doi:10.1016/j.jmb.2010.10.030 J. Mol. Biol. (2011) 406, 228–256

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

A New Clustering of Antibody CDR Loop Conformations

Benjamin North, Andreas Lehmann and Roland L. Dunbrack Jr*

Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA

Received 2 July 2010;
received in revised form
18 October 2010;
accepted 18 October 2010
Available online
28 October 2010

Edited by M. Sternberg

Keywords:
antibody structure;
canonical loop
conformations;
affinity propagation

Previous analyses of the complementarity-determining regions (CDRs) of antibodies have focused on a small number of "canonical" conformations for each loop. This is primarily the result of the work of Chothia and coworkers, most recently in 1997. Because of the widespread utility of antibodies, we have revisited the clustering of conformations of the six CDR loops with the much larger amount of structural information currently available. In this work, we were careful to use a high-quality data set by eliminating low-resolution structures and CDRs with high *B*-factors or high conformational energies. We used a distance function based on directional statistics and an effective clustering algorithm with affinity propagation. With this data set of over 300 nonredundant antibody structures, we were able to cover 28 CDR–length combinations (e.g., L1 length 11, or "L1–11" in our CDR-length nomenclature) for L1, L2, L3, H1, and H2. The Chothia analysis covered only 20 CDR-lengths. Only four of these had more than one conformational cluster, of which two could easily be distinguished by gene source (mouse/human; κ/λ) and one could easily be distinguished purely by the presence and the positions of Pro residues (L3–9). Thus, using the Chothia analysis does not require the complicated set of "structure-determining residues" that is often assumed. Of our 28 CDR-lengths, 15 have multiple conformational clusters, including 10 for which the Chothia analysis had only one canonical class. We have a total of 72 clusters for non-H3 CDRs; approximately 85% of the non-H3 sequences can be assigned to a conformational cluster based on gene source and/or sequence. We found that earlier predictions of "bulged" *versus* "nonbulged" conformations based on the presence or the absence of anchor residues Arg/Lys94 and Asp101 of H3 have not held up, since all four combinations lead to a majority of conformations that are bulged. Thus, the earlier analyses have been significantly enhanced by the increased data. We believe that the new classification will lead to improved methods for antibody structure prediction and design.

Introduction

Prediction of the three-dimensional structure of antibodies is an important step in improving their affinity, stability, and suitability as therapeutics. Given the conserved structure of the frameworks of the heavy-chain variable (VH) domain and the light-chain variable (VL) domain, much of the attention in structural bioinformatics has focused on complementarity-determining regions (CDRs) involved in binding antigens. Studies by Chothia, Lesk, Thornton, and others in the 1980s and 1990s centered around the idea of identifying a small number of "canonical structures" for six CDR loops [H1, H2, and H3 of the VH domain; L1, L2, and L3 of the VL domain] of various lengths.[1–4] The central

*Corresponding author. E-mail address: Roland.Dunbrack@fccc.edu.

Abbreviations used: CDR, complementarity-determining region; VH, heavy-chain variable; VL, light-chain variable; PDB, Protein Data Bank.

hypothesis, first stated in 1987,[1] was that "most of the hypervariable regions in immunoglobulins have one of a small discrete set of main-chain conformations that we call 'canonical structures,'" and that a small number of key residues could be used to predict to which conformational class a new CDR sequence might belong. In further studies, Al-Lazikani *et al.*,[2] Martin and Thornton,[3] Oliva *et al.*,[5] Wilmot and Thornton,[6] Shirai *et al.*,[7] and Kuroda *et al.*[8] defined canonical structures based on loop length and, in some cases, different conformations for certain loop lengths. Residues at some positions—in particular glycine, proline, aromatic residues, and hydrogen-bond donors and acceptors—were proposed to be responsible for differences in conformation. In their 1997 study, Chothia and coworkers found a total of 25 canonical classes due to the larger number of structures available.[2]

Chothia *et al.* used a manual clustering of antibody loops and sequences to define their canonical classes. Martin and Thornton in 1996 used a quantitative clustering approach for an automated classification scheme.[3] They performed a cluster analysis in internal coordinate space, followed by a postcluster merging of groups of structures in Cartesian coordinate space [using root-mean-square deviation (RMSD)] to classify the observed CDRs. In some instances, they observed that although a loop might be closer in sequence to one of the Chothia canonical classes, it structurally belonged to another. They note this as a limitation to the more sequence-based analyses of previous studies.

There have been a number of studies that focused specifically on the structural motifs found in the structurally diverse heavy-chain H3 CDR.[3,9–13] Morea *et al.* divided the H3 hypervariable region into a "torso" region and the "head" of the loop.[4] They found that the torso typically takes on one of two conformations, either bulged or extended β-sheet, and the possible conformations of the head region are then limited by the structure of the torso residues. Oliva *et al.* also divided H3 loops into groups based on structure.[5] They defined loop conformations using a geometric alphabet, as described by Wilmot and Thornton.[6] Shirai *et al.* identified, through inspection, a series of sequence–structure relationships that they then transformed into a set of rules to classify H3 structures.[7] In particular, they believed that the presence or the absence of salt bridges in the 'torso' region, as defined by Morea *et al.*, leads to either bulged or extended conformations in that region.[4] Kuroda *et al.* later revised their list of H3 sequence–structure rules with the availability of more H3 structures.[8]

For non-H3 loops, the most recent comprehensive analyses of their conformations were performed in 1996–1998. With the large increase in the number of available antibody structures, we decided to revisit the analysis of the conformations of antibody CDRs to see whether the canonical classes based on 17 structures[2] or fewer than 60 structures[3] have held up and whether new ones may be identified. In this article, we update the classification of all six CDR regions based on the current Protein Data Bank (PDB). We filtered out low-resolution structures, loops with high *B*-factors or high conformational energies, and redundant sequences. A total of 337 unique heavy chains and 311 unique light chains were used to construct a structural database of antibody loops. Unlike Chothia's analysis, we found it most intuitive to group CDRs into CDR type (L1, L2, etc.) and loop length. We refer to these as "CDR–length combinations" or simply "CDR-lengths" for short. For instance, a common loop length for CDR L1 is 11, and we designate this as "L1–11." We then applied clustering to the conformations of all loops of a particular CDR–length combination using an affinity propagation clustering method[9] with a dihedral-angle distance function. We found that most of the canonical conformations found by Chothia *et al.* occur in many of the 300+ antibody structures now available. We have identified a total of 72 clusters of conformations, most of which are observed in two or more antibody structures. We provide a detailed comparison of our results to previous antibody loop classifications based on smaller data sets.

Results

Data set

As described in Materials and Methods, we used manually curated multiple-sequence alignments to construct hidden Markov models (HMMs) of the VH and VL domains. We used these models to search the entire set of PDB sequences to identify all PDB chains with antibody variable domains. There were a total of 923 antibody PDB entries that contain at least one hypervariable loop with all backbone atom positions defined. Since the asymmetric units of many PDB entries contain more than one copy of the same antibody and since other PDB entries contain more than one antibody (anti-idiotypes), within those files were 1232 chains with a VH domain, 1304 chains with a VL domain, and 30 chains with both a VH domain and a VL domain within a single chain (scFv fragments). After low-resolution structures (>2.8 Å) and NMR structures had been excluded, there were 703 remaining entries comprising 882 VH domains, 953 VL domains, and 26 scFv chains.

We defined the CDRs differently from the most commonly used Kabat and Chothia schemes. We chose definitions such that the anchors of each loop (the residue immediately before or after the loop)

contained tightly clustered conformations relative to the framework, using structure alignments obtained by Honegger and Plückthun.[10] We also selected positions such that the N-terminal and C-terminal residues were opposite each other in the structure, whether they occurred in neighboring β-strands (CDR2 and CDR3) or in different β-sheets (CDR1). Where possible, we also chose definitions using homologous positions in the VL and VH chains.

The sequence motifs around our CDR starting and ending positions are shown in Fig. 1. We started with the positions immediately following the conserved cysteines of the intrachain disulfide bond and defined these as the N-terminal residues

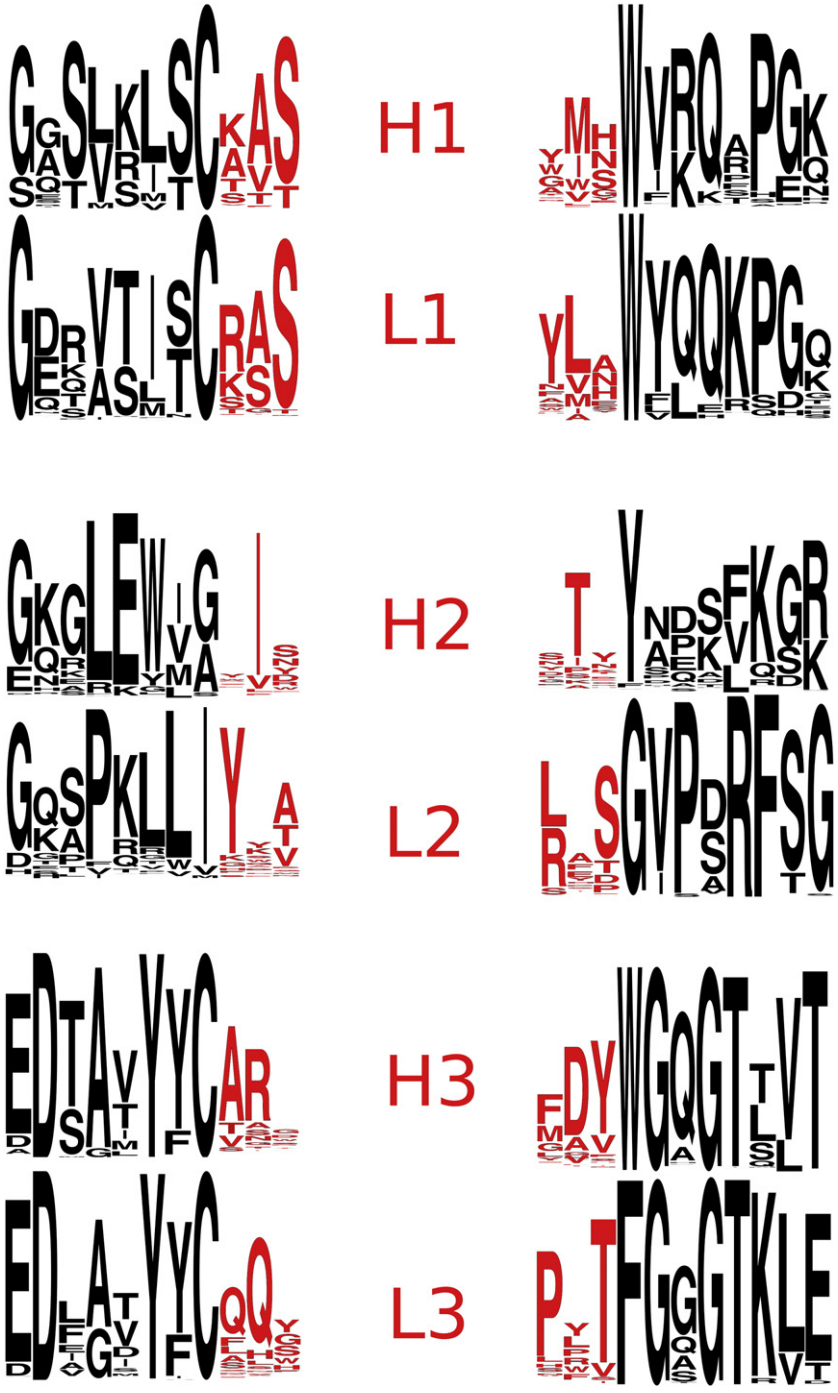

Fig. 1. CDR definitions used in this work. The sequence logos of each loop are shown with the first three residues and the last three residues of the CDR (red) and the flanking framework residues (black).

of H1, L1, H3, and L3. For both CDR1 and CDR3 loops, we chose the C-termini based on the C^{α} positions with the least variance across VL and VH domains, which also turned out to be at about the same depth in the structure as the N-termini. In these four cases, the C-termini were followed by conserved aromatic residues that are part of the hydrophobic core of each domain. We chose the L2 start site as the same one used by Chothia, since it occurs opposite the CDR1 sites that we had already chosen and at the end of a β-strand, and used this to put the H2 start site in the same position. We placed the end of H2 in a short β-strand immediately across the N-terminus of H2. However, this region in VL is not always a β-strand, and there is sequence and structural diversity for several more residues. We chose the L2 C-terminus that agrees with the Martin–Thornton definitions.[11] This L2 definition includes three more positions on its C-terminus than the H2 definition. Superpositions of VH and VL domains (from PDB entry 1MJU[12]) with the CDRs indicated are shown in Fig. 2.[14] Note that the N-terminus and the C-terminus of each loop are in homologous positions between the VH domain and the VL domain, with the exception of L2 (Fig. 2b).

We applied a number of criteria to filter out loops of uncertain or indeterminate conformation. These include loops with missing coordinates, backbone atoms with high *B*-factors, residues with cis residues that are not proline [including PDB entry 1OCW[13] (resolution, 2.0 Å) with 10 non-Pro cis residues, including 4 non-Pro cis residues in H1], and those with high backbone conformational energy, as determined by the Ramachandran probability distributions that we have recently published.[15] The remaining structures are highly redundant in sequence, since the structures of some antibodies have been determined multiple times. By representing each variable domain structure by the sequences of its six CDRs, we chose the structure with the highest resolution for each sequence. We also removed a small number of loops with conformations that are outliers with respect to all other structures, defined as having at least one backbone dihedral angle 90° away from every other structure in the data set. The number of loops for each CDR in the data set after each of these filters had been applied is shown in Table 1. The counts of the

Table 1. Count of structures, by CDR

	L1	L2	L3	H1	H2	H3
Starting count	1334	1334	1334	1262	1262	1262
Resolution >2.8 Å	351	351	351	349	349	349
Non-X-ray structures	4	4	4	5	5	5
B-factor>80	38	21	29	28	27	32
B-factor=0	30	30	30	0	0	0
Missing backbone atoms	6	1	5	14	4	30
ConfE≥9.5	17	33	22	4	5	23
Non-Pro cis	1	0	3	4	2	16
Identical CDR sequences	577	582	578	525	529	487
Outliers	9	1	1	3	4	13
Included	301	311	311	330	337	307

The first line of the table provides the number of initial CDR loops in the available PDB entries with antibody VL and/or VH domains. The subsequent lines are the numbers of CDR loops removed by each filtering criterion. The last line provides the resulting count of structures for each CDR.

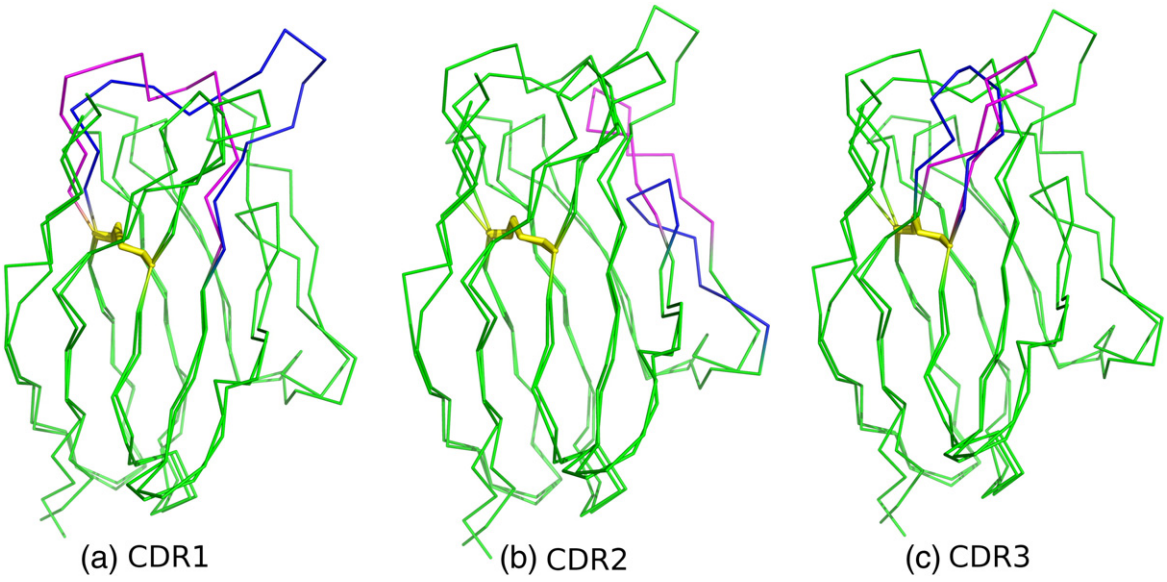

Fig. 2. CDRs based on our definitions. (a) L1 and H1. (b) L2 and H2. (c) L1 and H3. L1, L2, and L3, dark blue; H1, H2, and H3, magenta; disulfides, yellow. The structure is PDB entry 1Q9R.[14]

Table 2. Count of loops, by CDR and length

Loop	Number of structures	Number of unique sequences	Genes	Loop	Number of structures	Number of unique sequences	Genes
L1–10	22	17	κ Mo	H1–10	2	2	Camel
L1–11	136	96	κ	H1–12	1	1	Mo
L1–12	12	12	κ; λ Hu	H1–13	306	247	Hu, Mo, Camelid
L1–13	11	11	λ	H1–14	11	7	Mo
L1–14	18	12	λ	H1–15	9	7	Hu, Mo
L1–15	13	11	κ Mo	H1–16	1	1	Camelid
L1–16	68	50	κ				
L1–17	21	17	κ				
L2–8	306	165	κ,λ	H2–8	2	2	Hu, Camel
L2–12	4	2	λ	H2–9	81	61	Hu, Mo, Camelid
				H2–10	227	196	Hu, Mo, Camelid
				H2–12	26	22	Mo
				H2–15	1	1	Llama
L3–7	2	2	κ	H3–5	5	4	Hu, Mo
L3–8	22	19	κ	H3–6	3	3	Mo
L3–9	264	218	κ,λ	H3–7	33	18	Mo
L3–10	9	8	κ Mo	H3–8	5	5	Hu, Mo, Llama
L3–11	10	10	λ Hu	H3–9	26	25	Hu, Mo
L3–12	1	1	λ Mo	H3–10	28	24	Hu, Mo, Llama
L3–13	3	2	λ	H3–11	26	26	Hu, Mo, Llama
				H3–12	49	49	Hu, Mo, Camel
				H3–13	40	36	Hu, Mo, Llama
				H3–14	26	25	Hu, Mo, Camelid
				H3–15	9	8	Hu, Mo, Camel
				H3–16	23	23	Hu, Mo, Camelid
				H3–17	5	5	Hu, Mo, Camelid
				H3–18	7	7	Hu, Mo, Camelid
				H3–19	6	6	Hu, Mo, Camel
				H3–20	4	4	Hu, Mo, Camelid
				H3–21	6	6	Hu, Mo, Camel
				H3–22	1	1	Hu, Mo
				H3–24	3	3	Hu, Mo
				H3–26	2	1	Camel

For each CDR–length combination ("CDR-length"), the number of structures available after our filtering step and the number of unique sequences in each set are given. The genes were identified by comparing the CDR sequences to the V gene segments for each species available from the IMGT database (http://imgt.cines.fr)[16] and by finding the closest CDR sequences for CDR1 and CDR2 for each PDB chain. Some CDR-lengths occur only in specific genes and/or species. "Camelid" means that the CDR-length contains both llama and camel sequences.

different loop lengths for each CDR in the resulting data set are given in Table 2.

Affinity clustering of CDR loop conformations

We ran the affinity clustering algorithm for each combination of CDR, loop length, and cis–trans configuration separately. As an example of clustering, we show the Ramachandran distributions for the clusters of L1–12 in Fig. 3. This CDR-length comprises 12 structures with unique sequences, clustered into three conformations of sizes 5, 5, and 2. We divided the Ramachandran map into labeled regions as shown in Fig. 4 in order to label the clusters by conformation. In this definition, B is the β-sheet region, P is polyproline II, A is the α-helix, D is the δ region (near α-helix but at more negative values of ϕ), L is the left-handed helix, and G is the γ region (ϕ>0°, excluding the L and B regions). With these definitions, the median loop of cluster 1 (blue dots) has the conformation BPABPBPAADBB, cluster 2 (magenta dots) has the conformation BPABPPPLLPBB, and cluster 3 (green dots) has the conformation BPPAADAAPPBB. Cluster 1 differs from cluster 2 primarily at residues 8, 9, and 10, with conformations AAD and LLP, respectively.

The clustering results for CDRs L1, L2, L3, H1, and H2 are shown in Tables 3, 4, 5, 6, and 7, respectively. The clustering for the torso region of H3 loops is shown in Table 8. In each table, the results for each loop length are given; for each cluster, the structure count and percentage, the unique sequence count, the PDB entry for the median loop structure, the

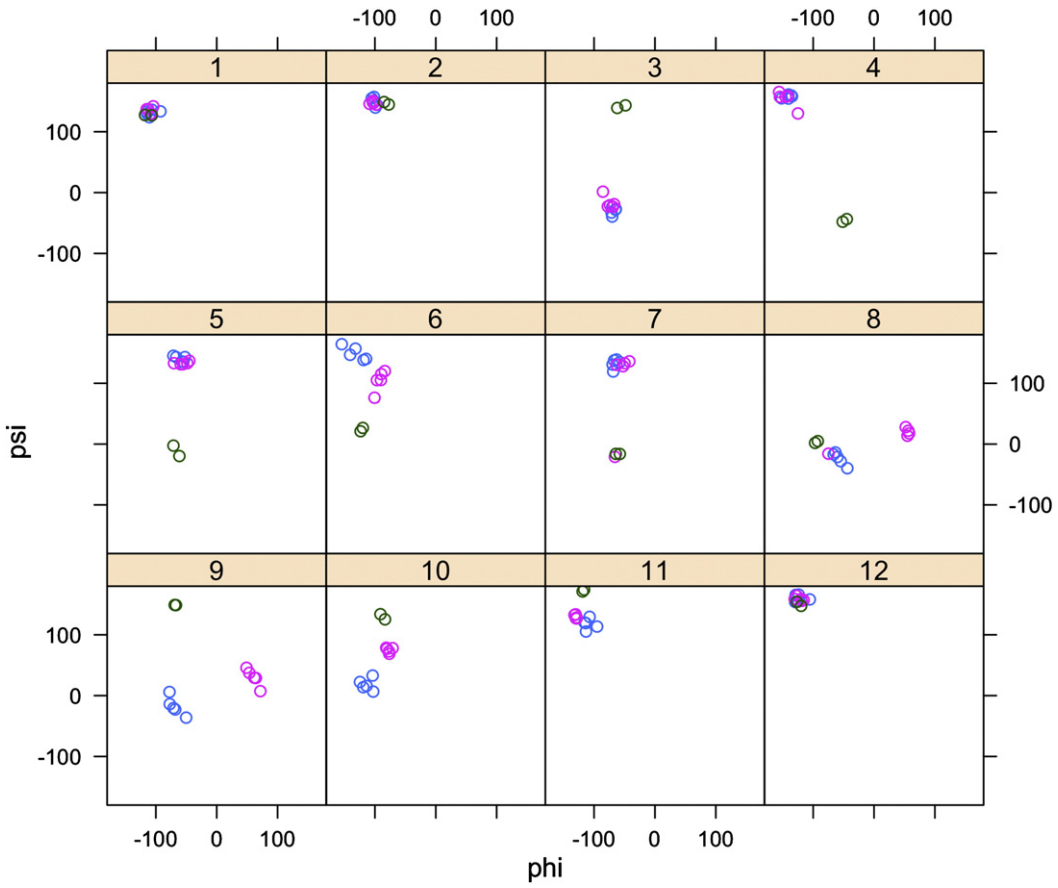

Fig. 3. Ramachandran maps of L1–12 clustering. The median loop of cluster 1 (blue dots) has the conformation BPABPBPAADBB, the median loop of cluster 2 (magenta dots) has the conformation BPABPPPLLPBB, and the median loop of cluster 3 (green dots) has the conformation BPPAADAAPPBB (see Fig. 4 for the definitions of Ramachandran regions).

consensus sequence, and the conformation of the median loop in terms of the Ramachandran conformations are given.

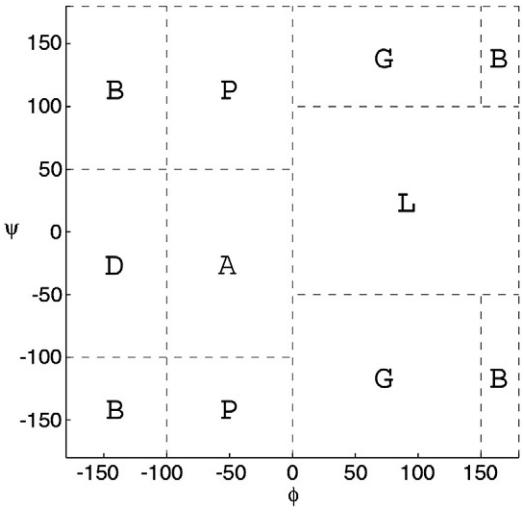

Fig. 4. Regions of the Ramachandran map.

Before we discuss the results of the clustering for each CDR, we can observe three different categories or types of antibody loop lengths, which indicate the degree to which a CDR-length conformation is predictable from its sequence:

(1) Type I: One-cluster CDR-lengths—For the first type, the loops of a certain CDR–length combination have one conformation that forms all of the structures or at least a large majority of the structures. When fed into the affinity algorithm, the result is a single conformational cluster or one large cluster and a small number of outlying conformations. The large cluster must be a fairly tight distribution. This CDR-length therefore has a predictable structure. We consider CDR-lengths to be of this type if there are at least 10 unique sequences, with more than 85% of the structures in the largest cluster of conformations.
(2) Type II: Predictable CDR-lengths—The second type of CDR–length combination has multiple possible structures, but each cluster is tightly grouped and significantly differs from the others in sequence. We include in this type some loops

Table 3. Clustering of CDR loop L1

Cluster	Number of structures	Percent loop length	Median PDB entry	Consensus sequence	Number of unique sequences	Percent unique sequences	Species	Gene	Loop conformation	Median angle (°)	Type	Comments
L1–10–1	20	91	1YQVL	sAsSSVsYmh	16	94	Mo	κ	BBABPBABBB	10	I	
L1–10–2	2	9	1AY1L	SASSSVSYmy	2	12	Mo	κ	BBABPB**PG**PB	21		
L1–11–1	76	56	1P7KL	rASQdisnyla	57	59	Hu Mo	κ	BPABPDGDPBB	9	II	F71, T71, G71
L1–11–2	55	40	1ZANL	rASqdIsnyLn	37	39	Mo	κ	BPABPD**LL**PBB	8		Y71
L1–11–3	5	4	1W72M	sgnnlgs-svh	5	5	Hu	λ	PB**PLA**A**AB**BPB	25		5_[ILA][GPS]
L1–12–1	5	42	35C8L	rASsSvSSsylh	5	42	Mo	κ	BPABPBPAADBB	9	III	Y71 (4/5); F71 (1/5)
L1–12–2	5	42	2FX7L	raSqSvssnylA	5	42	Hu Mo	κ	BPABPPP**LL**PBB	6		F71 (5/5)
L1–12–3	2	17	2OTUE	TLSsQHSTYTIE	2	17	Hu Mo	λ	BP**PAADA**A**PP**BB	5		All λ
L1–13–1	7	64	2A9ML	SGssSNIGnNyVs	7	64	Hu	λ	BBBAADAADBPBB	10	III	2_G; 5_[ST]
L1–13–2	4	36	1PEWA	TRSSGnIasNYVq	4	36	Hu	λ 6–57*01	PP**ABGB**AAABPBB	12		2_R; 5_G
L1–14–1	14	78	1NC2A	RSStGavTtsNYAn	8	67	Hu	λ	BBAAGPPBAAALPB	11	II	5_G
L1–14–2	4	22	1DCLB	TgtssdvgGynyVs	4	33	Mo	λ	BBBA**ADAA**A**BBB**BB	34		5_[SD]
L1–15–1	11	85	1EJOL	rASeSVDsyGnSfMn	9	82	Hu Mo	κ	PBABPDPBLLBPPBB	12	I	7_[DE][YSFN][YFD]
L1–15–2	2	15	1I7ZA	RASKSVSTSGYnYMH	2	18	Mo	κ	BPABPD**DAA**LBPPBB	14		7_STS
L1–16–1	68	100	2D03L	RSSqslvhsnGnTYLe	50	100	Hu Mo	κ	BBABPAPPAALPBPBB	10	I	
L1–17–1	21	100	1Q9RA	KSSQSLlnSrtrkNYLA	17	100	Hu Mo	κ	BBABPDPPAADLBPPBB	10	I	

Upper-case residues in the consensus sequence are those that represent more than 90% of the residues in the unique sequences. Lower-case residues represent the majority residue if it represents between 20% and 90% of the residues in the alignment. If the majority residue is less than 20% of the positions, then "-" is present. The loop conformation regions are defined in Fig. 4. After the largest cluster, those residues whose conformations differ from the largest cluster are shown in boldface. For this purpose, similar conformations, including (B,P), (A,D), and (L,G) conformations, were not differentiated. The median angle was calculated by averaging the distance in Eq. (2) of all the members of the cluster to the structure with the lowest average median distance to the others. The distance $D(i,j)$ was divided by two times the number of residues (to account for ϕ and ψ) and converted back into angles by inverting Eq. (2). It thus represents the average difference of ϕ or ψ from the median structure over all the loops in each cluster. Classes refer to the predictability of the CDR-length as described in the text. In the "Comments" column, "2_G" indicates that residue 2 is Gly, and "5_[ST]" indicates that residue 5 is either Ser or Thr.

Table 4. Clustering of CDR loop L2

Cluster	Number of structures	Percent loop length	Median PDB entry	Consensus sequence	Number of unique sequences	Percent unique sequences	Species	Gene	Loop conformation	Median angle (°)	Type	Comments
L2–8–1	290	95	1CR9L	Y-asnlas	159	96	Hu Mo	κ	BLLDPPPP	9	I	
L2–8–2	9	3	1FL5A	yaasnldS	8	5	Hu Mo	κ	BLLDPPP**A**	14		
L2–8–3	3	1	1I8KA	segNtlrP	2	1	Mo	κ,λ	B**P**L**LB**PPP	10		
L2–8–4	2	1	1ETZA	gGtnNRvp	2	1	Mo	κ,λ	B**GD**DBPPP	13		
L2–8–5	2	1	2AEPL	YsaSyRyS	2	1	Hu Mo	κ	**DBA**DBPPP	27		
L2–12–1	2	50	2H32A	RYFSQSDKSQGP	1	50	Hu	preB	BBDBAABBBPPA	18	III	Pre-B-cell receptor
L2–12–2	2	50	2OTUC	ELKKDGSHSTGD	1	50	Mo	λ 3*01	BB**PA**A**L**PBBPP**P**	5		

See notes under Table 3.

whose conformational clusters are easily predicted by the identity of certain framework residues, even if the loops in the different clusters do not have significantly different sequences. To belong to this type, loops had to have at least four unique sequences in each of the larger clusters, two or more clusters, and membership that was more than 85% predictable by the sequence of the loop (or by the identity of certain framework residues; see the text).

(3) Type III: Unpredictable CDR-lengths—For some CDR-lengths, structure prediction is likely to be difficult or statistically uncertain. This may occur for a number of reasons. First, the affinity propagation procedure may put most structures into a small number of highly dispersed clusters or into a large number of very small clusters. Second, there may be too few structures to have much confidence in the clustering. In some cases, it may be possible to suggest a sequence motif that determines the cluster, but the data are insufficient to do this with confidence. For other CDR-lengths, the structures may be well clustered into discrete conformations, but there is little systematic variability in their sequences. For these CDR-lengths, structure prediction for loops of unknown structure may depend on interactions with the other CDRs, antigens or the framework.

We discuss each CDR in turn.

L1

With our definitions, L1 can have loop lengths from 10 to 17 residues. The majority of L1 loops are of length 11 or 16, with 57 and 50 unique sequences, respectively. The results of the clustering analysis of L1 are shown in Table 3.

Several L1 lengths are of type I, which means that a single conformation strongly predominates. CDR length L1–10 is one of these, with 20 of 22 total structures (all mouse κ) belonging to a single conformation. The median conformations of the two are BBABPB***AB***BB *versus* BBABPB***PG***PB, which differ primarily in residue positions 7 and 8, involving a flip of the peptide bond between these residues. This is a common and relatively minor difference between two homologous structures. L1–16 also belongs to type I, with all 68 structures belonging to a single cluster. L1–17 is also a single-cluster CDR-length, with all 21 structures having a similar conformation. These loops have normalized average distances from their median structures of 10° per dihedral angle (see Table 3). These small values indicate tight clustering.

L1–11 belongs to type II, having three alternate conformations that are easily predictable by the sequence of the CDR or by the identity of certain framework residues. We refer to these clusters as L1–11–1, L1–11–2, and L1–11–3. We first looked at

Table 5. Clustering of CDR loop L3

Cluster	Number of structures	Percent loop length	Median PDB entry	Consensus sequence	Number of unique sequences	Percent unique sequences	Species	Gene	Loop conformation	Median angle (°)	Type	Comments
L3–7–1	2	100	1DFBL	qQynSYs	2	100	Hu Mo	κ	BPDADLP	18	III	
L3–8–1	15	68	2G5BG	lQyynlrT	13	68	Hu Mo	κ	BPDABGBB	10	III	
L3–8–2	4	18	1A7OL	qqfwrtpT	4	21	Mo	κ	BBD**BGP**PB	41		
L3–8–*cis*6-1	3	14	1E6OL	QqwnyPfT	2	11	Mo	κ	BPA**B**P**aL**P	4		All of Pro6
L3–9–*cis*7-1	219	83	1J1PL	qQgss-PlT	182	83	Hu Mo	κ	BBDABPpPB	10	II	93% of Pro7
L3–9–1	22	8	1F4XL	alw-snhwv	17	8	Hu Mo	κ,λ	BB**PBLLB**PB	37		88% of non-Pro; All all of λ
L3–9–2	12	5	1KCSL	qQsth-ppT	12	6	Hu Mo	κ	BBDAB**AP**PB	20		
L3–9–*cis*7-2	8	3	1G7IA	QHfwsTPrT	7	3	Hu Mo	κ	BPD**PG**BpPB	4		
L3–9–*cis*7-3	2	1	1L7IL	qQyyiyPyT	2	1	Hu Mo	κ	BBDABP**aL**P	21		
L3–9–*cis*6-1	1	<1	2FBJL	QQWTYPLIT	1	<1	Mo	κ	BB**BB**P**dB**BB	—		Only Pro6
L3–10–1	6	67	3B5GB	qsydss-svv	5	63	Hu	λ	BBBPAALPPB	40	III	All of non-Pro
L3–10–*cis*8-1	2	22	1I7ZC	lysrefPPwT	2	25	Mo	κ	BBBBA**BPp**BB	36		All Pro6,7
L3–10–*cis*7,8-1	1	11	1JGUL	SQSTHVPPLT	1	13	Mo	κ	BB**DABBpp**PB	—		All Pro6,7
L3–11–1	9	90	1RZFL	aawdssldavv	9	90	Hu	λ	BBPBAADLBPB	12	I	2 with Pro (position 9.10)
L3–11–*cis*7-1	1	10	2NXYC	QQYNNWPPRYT	1	10	Hu	κ	BP**DAPBpP**BPB	—		Pro7.8
L3–12–1	1	100	3C2AL	ATWDSGLSADWV	1	100	Hu	λ	BBBPAPADLPPB	—	III	
L3–13–1	3	100	2OTUG	aawDdsrggpdwV	2	100	Hu Mo	λ	BBBBPPAABPBBB	41	III	

See notes under Table 3. "All non-Pro" means that all loops in the cluster do not contain proline. "All of non-Pro" means that all non-Pro loops in the CDR loop length are in this cluster. Lower-case letters in the "Loop Conformation" column indicate cis residues.

Table 6. Clustering of CDR loop H1

Cluster	Number of structures	Percent loop length	Median PDB entry	Consensus sequence	Number of unique sequences	Percent unique sequences	Species	Loop conformation	Median angle (°)	Type
H1–10–1	2	100	1KXQF	aAStYTdtvG	2	100	Camel	BPABPBABBB	11	III
H1–12–1	1	100	1GHFH	KLWYTFTDYGMN	1	100	Mo	BBBBPAAABPBB	—	III
H1–13–1	267	87	1UYWM	kaSGftftdyymh	213	86	Hu, Mo	PPBLBPAAABPBB	13	I
H1–13–2	7	2	1C5DB	kaSgfnitdyyis	7	3	Hu, Mo	PP**DA**BPAADBPBB	23	
H1–13–3	5	2	1U0QA	kASGytFttyamn	2	1	Hu, Mo	PBP**GDA**AADBPBB	30	
H1–13–4	4	1	1IC4H	avsGfsfsgyyws	5	2	Hu, Mo	PBBLBPA**PL**PBBB	25	
H1–13–5	4	1	1MVFA	aASGftysinymg	4	2	Hu, Mo	BPBGP**A**AA**PA**BBB	17	
H1–13–6	4	1	2P45B	AaSGykytnycmG	4	2	Camel	BPBLB**ABBPA**BBB	29	
H1–13–7	3	1	1DQDH	svtGdsiTSgywn	4	2	Mo	BBBLBPAA**BG**BBB	14	
H1–13–8	3	1	1HCVA	kaSGytfttydmg	3	1	Mo	PBBG**L**B**G**AABBBB	38	
H1–13–9	3	1	1KXVD	AaSGnTlstydmg	3	1	Camel, Llama	BBPLB**A**A**PB**BPBB	34	
H1–13–10	2	1	1RHHB	KASGGTFSmYgfn	3	1	Hu	BP**AAL**BA**G**DPBBB	13	
H1–13–11	2	1	1UM5H	kASeyTltsylfq	2	1	Mo	BP**AB**P**DLPP**BBBB	32	
H1–13–*cis*9-1	2	1	1JTPA	AASGYTIGPYCMG	1	<1	Camel	BPBLP**DBGpA**BBB	24	
H1–14–1	11	100	1ORSB	TVTGYSITsgYaWn	7	100	Mo	BBBLBPADABGBBB	8	III
H1–15–1	9	100	2HWZH	sfSGFSlstsgmgVg	7	100	Hu, Mo	BBBLBBAAPPLPBBB	16	III
H1–16–1	1	100	1QD0A	AASGRAASGHGHYGMG	1	100	Llama	PBBGPBABGDLBBPBB	—	III

See notes under Table 3.

Table 7. Clustering of CDR loop H2

Cluster	Number of structures	Percent loop length	Median PDB entry	Consensus sequence	Number of unique sequences	Percent unique sequences	Species	Loop conformation	Median angle (°)	Type	Comments
H2–8–1	2	100	1F2XK	tilgGSty	2	100	Hu	BBBGAPBB	35	III	
H2–9–1	77	95	1KIPB	yIwysGsty	57	93	Hu, Mo	BBPAALPBB	10	I	6_[GD]
H2–9–2	2	2	1JGUH	sIyngfrih	2	3	Mo	BBB**LLD**PBB	16		6_[FV]
H2–9–3	2	2	1OSPH	yIrygGgtY	2	3	Mo, Camel	BBB**PL**LPBB	28		6_G
H2–10–1	155	68	2BDNH	-Iypgng-t-	131	67	Hu, Mo	BBPAADLPBB	12	II	R71 (8/155)
H2–10–2	42	19	1SEQH	-Issgggnty	40	20	Hu, Mo	BBPAA**LA**BBB	12		R71 (38/42)
H2–10–3	11	5	2Q76D	eIlPGsgstn	9	5	Hu, Mo	BBB**PG**ALPBB	14		
H2–10–4	7	3	1DSFH	tIssgGgytn	7	4	Mo	BBPP**LLA**BBB	17		
H2–10–5	3	1	2P45B	AisgGGtyih	2	1	Mo, Camel	BBPAA**LBL**PB	8		
H2–10–6	3	1	1OAQH	ridpnGggTk	3	2	Hu, Mo	BBPA**PL**LPBB	11		
H2–10–7	2	1	1INDH	TtlsGggfTf	2	1	Hu, Mo	BBP**PG**D**P**PBB	24		
H2–10–8	2	1	1UWEH	gIdPhnGGga	2	1	Hu, Mo	BBBAAD**GL**PB	23		
H2–10–9	2	1	1UWGY	gIdphnggpv	2	1	Hu, Mo	BBBAD**BPG**PB	22		
H2–12–1	26	100	1Q9RB	eIRnKannytTe	22	100	Mo	BBPPAADLLPBB	10	I	
H2–15–1	1	100	1I3UA	TIGRNLVGPSDFYTR	1	100	Llama	BBPABPDBADPPPBB	—	III	

See notes under Table 3.

Table 8. Clustering of CDR loop H3 anchors

Cluster	Number of structures	Percent H3 anchors	Median PDB entry	Consensus sequence	Number of unique sequences	Percent unique sequences	Loop conformation	Median angle (°)	Comments
H3–anchor-1	204	65	1UYWH	aR- yfdy	169	67	BPP BPAB	21	Bulged
H3–anchor-2	35	11	2J88H	ary dfdY	32	13	BBB **A**B**B**B	30	Nonbulged
H3–anchor-3	25	8	1NQBA	arg yfdy	23	9	BPP **A**BAB	23	Bulged
H3–anchor-4	24	8	1XGUB	anw dgDy	10	4	BP**A AL**AB	7	H3–7 only
H3–anchor-5	12	4	1IL1A	vr- -rdY	12	5	BPB PP**B**B	33	Nonbulged
H3–anchor-6	6	2	1KTRH	as- sfay	5	2	BB**L LL**AB	16	H3–7 only
H3–anchor-7	4	1	1HILB	ARr gfdy	4	2	BPP **GBBL**	26	
H3–anchor–*cis*4-1	2	1	1CE1H	ARe PfDY	2	1	BP**A pL**AB	38	

See notes under Table 3. Anchors are defined as the first three residues of the CDR and the last four residues of the CDR.

the sequence logos[17] derived from the unique sequences in each cluster to determine if sequence can differentiate the clusters; these are shown in Fig. 5. Cluster L1–11–3 has a very different amino acid distribution at positions 5 and 6, where clusters L1–11–1 and L1–11–2 have [SDNE][IV], while L1–11–3 has [ILA][GPS]. The L1–11–3 sequences all come from human Vλ chains, while L1–11–1 and L1–11–2 have very similar amino acid distributions coming from human and mouse Vκ chains. As has been noted by Al-Lazikani *et al.* based on only four structures, the structural difference between L1–11–1 and L1–11–2 is due to a difference in the framework at position 71[2] (Chothia numbering; 18 residues prior to the start of CDR L3; residue 89 in the Honegger–Plückthun numbering system[10]). When the residue at position 71 is Phe, 63 of 67 such structures (94%) are in cluster L1–11–1. All 8 structures with Thr at position 71 and both structures with Gly at position 71 are in L1–11–1. Among 50 structures with Tyr at position 71, 48 of them (96%) are in cluster L1–11–2. Loops in cluster L1–11–1 form a hydrogen bond from the carboxyl oxygen of residue 7 of the CDR to the amide hydrogen atom of residue 68 (21 residues prior to L3). In loops belonging to cluster L1–11–2, the orientation of the amide bond between residue 7 and residue 8 of the CDR is reversed. This directs the amide hydrogen atom of residue 8 towards the hydroxyl oxygen atom of the tyrosine residue at position 71, forming a hydrogen bond. These interactions are shown in Fig. 6.

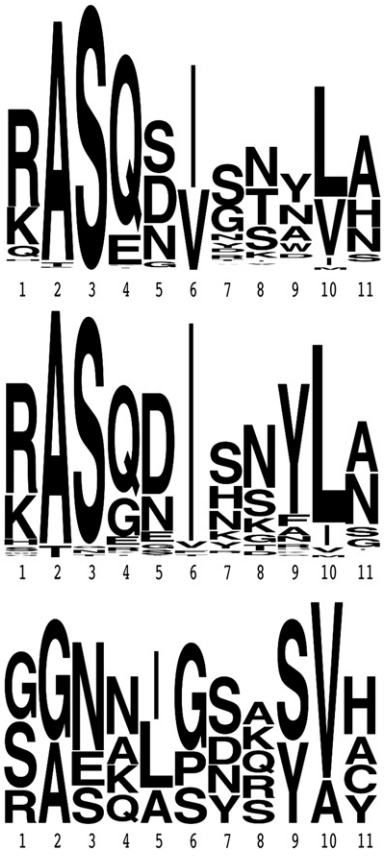

Fig. 5. Sequence logos for the three clusters of L1–11–1, L1–11–2, and L1–11–3 (top to bottom). The logos were drawn with the program WebLogo.[17]

The remaining L1 lengths only have a small number of available structures and sequences, including L1–12 (12 structures, 12 sequences, 3 clusters), L1–13 (11 structures, 11 sequences, 2 clusters), L1–14 (18 structures, 12 sequences, 2 clusters), and L1–15 (13 structures, 11 sequences, 2 clusters). Even here, though, there are some residues that differentiate these clusters; however, because of the small number observations, we cannot be confident that these features will always be predictive. We therefore classify them as being of type III. For instance, cluster L1–12–3 (mouse Vλ) has sequences very different from those of L1–12–1 (mouse Vκ) and L1–12–2 (human and mouse Vκ). Four of five L1–12–1 members have Tyr71, while all five L1–12–2 members have Phe71. The two clusters of L1–13 (both human Vλ) are easily distinguishable by sequence at positions 2 and 5, with the first five residues of L1–13–1 having the sequence motif [ST]**G**[ST][SAT]**[ST]** and with L1–13–2 having the sequence motif T**R**S**S**G. The Gly at position 5 of L1–13–2 presumably allows the γ conformation for this residue (ϕ,ψ = +70°,+160°). The two clusters of L1–14 have quite different sequences;

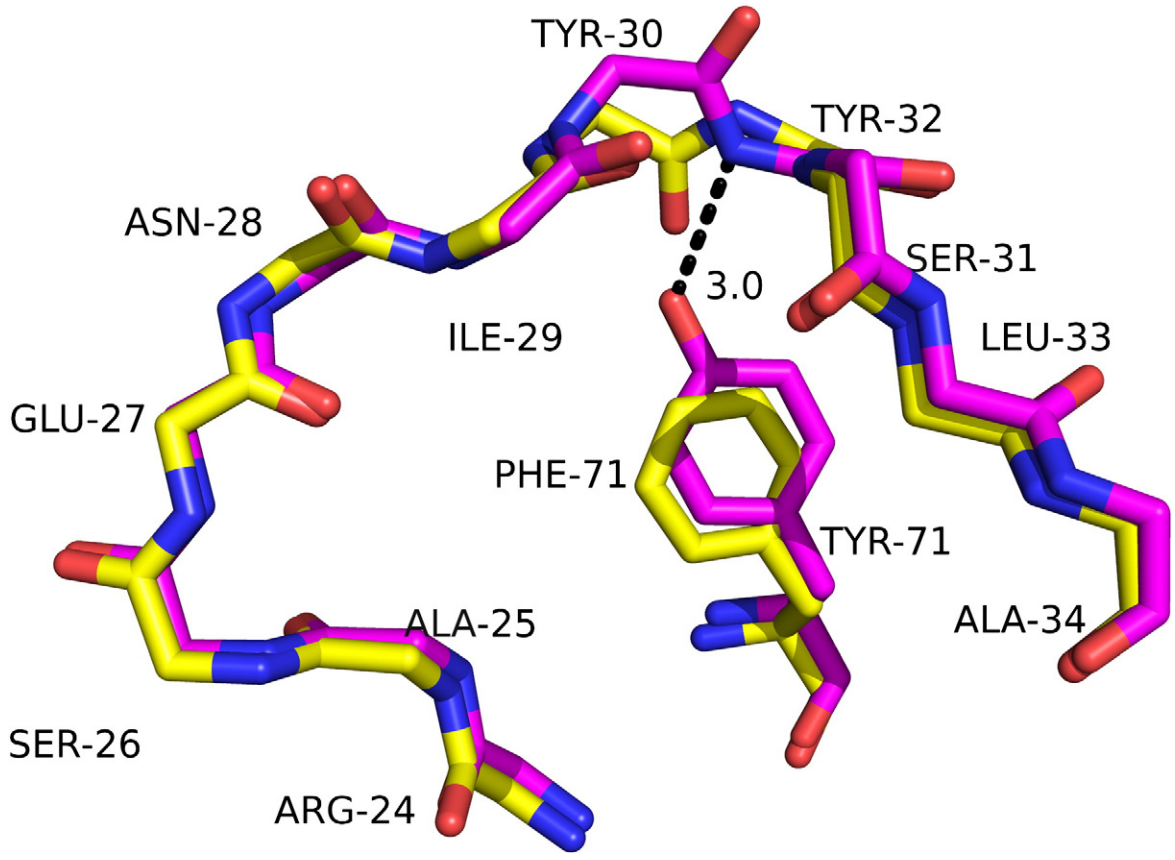

Fig. 6. The median structures of clusters L1–11–1 (yellow) and L1–11–2 (magenta). The hydrogen bond of Tyr71 to the NH of residue 8 in cluster 2 is shown. The sequence and residue numbering given are from the median structure of L1–11–1 (PDB chain 1P7KL).[18]

the human sequences in cluster L1–14–1 have the consensus sequence RSStGavTtsNYAN (completely conserved residues are in upper case), and the mouse sequences in L1–14–2 have the consensus sequence TgtssnvgGynyVs. The Gly at position 5 of L1–14–1 presumably favors the γ conformation for this residue. Finally, cluster L1–15–2 has only two mouse Vκ members that differ by only one residue from each other. The conformations of L1–15–1 and L1–15–2 differ at positions 7–9 with sequences [DE][YSFN][YFD] and STS, respectively.

L2

The results of the clustering analysis for L2 are shown in Table 4. L2 loops of known structure only come in two lengths: L2–8 and L2–12. There are 308 structures for L2–8; among these, 290 (94%) consisting of 159 unique sequences belong to the majority cluster with a conformation of BLLDPPPP. The next most common cluster, with 9 structures, has a median structure with a conformation of BLLDPPPA, which varies from the main conformation only at the last residue. There are also three additional very small clusters. We consider L2–8 to be of type I, i.e., effectively having only one conformation.

L2–12 contains only four structures in two clusters, each with only a single unique sequence. The first is the structure of the human pre B-cell receptor, while the second is a mouse Vλ structure. With so few sequences, this loop is of type III.

L3

The results of the clustering analysis for L3 are shown in Table 5. L3 loops come in lengths 7 through 13, and 85% of L3 loops are of length 9. The largest cluster of L3–9, representing 83% of this CDR-length, is one that contains a *cis*-proline at position 7, which we designate as L3–9–*cis*7-1. There are two additional very small clusters with *cis*7, two clusters that are all-trans, and one cluster that has *cis*6. The structure of an L3–9 loop can be predicted fairly well merely by the positions of the proline residues, if any. If all L3–9 loops with Pro7 are predicted to be in cluster L3–9–*cis*7-1, then this prediction is correct 219 of 235 times or 93.2% of the time (and 93.8% for unique sequences). Of the remainder, 10 are in the

other *cis*7 clusters, and 6 are in all-trans cluster L3–9–2. If Pro is entirely absent from L3–9, then 22 of 25 (or 88%) are in cluster L3–9–1. L3–9 is therefore of type II and is generally predictable in structure. See Fig. 7 for the superpositions of the representative structures of each of the four largest clusters.

There are three additional CDR-lengths for L3 that contain more than one cluster, and all three are of type III (i.e., having small numbers): L3–8, L3–10, and L3–11. All three L3–8 loops with Pro at position 6 belong to the L3–8–*cis*6-1 cluster. There are two all-trans clusters, which have no distinguishing sequence features from each other. For L3–10, all loops with no prolines belong to the all-trans cluster L3–10–1. The two clusters L3–10–*cis*8-1 and L3–10–*cis*7,8-1 both contain two prolines at positions 7 and 8. The single L3–11–*cis*7-1 structure has Pro at positions 7 and 8, while none of the all-trans L3–11–1 structures does.

Three loop lengths(L3–7, L3–12 and L3–13) have only one conformation and one or two unique sequences, and are therefore of type III. The latter two CDR-lengths are λ sequences.

H1

The results of the clustering analysis for H1 are shown in Table 6. CDR H1 comes in lengths 12 through 16 and also in length 10. The shortest and longest H1 sequences come from camelid antibodies. CDR-length H1–13 represents 92% of the H1 loops and is dominated by a single conformation. Cluster H1–13–1 comprises 267 of 306 structures (or 87%), with a conformation of PPBLBPAAABPBB and a minimum normalized median angle of 13° (see Table 5). It is therefore of type I. The remaining 39 structures are distributed over 11 different clusters with a wide range of possible structures. No obvious sequence differences exist among them, except that three of them occur only for camelid antibodies. The other CDR-lengths for H1 all exist in single clusters. However, they each contain fewer than 10 unique sequences; therefore, these CDR lengths are of type III.

H2

The results of the clustering analysis for H2 are shown in Table 7. For H2, there are two common loop lengths (H2–9 and H2–10), each with multiple clusters, as well as three loop lengths with only one cluster each (H2–8, H2–12, and H2–15). For H2–9, 77 of 81 structures (or 95%) belong to cluster H2–9–1, with a minimum normalized median angle of 10° (see Table 7). It is therefore of type I. All of the H2–9 human sequences are in H2–9–1. Clusters H2–9–1 and H2–9–3 both have an L conformation at position 6, while cluster H2–9–2 has a D conformation. Consistent with this, H2–9–1 and H2–9–3 have mostly Gly at this position (and a few Asp in H2–9–1), while H2–9–2 has Phe and Val.

CDR-length H2–10 represents 67% of all H2 loops. It is grouped into two large clusters (68% and 19% of structures) and seven much smaller clusters. We examined the sequence logos for the top four clusters and found that there are different patterns of the positions of Gly and Pro in the middle of the loop at several positions, as shown in Fig. 8. There are left-handed L or G conformations at positions 7, 6, 5, and 5+6 for the top four clusters, respectively.

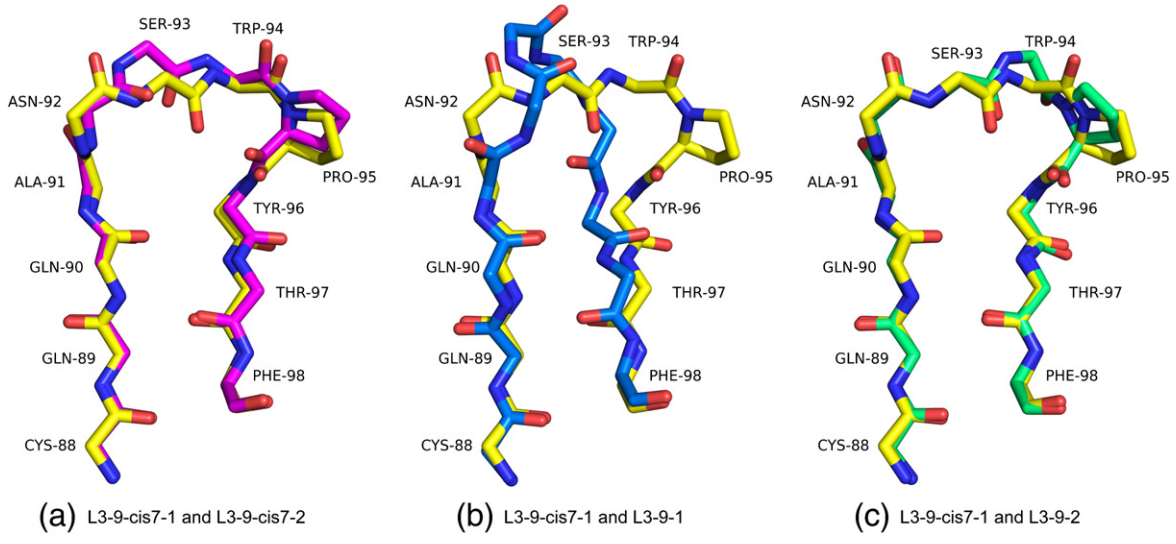

Fig. 7. The median structures of the largest clusters of L3–9. (a) L3–9–*cis*7-1 (yellow)+L3–9–*cis*7-2 (magenta). (b) L3–9–*cis*7-1 (yellow)+L3–9–1 (blue). (c) L3–9–*cis*7-1 (yellow)+L3–9–2 (green). The sequence of L3–9–*cis*7-1 from PDB chain 1J1PL is marked.[19]

Fig. 8. Sequence logos for clusters H2–10–1, H2–10–2, H2–10–3, and H2–10–4 (top to bottom, respectively).

No one position was completely predictive, so we created hidden Markov models with HMMER[20] based on the unique sequences in each cluster and then assigned each loop to the cluster with which it scored the highest. For cluster H2–10–1 with a conformation of BBPAADLPBB, 130 of 155 structures (or 84%) are predicted correctly. For cluster H2–10–2 with a conformation of BBPAALABBB, 30 of 42 (or 71%) structures are correctly predicted to be in the cluster. H2–10–3 and H2–10–4 are not as well predicted but are much smaller in population. H2–10–3, with a conformation of BBBPGALPBB, has 6 of 11 structures predicted correctly. Finally, H2–10–4, with a conformation of BBPPLLABBB, has only 2 of 7 structures predicted correctly. Overall, however, the scores of the loop sequences of H2–10 against the HMMs of its clusters are good at predicting the cluster membership of the sequences.

Additionally for H2, Tramontano *et al.* noted the effect of framework residues on determining the conformation of the loop, particularly the identity of residue 71[21] (Chothia numbering; 25 residues before the start of H3; Honegger and Plückthun[10] number 82). Using our CDR definitions, they analyzed H2–9, H2–10, and H2–12 (their lengths 3, 4, and 6); however, in 1990, they had only 2, 3, and 2 structures, respectively. We decided to investigate this to see if it holds up with a much larger data set. For H2–9, they found only one conformation regardless of position 71 (Val and Arg). We also found effectively only one conformation (H2–9–1 = 77 of 81 structures). Position 71 was not helpful in distinguishing H2–9–2 and H2–9–3 (data not shown) from H2–9–1. For H2–10, Tramontano *et al.* found two conformations: two structures with Arg71 similar to our cluster H2–10–2 and one structure with Ala similar to our cluster H2–10–1. In Table 9, we show a contingency table for H2–10 with the different residues at position 71 in columns and with the different clusters of H2–10 in rows. We have a total of 227 structures and 196 unique sequences for H2–10; we also have nine conformational clusters instead of just two, although only the first two clusters are highly populated. If we predict the cluster to which a structure belongs merely from position 71, we would assign the cluster with the highest number in each column of Table 9. For example, if position 71 is Ala, we would predict cluster H2–10–1, and we would get 67 correct assignments and 13 incorrect assignments. If position 71 is Arg, we would predict cluster H2–10–2 and get 38 of 58 assignments correct. If we add the largest numbers in each cluster, we would correctly predict 186 of the loops (or 80%), comparable to the hidden Markov models discussed above (78% of the loops in clusters 1–4). As the table shows, the major

Table 9. Residue 71 and H2–10 contingency table

Cluster	A	D	I	L	Q	R	S	T	V	Number in cluster
1	**67**	0	**1**	**23**	**4**	8	**4**	**3**	**45**	155
2	2	**1**	0	0	1	**38**	0	0	0	42
3	9	0	0	0	1	0	0	0	1	11
4	2	0	0	0	0	5	0	0	0	7
5	0	0	0	0	0	3	0	0	0	3
6	0	0	0	0	1	1	0	0	1	3
7	0	0	0	0	0	2	0	0	0	2
8	0	0	0	0	0	0	0	0	2	2
9	0	0	0	0	0	1	0	0	1	2
Residue total	80	1	1	23	7	58	4	3	50	227

The number of the CDR structures of H2–10 with each residue type in each cluster is given below the residue type. The most common cluster for each residue type is given in boldface.

determinant is whether the residue at position 71 is a small hydrophobic residue (A, I, L, or V), a small polar residue (S and T), or Q, in which case the loop mostly belongs to cluster H2–10–1 (143 of 161 times, or 90%); if the residue is R or D, then the residue belongs to cluster H2–10–2 (39 of 59 times, or 66%). Superpositions of the median structure of cluster H2–10–1 with clusters 2, 3, and 4 are shown in Fig. 9. Both clusters H2–10–2 and H2–10–4 have Arg at position 71 and a hydrogen bond to the carbonyl oxygen of residue 3 of the CDR.

Finally, for H2–12, all 26 structures belong to a single tight cluster with a minimum normalized median angle of 8°, therefore qualifying this loop as type I while the two very small population CDR lengths H2–8 and H2–15 also have only one cluster (type III).

H3

The known loop structures for H3 are very diverse in length, ranging from 5 to 26 residues, with the majority (86%) between 7 and 16 residues. The shorter loops can be clustered well, but these are low in population (Table 2). The longer loops form a few large clusters with higher self-similarity values, but the clusters have very large distances to the median. Some clusters have residues in different bins of the Ramachandran map (e.g., A and L regions). At low self-similarity, the number of clusters becomes very large, and the cluster sizes become rather small. They are therefore not likely to have predictive value.

Because of these difficulties, a number of analyses have split H3 into a "torso" or anchor region corresponding to its N-terminal and C-terminal ends and a "head" or apex region at the turn of the loop,[3,22] dividing the torso region into two groups: "bulged" and "nonbulged,"[4] or "kinked" and "extended."[23] We performed affinity propagation clustering on a set of seven residues comprising the first three residues of H3 (in red for the N-terminal region in Fig. 1) and the last four residues of H3 (in red for the C-terminal region in Fig. 1 plus one more to the left). The clustering results for these seven-residue discontinuous peptides are shown in Table 8. For the H3 torso clustering, a total of eight clusters are apparent. Cluster H3–anchor-1 covers about two-thirds of the structures, and the top four clusters cover about 95%. The first three clusters are shown in Fig. 10. Contingency tables on individual residue positions did not demonstrate the predictability of the H3 torso clusters (data not shown) much beyond the 65% in the first cluster.

We examined the distribution of these clusters for H3 loops of different lengths. The results are shown in Table 10. We included H3–7 loops in the H3–anchor clustering, even though these would not be expected to cluster well with the torso regions of the longer loops. Indeed, these loops clustered predominantly into three clusters, separately from the others: H3–anchor-4, H3–anchor-6, and a cluster with *cis*4. A small number of H3–7 structures were placed in cluster H3–anchor-1. Interestingly, for the other lengths, the distribution is somewhat dependent on length. For H3–8 (only 5 structures), 2, 1, and 2 of the structures are in H3–anchor-1, H3–anchor-2, and H3–anchor-5, respectively. H3–9 (26 structures) is the only H3 CDR-length for which the nonbulged H3–anchor-2 cluster predominates. For H3 lengths from 10 to 14, 74–79% of structures

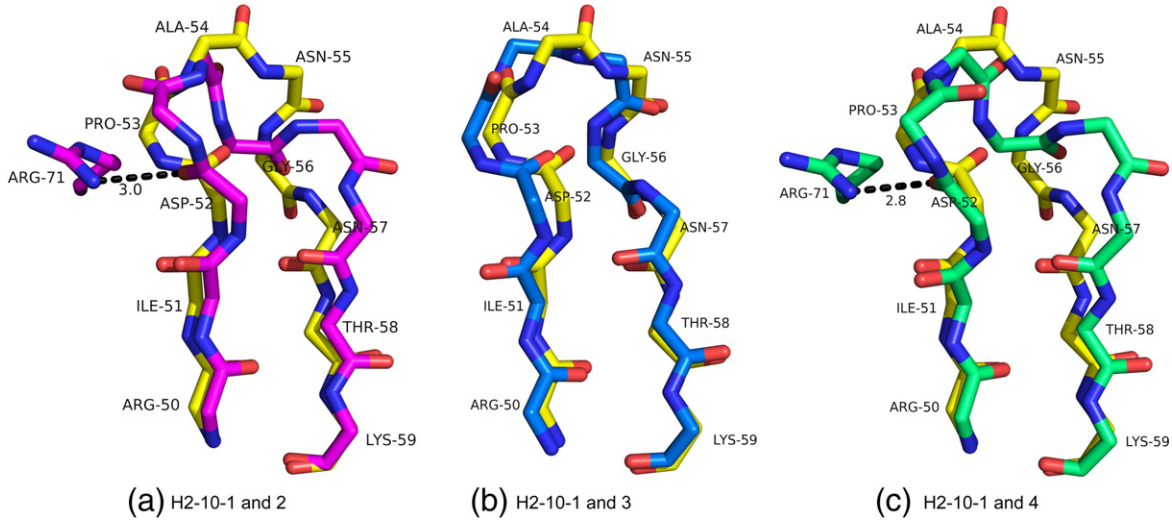

Fig. 9. The median structures of the largest clusters of H2–10. (a) Cluster H2–10–1 (yellow) and cluster H2–10–2 (magenta). (b) Cluster H2–10–1 (yellow) and cluster H2–10–3 (blue). (c) Cluster H2–10–1 (yellow) and cluster H2–10–4 (green). The side chains of Arg71 of clusters H2–10–2 and H2–10–4 are shown. This residue is Ala in clusters 1 and 3 (not shown).

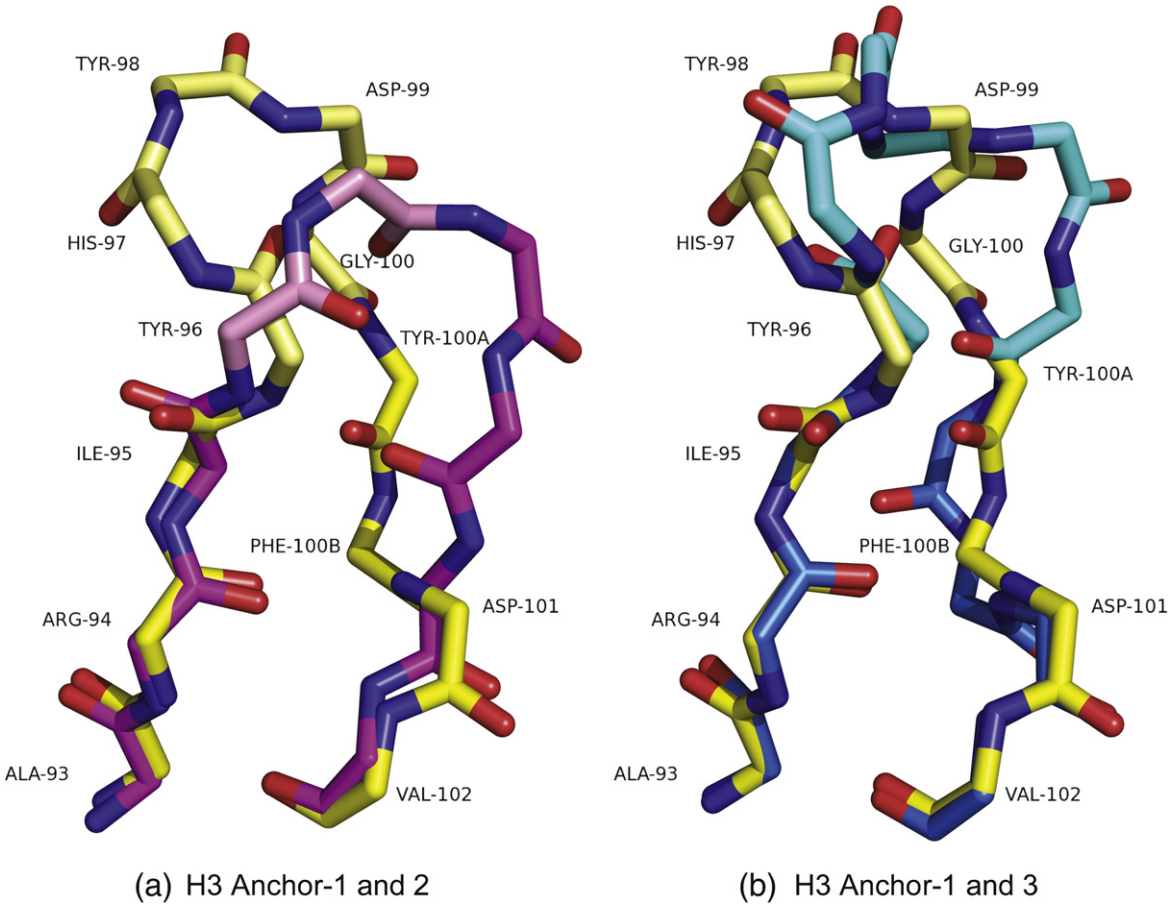

Fig. 10. The median structures of the H3–anchor regions. (a) Cluster H3–anchor-1 (yellow) and cluster H3–anchor-2 (magenta). (b) Cluster H3–anchor-1 (yellow) and cluster H3–anchor-3 (blue/green). Cluster H3–anchor-1 and cluster H3–anchor-3 are bulged; cluster H3–anchor-2 is nonbulged.

belong to H3–anchor 1. However, for lengths 15 and 16, 92% of the structures belong to H3–anchor-1, while the remainder are in cluster H3–anchor-5. For H3 loops longer than 16, 71% belong to H3–anchor-1, while the remainder belong to H3–anchor-3. These frequencies are consistent across loop lengths from 17 to 26 (data not shown).

Comparison to Chothia and Martin–Thornton clustering

There are several previous studies on the categorization of antibody loop structures.[1,4,5,7] The clustering results in this study recapitulate many of the canonical conformations found by both

Table 10. H3–anchor cluster frequencies (%) for each H3 loop length

Loop length	Count	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster *cis*4-1
7	34	9	—	—	**68**	—	18	—	3
8	5	**40**	20	—	—	**40**	—	—	—
9	26	35	**65**	—	—	—	—	—	—
10	29	**79**	21	—	—	—	—	—	—
11	27	**71**	11	19	—	—	—	—	—
12	37	**74**	6	12	—	6	—	2	—
13	40	**78**	10	5	—	—	—	8	—
14	28	**75**	4	7	—	14	—	—	—
15	13	**92**	—	—	—	8	—	—	—
16	24	**92**	—	—	—	8	—	—	—
>16	34	**71**	—	29	—	—	—	—	—

For each loop length of H3, the number of structures ("Count") and the percentage in each cluster are given. The most prevalent cluster percentage for each loop length is shown in boldface.

Al-Lazikani *et al.*[2] and Martin and Thornton.[11] However, our conformational clustering approach and more recent structure database have produced a few significant differences with the results of Al-Lazikani *et al.* and Martin and Thornton. The correspondences between our clustering and those of Al-Lazikani *et al.* and Martin and Thornton are given in Tables 11, 12, and 13.

We used the 1997 study of Al-Lazikani *et al.*[2] to define the Chothia canonical conformations, since this is the most recent and most comprehensive of their previous analyses of antibody CDR structures.[1,21,24–26] Chothia *et al.* designated canonical classes for each CDR by integers (1, 2, 3, etc.) regardless of the length of the loop and in no particular order. Different designations might be loops of different lengths or loops of the same length but of different conformations. CDRs of λ light chains were analyzed and numbered separately from κ chains; following Martin and Thornton, we call them 1λ, 2λ, etc. Some classes were broken down into subclasses, usually because of a flip of a two-amino-acid segment within the loop between one structure and another. They designated these as A, B, etc., and we append these to the Chothia class name (e.g., L1–2A and L1–2B). For each canonical

Table 11. Chothia canonical conformations

Chothia	Number	This work	PDB chains
L1–1	1	L1–10–1	2FBJL
L1–2A	2	L1–11–1	1IGML 1FVCA
L1–2B	2	L1–11–2	1FGVL 1VFAA (7)
L1–3	2	L1–17–1	1HILA 2IMMA
L1–4	3	L1–16–1	1FLRL (14) 2CGRL (10) 1TETL (9)
L1–5	5	L1–15–1	"59.1" 1ACYL (15) 1AI1L (17) ***"50.1"1GGIM (31)1GGCL (30) "40±50" 1IBGL (36)***
L1–6	***1***	***L1–12–3***	***"1F7" 1FIGL (52; 3 Å resolution)***
L1–1λ	2	L1–13–1	2FB4L 2RHEA
L1–2λ	***1***	***L1–14–2***	***7FABL (43)***
L1–3λA	2	L1–14–1	1INDL 1GIGL (9)
L1–3λB	1	L1–14–1	1MFAL
L1–4λ	1	L1–11–3	8FABA (Chothia length incorrect)
L2–1	16	L2–8–1	1FGVL 1FLRL 1HILAL 1IGML 1INDL 1MFAL 2FB4L 2IMMA 1FVCA (14) 1VFAA (7) 1TETL (9) 2CGRL (7) 2RHEA (9) 8FABC (9)
		L2–8–2	2FBJL 1GIGL (14)
L3–1	9	L3–9–*cis*7-1	1FGVL 1HILA 1IGML 2IMMA 1FVCA (9) 1FLRL (9) 1TETL (11) 2CGRL (12)
		L3–9–*cis*7-2	1VFAA (4)
L3–2	1	L3–9–*cis*6-1	2FBJL
L3–3	3	L3–8–2	"HyHel5" 1YQVL 2IFFL ***(50) 1BQLL (47)***
L3–4	1	L3–7–1	"3D6" 1DFBL
L3–5	1	L3–10–*cis*8-1	"AN02" 1BAFL
L3–1λA	2	L3–9–1	1INDL (21) ***1GIGL (35)***
L3–1λB	***1***	***L3–9–1***	***7FABL (46)***
L3–1λC	1	L3–9–1	1MFAL
L3–2λ	2	L3–11–1	2FB4L 2RHEA
H1–1	15	H1–13–1	1FGVH 1FVCB 1GIGH 1HILB 1IGMH 1INDH 1MFAH 2FB4H 2FBJH 7FABH 8FABH 1VFAB (14) 1TETH (12) 2CGRH (11)
		H1–13–2	***1FLRH (29; B-factor=88)***
H1–2	1	H1–14–1	"AN02" 1BAFH
H1–3	4	H1–15–1	"59.1" 1ACYH (22) 1AI1L (18) "50.1" 1GGIH (15) 1GGCH (18)
H2–1	3	H2–9–1	7FABH 1GIGH 1VFAB (5)
H2–2A	4	H2–10–1	1FVCD 1FGVH 1MFAH (10) 1TETH (19)
H2–2B	1	H2–10–3	2CGRH (16)
H2–3A	4	H2–10–2	1HILB 2FB4H 2FBJH 8FABD
H2–3B	1	H2–10–7	1INDH
H2–3C	1	H2–10–2	1IGMH
H2–4	1	H2–12–1	1FLRH (20)

The canonical classes defined by Al-Lazikani *et al.* are given with the number of PDB structures listed in their study; PDB entries are listed in the last column.[2] If the CDRs of the PDB entries given by Al-Lazikani *et al.* are in our data set, then their cluster is given in column 3. If these PDB entries are not in our data, then the distance to the nearest cluster is given in parentheses after the PDB chain. If the distance to the nearest cluster median is larger than 25°, then the PDB and our cluster are shown in italic boldface. These are uncertain assignments. For several canonical classes, Al-Lazikani *et al.* only give the antibody name, and we show the name in quotation marks. The PDB entries currently in the PDB for that antibody are listed. When all PDBs for the conformation or cluster in Column 1 are more than 25 degrees from one of our cluster members, then the entire line is shown in italic-boldface.

Table 12. Conformational clusters of Martin and Thornton

Martin	Number	This work	PDB chains
L1–10A	4	L1–10–1	2FBJL 1FORL 1BAFL (8) 1YQV (0)
L1–11A	22	L1–11–1	1IGML 1FVCA 1DFBL 1IVLA 1MLBA 1FVDA 1IGCL 1WTLA 1REIA (19)
		L1–11–2	1FGVL 1FAIL 1IKFL 1JHLL 1MAML 1VFAA (7)+5 more; ***1BBJA (40) 3HFML (46)***
L1–11B	1	L1–11–3	8FABA
L1–12A	***1***	***L1–12–3***	***1FIGL (52; 3 Å resolution)***
L1–13A	2	L1–13–1	2FB4L 2RHEA
L1–14A	***1***	***L1–14–2***	***7FABL (43)***
L1–14B	3	L1–14–1	1GIGL 1MFAL 1INDL
L1–14C	***1***	***L1–14–2***	***1MCWW (58; 3.5 Å resolution)***
L1–14D	***1***	***L1–14–2***	***2MCG1 (69; ConfE=13.1)***
L1–14E	***1***	***L1–14–2***	***1MCWM (28; 3.5 Å resolution)***
L1–14F	***1***	***L1–13–1***	***4BJLB (22; ConfE=9.7) (length incorrect in the Martin–Thornton study)***
L1–15A	***1***	***L1–15–1***	***1GGIL (31; 2.8 Å resolution)***
L1–15B	2	L1–15–1	1ACYL (15) ***1IBGL*** (36)
L1–16A	9	L1–16–1	1LMKA 1TETL (9) 2CGRL (10) 1FPTL (19) 1IGIL (12) 1RMFL (14) ***1CGSL (28) 1DBBL (26) 1IGFL (27)***
L1–16B	2	L1–16–1	4FABL ***1NBVL (44; ConfE=9.8)***
L1–16C	1	L1–16–1	2JELL (10)
L1–17A	4	L1–17–1	1HILA ***1BBDL*** (17) 1FRGL (12) 1MCPL (22)
L2–7A	55	L2–8–1	1FGVL 1FLRL 1HILA 1IGML 1INDL 1MFAL+38 more
		L2–8–2	2FBJL 1BAFL (12) 1GIGL (14) 1BBJL (25)
		L2–8–4	1IGCL 1RMFL (18) 1MCPL (19) ***3HFML*** (28)
		L2–8–5	***4BJLB (33; ConfE=10.7) 1FPTL (45; 3 Å resolution) 1FIGL (48; 3 Å resolution)***
L2–7B	***1***	***L2–8–2***	***6FABL (65; ConfE=9.8)***
L3–7A	1	L3–7–1	1DFBL
L3–8A	1	L3–8–2	1YQVL
L3–8B	1	L3–8–1	1EAPA (14)
L3–9A	40	L3–9–*cis*7-1	1FGVL 1HILA 1IGML 1FVCA 1TETL+30 more; ***1BBDL (25) 1CGSL (30)***
		L3–9–*cis*7-2	1VFAA (4) ***1BBJA (40)***
		L3–9–2	2GFBM (9)
L3–9B	1	L3–9–*cis*6-1	2FBJL
L3–9C	2	L3–9–1	1MFAL (23) ***7FABL (46)***
L3–9D	2	L3–9–1	1INDL (21) 1GIGL (19)
L3–9E	1	***L3–9–cis7-1***	***1FIGL (63; 3.0 Å resolution)***
L3–9F	1	***L3–9–1***	***8FABA (54; ConfE=9.8)***
L3–10A	1	***L3–10–cis8-1***	***1BAFL (56; 2.9 Å resolution)***
L3–10B	1	***L3–10–1***	***1MCWW (56; 3.5 Å resolution)***
L3–10C	1	***L3–10–1***	***2MCG1 (53; ConfE=9.9)***
L3–10D	1	***L3–10–1***	***1MCWM (50; 3.5 Å resolution)***
L3–11A	2	L3–11–1	2FB4L 2RHEA
L3–11B	1	***L3–11–1***	***4BJLB (24; ConfE=9.7)***
H1–10A	42	H1–13–1	1FGVH 1FVCH 1GIGH 1HILB+35 more; ***1CGSH (29)*** ***3HFMH (43; 3 Å resolution) 4FABH (37; ConfE=9.8)***
H1–10B	***1***	***H1–13–8***	***1IGIH (51)***
H1–10C	***1***	***H1–13–10***	***1NBVH (55; ConfE=10.4)***
H1–10D	***1***	***H1–13–7***	***1FIGH (41; 3.0 Å resolution)***
H1–11A	***1***	***H1–14–1***	***1BAFH (33; 2.9 Å resolution)***
H1–12A	2	H1–15–1	1GGIH (15) 1ACYL (22) (typo in Table 6 indicates 10A)
H2–9A	8	H2–9–1	7FABH 1GIGH 1IBGH 1VFAB (5) 1BAFH (14) 1ACYH (18) 1GGIJ (17)

Table 12 (*continued*)

Martin	Number	This work	PDB chains
		H2–9–3	***3HFMH (33)***
H2–10A	21	H2–10–1	1FVCD 1FGVG 1FOR 1FVD 1MLB 1FAI 1JHL+10 more
		H2–10–3	1LMKA 1YQVH (14) 1CGSH (22) 2CGRH (16)
H2–10B	11	H2–10–2	1HILB 2FB4H 2FBJH 8FABD 1IGMH 1DFBH+5 more
H2–10C	2	H2–10–7	1INDH ***1BBJB*** (37)
H2–10D	***1***	***H2–10–7***	***1RMFH (73; 2.8 Å resolution)***
H2–10E	***1***	***H2–10–1***	***6FABH (44; outlier)***
H2–10F	***1***	***H2–10–1***	***1FIGH (50; 3.0 Å resolution)***
H2–12A	***2***	***H2–12–1***	***1MCPH (45; ConfE=9.6) 1MAMH (48; outlier)***
H2–12B	2	H2–12–1	4FABH (24) 1NBVH (18)

The annotations for the clusters given by Martin and Thornton are the same as those in Table 11. Some loops were removed from our data sets due to low resolution or large conformational energies. When all PDBs for the conformation or cluster in Column 1 are more than 25 degrees from one of our cluster members, then the entire line is shown in italic-boldface.

class, they provided one or more PDB entries that fit that class and the CDR sequences of those loops and their ϕ,ψ values. For some loops, they provided only the names of antibodies, and we located the corresponding PDB entries from these names. Their clustering, based on a total of 17 high-resolution structures, was performed manually and visually, not computationally.

Martin and Thornton performed clustering in dihedral-angle space (using vectors of sines and cosines), similar to the one performed here, followed by the merging of clusters based on coordinate RMSD.[11] They designated their clusters by CDR, length, and letters for each different conformation (viz. L1–11A, L1–11B, L1–12A, etc.). They provided PDB entries for a representative of each cluster, as well as a table of assignments of their clusters to 57 PDB entries.

Our CDR definitions differ somewhat from those of Chothia *et al.* and Martin and Thornton. A comparison of these definitions applied to sample κ, λ, and heavy-chain sequences is given in Fig. 11. For Chothia, we use the example sequences given in the study by Al-Lazikani *et al.* These are the regions within the Kabat-defined CDRs that they observe to vary in conformation, usually with one extra amino acid on each end for good measure. The regions described in Al-Lazikani *et al.* do not always coincide with what others take to be the "Chothia definitions" of the CDRs.[10,27] As shown in Fig. 11, Chothia *et al.* defined their κ and λ CDR1s differently from each other. Their κ definition is two amino acids shorter on both the N-terminus and the C-terminus of our L1 definition. Their λ definition is only one amino acid shorter on each end. Their L2 definition is three residues shorter than ours on the C-terminus, and their L3 definition is one residue shorter than ours on the N-terminus. Similarly to L1 κ, our H1 definition is two residues longer on both ends than the Chothia definition, as is our H2 definition. Martin and Thornton used the same CDR definitions as we do for L1, L3, and H2. Their L2 begins one residue after ours, and their H1 begins three residues after ours (ours begins as our L1 does immediately after Cys, while theirs begins after Cys-Xxx-Xxx-Xxx).

For both Chothia and Martin–Thornton, we used the PDB entries given in their studies to match their clusters to ours. In many cases, the same PDB chains are present in our filtered data, and we can make a one-to-one correspondence. In some cases, we excluded some PDB entries or particular loops because of low resolution, high *B*-factors, high conformational energies, or removal of redundant sequences. In these cases, we calculated our distance function D between the loop in the PDB entry cited by either study and the median of our clusters for the same CDR and same length. We normalized D by two times the number of residues in the loop (to account for ϕ and ψ) and then inverted Eq. (2) to calculate an average difference in ϕ or ψ (in degrees).

The results of these comparisons are given in Table 11 for the Chothia data and in Table 12 for the Martin–Thornton data. The tables provide some or all of the PDB entries mentioned in these studies for each of their loop designations. If the chain is listed along with our cluster designation, then that loop is in our clustering data and is present in that loop cluster. If a distance is given in parentheses after the PDB chain, then that is the mean absolute difference in ϕ and ψ angles from the median of our loop cluster. In some cases, this distance is larger than 25°, and we list these in italic boldface. These correspondences are then less certain and may be the result of the low resolution or high *B*-factors of that loop in the PDB. This is noted in some cases.

Al-Lazikani *et al.* listed 25 canonical classes over 20 CDR–length combinations in their 1997 study (Table 11); if we consider their alternate conformations within a class as separate classes, then there are 32 classes. It should be noted that only 3 of these 32

Table 13. Clusters in this work and those in Chothia *et al.* and Martin–Thornton

This work	Number	Chothia *et al.*	Martin and Thornton	Number in this work	Number in Chothia *et al.*	Number in Martin and Thornton
L1–10–1	20	L1–1 (1/1)	L1–10A (4/4)	2	1	1
L1–10–2	2	—	—			
L1–11–1	76	L1–2A (2/2)	L1–11A (9/22)	3	2 (3)	2
L1–11–2	55	L1–2B (2/2)	L1–11A (13/22)			
L1–11–3	5	L1–4λ (1/1)	L1–11B (1/1)			
L1–12–1	5	—	—	3	1	1
L1–12–2	5	—	—			
L1–12–3	2	***L1–6 (1/1)***	***L1–12A (1/1)***			
L1–13–1	7	L1–1λ (2/2)	L1–13A (2/2)	2	1	1
L1–13–2	4	—	—			
L1–14–1	14	L1–3λA (2/2) 3λB (1/1)	L1–14B (3/3); ***L1–14F (1/1)***	2	2 (3)	6
L1–14–2	4	***L1–2λ (1/1)***	***L1–14A (1/1)***; ***14C (1/1)***; ***14D (2/2)***; ***14E (1/1)***			
L1–15–1	11	L1–5 (2/5; ***3/5***)	L1–15B (2/2); ***L1–15A (1/1)***	2	1	2
L1–15–2	2	—	—			
L1–16–1	68	L1–4 (3/3)	L1–16A (3/6; ***3/6***), L1–16B (1/2; ***1/2***), L1–16C (1/1)	1	1	3
L1–17–1	21	L1–3 (2/2)	L1–17A (4/4)	1	1	1
L2–8–1	290	L2–1 (14/16)	L2–7A (44/55)	5	1	1
L2–8–2	9	L2–1 (2/16)	L2–7A (4/55)			
L2–8–3	3	—	—			
L2–8–4	2	—	L2–7A (4/55)			
L2–8–5	2	—	***L2–7A (3/55)***			
L2–12–1	2	—	—	2	—	—
L2–12–2	2	—	—			
L3–7–1	2	L3–4 (1/1)	L3–7A (1/1)	1	1	1
L3–8–*cis*6-1	3	—	—	3	1	1
L3–8–1	15	—	L3–8B (1/1)			
L3–8–2	4	L3–3 (1/3; ***2/3***)	L3–8A (1/1)			
L3–9–*cis*6-1	1	L3–2 (1/1)	L3–9B (1/1)	6	3 (4)	6
L3–9–*cis*7-1	219	L3–1 (8/9)	L3–9A (35/40; ***2/40***); ***L3–9E (1/1)***			
L3–9–*cis*7-2	8	L3–1 (1/9)	L3–9A (1/40; ***1/40***)			
L3–9–*cis*7-3	2	—	L3–9A			
L3–9–1	22	L3–1λA (1/2; ***1/2***), 1λC (1/1)	L3–9C (1/1; ***1/1***), D (2/2), ***F (1/1)***			
L3–9–2	12	—	***L3–9A (1/40)***			
L3–10–*cis*7,8-1	1	—	—			
L3–10–*cis*8-1	2	L3–5 (1/1)	***L3–10A (1/1)***	3	1	4
L3–10–1	6	—	L3–10B (1/1); 10C (1/1); 10D ***(1/1)***			
L3–11–*cis*7-1	1	—	—	2	1	1
L3–11–1	9	L3–2λ (2/2)	L3–11A (2/2)			
L3–12–1	1	—	—	1	—	—
L3–13–1	3	—	—	1	—	—
H1–10–1	2	—	—	1	—	—

Table 13 (*continued*)

This work	Number	Chothia *et al.*	Martin and Thornton	Number in this work	Number in Chothia *et al.*	Number in Martin and Thornton
H1–12–1	1	—	—	1	—	—
H1–13–*cis*9-1	2	—	—	12	1	4
H1–13–1	267	H1–1 (14/15)	H1–10A (39/42; ***3/42***)			
H1–13–2	7	H1–1 (1/15)	—			
H1–13–3	5	—	—			
H1–13–4	4	—	—			
H1–13–5	4	—	—			
H1–13–6	4	—	—			
H1–13–7	3	—	***H1–10D (1/1)***			
H1–13–8	3	—	***H1–10B (1/1)***			
H1–13–9	3	—	—			
H1–13–10	2	—	***H1–10C (1/1)***			
H1–13–11	2	—	—			
H1–14–1	11	H1–2 (1/1)	***H1–11A (1/1)***	1	1	1
H1–15–1	9	H1–3 (4/4)	H1–12A (2/2)	1	1	1
H1–16–1	1	—	—	1	—	—
H2–8–1	2	—	—	1	—	—
H2–9–1	77	H2–1 (3/3)	H2–9A (8/9)	3	1	1
H2–9–2	2	—	—			
H2–9–3	2	—	H2–9A (1/9)			
H2–10–1	155	H2–2A (4/4)	H2–10A (17/21); ***H2–10E (1/1)***; ***H2–10F (1/1)***	9	2 (5)	5
H2–10–2	42	H2–3A (4/4), 3C (1/1)	H2–10B (11/11)			
H2–10–3	11	H2–2B (1/1)	H2–10A (4/21)			
H2–10–4	7	—	—			
H2–10–5	3	—	—			
H2–10–6	3	—	—			
H2–10–7	2	H2–3B (1/1)	H2–10C (2/2)			
H2–10–8	2	—	—			
H2–10–9	2	—	—			
H2–12–1	26	H2–4 (1/1)	H2–12B (2/2); ***H2–12A (2/2)***	1	1	2
H2–15–1	1	—	—	1	—	—

Our clusters are listed in column 1, and the number of structures in each is given in column 2. The Chothia canonical classes closest to our cluster are given. The number of each canonical class closest to that cluster is given in parentheses along with the total number of members of that canonical class. Thus, "(3/4)" means that three of four members of a Chothia canonical class are closest to our cluster listed on that line. Uncertain assignments are given in italic boldface. The same procedure was used for the Martin–Thornton data in column 4. In the last three columns, the number of clusters in this work, the number of canonical classes for Chothia *et al.* (subclasses counted separately in parentheses), and the number of clusters in the Martin–Thornton analysis are given for each CDR–length combination.

classes were based on more than five structures in the PDB, and 15 of 32 (nearly half) were based on only one structure. For most of the canonical classes, we can make a clear one-to-one assignment to our clusters via the PDB chains given by Chothia *et al.* For instance, their L1–2A, L1–2B, and L1–4λ clusters are our L1–11–1, L1–11–2, and L1–11–3 clusters. As noted above, L1–11–3 is easily distinguishable by sequence from L1–11–1 and L1–11–2, while L1–11–1 and L1–11–2 differ from each other because of the residue at position 71 of VL.

In three cases, the PDB chains given by Chothia *et al.* for a canonical class fall into more than one of our clusters. This happens for their largest clusters: L2–1, L3–1, and H1–1. In all three cases, most of the structures given by Chothia *et al.* fall into one of our clusters, while a small number fall into another cluster. Since our loops were longer in these cases, the structural differences may occur outside of the region analyzed by Chothia *et al.* In four cases, the structures in more than one Chothia canonical class for a given CDR fall into one of our clusters. This occurs for the

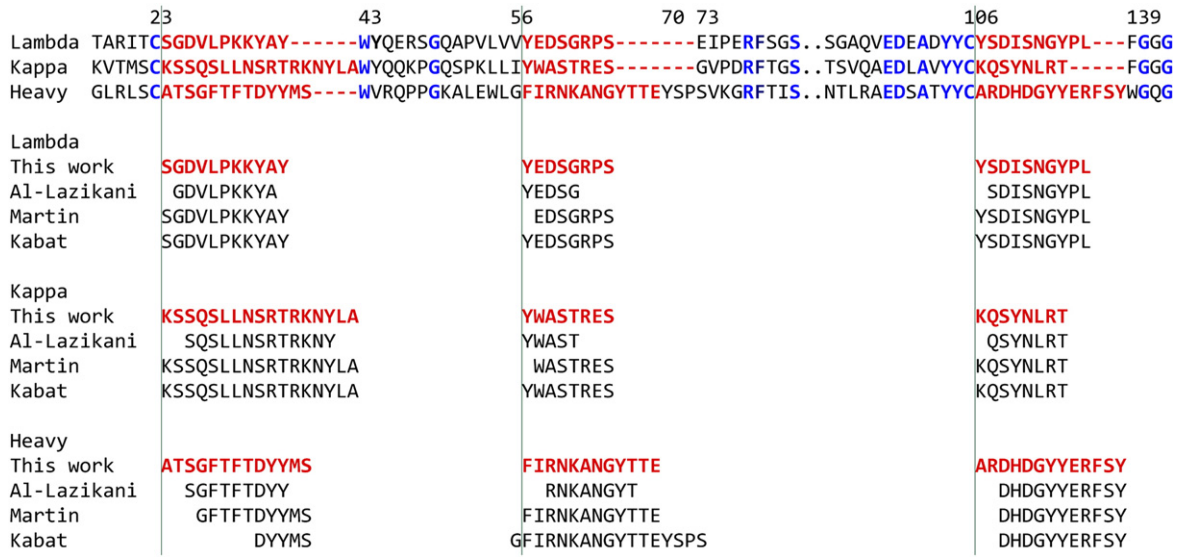

Fig. 11. Comparison of our CDR definitions with those of Al-Lazikani *et al.*[2] and Martin and Thornton,[3] with the numbering scheme proposed by Honegger and Plückthun.[10]

subclasses L1–3λA and B, L3–1λA and B, H2–3A, and H2–3C, which we put into single clusters.

There are also a few cases where the Chothia representatives do not appear in our data set and are relatively far away from our median structures. In these cases, the assignments to our clusters are uncertain. For instance, their L1–6 is a low-resolution (3 Å) structure that is 52° away from our L1–12–3 cluster. Their L1–2λ cluster is far away (43°) from its closest neighbor in our data, the L1–14–2 cluster, although its sequence (TGSSSNIGAGHNVK; PDB entry 7FABL) clearly fits our L1–14–2 pattern. Their L3–1λB structure from PDB entry 7FAB is also not very close (46°) to our L3–9–1 cluster.

Interestingly, only 4 of 20 Chothia CDR–length combinations comprise more than one canonical class: L1–11 (L1–2A,B and L1–4λ); L1–14 (L1–2λ and L1–2λA,B); L3–9 (L3–1, L3–3, and L3–1λA,B,C); and H2–10 (H2–2A,B and H2–3A,B,C).

The Martin–Thornton clusters are listed in Table 12. Their study listed 49 clusters for L1, L2, L3, H1, and H2. Only eight of these clusters (15%) are observed in five or more PDB entries, and 28 of them (57%) are observed in only one PDB entry. Twenty-two of the latter are far away from any of the median structures of our clusters, and these are highlighted in Table 12. Whether they are of low resolution or high conformational energies is noted, thus lending some doubt on whether they should be listed as separate clusters. These include L1–14C,D,E,F, L3–9E,F, H1–10C,D, and H2–10D,E,F. In some cases, the Martin–Thornton clusters are divided into more than one cluster in our analysis in part because our loops are sometimes longer (L2 and H2) or because of the RMSD step used by Martin and Thornton. For instance, their L1–11A is split about evenly between our L1–11–1 and L1–11–2 clusters. Martin and Thornton merged the two structures into the same cluster due to the small RMSD difference between the main-chain atoms of the two structures. Our algorithm keeps the two clusters separate due to the large difference in ϕ and ψ angles at loop positions 7 and 8. Chothia *et al.* listed them as A and B conformations of the same canonical class. Much of the Martin–Thornton L2–7A cluster corresponds to our L2–8–1, although several structures are members of or are closer to our L2–8–2, L2–8–4, and L2–8–5 clusters. Similarly, their L3–9A cluster corresponds to our L3–9–*cis*7-1, but one of their cluster members is an all-trans structure corresponding to our L3–9–2. We also split their H2–9A and H2–10A clusters.

We examined the CDR–length combinations in the Martin–Thornton analysis and found that, effectively, only six of them have more than one conformational cluster that can be validated with our data: L1–11, L1–14, L3–8, L3–9, L3–11, and H2–10 (in our definitions). Several other CDR-lengths have multiple clusters in the Martin–Thornton analysis but rely on very low resolution structures or structures with high conformational energy. For instance, their L3–10 loops consist of four clusters, but all of these are of low resolution or high conformational energy.

Finally, we examine the results the other way around by listing our clusters in Table 13 along with the number of PDB chain loops that overlap with the Chothia and Martin–Thornton data. We have a total of 72 clusters, each of which has at least two members, since we removed singleton outliers, except when there was only one structure for a given CDR and a given length (e.g., H2–15–1) or cis–

trans configuration. Thirty-one of our clusters have five or more members.

Forty-one of our clusters do not have a corresponding canonical class in the Chothia analysis. Thus, we have more than twice as many clusters as those present in the Chothia analysis. Many of these are for CDR-lengths not present in the PDB available to Chothia *et al.* These include L2–12, L3–12, L3–13, H1–10, H1–12, H1–16, H2–8, and H2–15. In a small number of cases, our clusters comprise more than one Chothia canonical class, usually when there are small differences in structure (e.g., L3–1λA and L3–1λC are both in our L3–9–1).

Thirty-two of our clusters do not have a corresponding cluster in the Martin–Thornton analysis, and an additional 10 have only distant relationships to their clusters (in italic boldface in Table 13), for a total of 42. Some loop lengths were not represented in their data set (mostly the same as those not present in the Chothia data), since the analyses were performed at around the same time (1996–1997). Some of our clusters comprise more than one Martin–Thornton cluster; however, in almost all cases, these consist of conformations that are quite distant from our median structures and were therefore excluded from our data set, often due to low resolution or high conformational energy.

Comparison of H3 torso analysis to Morea *et al*

Morea *et al.* presented rules for the prediction of the bulged and nonbulged conformations of the torso based on the residue types at positions 94 and 101 in the Chothia numbering (Honegger–Plückthun numbers 108 and 137, respectively).[4] These are positions 2 and 6 of the seven-residue segments shown in Table 8. Bulged conformations are those with –AB conformations for the last two residues of the loop in our definition, predominantly cluster H3–anchor-1. Nonbulged conformations have –BB conformations consisting predominantly of cluster H3–anchor-2. In the analysis of Morea *et al.*, bulged torsos have either lysine or arginine at position 94, while at position 101, aspartic acid is usually (but not always) present. For our data, we summarize the number of structures with Lys/Arg94 and Asp101 present or absent and the state of the loop as bulged or nonbulged in a contingency table shown in Table 14.

In accordance with Morea *et al.*, if position 94 is Lys/Arg and position 101 is Asp, the structure is bulged. One hundred fifty-five structures have this sequence and end in the bulged conformation –AB, while 11 have that sequence but are not bulged and so are counterexamples to the rules of Morea *et al.* In accordance with Morea *et al.*, if Lys/Arg is present at residue 94 but Asp is absent at position 101, the structure should still be bulged. This is true for 36 of the structures with that sequence, but is not true for the remaining five structures. If Lys/Arg is not present at residue 94 but Asp is present at 101, the structure is supposed to be nonbulged. However, we find 39 bulged examples and only 16 nonbulged structures. Finally, in their study, no structures lacking both the Lys/Arg at position 94 and the Asp at position 101 were observed. In our data set, there are 44 examples, of which 27 are bulged and 17 are not. Six structures do not seem to fit either the bulged conformation or the nonbulged conformation and thus are not considered. Thus, regardless of Lys/Arg or other residues at position 94, or Asp or other residues at position 101, the majority of the H3 torso structures are bulged. However, with Lys/Arg at position 94, 92% of the structures are bulged. Without Lys/Arg, 67% of the structures are bulged.

Table 14. Chothia rules for bulged or nonbulged H3 torso

Lys/Arg94	Yes	Yes	No	No	
Asp101	Yes	No	Yes	No	Row total
Bulged	155	36	39	27	257
Nonbulged	11	5	16	17	49
Column total	166	41	55	44	306

The table does not include six loops with neither bulged nor nonbulged conformation.

Discussion

In this work, we have revisited the problem of clustering the structures of the six CDR loops of antibodies. A thorough analysis such as this has not been accomplished since the work of Chothia *et al.* and Martin and Thornton in 1996–1997. The number of antibody structures is at least 5-fold larger now than it was then (and 15-fold larger than the set used by Chothia). Because of this, we have been able to remove questionable structures (those of low resolution or high-energy backbone conformations) and outliers with respect to all other structures of the same CDR and length. Nearly all of our clusters are represented by more than one structure, unless they are the only representative of a given CDR, length, and cis–trans configuration.

Two interesting questions arise from the present analysis: (1) To what extent are the CDR structures predictable from sequence? (2) Have earlier analyses—in particular the Chothia "canonical classes," which are widely used as the standard set of conformations for antibody modeling and analysis—held up?[22,28,29]

With the large data set, we have made a start on developing predictive methods for the prediction of CDR loop structures. For many CDR-lengths, there is one cluster that represents all or at least a large majority of the available structures. As a matter of structure prediction, it may be safe in most cases to

use the predominant cluster, unless specific residues that argue against that cluster and/or for another one are present. We have tried to annotate the obvious differences in Tables 3–7. In some cases, there are obvious structure-determining residues (e.g., Pro in L3–9) or large sequence differences (mouse/human or κ/λ sequences) that make identification of the appropriate cluster relatively easy. Additional work will be needed to turn the statistical analysis of the CDR conformations into a structure prediction method. However, we have attempted to classify the available CDR–length combinations into different types, depending on whether they exhibited only one effective conformation ("type I"), two or more conformations that are largely predictable based on the sequence of the CDRs and/or certain framework positions ("type II"), or one or more conformations that are based on insufficient data ("type III"). We have a total of 1202 sequences of CDRs covering L1, L2, L3, H1, and H2. Among these, 600 (50%) fall into CDR-lengths classified as type I, and 522 (43%) fall into CDR-lengths classified as type II. Only 80 (7%) fall into type III CDR-lengths. Not all of our type I and type II CDR-lengths are 100% predictable, since we required only a minimum of 85% predictability for these definitions. At the same time, the majority of the type III CDR-lengths seem to be predictable based on gene source or even sequence; however, because of small numbers, we cannot be sure that this will hold up given additional structures. Thus, we estimate that at least 85% of the non-H3 CDR structures can be easily predicted based on gene source (mouse or human; κ or λ, etc.) and sequence of the CDR or the identity of certain framework positions.

There are some remaining challenges for some loop lengths that are highly variable in structure and sequence, with no obvious patterning at present. Some of our clusters have high variance, possibly indicating the need to divide the set into a larger number of clusters (with lower self-similarity in the affinity propagation algorithm). With or without that step, methods for choosing the best loops from the clusters for a target antibody sequence will need to be developed.

Our second question is "Has the Chothia analysis of the 1980s and 1990s held up over time?" Chothia *et al.* provided canonical classes for 20 CDR–length combinations for the loops, excluding H3. In our analysis, we have a total of 28 CDR-lengths, although many of those that Chothia *et al.* did not have are rare in the PDB, even to date. Of the 20 CDR-lengths in the Chothia analysis, only four have more than one canonical class. The canonical classes of three of these four CDR-lengths are relatively easy to identify, since they are either mouse/human (L1–14) or κ/λ (L1–11), or dependent on the presence and the position of proline residues (L3–9). H2–10 has some important structure-determining residues, including the framework VH residue 71. Thus, the idea that a complicated set of structure-determining residues is needed to utilize the Chothia canonical classes for structure prediction[28] is largely not true, since almost all of their classes (except for the cases of L3–9 and H2–10) can be differentiated based purely on CDR, loop length, and gene source, and L3–9 is easily predictable based on the presence of proline at certain positions. For the cases of L1–14 and L1–11, we have the same clusters as the Chothia canonical classes. For L3–9 and H2–10, we have more clusters than the canonical classes.

While Chothia had only 4 CDR-lengths with more than one canonical class and 16 CDR-lengths with only one class, we have 15 CDR-lengths with more than one cluster and 13 CDR-lengths with only one cluster. Chothia *et al.* had a total of 25 canonical classes, and we have 72 clusters. Of the 16 Chothia CDR-lengths with one class, we now have 10 of these split into more than one cluster. In this sense, the Chothia analysis has not held up with the large increase in the number of structures.

Antibodies present a unique opportunity for the development of methods for protein structure prediction. Loop modeling, in particular, is a challenging problem, and the structures of over 300 loops at each of the six different CDRs provide a unique data set for defining sequence–structure relationships within the context of highly similar core structures. While it is common practice to borrow information from multiple templates in loop modeling,[30,31] it is possible that a similar clustering approach may be effective in other protein families for which there are many structures.

Materials and Methods

Hidden Markov models of the V domains of heavy and light chains

PSI-BLAST[32] was used to search a database of all sequences in the PDB (the nonredundant sequence file *pdbaanr* available on our PISCES Web site),[33,34] using the variable domain regions of the antibody structure in PDB entry 1Q9R.[14] Only sequences above a 35% identity and *E*-values better than 1.0×10^{-20} were kept, such that only antibody domains remained (e.g., excluding T-cell receptors and other Ig sequences). The resulting heavy-chain and light-chain sequences were culled at 90% identity using the PISCES server. Multiple-sequence alignments of the heavy-chain sequences and light-chain sequences were determined separately with CLUSTAL W[35] and then manually culled and edited. These alignments were then used to create heavy-chain-specific and light-chain-specific hidden Markov models, using the program HMMER.[36] Profile HMM is a statistical model of a multiple-sequence alignment of a protein family,[37] including position-specific insertion probabilities. This makes them well suited for determining the positions of the

CDRs, which occur at well-defined positions within the variable domain sequences and vary in length.

These HMMs were used to search for *pdbaa* (the set of all protein sequences in the PDB, including redundancy) available from our PISCES server†. The cutoff values for HMMER scores and *E*-values were chosen such that when searching for *pdbaa* protein sequences, only antibody heavy-chain and light-chain sequences scored better than the cutoffs. Sequences found by both HMMs were assigned to the one with the higher score and the smaller *E*-value. Both κ and λ light chains score better than the cutoffs for the light-chain HMM. These profile HMMs—one for the heavy chain and one for the κ light chain—were further utilized to identify specific conserved framework positions before and after each CDR.

Defining CDRs

Consistent definitions for the CDRs are required for this study. Kabat *et al.* derived CDR definitions and a residue numbering scheme based purely on antibody sequence information.[38] Chothia and Lesk[1] and Chothia *et al.*[24] defined CDRs from the earliest structures of antibodies and used this scheme (with some variations) in their studies of canonical CDR conformations. Martin *et al.*[39] presented a modified version of Chothia's CDR definitions‡, which is used in their SACS database.[27] Honegger and Plückthun performed a multiple-structure alignment of 16 variable chains and analyzed the variation in C^{α} positions in order to define a consistent numbering scheme for VH and VL chains, as well as for T-cell receptor α, β, γ, and δ chains.[10] They did not strictly define the boundaries for the CDRs.

In defining the boundary positions for the CDRs, we had several criteria in mind. First, we wanted positions with little structural variability across antibodies. Second, where possible, we wanted positions across from each other in the β-sheet framework (i.e., extending equal lengths into the framework). Third, we wanted definitions that were more or less symmetric between the VH domain and the VL domain. The sequence logos for the positions we have chosen as boundaries between the CDRs and the surrounding framework regions are shown in Fig. 1.

The Kabat, Chothia, and Martin CDRs for L3 begin one residue after the Cys residue that forms one end of the intrachain disulfide bond.[39] This position is also the last position before the CDR3s with less than 0.5 Å variability in the data of Honegger and Plückthun. We therefore used the residue after the second Cys in the disulfide bond to begin both CDR L3 and H3. To identify the C-terminal end of the CDR3s, we examined the structure and chose the residue across from the Cys+1 to be the last residue of the CDR. The framework residue following this position is also the first of that framework region to have less than 0.5 Å variability in the Honegger–Plückthun data. These positions are easy to identify visually in both light-chain and heavy-chain sequences. The motif that follows CDR H3 in the VH chain is almost always WG[X]G, where X is usually Q, E, K, H, or P. The motif that follows CDR L3 in the VL chain is almost always FG[X]G, where X is G, A, S, Q, or T.

With one end of the disulfide bond already defined as the boundary residue just before the CDR3s, we decided to place the H1 and L1 boundaries between the first Cys in the disulfide and the residue immediately following. This is also the definition of L1 by Kabat *et al.*, Chothia, and Martin and Thornton, although their H1 definitions start five residues (Chothia and Martin–Thornton) or eight residues (Kabat *et al.*) later. Notably, here, the Cys residue is also the last residue with a <0.5 Å variability identified by Honegger and Plückthun. As with the H3 and L3, we defined the end of the H1 and L1 CDRs as the residues immediately adjacent in the neighboring strand to Cys+1. In both VH and VL, this is a position before a highly conserved tryptophan. Both the last residue of the CDR and the first residue after the CDR have low variabilities in the CA position (0.5 Å), while residues in the CDR have high variability. The motif that follows H1 is W[VIF][RK][QK], while the motif that follows L1 is usually W[YFLV][QL][QEH].

Chothia and Martin–Thornton defined the first residue of H2 as following a hydrophobic stretch of amino acids in the preceding β-sheet strand, usually with the sequence LLI or WIG. This first residue is also located immediately across in the neighboring β-sheet strand from the last residue of H1 that we defined above, thus making the definition of the first residue of H2 (and L2) symmetric in the β-sheets. With this definition, the last residue of the framework before H2 or L2 is also the last in this region with a low variability in the C^{α} position.

The C-terminal ends of H2 and L2 are somewhat more difficult to define. H2 and L2 connect adjacent strands in a β-sheet, so we could define the end of the loop directly opposite the beginning residue. Indeed, this works well for H2, where this segment is a short β-sheet strand, in contrast to the VL chain in this region, which is usually all-coil. The existence and position of this short β-strand are well conserved, and the residue following it is a very-well-conserved Tyr residue (sometimes Phe) that packs against the other β-sheet. We therefore decided to let H2 end with the two β-sheet residues, just before this conserved Tyr. The sequence motif following H2 is [YF][NAVSG][PEQD][KDS]. The lack of a strong consensus of sequence here correlates with a somewhat higher structural variability in the Honegger–Plückthun data, but the variability does not reach a minimum until well into the next strand (and sheet), heading back toward the antigen binding site.

For L2, however, this region is not a β-strand, and the region after it exhibits quite a bit of structural variability. Martin, Kabat *et al.*, and Chothia all make the L2 definition three residues longer than the H2 definition described above. This sequence is also quite variable, and it therefore seems justified to make a nonsymmetric definition for the L2/framework boundary with respect to H2. The L2 region is almost always eight residues long in our definition, while the Martin/Kabat/Chothia definition begins one residue later and is therefore seven residues long. The sequence motif following L2 is usually G[VI]P[SA]. The CDR loops according to our definitions are shown in Fig. 2.

† http://dunbrack.fccc.edu/PISCES.php

‡ These definitions are summarized in http://www.bioinf.org.uk/abs/

The match states of residues just before and after the CDRs were identified within the HMMs. Thus, matching a query sequence against the HMM could readily identify the CDR boundaries by determining which residues in the sequence aligned with the match states identified with these boundaries. We compared our definitions with the results from the SACS database of 2009 (since updated). All discrepancies (considering the differences in CDR definitions) were examined visually in the structures, and the HMMs had correctly identified the CDR positions as we have defined them. The SACS database sometimes identifies Cys and Trp residues within the CDRs as the conserved Cys and Trp residues of the framework, thus defining the CDRs inconsistently. SACS was also missing some CDRs from single-chain antibodies and PDB entries containing different antibodies or idiotypes and their anti-idiotypes.

Filtering the data for poorly defined loop conformations

At the level of PDB entries, non-X-ray structures and structures with a resolution of worse than 2.8 Å were removed from the database. Several criteria were applied to remove specific CDR loops from the data set. First, any loop with missing backbone atoms was eliminated. Second, any loop with a backbone atom with a *B*-factor of 80 or higher was also excluded from the database to remove highly mobile loops. We also removed those with missing *B*-factors ($B=0$). Third, any loop with *cis*-peptide bonds for a residue other than proline was also excluded. We found only 12 such structures of antibodies. While non-Pro cis residues do exist, they are quite rare and—at least in some cases—are very likely to be the result of poor structure determination. Fourth, we used a set of Ramachandran probability densities that we recently developed to remove any loops with highly improbable backbone conformations.[15] These Ramachandran distributions§ are sequence dependent, with different distributions for a given residue type and its neighbor to the right or to the left. These can be simply combined into probability distributions for residue sequence triples. In a survey of loops in high-resolution structures, 98% of loops had energies of less than 9.5 per residue (derived from $-\log(p)$ in arbitrary units). Thus, we removed a small number of loops with very high conformational energies.

The resulting set of structures is highly redundant, with many examples of different crystal structures of the same antibody in the PDB. To take this into account, we compared the sequences of the six CDRs; for any set of structures with the same set of CDR sequences, we used the structure with the highest resolution. The effects of the steps to filter out poor structures and redundancy are shown in Table 1.

Clustering loop conformations for each CDR-length

With the data set culled from poor-quality structures, we clustered the loops by structure as follows. First, for each loop length, we examined dihedral-angle differences at each position of the loop. The loops were clustered by cis–trans configurations. For instance, L3–9 (CDR L3, length 9) loops often have a *cis*-proline at one or two positions. In this case, L3–9–allT (all-trans) loops were grouped together, as were L3–9–*cis*7 (*cis*-Pro at position 7) and L3–9–*cis*6,7 (*cis*-Pro at positions 6 and 7). We removed a small number of outliers defined as a loop with at least one ϕ or ψ more than 90° away from every other loop in that CDR–length group and cis–trans configuration.

Once loops had been sorted by CDR, length, and cis–trans configuration, in order to cluster the loops by structure, we required a distance function between any two loop structures. We chose to use a metric used in directional statistics[40] to calculate the distance between two angles. For two dihedral angles θ_1 and θ_2 of the same type and at the same residue position of two different structures, the distance between them is defined as:

$$D(\theta_1, \theta_2) = 2(1 - \cos(\theta_1 - \theta_2)) \quad (1)$$

This is the squared distance of the chord on a unit circle connecting the vectors ($\cos\theta_1$, $\sin\theta_1$) and ($\cos\theta_2$, $\sin\theta_2$). The distance between two loops i and j of the same CDR type and length N is defined as the sum of the distances between their backbone dihedral angles ϕ and ψ over the residues r of the loop:

$$D(i,j) = \sum_{r=1,N} D\left(\phi_r^i, \phi_r^j\right) + D\left(\psi_r^i, \psi_r^j\right) \quad (2)$$

For any set of loops, we used the affinity clustering algorithm proposed by Frey and Dueck to identify potential structural clusters.[9] The affinity clustering algorithm requires a similarity measure rather than a distance measure; in this method, this is set to the negative of the distance measure:

$$s(i,j) = -D(i,j) \quad (3)$$

for $i \neq j$. Self-similarities $s(i,j)$ are set to a constant whose value is the average value of the non-self-similarities:

$$s_{\text{self}} = \frac{2}{n(n-1)} \sum_i \sum_{j>i} s(i,j) \quad (4)$$

The self-similarity values can be scaled to produce more or fewer clusters, depending on the application.

Affinity propagation clustering is a message-passing algorithm that makes choices as to which structures should be associated with one another. Each data point has an *exemplar*, a data point that represents a particular cluster of points. Each iteration of the algorithm consists of a set of messages that are passed between all structures that determine the values of two quantities per structure pair: the *responsibility* $r(i,k)$, which reflects the accumulated evidence that k should be the exemplar for i, and the *availability* $a(i,k)$, which represents whether k would be a good exemplar for i (i.e., is similar to i and very similar to a large number of other data points).

§ http://dunbrack.fccc.edu/ndrd

At the start of the algorithm, all availabilities $a(i,k)$ are set to zero. For each iteration, the first step is that the responsibilities are assigned the following values:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k} \{a(i,k') + s(i,k')\} \quad (5)$$

Next, the availabilities are updated using different equations for non-self-availabilities ($i \neq k$) and self-availabilities. For the non-self-availabilities:

$$a(i,k) \leftarrow \min\left\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\} \quad (6)$$

while the self-availabilities are assigned these values:

$$a(k,k) \leftarrow \sum_{i' \neq k} \max\{0, r(i',k)\} \quad (7)$$

These update quantities—the values of the expressions in Eqs. (5), (6), and (7)—are then averaged with the value for the given quantity (either responsibility or availability) for the previous iteration of the algorithm. When the assignment of clusters has remained identical for four iterations of the algorithm and when the algorithm has run for at least 10 iterations, the algorithm terminates.

Once the algorithm converges, the exemplar of each structure i is the structure k that maximizes this quantity:

$$\text{Exemplar}(i) = \arg\max_{k} \{a(i,k) + r(i,k)\} \quad (8)$$

and the resulting clusters each consist of all structures that are connected, directly or otherwise, by an exemplar–member relationship.

The wide structural diversity of the H3 loop segments makes their clustering difficult. In particular, the tips of their loop regions vary greatly in length, sequence, and conformation. However, if the H3 regions longer than 10 residues in length are split into 'torso' regions that connect to the framework at the N-terminal and C-terminal ends of the H3 loop and 'head' regions corresponding to the tip of the loop, the 'torso' regions are less conformationally variable and may conform to a discrete set of conformations.[3,4,12]

Once the clustering is complete, clusters that are structurally similar to one another are merged. For each pair of clusters that are of the same loop type, length, and cis–trans configuration, the conformational distance between the median loops is calculated (see Eq. (3)) and then divided by the length of the loop. We inverted Eq. (2) to calculate the average absolute angular difference per residue. If the distance for each of the dihedral pairs is less than the distance of two angles differing by 65°, then the two clusters are merged. This value seemed to work well empirically at removing conformations that obviously did not cluster with the rest.

Conformational class definitions

In order to facilitate a comparison between different clusters, we assigned residue conformations based on the division of the Ramachandran map into regions. The different conformational classes include A (α-helix region), B (β-sheet region), P (for polyproline II), L (left-handed helix region), D (δ), and G (γ). These are shown with definitions in Fig. 4. These classes are then used to annotate the cluster based on the structure of the loop, with the lowest median distance to all other loops in its cluster. This step gives an easy means of comparing different clusters.

Acknowledgements

This work was supported by National Institutes of Health grants P20 GM76222 and R01 GM84453 (R.L.D., principal investigator) and National Institutes of Health training grant T32 CA009035.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2010.10.030

References

1. Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917.
2. Al-Lazikani, B., Lesk, A. M. & Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* **273**, 927–948.
3. Martin, A. C. R. & Thornton, J. M. (1996). Structural families in loops of homologous proteins: automatic classification, modeling, and application to antibodies. *J. Mol. Biol.* **263**, 800–815.
4. Morea, V., Tramontano, A., Rustici, M., Chothia, C. & Lesk, A. M. (1998). Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J. Mol. Biol.* **275**, 269–294.
5. Oliva, B., Bates, P. A., Querol, E., Aviles, F. X. & Sternberg, M. J. (1998). Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *J. Mol. Biol.* **279**, 1193–1210.
6. Wilmot, C. M. & Thornton, J. M. (1990). Beta-turns and their distortions: a proposed new nomenclature. *Protein Eng.* **3**, 479–493.
7. Shirai, H., Kidera, A. & Nakamura, N. (1999). H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Lett.* **455**, 188–197.
8. Kuroda, D., Shirai, H., Kobori, M. & Nakamura, H. (2008). Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins Struct. Funct. Bioinf.* **73**, *608–620.*
9. Frey, B. J. & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, **315**, 972–976.
10. Honegger, A. & Plückthun, A. (2001). Yet another numbering scheme for immunoglobulin variable

domains: an automatic modeling and analysis tool. *J. Mol. Biol.* **309**, 657–670.
11. Martin, A. C. & Thornton, J. M. (1996). Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J. Mol. Biol.* **263**, 800–815.
12. Ruzheinikov, S. N., Muranova, T. A., Sedelnikova, S. E., Partridge, L. J., Blackburn, G. M., Murray, I. A. *et al.* (2003). High-resolution crystal structure of the Fab-fragments of a family of mouse catalytic antibodies with esterase activity. *J. Mol. Biol.* **332**, 423–435.
13. James, L. C., Roversi, P. & Tawfik, D. S. (2003). Antibody multispecificity mediated by conformational diversity. *Science*, **299**, 1362–1367.
14. Nguyen, H. P., Seto, N. O., MacKenzie, C. R., Brade, L., Kosma, P., Brade, H. & Evans, S. V. (2003). Germline antibody recognition of distinct carbohydrate epitopes. *Nat. Struct. Biol.* **10**, 1019–1025.
15. Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M. I. & Dunbrack, R. L., Jr (2010). Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Comput. Biol.* **6**, e1000763.
16. Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F. *et al.* (2009). IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* **37**, D1006–D1012.
17. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190.
18. Schuermann, J. P., Henzl, M. T., Deutscher, S. L. & Tanner, J. J. (2004). Structure of an anti-DNA Fab complexed with a non-DNA ligand provides insights into cross-reactivity and molecular mimicry. *Proteins*, **57**, 269–278.
19. Yokota, A., Tsumoto, K., Shiroishi, M., Kondo, H. & Kumagai, I. (2003). The role of hydrogen bonding via interfacial water molecules in antigen–antibody complexation. The HyHEL-10–HEL interaction. *J. Biol. Chem.* **278**, 5410–5418.
20. Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inf.* **23**, 205–211.
21. Tramontano, A., Chothia, C. & Lesk, A. M. (1990). Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the VH domains of immunoglobulins. *J. Mol. Biol.* **215**, 175–182.
22. Whitelegg, N. R. & Rees, A. R. (2000). WAM: an improved algorithm for modelling antibodies on the WEB. *Protein Eng.* **13**, 819–824.
23. Shirai, H., Kidera, A. & Nakamura, H. (1996). Structural classification of CDR-H3 in antibodies. *FEBS Lett.* **399**, 1–8.
24. Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G. *et al.* (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877–883.
25. Tomlinson, I. M., Cox, J. P., Gherardi, E., Lesk, A. M. & Chothia, C. (1995). The structural repertoire of the human V kappa domain. *EMBO J.* **14**, 4628–4638.
26. Morea, V., Tramontano, A., Rustici, M., Chothia, C. & Lesk, A. M. (1997). Antibody structure, prediction and redesign. *Biophys. Chem.* **68**, 9–16.
27. Allcorn, L. C. & Martin, A. C. (2002). SACS—self-maintaining database of antibody crystal structure information. *Bioinformatics*, **18**, 175–181.
28. Morea, V., Lesk, A. M. & Tramontano, A. (2000). Antibody modeling: implications for engineering and design. *Methods*, **20**, 267–279.
29. Marcatili, P., Rosi, A. & Tramontano, A. (2008). PIGS: automatic prediction of antibody structures. *Bioinformatics*, **24**, 1953–1954.
30. Fernandez-Fuentes, N., Rai, B. K., Madrid-Aliste, C. J., Fajardo, J. E. & Fiser, A. (2007). Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics*, **23**, 2558–2565.
31. Chakravarty, S., Godbole, S., Zhang, B., Berger, S. & Sanchez, R. (2008). Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. *BMC Struct. Biol.* **8**, 31.
32. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of database programs. *Nucleic Acids Res.* **25**, 3389–3402.
33. Wang, G. & Dunbrack, R. L., Jr (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
34. Wang, G. & Dunbrack, R. L., Jr (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **33**, W94–W98.
35. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
36. Durbin, R. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK.
37. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
38. Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991). *Sequences of Proteins of Immunological Interest*, 5th edit. National Institutes of Health, Bethesda, MD.
39. Martin, A. C., Cheetham, J. C. & Rees, A. R. (1989). Modeling antibody hypervariable loops: a combined algorithm. *Proc. Natl Acad. Sci. USA*, **86**, 9268–9272.
40. Mardia, K. V. & Jupp, P. E. (2000). *Directional Statistics.* Wiley Series in Probability and Statistics. Wiley, London, UK.