Common Features of the Conformations of Antigen-Binding Loops in Immunoglobulins and Application to Modeling Loop Conformations

Anna Tramontano[1] **and Arthur M. Lesk**[2]
[1]Istituto di Ricerche di Biologia Molecolare, 00040 Pomezia, Roma, Italy and [2]Department of Haematology, University of Cambridge Clinical School, MRC Centre, Cambridge CB2 2QH, England

ABSTRACT **Using database screening techniques we have examined the relationship between antigen-binding loops in immunoglobulins, and regions of similar conformation in other protein families. The conformations of most antigen-binding loops are not unique to immunoglobulins. But in many cases, the geometrical relationship between the loop and the peptides flanking it differs between the immunoglobulins and other structures with the same loop. We assess model building by data base screening, compared with that based on canonical structures.** © 1992 Wiley-Liss, Inc.

Key words: **protein structure, modeling, immunoglobulins, loops, data base screening**

INTRODUCTION

We are interested in understanding the molecular basis of the immune response, and, more particularly, in relating sequence to structure in the antigen-binding sites of immunoglobulins. A test and application of this understanding is how well we can predict the structures of antigen-binding domains. From overall sequence divergence we can estimate the similarity of the frameworks of immunoglobulin variable domains.[1–2] Predictions of loop conformations could be based on general sequence-structure relationships in loops[3–7] and features specific to antigen-binding loops,[6,8–11] or alternatively we could try to import nonhomologous loops, using techniques developed by Jones and Thirup.[12] The goals of this study are first, to understand the relationship between the conformation, structural context, and stabilizing interactions of antigen-binding loops and loops of similar conformation in other protein families and, second, to see to what extent loops of similar conformation in other protein families can be identified and used in modeling antigen-binding sites.

Immunoglobulins (Igs) are composed of four chains containing variants of a basic folding unit (Fig. 1A). In IgGs the light chain contains a variable domain (V_L) and a constant domain (C_L), and the heavy chain contains a variable domain (V_H) and three constant domains (C_{H^1}, C_{H^2}, and C_{H^3}). The domains contain two β-sheets packed face to face, and the V_L and V_H domains pack together similarly in different IgGs.[13,14] The V_L and V_H domains contain six hypervariable loops, clustered together in space to form the antigen-binding site (Fig. 1B). Variations in sequence and structure of these regions give antibodies their great range of specificity and affinity.

Four of the antigen-binding loops—L2, L3, H2, and H3—are hairpins. L1 and H1 join one sheet of either domain to the other. For five of the six loops, there is a limited repertoire of "canonical structures," each stabilized by specific packing interactions, hydrogen bonding, or ability to assume special conformations, of a few particular residues.[6,8–11]

For database screening to be a useful tool for modeling antigen-binding loops, it must be shown, first that the loop conformations occur in other known protein structures, and second, that the relationship between the loop and the flanking peptides is similar in immunoglobulins and other proteins. In this work we studied the uniqueness to immunoglobulins of the conformations of the antigen-binding loops L1, L2, L3, H1, and H2, and of the relationship of the loops to their stems (the regions flanking the loop.) Where loops of similar conformation appear in immunoglobulins and other proteins, we compared their structural contexts and the interactions that stabilize their conformations. We also compared the loops identified by data base screening with the classification of loops according to canonical structures.

Many, but not all, of the loop conformations can be found in other proteins and, in some cases, the best-fitting regions come from structures other than immunoglobulins. In some cases we picked up standard hairpins,[3–7] but in others, the same loop appears in quite different structural contexts. However, there is great variability in the relationship of the loop to the stem.

Received February 26, 1991; revision accepted June 26, 1991.

Address reprint requests to Dr. Anna Tramontano, Istituto di Richerche di Biologia Molecolare, Via Pontina Km 30.600, 00400 Pomezia, Roma, Italy.

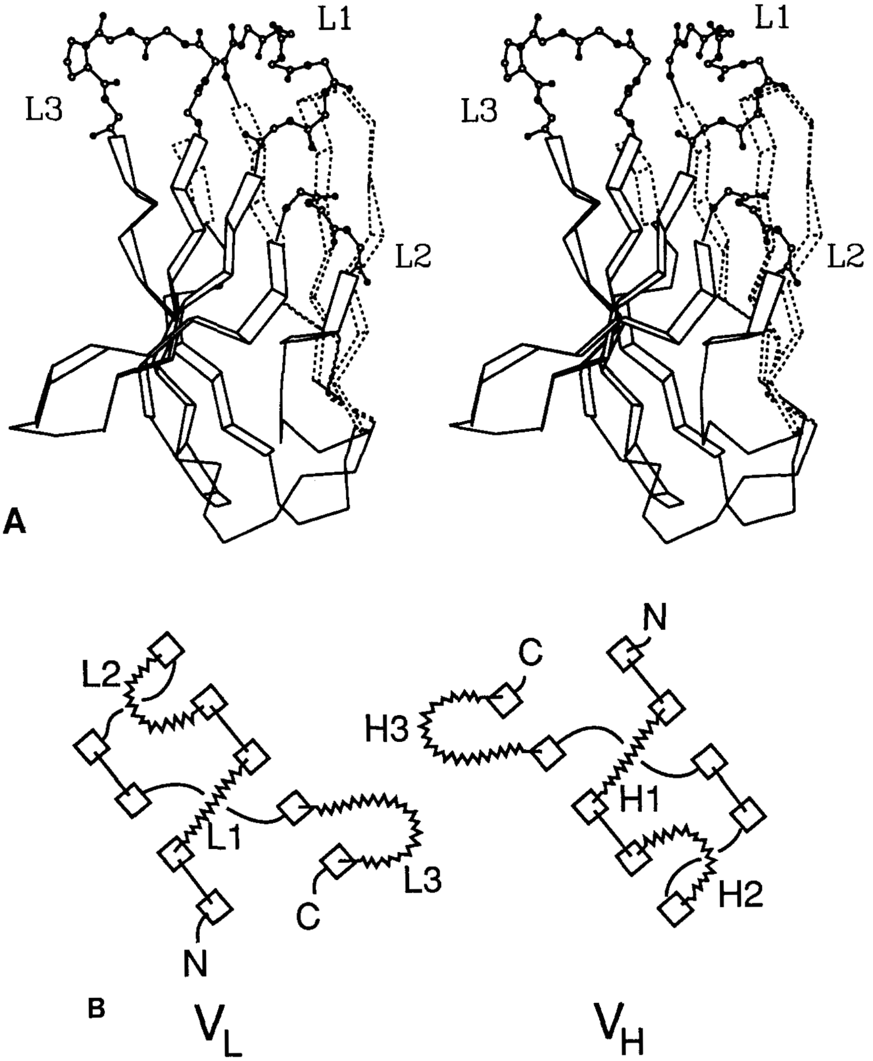

Fig. 1. **(A)** The structure of an immunoglobulin V domain (KOL V_L). Strands of β-sheet appear as ribbons. L1, L2, and L3 are the hypervariable loops. V_H domains and their hypervariable regions, H1, H2, and H3 have homologous structures. **(B)** A schematic representation of the hypervariable loops in antigen-binding sites of immunoglobulins, looking into the antigen-binding site.

MATERIALS AND METHODS

Protein structures used were distributed by the Protein Data Bank in the September 1990 release.[15] We studied immunoglobulin variable domains in the data bank solved at 2.7 Å resolution or better (see Table I). Table II contains the definitions of the loops, with residues in each light chain numbered sequentially from 1 and the residues in each heavy chain numbered sequentially from 301. A conversion to the residue numbering of Kabat et al.[24] appears in Table II.

The programs used[25] offer the following facilities: The user may specify within a selected structure one, two, or more regions of consecutive residues. If more than one region is specified, the length of the gap(s) between them must be specified; these need not be the same lengths as the gaps in the original structure. The program searches in the data base for regions in other structures matching the selected regions, with gaps of the specified lengths. The criterion for matching is a threshold on the value of the root mean square (rms) deviation Δ after optimal superposition. Optionally, the program will fit α-carbons only, main chain atoms, or all atoms (rarely a useful facility). The user may assign different weights to different residues. The algorithm is described in the appendix. The program reports the best fits found within a specified threshold of rms deviation Δ, sorted in order of increasing Δ. (If no fit within the threshold exists, the program reports the best fit found.)

We applied these search techniques in three ways.

1. To search for loops, we specified the main chain N, C_α, and C atoms of one of the antigen-binding

TABLE I. Immunoglobulin Variable Domains of Known Atomic Structure in the Protein Data Bank,[15] September 1990 Release, Determined at a Resolution of 2.7 Å or Better

	Chain type		Reference	Protein Data Bank designation
Molecule	L	H		
Fab'NEWM	λI	γII	16	3FAB
Fab KOL	λI	γIII	17	2FB4
V_L RHE	λI		18	2RHE
Fab McPC603	κ	γIII	19	1MCP
Fab J539	κ	γIII	20	2FBJ
Fab HyHEL-5	κ	γII	21	2HFL
Fab 4-4-20	κ	γII	22	4FAB
V_L REI	κ		23	1REI

loops in a known structure. Given the short length (3 residues) of the L2 loop, in this case we included one residue on either side of the loop and searched for five-residue fragments.

2. To study the relationship between the loop and its "stem," we specified two sets of residues, one starting from the residue preceding a loop and extending "backwards" for N residues, and the other starting from the residue following the loop and extending "forward" for N residues. This search identified regions from the data base that matched the stems of the loops in structure and spacing in the sequence, even if the intervening region (corresponding to the loop itself) did not match in structure.

In the stem searches we used C_α atoms only, and tested different possible values of the parameters: we used 3 and 4 for the length N of the flanking regions and assigned each residue a weight according to its distance from the loop: a residue adjacent to the loop has weight 1.0, the next residue has weight x, the next x^2 etc., where we explored $x = 0.3$, 0.5, 0.8, and 1.0 ($x = 1.0$ corresponds to uniform weights.) The search with four-residue flanking regions and a ratio of $x = 0.8$ gave the best results.

When well-fitting stems were identified, we fit intervening residues using all main chain atoms—N, C_α, C, O—with equal weights.

3. To search for loop and stem together, we defined a region by extending the loop three or four residues in both directions.

RESULTS

The structures of 41 hypervariable regions from eight different proteins are known from crystal structures solved to a resolution of 2.7 Å or better (see Tables I and II). The Fab fragments have six loops except for NEWM, from which L2 is deleted; the Bence–Jones proteins V_L REI, and V_LRHE contain light chains only.

Searches for Loop Regions

Table III shows the results of the searches for the hypervariable loops. We report the five best fits to the loop within the family of immunoglobulins and the five best fits in nonimmunoglobulin structures. In most cases many other regions of comparable rms deviation were also found. We also found fits to nonhomologous regions in immunoglobulin structures, but will not discuss these here.

In most cases, the best fit is to the homologous region of another immunoglobulin. However, in some cases no homolog of a loop with the same length exists in other immunoglobulins of known structure. As more and more immunoglobulin structures are determined, such cases will become rarer. In all but three cases, a loop of similar conformation exists in a protein foreign to the immunoglobulin family. The three exceptional cases are the L1 loops of McPC603 and 4–4–20, which are unusually long, and the L1 loop of NEWM.

Searches for Regions Flanking Loops, or Stems

Table IV contains the results of the searches for the stems. In a few cases, a low value of Δ for the stem is associated with a low value of Δ for the intervening region; for example, for L3 of NEWM, residues 259–264 of penicillopepsin (2APP) fit with $\Delta_{stem} = 0.4$ Å and $\Delta_{loop} = 0.7$ Å. In other cases, there is a good fit to the stem but a poor fit to the loop; for example, for L3 of McPC603, residues 258–271 of pepsinogen (1PSG) fit with $\Delta_{stem} = 0.5$ Å, but $\Delta_{loop} = 4.4$ Å.

DISCUSSION

Uniqueness of Hypervariable Loops to the Immunoglobulin Family

The conformations of short hairpins tend to follow general rules based on the sequences of residues in the loop.[3–7] These observations and the results of Jones and Thirup[11] suggest that the hairpin loops in immunoglobulins, except when unusually long, ought not to be expected to be unique. The cases of V_κ, L3 loops, V_H H1 loops and H3 from HyHEL-5 were described in ref. 10. However, it was not clear whether L1 and H1 would be unique to β-sheet proteins with the immunoglobulin topology.

V_λ L1 loops

The L1 loops of V_λ domains have the unusual feature, among regions bridging the sheets of parallel double-β-sheet proteins, of penetrating deeply between the sheets.[26] A large hydrophobic sidechain at position 30 points into the core of the molecule, and is packed in a cavity formed by framework residues 25, 33, and 71.

The cytochrome subunit of the photoreaction center from *Rhodopseudomonas viridis* is a primarily

TABLE II. Residues Defining Antigen-Binding Loops, and Their Lengths

	Residues length								
		L1			L2			L3	
NEWM	25	34	10		deleted		86	91	6
KOL	25	33	9	51	53	3	92	99	8
RHE	25	33	9	51	53	3	92	99	8
McPC603	26	38	13	56	58	3	97	102	6
J539	26	31	6	49	51	3	90	95	6
HyHEL-5	26	31	6	49	51	3	90	94	5
4–4–20	26	37	12	55	57	3	96	101	6
REI	26	32	7	50	52	3	91	96	6
Kabat numbering[24]	26	32		50	52		91	98	
		H1			H2			H3	
NEWM	326	332	7	353	355	3	399	405	7
KOL	326	332	7	352	357	6	400	414	15
McPC603	326	332	7	353	358	6	402	410	9
J539	326	332	7	353	356	4	400	406	7
HyHEL-5	326	332	7	353	356	4	399	403	3
4-4-20	336	332	7	353	358	6	402	406	5
Kabat numbering[24]	26	32		52a	55		95	100	

α-helical protein containing four heme groups.[27] A region of this subunit is similar in conformation to the V_λ L1 loops of RHE and KOL. The rms deviation of all N, C_α, C, O atoms is 0.8 Å. Figure 2A shows the superposition of the L1 loop of RHE with this region of the reaction center cytochrome. Corresponding to the deeply packed residue Ile-30 in RHE there is an inward-pointing Phe side chain in the cytochrome. Figure 2B shows the structural role of the region in the reaction centre cytochrome. It is part of a long turn arching over an α-helix, not entirely unlike a bracket holding a pipe against a wall. This helix is bound to the heme group.

These loops are members of a general class of loop, characterized by rather long end-to-end distances, stabilized by packing of a large, hydrophobic, inward-pointing residue.[10]

The H1 loops of V_γ domains also connect two different β-sheets. The similarity of the most common H1 conformation to a region in *Chironomus* erythrocruorin was described in ref. 10. The H1 loops of NEWM and HyHEL-10 have a distorted version of this conformation.[9] Figure 3A shows the superposition of H1 from NEWM (residues 326–332) with residues 42–48 of actinoxanthin (1ACX).[28] However, the structural context is completely different: in actinoxanthin this region is a rather extended bridge between two domains (Fig. 3B).

Two of the H2 loops present interesting features. Figure 4A shows the superposition of McPC603 H2 with residues 276–281 of *Alcaligenes denitrificans* azurin (2AZA).[29] The conformation of this loop is unusual because it is long: in McPC603 it is part of a 10-residue hairpin. In McPC603 Lys 357 is in a $\phi > 0$, $\psi > 0$ conformation. (It is interesting that this Lys arises by somatic mutation from a germ-line gene that codes for a Gly at this position.) In this azurin loop the corresponding residue, Asp-280, is also in a $\phi > 0$, $\psi > 0$ conformation, but so are 276 Gly and 281 Tyr. These regions of similar structure in McPC603 and azurin both contain a tyrosine in the last position. In both structures the ring of the tyrosine is approximately parallel to the plane of a peptide, a juxtaposition of polarizable unsaturated groups that might provide a favourable stacking interaction (Fig. 4B and C).

Figure 5 shows the superposition of HyHEL-5 H2 and residues 167–172 of garden pea lectin (2LTN).[30] In the middle residues of these turns, HyHEL-5 has the sequence Pro-Gly-Ser-Gly. This region adopts a conformation quite close to that expected for a four-residue X-X-X-G turn. Pea lectin has the sequence Ala-Ala-Tyr-Asn, with the asparagine in a $\phi > 0$, $\psi > 0$ conformation. This example shows that the presence of a glycine at a particular position is not an essential requirement for a conformation of a loop in which a residue has the α_L conformation (see also ref. 31). This makes it more difficult to use sequence cues to select loops of proper conformation from a set of choices spanning equivalent endpoints.

Uniqueness to the Immunoglobulin Family of the Relationship Between Loop and Stem

The use of database searches for model building will produce acceptable structures only if there is good structural similarity for the entire region spanning the loop and the stem. We searched the data bank for structures consisting of the mainchain atoms (N, C_α, C; carbonyl oxygens omitted) of the antigen-binding loops (Table II) extended by four residues at both ends. Good fits to these extended loops are common among the homologous regions

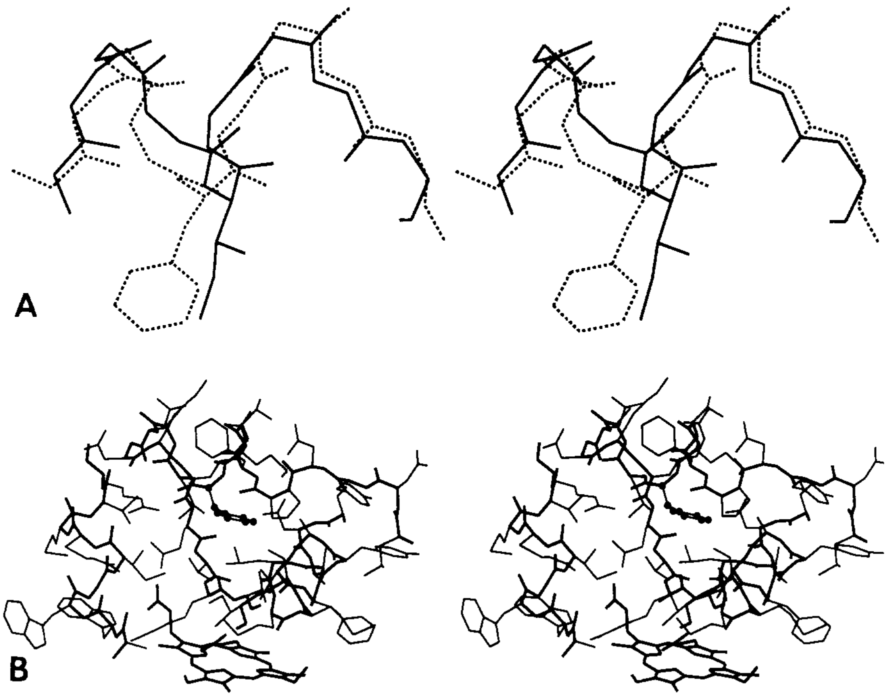

Fig. 2. Superposition of L1 hypervariable regions from V_λ RHE, and a region of similar conformation from the cytochrome from the reaction center of *Rhodopseudomonas virdis* (1PRC). Recall that L1 is not a hairpin but links strands from different sheets within a domain. **(A)** Superposition of the backbones, showing the corresponding inward-pointing side chains of that stabilize the conformations of the regions: V_λ RHE, residues 25–33 (solid lines); cytochrome, residues 189–197 (broken lines). **(B)** Structural context of this region in the reaction center cytochrome. Backbone, and heme group, drawn in bolder lines.

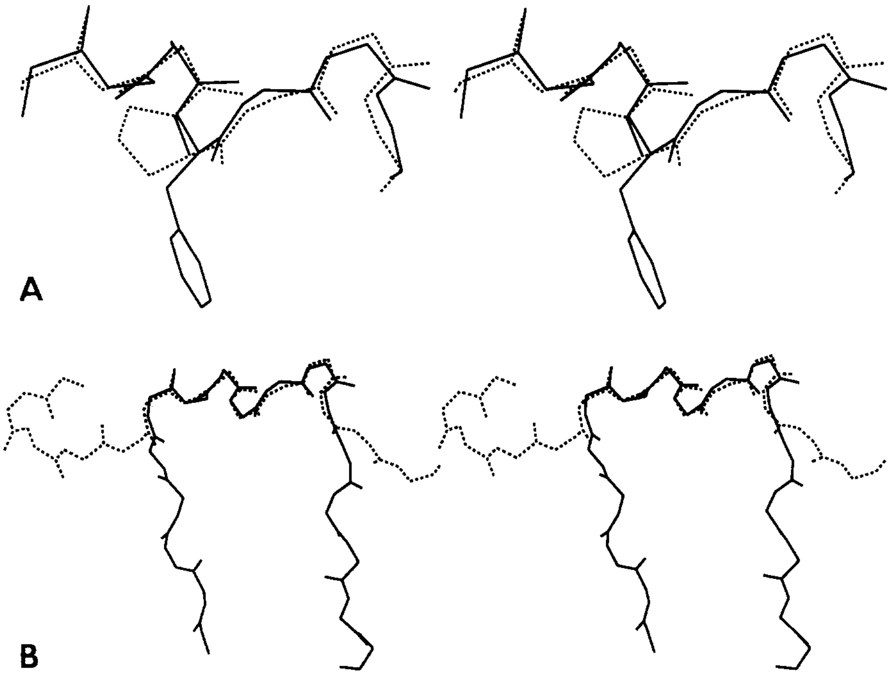

Fig. 3. **(A)** Superposition of the H1 loop of NEWM (solid) and residues 42–48 of actinoxanthin (1ACX) (broken). **(B)** The residues flanking the regions of similar conformation, showing how the structures diverge outside the limited region.

TABLE III. Best Fits to Hypervariable Loops in Immunoglobulins Found by Screening the Data Base*

	L1				L2				L3			
2RHE		25	SATDIGSNS			50	YYNDL			92	WNDSLDEP	
	2FB4	25	TSSNIGSST	0.24	2FB4	50	YRDAM	0.14	2IG2	92	WDVSLNAY	0.58
					2FBJ	48	YEISK	0.14				
					1REI	49	YEASN	0.26				
					2MCP	55	YGAST	0.35				
					4FAB	54	YKVSN	0.37				
	1PRC	C189	PFTMFLAND	0.80	2GD1	300	IDGKM	0.57	4PTI	23	YNAKAGLC	0.39
					1LDX	12	VAQQS	0.65	2PAB	8	LDAVRGSP	0.46
					3FXN	56	MGDEV	0.67	6API	212	HCKKLSSW	0.46
					2ALP	11	INNAS	0.69	4APE	242	SSSSVGGY	0.50
					8CAT	41	VGPRG	0.75	1CLA	96	FHQETETF	0.57
2FB4		25	TSSNIGSST			50	YRDAM			92	WDVSLNAY	
	2RHE	25	SATDIGSNS	0.24	2FBJ	48	YEISK	0.10	2RHE	92	WNDSLDEP	0.64
					2RHE	50	YYDNL	0.14				
					1REI	49	YEASN	0.24				
					4FAB	54	YKVSN	0.32				
					2MCP	55	YGAST	0.32				
	1PRC	C189	PFTMFLAND	0.82	2GD1	300	IDGKM	0.52	1CLA	96	FHQETETF	0.25
					2ALP	11	INNAS	0.63	5PTI	23	YNAKAGLC	0.29
					1LDX	12	VAQQS	0.66	1LZ1	45	YNAGDRST	0.30
					3P2P	25	YGCYC	0.68	5PEP	314	FDRANNKV	0.33
					3FXN	56	MGDEV	0.69	1PRC	H173	VDRSEHYF	0.33
1MCP		26	SQSLLNSGNQKNF			55	YGAST			97	DHSYPL	
					2HFL	48	YDTSK	0.26	3HFM	91	SNSWPY	0.37
					2FB4	50	YRDAM	0.35	1REI	91	YQSLPY	0.40
					2RHE	50	YYNDL	0.37	4FAB	96	STHVPW	0.62
					2FBJ	48	YEISK	0.37				
					1REI	49	YEASN	0.39				
					2GD1	300	IDGKM	0.57	2TBV	255	VSSLPA	0.73
					3P2P	25	YGCYC	0.62	1CAC	131	VQQPDG	0.77
					8CAT	41	VGPRG	0.65	3XIA	205	LERPEL	0.77
					2ALP	11	INNAS	0.68	2CPP	85	CPFIPR	0.79
					4I1B	21	SGPYE	0.72	2CDV	12	DKTKQP	0.81
3FAB		25	SSSNIGAGNH				deleted			86	YDRSLR	
									2FBJ	90	WTYPLI	0.86
									2APP	259	SISGYT	0.34
									2GRS	124	EVSGKK	0.34
									2APR	280	EFQGQC	0.36
									4PEP	90	QVGGIS	0.36
									1LDX	206	NNAGVL	0.36
1REI		26	SQDIIKY			49	YEASN			91	YQSLPY	
	1F19	26	SQDISNY	0.51	2FBJ	48	YEISK	0.19	3HFM	91	SNSWPY	0.23
	3HFM	26	SQSIGNN	0.58	2FB4	50	YRDAM	0.24	1MCP	97	DHSYPL	0.40
					2RHE	50	YYNDL	0.27	4FAB	96	STHVPW	0.57
					4FAB	54	YKVSN	0.31				
					2MCP	55	YGAST	0.37				
	1PRC	M97	KAQYGMG	0.59	2GD1	300	IDGKM	0.42	2CDV	12	DKTKQP	0.66
	1L31	51	GRNCNGV	0.66	2ALP	11	INNAS	0.51	1CAC	131	VQQPDG	0.72
	3GPD	15	DEVVSDD	0.73	5PEP	9	YLDTE	0.64	1FX1	95	SSYEYF	0.75
	4ATC	49	FEASTRT	0.81	2BP2	25	YGCYC	0.65	3CA2	167	IKTKGK	0.75
	1HMG	571	SEVEGRI	0.82	2FD1	11	CKYTD	0.71	1PHH	132	LQGERP	0.77
2FBJ		26	SSSVSS			48	YEISK			90	WTYPLI	
	2HFL	26	SSSVNY	0.48	2FB4	50	YRDAM	0.10	3FAB	86	YDRSLR	0.86
					2RHE	50	YYNDL	0.10				
					1REI	49	YEASN	0.19				
					4FAB	54	YKVSN	0.28				
					2MCP	55	YGAST	0.34				

(continued)

TABLE III. Best Fits to Hypervariable Loops in Immunoglobulins Found by Screening the Data Base* (Continued)

	L1				L2				L3			
	2GPD	118	SAPSAD	0.31	2GD1	300	IDGKM	0.49	2RHE	67	KSGTSA	0.50
	1GCR	19	SSDCPN	0.40	2ALP	11	INNAS	0.61	2PLV	417	TLGNST	0.51
	2PKA	57	FENENT	0.51	1LDX	12	VAQQS	0.68	1CMS	278	QDQGFC	0.53
	8CAT	172	HLKDPD	0.56	4FAB	155	IDGSE	0.70	3BCL	99	AVGSFA	0.55
	3RP2	61	RKAEST	0.57	3P2P	25	YGCYC	0.71	1PHH	277	QHGRLF	0.56
2HFL		26	SSSVNY			48	YDTSK			90	WGRNP	
	2FBJ	26	SSSVSS	0.48	2MCP	55	YGAST	0.22				
					2FB4	50	YRDAM	0.36				
					2RHE	50	YYNDL	0.38				
					2FBJ	48	YEISK	0.40				
					1F19	49	YYTSR	0.41				
	2GCR	19	SSDHSN	0.27	3P2P	25	YGCYC	0.55	3CPP	393	SGIVS	0.28
	2GCR	106	TEDCSS	0.41	8CAT	41	VGPRG	0.60	1HKG	209	FGSGV	0.36
	2RSP	22	SHPVKQ	0.42	1PRC	H6	LAQHL	0.61	2TAA	389	DDTTI	0.53
	2GPD	118	SAPSAD	0.44	4I1B	21	SGPYE	0.62	1LYM	53	YGILQ	0.56
	2PKA	57	FENENT	0.50	2GD1	300	IDGKM	0.66	2LTN	228	TGAEY	0.57
4FAB		26	SQSLVHSQGNTY			54	YKVSN			96	STHVPW	
					2FBJ	48	YEISK	0.28	3HFM	91	SNSWPY	0.48
					1REI	49	YEASN	0.31	1REI	91	YQSLPY	0.57
					2FB4	50	YRDAM	0.32	1MCP	97	DHWYPL	0.62
					2RHE	50	YYNDL	0.37				
					2MCP	55	YGAST	0.45				
					2GD1	300	IDGKM	0.38	3CA2	135	VQQPDG	0.70
					2ALP	11	IINAS	0.58	5LDH	219	DNDSEN	0.73
					5PEP	9	YLDTE	0.62	1FX1	95	SSYEYF	0.73
					7CAT	41	VGPRG	0.69	4XIA	703	EQLEHG	0.75
					3P2P	25	YGCYC	0.70	3P2P	56	KNLSGC	0.76

	H1				H2			
1MCP		326	GFTFSDF			353	NKGNKY	
	2HFL	326	GYTFSDY	0.21	4FAB	353	NKPYNY	0.78
	2FBJ	326	GFDFSKY	0.34				
	2FB4	326	GFIFSSY	0.37				
	4FAB	326	GFGFSDY	0.50				
	3FAB	326	GTSFDDY	0.71				
	1ECO	111	HTDFAGA	0.56	2AZA	67	GLAQDY	0.59
	1ETU	161	DFPGDDT	0.64	1PRC	H188	SIRYGN	0.65
	2GD1	196	ARAAAES	0.65	3WGA	133	GGDAGG	0.67
	2MEV	109	KQDYSFC	0.67	1PRC	C177	AKYTAY	0.69
	1GOX	178	LTLKNFE	0.68	2CTS	40	MMYGGM	0.70
3FAB		326	GTSFDDY			352	FYHGT	
	2FBJ	326	GFDFSKY	0.62	2FBJ	352	HPDSG	0.60
	2FB4	326	GFIFSSY	0.68	3HFM	352	SYSGS	0.68
	2HFL	326	GYTFSDY	0.71	2FB4	352	WDDGS	0.68
	1ACX	42	ACNPATA	0.60	2OVO	25	GSDNK	0.15
	1PRC	C186	NYDPFTM	0.61	3WGA	148	SAGGS	0.15
	1FDH	160	KVNVEDA	0.63	1TGS	250	GTDGI	0.16
	1ETU	161	DFPGDDT	0.64	2ALP	149	TSAGQ	0.17
	5ADH	163	ASPLEKV	0.65	2CYP	225	SKSGY	0.19
2FBJ		326	GFDFSKY			352	HPDSGT	
	2FB4	326	GFIFSSY	0.19	2FB4	352	WDDGSD	0.29
	2MCP	326	GFTFSDF	0.32				
	4FAL	326	GFTFSDY	0.41				
	2HFL	326	GYTFSDY	0.46				
	3FAB	326	GTSFDDY	0.62				

(continued)

TABLE III. Best Fits to Hypervariable Loops in Immunoglobulins Found by Screening the Data Base* (Continued)

			H1				H2	
	1ECN	111	HTDFAGA	0.54	3HLA	127	KEDLRS	0.50
	1ETU	161	DFPGDDT	0.58	3SGB	44	NSARTT	0.51
	2YHX	163	?KLISAM	0.59	3FXN	7	SGTGNT	0.51
	2MEV	109	KQDYSFC	0.60	1CMS	158	DRNGQE	0.54
	2GBP	109	GTDSKES	0.61	1WSY	617	NSIGRA	0.63
2HFL		326	GYTFSDY			352	LPGSGS	
	1MCP	326	GFTFSDF	0.21	4FAB	353	NKPYNY	0.85
	2FBJ	326	GFDFSKY	0.46	2MCP	353	NKGKNY	1.02
	2FB4	326	GFIFSSY	0.47				
	4FAB	326	GFTFSDY	0.53				
	3FAB	326	GTSFDDY	0.71				
	4GPD	595	GRGAAQN	0.59	2LTN	167	NAATNV	0.32
	1ETU	161	DFFGDDT	0.63	3PEP	314	DRANNK	0.33
	1ECN	111	HTDFAGA	0.65	1TGN	76	NSNTLN	0.34
	2MEV	109	KQDYSFC	0.61	1TRM	95	DRKTLN	0.36
	1GOX	178	LTLKNFE	0.70	2FB4	93	DVSLNA	0.36
2FB4		326	GFIFSSY			352	WDDGSD	
	2FBJ	326	GFDFSKY	0.19	2FBJ	352	HPDSGT	0.29
	2MCP	326	GFTFSDF	0.34	3HFM	397	NWDGDY	0.50
	4FAB	326	GFTFSDY	0.44				
	2HFL	326	GYTFSDY	0.47				
	2GBP	109	GTDSKES	0.55	3SGB	44	NSARTT	0.45
	1ECN	111	HTDFAGA	0.58	3FXN	7	SGTGNT	0.54
	1GOX	178	LTLKNFE	0.58	1FX1	9	STTGNT	0.60
	2MEV	109	KQDYSFC	0.60	3HLA	127	KEDLRS	0.64
	2AAT	297	NPPAHGA	0.63	1WSY	410	TGACQH	0.66
4FAB		326	GFTFSDY			353	NKPYNY	
	2FBJ	326	GFDFSKY	0.41	2MCP	352	NKGNKY	0.78
	2MCP	326	GFTFSDF	0.42	2HFL	353	LPGSGS	0.85
	2FB4	326	GFIFSSY	0.44				
	2HFL	326	GYTFSDY	0.53				
	2MEV	109	KQDYSFC	0.60	3WGA	4	GEQGSN	0.35
	2GD1	196	ARAAAES	0.61	1AZA	67	GLAQDY	0.35
	1ETU	161	DFPGDDT	0.61	2PLV	577	VDYLLG	0.44
	2YHX	163	?KLISAM	0.62	2APP	239	DSNAGG	0.55
	2PFK	1024	MCDVDEL	0.64	1PHH	89	KRLSGG	0.57

*For each loop, the initial residue number and sequence of the parent loop are given; underneath them we list up to five homologous loops of similar conformation in other immunoglobulin structures, then up to five homologous loops of similar conformation from proteins of other families. For each loop found, the initial residue, the sequence, and the root mean square deviation in atomic position of N, C_α, and C atoms are given.

within the immunoglobulin family, but otherwise appear to be rare, except for the short L2 and H2 hairpin loops. Two examples from unrelated proteins are J539(2FBJ) L3 (residues 87–98) and residues 13–24 of α-amylase inhibitor (1HOE)[31] and HyHEL-5 (2HFL) L3 (residues 87–97) and residues 369–379 of glutathione reductase (3GRS).[32] The case of a good fit of the stem but a poor fit of the loop occurs in a number of entries in Table IV, including J539 L3—1MCP L3 and KOL H2—HyHEL-5 H2 for examples involving homologous loops within the immunoglobulin family and McPC603 H2—2GCR for an immunoglobulin loop and a loop from another protein family.

Applications to Model Building

The entries in Table IV confirm results of Jones and Thirup[12] and others that there is not a secure correlation between a low rms deviation of the stem and a low rms deviation of the loop. Therefore we cannot identify the best-fitting loop from the best-fitting stem. This applies both to homologous loops within the immunoglobulin family and the regions from other families.

Within the immunoglobulin family, this conclusion illuminates the relationship between variations in the structure of the framework and the canonical structure of the loops. The choice of canonical struc-

TABLE IV. Results of Screening the Data Base for the Stems of Hypervariable Loops of Immunoglobulins*,†

L1					L2					L3				
2RHE	2FB4	21–37	0.16	0.28	2RHE	IREI	46–56	0.32	0.28	2RHE	2FB4	88–103	0.40	0.75
(21–37)					(47–57)	2MCP	52–62	0.39	0.96	(88–103)				
						1F19	45–56	0.39	0.90					
						2FB4	47–57	0.39	0.12					
						4FAB	51–61	0.43	0.51					
	1HLA	202–218	0.61	4.16		2SNS	80–90	0.57	1.17		3CNA	24–39	0.54	2.61
	2PAB	6–22	0.72	4.06		5CHA	372–382	0.76	2.83		1GP1	152–167	0.57	2.11
	2PLV	817–833	0.86	4.36		1RHD	214–224	0.78	1.41		2CNA	50–65	0.59	1.53
	2TAA	407–423	0.87	4.49		2RHV	435–445	0.85	0.15CA		1BMV	1028–1043	0.59	1.23
	4MDH	287–303	0.89	4.19		1GD1	444–454	0.86	1.31		5LDH	289–304	0.60	2.51
2FB4	2RHE	21–37	0.16	0.28	2FB4	2RHE	47–57	0.39	0.12	2FB4	2RHE	88–103	0.40	0.74
(21–37)					(47–57)	1MCP	52–62	0.46	0.94	(88–103)				
						1REI	46–56	0.47	0.22					
						4FAB	51–61	0.49	0.47					
						3HFM	46–56	0.50	0.95					
	1HLA	202–218	0.64	4.21		2SNS	80–90	0.61	1.14		5LDH	289–304	0.43	2.36
	2PAB	6–22	0.80	4.08		1GD1	444–454	0.62	1.29		2PAZ	5–20	0.59	1.75
	2TAA	407–423	0.83	4.39		1RHD	214–224	0.81	1.39		3DFR	141–156	0.60	2.19
	2PLV	817–833	0.91	4.36		2RHV	435–445	0.83	1.34		1GP1	5–20	0.61	1.26
	4MDH	787–803	0.94	4.30		5CHA	135–145	0.97	2.77		4SGB	312–327	0.74	4.16
3FAB						Not present				3FAB	2FBJ	86–99	0.53	0.99
(21–38)										(82–95)	1REI	87–100	0.60	2.10
											1F19	87–100	0.62	1.71
											1MCP	93–106	0.65	2.25
											4FAB	92–105	0.68	2.23
	1PSG	287–304	0.94	4.41							2APP	255–268	0.36	0.68
	2STV	161–178	0.96	4.20							1PHH	138–151	0.44	1.07
	5TLN	121–138	0.99	3.86							1SN3	37–50	0.46	0.92
											3GRS	252–265	0.46	1.20
											8CAT	679–692	0.46	1.42
2FBJ	2HFL	22–35	0.28	0.54	2FBJ	2HFL	45–55	0.25	0.92	2FBJ	1MCP	93–106	0.22	1.96
(22–35)					(45–55)	1REI	46–56	0.28	0.20	(86–99)	1REI	87–100	0.23	1.88
						1MCP	52–62	0.37	0.97		1F19	87–100	0.28	1.56
						3HFM	46–56	0.39	0.94		3HFM	87–100	0.28	1.94
						4FAB	51–61	0.45	0.43		4FAB	92–105	0.33	2.06
	1CMS	287–300	0.86	1.83		2SNS	80–90	0.74	1.16		1PSG	258–271	0.47	1.03
						1GD1	44–54	0.80	1.29		1SN3	37–50	0.51	1.19
						5CHA	135–145	0.87	2.74		1NXB	26–39	0.53	1.11
						1CPB	35–45	0.91	1.53		1GCR	132–145	0.56	1.41
						2RHV	435–445	0.91	1.35		1ETU	66–79	0.58	1.02
1MCP					1MCP	1RE1	153–163	0.28	0.94	1MCP	2FBJ	86–99	0.22	1.96
(22–42)					(52–62)	4FAB	51–61	0.33	1.03	(93–106)	1REI	87–100	0.29	3.79
						2FBJ	45–55	0.37	0.97		4FAB	92–105	0.34	3.82
						3HFM	46–56	0.37	0.69		1F19	87–100	0.36	3.56
						2RHE	47–57	0.42	0.95		3HFM	87–100	0.39	0.45
	1TNF	317–337	0.64	3.22		1GD1	444–454	0.65	0.70		1PSG	258–271	0.53	4.37
	4RHV	712–732	0.68	4.72		2SNS	80–90	0.66	1.12		1GCR	43–56	0.60	4.16
	2PAB	118–138	0.70	6.17		1RHD	214–224	0.84	0.91		1SN3	37–50	0.61	4.42
	3APR	104–124	0.71	4.04		2RHV	435–445	0.88	1.10		1NXB	26–39	0.62	4.54
	2GLS	30–50	0.82	3.80		5CHA	135–145	0.91	2.71		1HOE	12–25	0.63	4.37
2HFL	2FBJ	22–35	0.28	0.54	2HFL	2FBJ	45–55	0.25	0.92	2HFL				
(22–35)					(45–55)	1REI	46–56	0.36	0.89	(86–98)				
						3HFM	46–56	0.46	0.82					
						1MCP	52–62	0.47	0.25					
						4FAB	51–61	0.51	0.97					

(continued)

TABLE IV. Results of Screening the Data Base for the Stems of Hypervariable Loops of Immunoglobulins*,† (Continued)

		L1					L2					L3		
	1CMS	287–300	0.92	2.04		2SNS	80–90	0.78	1.21		4APE	313–325	0.63	0.63
	1CTX	54–67	0.94	2.73		1GD1	444–454	0.90	0.73		3BCL	11–23	0.64	1.56
	2PLV	897–910	0.99	2.58		2RHV	435–445	0.92	1.20		6ACN	321–333	0.64	2.30
						5CHA	372–382	0.93	2.75		3CPP	53–65	0.64	1.71
						1RHD	214–224	0.95	0.92		3GRS	368–380	0.69	0.99
4FAB		—			4FAB	1REI	46–56	0.24	0.40	4FAB	2MCP	93–106	0.31	0.89
22–41					(51–61)	3HFM	46–56	0.32	1.06	(92–105)	2FBJ	86–99	0.33	2.06
						1MCP	52–62	0.33	1.03		1REI	87–100	0.37	0.91
						1F19	46–56	0.42	1.04		3HFM	87–100	0.38	0.90
						2RHE	47–57	0.43	0.51		1F19	87–100	0.42	1.41
	2PLV	137–156	0.90	5.13		2SNS	80–90	0.46	1.31		1NXB	26–39	0.53	2.02
	3DFR	5–24	0.97	2.55		1RHD	214–224	0.69	1.48		2TBV	193–206	0.55	2.15
	2SOD	82–101	0.98	5.64		2GD1	44–54	0.73	1.31		1PSG	258–271	0.56	2.04
						2RHV	435–445	0.85	1.46		1SN3	37–50	0.59	1.99
						1ALP	115–125	0.98	1.69		1ACX	32–45	0.61	2.10
1REI	3HFM	22–36	0.29	1.08	1REI	4FAB	51–61	0.29	0.40	1REI	2FBJ	86–99	0.23	1.88
(22–36)	1F19	22–36	0.68	0.81	(46–56)	3HFM	45–56	0.30	0.92	(87–100)	1F19	(87–100)	0.24	1.52
						2FBJ	45–55	0.31	0.20		1MCP	93–106	0.29	0.42
						2RHE	47–57	0.32	0.28		3HFM	87–100	0.34	0.28
						1MCP	52–62	0.32	0.94		4FAB	92–105	0.40	0.91
	2MEV	130–144	0.64	2.38		2SNS	80–90	0.55	1.13		1PSG	258–271	0.56	1.96
	1HOE	22–36	0.74	2.24		1GD1	44–54	0.78	1.22		1NXB	26–39	0.60	1.89
	4TLN	102–116	0.82	2.48		1RHD	214–224	0.78	1.38		1GCR	43–56	0.63	1.89
	2RS3	133–147	0.86	1.66		1RHV	438–448	0.82	1.31		1SN3	37–50	0.63	1.96
	2AIT	22–36	0.86	2.09		2PLV	578–588	0.90	1.45		1HOE	12–25	0.64	1.82
		H1					H2							
2FB4	2FBJ	322–336	0.23	0.23	2FB4	2HFL	348–361	0.25	2.05					
(322–336)	1MCP	322–336	0.37	0.44	(348–361)	2FBJ	348–361	0.28	0.32					
	3HFM	322–336	0.44	1.31		1F19	348–361	0.61	2.17					
	2HFL	322–336	0.56	0.57										
	3FAB	322–336	0.58	0.95										
	2TBV	672–686	0.56	2.84		2BCL	80–93	0.21	2.18					
	2PLV	818–832	0.78	3.00		2CNA	113–126	0.31	1.84					
	1RHV	646–660	0.78	2.83		3PEP	310–323	0.33	1.95					
	2TAA	408–422	0.86	2.83		3HLA	100–113	0.35	1.32					
						1BMV	1029–1042	0.35	1.87					
3FAB	1MCP	322–336	0.35	1.02	3FAB		—							
(322–336)	2HFL	322–336	0.38	1.00	(349–359)									
	1F19	322–336	0.50	1.29										
	2FBJ	322–336	0.52	0.87										
	4FAB	322–336	0.57	0.98										
	2TBV	64–78	0.54	2.69		2GD1	297–307	0.38	1.33					
	1RHV	646–660	0.67	3.01		2APP	72–82	0.41	0.26					
	2HLA	325–339	0.70	2.26		1THI	91–101	0.57	1.22					
	2TAA	408–422	0.84	2.75		1DPI	372–382	0.63	0.41CA					
	1PTE	165–179	0.96	3.43CA		2CNA	146–156	0.63	0.69					
2FBJ	2FB4	322–326	0.23	0.23	2FBJ	2FB4	349–360	0.31	0.23					
(332–336)	2MCP	322–326	0.26	0.42	(349–360)	2HFL	349–360	0.46	2.03					
	3HFM	322–326	0.28	1.19		1F19	349–360	0.59	1.65					
	4FAB	322–326	0.49	0.56										
	3FAB	322–326	0.52	0.87										
	2TBV	672–686	0.54	2.22		1BMV	1030–1041	0.28	1.66					
	1RHV	646–660	0.70	2.85		2LTN	164–175	0.29	1.68					
						1CMS	308–319	0.46	1.65					
						2RUB	72–83	0.50	2.59					
						6HIR	28–39	0.53	1.38					

(continued)

TABLE IV. Results of Screening the Data Base for the Stems of Hypervariable Loops of Immunoglobulins*,† (Continued)

		H1					H2		
1MCP	2FBJ	322–336	0.28	0.42	1MCP	4FAB	349–362	0.51	1.11
(322–336)	3FAB	322–336	0.35	1.02	(349–362)				
	2FB4	322–336	0.37	0.44					
	3HFM	322–336	0.41	1.16					
	2HFL	322–336	0.44	0.30					
	2TBV	64–78	0.57	2.92		2GCR	4–17	0.25	2.00
	1RHV	646–660	0.73	2.87		2ENL	20–33	0.28	1.76CA
	2TAA	408–422	0.95	3.00		1NXB	26–39	0.31	1.72
						1CHG	131–144	0.32	2.94
						2PAB	212–225	0.33	1.93
2HFL	1F19	322–336	0.31	1.25	2HFL	2FB4	349–360	0.32	2.03
(322–336)	3FAB	322–336	0.38	1.00	(349–360)	2FBJ	349–360	0.46	2.00
	1MCP	322–336	0.44	0.30		1F19	349–360	0.46	1.68
	2FB4	322–336	0.56	0.57					
	2FBJ	322–336	0.58	0.56					
	2TBV	64–78	0.58	2.99		2LTN	164–175	0.55	0.69
	1RHV	646–660	0.67	2.87		1BMV	1030–1041	0.55	0.84
	2TAA	408–422	0.85	2.97		5HIR	28–39	0.56	1.64
						451C	51–62	0.57	1.58
						3CNA	52–63	0.61	0.89
4FAB	2MCP	322–336	0.49	0.62	4FAB	1MCP	349–362	0.51	1.11
(322–336)	2FBJ	322–336	0.49	0.56	(349–362)				
	3HFM	322–336	0.51	1.46					
	3FAB	322–336	0.57	0.98					
	2FB4	322–336	0.62	0.54					
	2HLA	325–339	0.75	2.11		2MEV	314–327	0.23	2.30
	2TBV	64–78	0.76	2.82		3BCL	248–261	0.24	1.75
	1RHV	646–660	0.83	3.14CA		1HMG	1323–1336	0.26	1.65
	2TAA	408–422	1.00	2.87		4SBV	446–459	0.27	2.21
						6API	317–330	0.38	1.76

*For each loop, the limits given in parentheses under the name of the parent molecule include the four flanking residues on the N-terminal side of the loop, the loop itself, and the four flanking residues on the C-terminal side of the loop. The searches probed the data base with C_α atoms of the eight flanking residues, assigning on each side the weight 1.0 to the residue closest to the loop, 0.8 to the next residue, 0.64 to the next, and 0.512 to the fourth residue, farthest from the loop.

†For each structure with structural similarity in the stems, we report the residue range identified, including flanking and intervening residues, the weighted root mean square deviation of the C_α atoms of the stem residues, and the root mean square deviation of the mainchain atoms N, C_α, C, O of the loop residues themselves. For example, the residues flanking the L1 loop of 2RHE, 21–25 and 34–37 have a weighted rms deviation of 0.16 Å from the residues 21–25 and 34–37 of 2FB4. For the loops themselves, residues 26–33, the rms deviation of all mainchain atoms N, C_α, C, O is 0.28 Å.

ture depends on the presence of specific amino acids at specific positions in the sequences of the loop and the framework. If the framework were constant in structure, the structure of the stem would be entirely noncommittal about the canonical structure of the loop. However, because the framework residues that form the stems of the loops do vary in structure to some extent, one *might* observe a correlation between the details of the structures of the stems and the conformation of the loop.

The results in Table IV do not reveal any such correlation, however. For V_κ L3 loops, there are three known canonical structures.[9] The "stem" searches for the L3 loops of McPC603, J539, and REI identified an immunoglobulin with a different canonical structure of L3 as the best stem fit. For 4–4–20, the correct canonical structure was identified. In most cases the weighted rms deviations of the stems are rather similar in value. The stem searches do not reliably indicate the correct canonical structure. Conversely, a canonical structure of an antigen-binding loop does not induce—or require—a specific adjustment of the mainchain of the framework.

For H2 the situation is similar. Four canonical structures are known.[9] For McPC603, J539, and 4–4–20, an immunoglobulin with the correct canonical structure has the best stem fit, but for KOL and HyHEL-5, an immunoglobulin with a different canonical structure has the best stem fit.

There have now been a number of studies aimed at predicting the structure of loops by first generat-

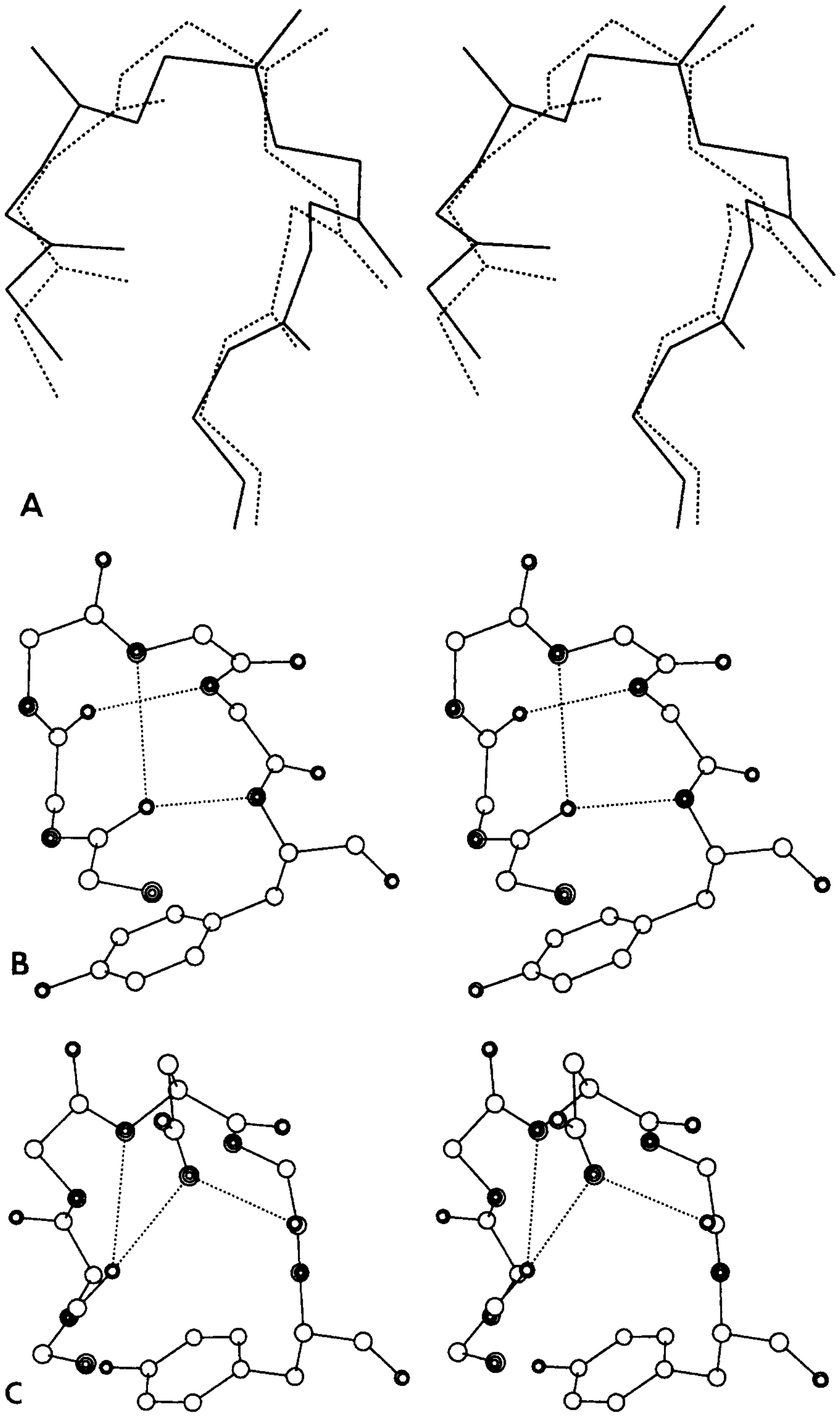

Fig. 4. **(A)** Superposition of McPC603 H2 (solid) with residues 276–281 of *Alcaligenes denitrificans* azurin (2AZA) (broken). **(B)** H2 loop of McPC603. **(C)** Region of similar conformation in azurin.

ing a set of candidate loops and then attempting to select one, on the basis of conformational energy estimates and/or accessible surface area.[33–38] The candidate loops may be generated by saturating conformational space or by data base searching. The results presented here show that, although in most cases loops of the desired conformation exist in nonhomologous proteins, it will not in general be possi-

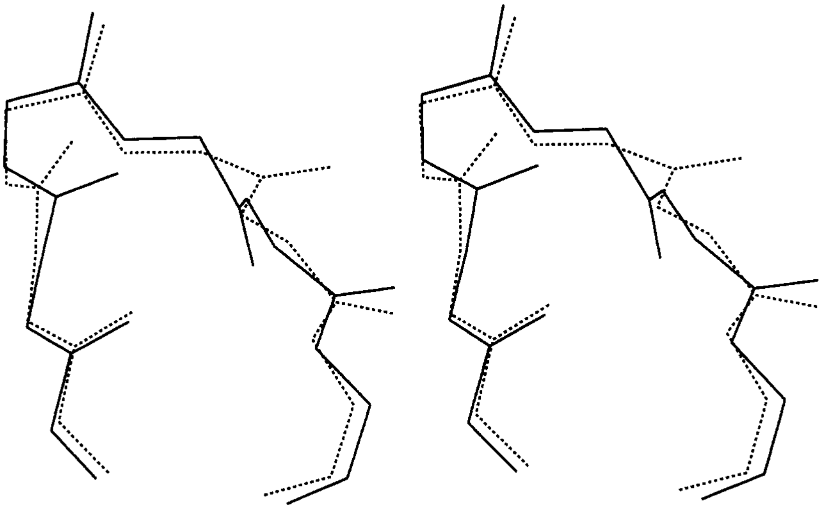

Fig. 5. Superposition of HyHEl-5 H2 (solid) and residues 167–172 of garden pea lectin (2LTN) (broken).

ble to identify them soley by data base screening for the local region. This is because of the differences in conformation in the flanking residues, or, if one remains within the set of homologous immunoglobulin loops, because the stems do not distinguish the correct canonical structure.

CONCLUSIONS

We have elucidated the structural relationships between antigen-binding loops L1, L2, L3, H1, and H2 and regions of similar conformation in other proteins. Most but not all of the antigen-binding loops appear in other protein families, even some with very unusual structural features such as the L1 loop of V_λ domains. However, the structural contexts of the regions of similar structure are often quite different. A good fit of an antigen-binding loop usually does not extend to the residues flanking the loops, and vice versa. This precludes there being any simple and general way to apply data base search methods to modeling antigen-binding loops in immunoglobulins of unknown structure.

ACKNOWLEDGMENTS

A.M.L. thanks the Kay Kendall Foundation for generous support.

REFERENCES

1. Lesk, A.M., Chothia, C. The response of protein structures to amino acid sequence changes. Phil. Trans. R. Soc. (London) 317:345–356, 1986.
2. Chothia, C., Lesk, A.M. Relationship between the divergence of sequence and structure in proteins. EMBO J. 5: 823–826, 1986.
3. Sibanda, B.L., Thornton, J.M. β-Hairpin families in globular proteins. Nature (London) 316:170–174, 1985.
4. Efimov, A.V. Standard conformations of polypeptide chains in irregular regions of proteins. Mol. Biol. (USSR) 20:208–216, 1986.
5. Milner-White, E.J., Poet, R. Four classes of β-hairpins in proteins. Biochem. J. 240:289–292, 1986.
6. Chothia, C., Lesk, A.M. Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. 196: 901–918, 1987.
7. Wilmot, C.M., Thornton, J.M. Analysis and prediction of the different types of β-turns in proteins. J. Mol. Biol. 203: 221–232, 1988.
8. Chothia, C., Lesk, A.M., Levitt, M., Amit, A.G., Mariuzza, R.A., Phillips, S.E.V., Poljak, R. The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. Science 233:755–758, 1986.
9. Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., Colman, P.M., Spinelli, S., Alzari, P.M., Poljak, R.J. The conformations of immunoglobulin hypervariable regions. Nature (London) 342: 877–883, 1989.
10. Tramontano, A., Chothia, C., Lesk, A.M. Structure determinants of the conformations of medium-sized loops. Proteins 6:382–394, 1989.
11. Tramontano, A., Chothia, C., Lesk, A.M. Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the V_H domains of immunoglobulins. J. Mol. Biol. 215:175–182, 1990.
12. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5:819–822, 1986.
13. Poljak, R.J., Amzel, L.M., Chen, B.L., Phizackerley, R.P., Saul, F. Structural basis for the association of heavy and light chains and the relation of subgroups to the conformation of the active site of immunoglobulins. Immunogenetics 2:393–394, 1975.
14. Chothia, C., Novotny, J., Bruccoleri, R., Karplus, M. Domain association in immunoglobulin molecules: The packing of variable domains. J. Mol. Biol. 186:651–663, 1985.
15. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein databank: A computer-based archival file for macromolecular structure. J. Mol. Biol. 112:535–542, 1977.
16. Saul, F.A., Amzel, L.M., Poljak, R.J. Preliminary refinement and structural analysis of the Fab fragment from the human immunoglobulin New at 2.0 Å. J. Biol. Chem. 253: 585–597, 1978.
17. Marquart, M., Deisenhofer, J., Huber, R., Palm, W. Crystallographic refinement and atomic models of the intact immunoglobulin molecule Kol and its antigen-binding fragment at 3.0 Å and 1.9 Å resolution. J. Mol. Biol. 141: 369–391, 1980.
18. Furey, W., Jr., Wang, B.C., Yoo, C.S., Sax, M. Structure of a novel Bence-Jones protein (Rhe) fragment at 1.6 Å resolution. J. Mol. Biol. 167:661–692, 1983.

19. Segal, D.M., Padlan, E.A., Cohen, G.H., Rudikoff, S., Potter, M., Davies, D.R. The three dimensional structure of a phosphocholine-binding mouse immunoglobulin Fab and the nature of the antigen binding site. Proc. Natl. Acad. Sci. U.S.A. 71:4298–4302, 1974.
20. Suh, S.W., Bhat, T.N., Navia, M.A., Cohen, G.H., Rao, D.N., Rudikov, S., Davies, D.R. The galactan-binding immunoglobulin FabJ539: An x-ray diffraction study at 2.6 Å resolution. Proteins 1:74–79, 1986.
21. Sheriff, S., Silverton, E.W., Padlan, E.A., Cohen, G.H., Smith-Gill, S.J., Finzel, B.C., Davies, D.R. Three-dimensional structure of an antibody-antigen complex. Proc. Natl. Acad. Sci. U.S.A. 84:8075–8079, 1987.
22. Herron, J.N., He, X., Mason, M.L., Voss, E.W., Jr., Edmundson, A.B. Three-dimensional structure of a fluorescein-Fab complex crystallized in 2-methyl-2,4-pentanediol. Proteins 5:271–280, 1989.
23. Epp, O., Lattman, E.E., Schiffer, M., Huber, R., Palm, W. The molecular structure of a dimer composed of the variable portions of the Bence-Jones protein REI refined at 2.0 Å resolution. Biochemistry 14:4943–4952.
24. Kabat, E.A., Wu, T.T., Reid-Miller, M., Perry, H.M., Gottesman, K.S. "Sequences of Proteins of Immunological Interest," 4th ed. Bethesda, MD: National Institutes of Health, 1987.
25. Lesk, A.M. Integrated access to sequence and structural data. In: "Biosciences: Perspectives and User Services in Europe." Saccone, C. ed. Bruxelles: EEC, 1986:23–28, and references contained therein.
26. Lesk, A.M., Chothia, C. The evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domains. J. Mol. Biol. 160:325–342, 1982.
27. Deisenhofer, J., Epp, O., Miki, K., Huber, R., Michel, H. Structure of the protein subunits in the photosynthetic region centre of *Rhodopseudomonas viridis* at 3 angstroms resolution. Nature (London) 318:618–624, 1985.
28. Pletnev, V.Z., Kuzin, A.P., Trakhanov, S.D., Kostetsky, P.V. Three-dimensional structure of actinoxanthin IV. A 2.5-angstrom resolution. Biopolymers 21:287–300, 1982.
29. Baker, E.N. Structure of azurin from *Alcaligenes dentrificans*. Refinement at 1.8 angstroms resolution and comparison of the two crystallographically independent molecules. J. Mol. Biol. 203:1071–1095, 1988.
30. Einspahr, H., Parks, E.H., Suguna, K., Subramanian, E., Suddath, F.L. The crystal structure of pea lectin at 3.0-angstroms resolution. J. Biol. Chem. 261:16518–16527, 1986.
31. Pflugrath, J.W., Wiegand, G., Huber, R., Vertesy, L. Crystal structure determination, refinement and the molecular model of the alpha-amylase inhibitor HOE-467A. J. Mol. Biol. 189:383–386, 1986.
32. Karplus, P.A., Schulz, G.E. Refined structure of glutathione reductase at 1.54 angstroms resolution. J. Mol. Biol. 195:701–729, 1987.
33. de la Paz, P., Sutton, B.J., Darsley, M.J., Rees, A.R. Modeling of the combining sites of three antilysozyme monoclonal antibodies and of the complex between one of the antibodies and its epitope. EMBO J. 5:415–425, 1986.
34. Moult, J., James, M.N.G. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1:146–163, 1986.
35. Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L., Levinthal, C. Predicting antibody hypervariable loop conformations. II. Minimizing and molecular dynamics studies of McPC603 from many randomly generated loop conformations. Proteins 1:342–362, 1986.
36. Bruccoleri, R.E., Haber, E., Novotny, J. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. Nature (London) 335:564–568, 1988.
37. Martin, A.C.R., Cheetham, J., Rees, A.R. Modeling antibody hypervariable loops: A combined algorithm. Proc. Natl. Acad. Sci. U.S.A. 86:9628–9272, 1989.
38. Summers, N.L., Karplus, M. Modeling of globular proteins. A distance-based data search procedure for the construction of insertion/deletion regions and Pro↔non-Pro mutations. J. Mol. Biol. 296:991–1016, 1990.
39. Lesk, A.M. Detection of three-dimensional patterns of atoms in chemical structures. Commun. Assoc. Comp. Mach. 22:219–224, 1979.

APPENDIX: SCREENING DATABASES OF STRUCTURES FOR PRESCRIBED COMBINATIONS OF SEGMENTS

We describe a procedure for efficient searching for structures similar to prescribed oligopeptides in a database of protein structures. Given an oligopeptide S, with $S(i)$, $i = 1, ..., n$ representing the mainchain (or C_α) coordinates of the ith residue; $w_i > 0$ a set of weighting coefficients and a data bank of coordinates of protein structures P_j, with $P_j(i)$ representing the mainchain (or C_α) coordinates of the ith residue of protein j; we wish to identify proteins containing sets of consecutive residues $P_j\,(i)$, $i = k, ..., k + n$ such that the root mean square deviation after optimal superposition:

$$\Delta = \begin{matrix}\text{Rotations } R\\ \text{translations } t\end{matrix} \quad \{\sum_i w_i \,|\, S(i) - [R\,P_j\,(k + i - 1) - t]|^2\}^{1/2}$$

is small. Here R is a proper rotation matrix, t a translation vector, and the quantity minimized is the weighted sum of the deviations of corresponding atoms (C_α or all mainchain atoms) after the rotation and translation have been applied to the atoms in the protein P_j. (A generalization to weighting schemes in which different atom types are given different weights—for example, to lower the weight associated with the main chain oxygen atoms—presents no difficulties.)

We note that if all weights $w_i = 1$, the task corresponds to searching the data bank for segments similar to a given set of consecutive residues without gaps. If the sequence of weights contain stretches of zeros—e.g., w_i = 1,1,1,1,0,0,0,0, 1,1,1,1—the task is that of finding a 12-residue segment in the data bank such that the first four and last four residues match the structure of a protein in the data bank, but the middle four residues do not enter the calculation and indeed need not even be specified in the probe structure S. It is this case that is useful in trying to build a loop spanning a gap in the probe structure.

Although the individual superposition calculations are straightforward, it is useful to try to improve the efficiency of the method by a prescreening of the database so that no superposition calculations are performed unless there is a good chance that Δ will be low. Jones and Thirup[12] did this by creating a separate representation of the structures in terms of inter-C_α distances (compare Lesk[39]). Here we suggest an alternative, which is convenient because it does not require a separate representation of the data base, and gives adequate performance.

To simplify the notation, suppose we wish to superpose two sets of atoms: x_i, $i = 1, ..., n$ and y_i, $i = 1, ..., n$. To each pair of corresponding atoms we assign a weight $w_i \geq 0$. We assume without loss of generality that the (weighted) mean positions of the two sets of atoms coincide so that the translation vector t in the optimal superposition is zero.

$$\text{Let } y_i' = Ry_i \text{ and } \Delta^2 = \sum_i w_i \,|\, x_i - y_i' \,|^2.$$

We are willing to set a threshold $D > 0$ such that we wish to identify segments only if $\Delta \leq D$. Given D, how can we decide quickly whether a given pair of sets of atoms, such as x_i and y_i, can be superposed with $\Delta < D$, or equivalently of course, that $\Delta^2 \leq D^2$?

Observe that

$$\Delta^2 = \sum_i w_i \,|\, x_i - y_i' \,|^2$$

$$= \sum_i w_i \,[|x_i|^2 + |y_i'|^2 - 2\, x_i \cdot y_i']$$

$$\geq \sum_i w_i \,[|x_i|^2 + |y_i'|^2 - 2|x_i|\,|y_i'|]$$

But $|y_i'| = |Ry_i| = |y_i|$ because R is an orthogonal matrix.

Therefore

$$\Delta^2 \geq \sum_i w_i \,[|x_i|^2 + |y_i|^2 - 2|x_i|\,|y_i|] = \sum_i w_i \,[|x_i| - |y_i|]^2.$$

Note that the lower bound to Δ is independent of R.

This inequality provides the basis for a screening method. We accumulate successive terms in the sum on the right hand side. If for any i the partial sum exceeds D^2, we reject y_i as a potential "fit" to x_i with $\Delta \leq D$.

Obviously, the power of this procedure depends on the value of the threshold we impose. In our calculations of loop and stem fits presented in this paper, we set a threshold of $D = 0.75$ Å rms deviation. Under these conditions, the prescreening procedure rejected 99% of the possible oligopeptides.