparison), has ASD-S and ASD-G of 47.8 and 51.2, respectively; these are lower than the ASD-S and ASD-G (65.5 and 57.1, respectively) for position 18, which has a Wu-Kabat variability of 8.9. Of course, the sequence and structural variabilities depend on the particular sequences being compared. For example, in a comparison of all the sequences available at the time, Kabat *et al.* (14) obtained Wu-Kabat variabilities of 22, 29 (or 31 depending on the identity of the Asx and Glx), 20 (or 22), and 13 (or 17) for positions 13, 18, 60, and 105, respectively. The sequences used in this paper were chosen to minimize structural similarity due to genetic influences.

The results summarized in Table 3 show that the average sequence and structural variabilities for interior residues are significantly lower than those for exterior residues. Moreover, the variabilities for C_L and the framework regions of V_L and V_H are comparable, with the exception of the residues in contact with the homologous domain. In V_H, the contact residues have variabilities lower than those computed for the interior residues; in C_L and V_L, the variabilities for the contact residues are much higher and are close to the average for all the residues. The results in Table 3 further show that the average variabilities computed for the HV regions are higher than those for C_L or for the framework regions of V_L or V_H.

The comparison of V_L framework sequences (Table 2) shows that 11 positions—namely, 7, 13, 15, 18, 42, 43, 60, 77, 83, 85, and 105—have nonconservative replacements. Of these, seven are exposed to solvent. One interior position, residue 13, has nonconservative replacements. Although this residue is interior in M603, a rotation about the C_α—C_β bond could easily swing the side group to the exterior. It is interesting that many rabbit κ chains have Glu at this position and these almost invariably have Pro at position 14 (14). The presence of Pro at 14 would cause a slight restructuring of the polypeptide backbone which will facilitate the exposure of Glu at 13 to solvent. One contact position, residue 43, has nonconservative replacements. In addition, residue 85, which in M603 is in the interdomain interface but is not in contact with V_H, shows structural hypervariability. One partly exposed, partly buried residue, position 83, shows structural hypervariability. Position 77, which has nonconservative substitutions, is in a region homologous to the He region in heavy chains.

The comparison of V_H framework sequences (Table 2) shows that nine positions have nonconservative replacements. Eight of these are the exposed positions 13, 16, 19, 23, 73, 75, 77, and 108, with position 16 showing structural hypervariability. One position, 89, whose side group is in the interdomain interface but is not in contact with V_H, has nonconservative replacements.

The H1, He, L2, and L3 regions being compared here have the same lengths as the corresponding regions in M603 Fab (11) so that structural classifications, based on the M603 structure, are available for these regions (6). The structural classifications for the L1 region 11 residues long are based on the model of the rabbit BS-5 antibody (5) whose L1 was built to approximate as closely as possible the known L1 structure of protein REI (21) which has the same length. No V_H structure with 17 residues in H2 has been elucidated by x-ray diffraction. HV residues, which in M603 Fab are in contact with the homologous domain (11), will be considered simply as exterior in this discussion.

The comparisons of the HV regions (Table 2) show that, by and large, structural variability at a given position is correlated with whether the residue at that position is buried in the domain interior or not. This is seen in the comparison of H1 sequences 5 residues long. Two positions, 32 and 34, are buried and the replacements at both positions are conservative. In contrast, the exposed positions 31, 33, and 35 show high structural variability. Furthermore, positions 33 and 35 show high Wu-Kabat variability and structural hypervariability. In M603 Fab, the residues at heavy chain positions 33 and 35 are intimately involved in phosphocholine binding (11, 22).

The comparison of L3 sequences 9 residues long shows high structural variability for eight of the nine positions. Position 97, which is mainly buried in the domain interior, has conservative

replacements. Four positions, 92, 93, 94, and 96, show structural hypervariability. The residue at position 96 appears to be involved in the binding of phosphocholine in M603 (22). Ligand binding studies of NEWM Fab implicate the residue at position 91 in the interaction with the vitamin K_1OH ligand (12).

The comparison of L2 sequences 7 residues long reveals only one position, 50, showing structural hypervariability. In L2, the residue at position 50 is nearest to the tip of the Fab, closest to the other HV regions, and, by virtue of its location, has the greatest potential involvement in antigen binding. Three positions, 52, 54, and 56, have conservative replacements. Here, no correlation is apparent between structural variability and whether the residue is buried or not because all the residues in L2 are either mainly or completely exposed to solvent.

Structural hypervariability is not observed in any He position in the comparison of sequences from proteins with different specificities. On the contrary, with the exception of position 83, all He positions have conservative replacements.

The comparison of L1 sequences 11 residues long shows five positions, 25–27, 29, and 33, with conservative replacements and two, 30 and 32, displaying structural hypervariability. On the basis of the structural classifications from the model of BS-5 antibody (5), both structurally HV positions are completely exposed to solvent and two positions with conservative replacements, 25 and 33, are completely buried in the domain interior. Four of the five positions with nonconservative replacements are on the exterior.

The comparison of H2 sequences 17 residues long shows seven positions with conservative replacements, 51, 59, 60, and 62–65, and four showing structural hypervariability, 50, 52, 52a, and 53. On comparing the only two V_H domains whose structures have been elucidated by x-ray diffraction (11, 23) and that have different lengths in H2, it would appear that the initial segments (including at least the first three residues) and the terminal segments (including at least the last six residues) of the two known H2 structures are very similar, the difference being mainly in the bend region where the two proteins differ by three residues. If the initial and terminal segments of the H2 regions being compared here are also similar structurally to the two H2 regions already solved, then there appears to be a correlation between the structural hypervariability observed in some H2 positions and their potential involvement in specificity determination. For example, the residues at positions 50, 52, and 58 in M603 Fab are apparently crucial in producing the complementarity between the binding cavity and phosphocholine (22). In the present analysis, the replacements at positions 50 and 52 show structural hypervariability and those at position 58 are nonconservative but with variability values just below the minimal measure of hypervariability.

CONCLUSION

The results obtained indicate that, in immunoglobulins also, structural variability is correlated with whether a residue is in the protein interior or on the exterior. Furthermore, structural hypervariability at certain positions appears to be correlated with the involvement of these positions in ligand binding. This suggests that the technique presented might be useful in predicting which residues are potential ligand-contacting residues. One would predict from the results, for example, that the residues at positions 30 and 32 in light chains with L1 regions 11 residues long probably will be involved in antigen binding and may even be crucial in providing specificity. Similarly, the residues at heavy chain positions 50 and 52–53 could also be involved in antigen binding in antibodies with 17 residues in their H2.

The results of the comparison of He region sequences, although based on only 10 sequences, suggest that this region may not be involved in specificity determination. Topographically, He is contiguous with the main HV surface, being adjacent to the terminal segment of H2 (4). The polypeptide segment containing He, however, is a bend physically located at the carboxyl end of V_H near the switch region. At this location, He is far removed from the main binding site region and may only play an ancillary role in the binding of ligand. Alternatively, He may not be involved in binding at all and the observed sequence variability may reflect some other function for He—for example, in recognition. A similar role is conceivable for other positions that show high structural variability but are outside the complementarity-determining regions.

I thank Drs. David R. Davies and Elvin A. Kabat for comments and discussions.

The costs of publication of this article were defrayed in part by the payment of page charges from funds made available to support the research which is the subject of the article. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

1. Wu, T. T. & Kabat, E. A. (1970) *J. Exp. Med.* **132**, 211–250.
2. Kabat, E. A. & Wu, T. T. (1971) *Ann. N.Y. Acad. Sci.* **190**, 382–393.
3. Capra, J. D. & Kehoe, J. M. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 845–848.
4. Davies, D. R., Padlan, E. A. & Segal, D. M. (1975) *Ann. Rev. Biochem.* **44**, 639–667.
5. Davies, D. R. & Padlan, E. A. (1977) in *Antibodies in Human Diagnosis and Therapy*, eds. Haber, E. & Krause, R. M. (Raven Press, New York), pp. 119–132.
6. Padlan, E. A., Davies, D. R., Pecht, I., Givol, D. & Wright, C. (1976) *Cold Spring Harbor Symp. Quant. Biol.*, in press.
7. Margolies, M. N., Cannon, L. E., III, Strosberg, S. D. & Haber, E. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 2180–2184.
8. Kabat, E. A., Wu, T. T. & Bilofsky, H. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 4471–4473.
9. Kabat, E. A., Wu, T. T. & Bilofsky, H. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 617–619.
10. Padlan, E. A. & Davies, D. R. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 819–823.
11. Segal, D. M., Padlan, E. A., Cohen, G. H., Rudikoff, S., Potter, M. & Davies, D. R. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 4298–4302.
12. Amzel, L. M., Poljak, R. J., Saul, F., Varga, J. M. & Richards, F. F. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1427–1430.
13. Edmundson, A. B., Ely, K. R., Girling, R. L., Abola, E. E., Schiffer, M., Westholm, F. A., Fausch, M. D., & Deutsch, H. F. (1974) *Biochemistry* **13**, 3816–3827.
14. Kabat, E. A., Wu, T. T. & Bilofsky, H. (1976) *Variable Regions of Immunoglobulin Chains: Tabulations and Analyses of Amino Acid Sequences* (Bolt Beranek & Newman, Inc., Cambridge MA).
15. Kabat, E. A., Padlan, E. A. & Davies, D. R. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 2785–2788.
16. Sneath, P. H. A. (1966) *J. Theor. Biol.* **12**, 157–195.
17. Grantham, R. (1974) *Science* **185**, 862–864.
18. Dayhoff, M. O., Eck, R. V. & Park, C. M. (1972) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, pp. 89–99.
19. Chen, K. C. S., Kindt, T. J. & Krause, R. M. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1995–1998.
20. Novotny, J. & Franek, F. (1975) *Nature* **258**, 641–643.
21. Epp, O., Colman, P., Fehlhammer, H., Bode, W., Schiffer, M. & Huber, R. (1974) *Eur. J. Biochem.* **45**, 513–524.
22. Padlan, E. A., Davies, D. R., Rudikoff, S. & Potter, M. (1976) *Immunochemistry* **13**, 945–949.
23. Poljak, R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizackerley, R. P. & Saul, F. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 3305–3310.