doi:10.1016/S0022-2836(02)01222-6 *J. Mol. Biol.* (2003) **325**, 337–354

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Analysis of the Antigen Combining Site: Correlations Between Length and Sequence Composition of the Hypervariable Loops and the Nature of the Antigen

Abigail V. J. Collis, Adam P. Brouwer and Andrew C. R. Martin*

School of Animal and Microbial Sciences, University of Reading Whiteknights, P.O. Box 228 Reading RG6 6AJ, UK

**Corresponding author*

It has long been suggested that the overall shape of the antigen combining site (ACS) of antibodies is correlated with the nature of the antigen. For example, deep pockets are characteristic of antibodies that bind haptens, grooves indicate peptide binders, while antibodies that bind to proteins have relatively flat combining sites. In 1996, MacCallum, Martin and Thornton used a fractal shape descriptor and showed a strong correlation of the shape of the binding region with the general nature of the antigen.

However, the shape of the ACS is determined primarily by the lengths of the six complementarity-determining regions (CDRs). Here, we make a direct correlation between the lengths of the CDRs and the nature of the antigen. In addition, we show significant differences in the residue composition of the CDRs of antibodies that bind to different antigen classes. As well as helping us to understand the process of antigen recognition, autoimmune disease and cross-reactivity, these results are of direct application in the design of antibody phage libraries and modification of affinity.

Keywords: antibodies; antigen combining site; complementarity-determining regions; sequence composition

Introduction

It is well known that the immune system is able to generate a repertoire of antibodies capable of recognizing an almost infinite range of molecules, from small organic haptens, through sugars, lipids, peptides and nucleotides to intact proteins. Despite this amazing ability of the antibody to bind such a range of compounds, its variability is essentially confined to just six hypervariable loops or complementarity-determining regions (CDRs) that form the antigen combining site (ACS). The β-sheet framework that supports the six CDRs is remarkably invariant in sequence and structure, although certain residues in the framework make critical packing interactions directly or indirectly with the CDRs and can influence their conformation.[1,2] Sequence variability is achieved through a mixture of gene selection, splice variation and somatic hypermutation. The CDRs are normally defined by the early analysis of sequence hypervariability performed by Wu & Kabat.[3] However, for analysis and modelling purposes, the structural loop definition described by Chothia[1] is often more appropriate.

The Fv fragment of the antibody (the smallest fragment able to bind antigen in normal antibodies) consists of two chains (heavy and light) totalling some 230 amino acid residues, of which about 70 form the CDRs. Perhaps surprisingly, the six CDRs generally adopt only a limited set of canonical backbone conformations,[1] defined by their length and the presence of certain key structurally determining residues.

Many groups have analyzed the structure of antibodies and the recognition process.[4–8] There has been a long-held belief that the shape of the ACS varies with the general nature of the antigen. For example, deep pockets are characteristic of antibodies that bind haptens, grooves indicate peptide binders, while antibodies that bind to proteins have relatively flat combining sites. In 1996, we set out to confirm this objectively.[9] We analyzed the topography of the combining site using a fractal-based measure of convexity/concavity.[10] We showed that the topography of the residues which form the actual antibody–antigen interface could

Abbreviations used: ACS, antigen combining site; CDR, complementarity-determining region.

E-mail address of the corresponding author: a.c.r.martin@reading.ac.uk

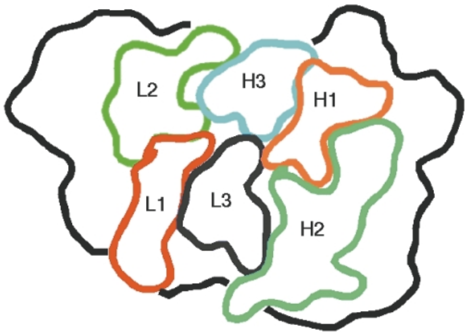

Figure 1. A representation of the antigen combining site.

Table 2. Numbers of antibodies with known antigens extracted from Kabat

Antigen type	Number of antibodies identified	Number with complete loop length information	Number with complete sequence information
Hapten	255	253	242
Lipid	68	67	67
Nucleotide	267	255	252
Peptide	46	46	45
Protein	849	841	843
Sugar	79	77	78
Virus	224	224	223
Total	1788	1763	1750

The number of antibodies identified (column 2) is the total number classified; the number with complete sequence information (column 4) is the number remaining after stripping unknown amino acids and rejecting antibodies where no sequence information was available for one or more of the contact CDRs; the number with complete loop length information (column 3) is the number remaining after stripping all those with zero-length Kabat CDRs, but with no treatment of unknown residues. In general, this number will be higher than the number with complete sequence information. However in some cases it is smaller because of the difference in CDR definitions. In all, 87 antibodies remained unclassified either because they could not be classified or because they were cross-reactive binding to multiple antigen classes.

indeed be correlated with the nature of the antigen: hapten binders had the most concave interface, followed by medium and large antigens. However, when the complete combining site was considered, the variation in the binding region (normally at the centre of the combining site) was masked by the overall convex nature of the CDRs at the apex of the antibody molecule, although general trends were still clear. Finally, we went on to analyze which residue positions in the CDRs are primarily involved in interactions with different classes of antigen and thus proposed a "contact definition' for the CDRs based on overall likelihood of residues being in contact with the antigen.

The gross shape of the ACS is defined primarily by the lengths of the six CDRs. This is true because length is the primary determining factor for the canonical conformations of the CDRs. Almagro's group analyzed which combinations of canonical classes are observed and found that only a limited set of the possible combinations occur.[11] They went on to correlate these combination classes with the type of antigen.[12] More simply, length is directly responsible for the topography of the ACS because the bulk of the CDRs will be responsible for opening up pockets or grooves in the antibody surface, or filling in the spaces to form a flat combining site or, indeed, leading to protruberences from the surface.

It is clear from Figure 1 that gross changes in the topography of the binding site can be correlated with the lengths of the CDRs. For example, long CDR-L1, CDR-L2, CDR-H1 and CDR-H2 with short CDR-L3 and CDR-H3 would result in a groove. A long CDR-L1 and CDR-H2 with a medium-to-long CDR-H2 and short CDR-L3 would result in a pocket. By filling in the middle of the binding site using CDR-L3 and CDR-H3, a flat binding site can be obtained, while a very long CDR-H3 can result in a protruberence from the surface.

In practice, some of the CDRs vary in length more than others, as shown in Table 1. Thus, in the majority of antibodies, it is only the lengths of CDRs L1, L3, H2 and H3 (and to a smaller extent

Table 1. Observed lengths of the CDRs in the Protein Databank (PDB) and the Kabat sequence database as of February 2002

	Observed range	
CDR	PDB	Kabat
L1	10–17	10–17 (7–17)
L2	7[a]	7 (2–11)
L3	7–13	7–13 (1–18)
H1	5–7 (3–7)[b]	5–7 (1–8)
H2	16–19	16–19 (10–23)
H3	3–19	4–20 (1–30)

In the case of the Kabat data, the common observed range is shown first (at least 100 observations of the maximum and minimum lengths) with the total range observed shown in parentheses.

[a] Antibody structure New (PDB code 7fab[35]) has an unusual deletion at the junction of CDR-L2 and framework region 3 such that only four residues are from the CDR-L2 loop.

[b] Only two antibodies (PDB codes 1jgl[36] and 1ghf[37]) have CDR-H1s shorter than five residues.

Table 3. Distribution of antibodies analyzed here between species

	Human	Mouse	Other
Hapten	6 (2.4)	247 (97.6)	0 (0)
Nucleotide	37 (14.5)	218 (85.5)	0 (0)
Protein	245 (29.1)	551 (65.5)	45 (5.4)
Virus	95 (42.4)	128 (57.1)	1 (0.4)
Lipid	17 (25.4)	49 (73.1)	1 (1.5)
Peptide	5 (10.9)	39 (84.8)	2 (4.3)
Sugar	12 (15.6)	59 (76.6)	6 (7.8)
Total	417 (23.7)	1291 (73.2)	55 (3.1)

Values within parentheses are percentages.

Table 4. Mean lengths of the CDRs for antibodies binding to each of the seven classes of antigen

	CDR																	
	L1			L2			L3			H1			H2			H3		
Antigen class	Human	Mouse	All	Human	Mouse	All	Human	Mouse	All	Human	Mouse	All	Human	Mouse	All	Human	Mouse	All
Virus	12.09	13.35	12.82	7.00	7.00	7.00	9.17	8.90	9.01	5.21	5.15	5.17	17.22	16.95	17.06	16.53	10.07	12.86
Protein	12.49	13.41	13.06	7.00	7.01	7.00	9.44	8.84	9.05	5.18	5.09	5.12	16.91	17.08	17.02	12.86	9.99	10.90
Nucleotide	12.00	13.76	13.51	7.00	7.00	7.00	9.16	8.72	8.78	5.24	5.05	5.08	16.76	17.13	17.07	12.59	10.44	10.76
Lipid	13.71	12.71	12.97	7.00	7.00	7.00	10.35	9.14	9.48	5.12	5.14	5.13	16.94	16.78	16.81	11.71	10.45	10.82
Peptide	11.20	14.69	14.15	7.00	7.00	7.00	9.20	8.92	8.96	5.00	5.23	5.20	17.20	16.90	16.93	12.00	8.74	9.30
Sugar	13.08	12.22	12.26	7.00	7.00	7.00	9.67	8.98	9.21	5.33	5.00	5.05	17.08	17.51	17.32	10.83	8.25	9.25
Hapten	13.00	12.35	12.37	6.67	7.00	6.99	10.33	8.88	8.91	5.00	5.09	5.09	16.67	16.80	16.80	10.17	8.50	8.54

The antigens are arranged in approximate order of decreasing size. Means are shown separately for human and mouse antibodies, and for all antibodies together.

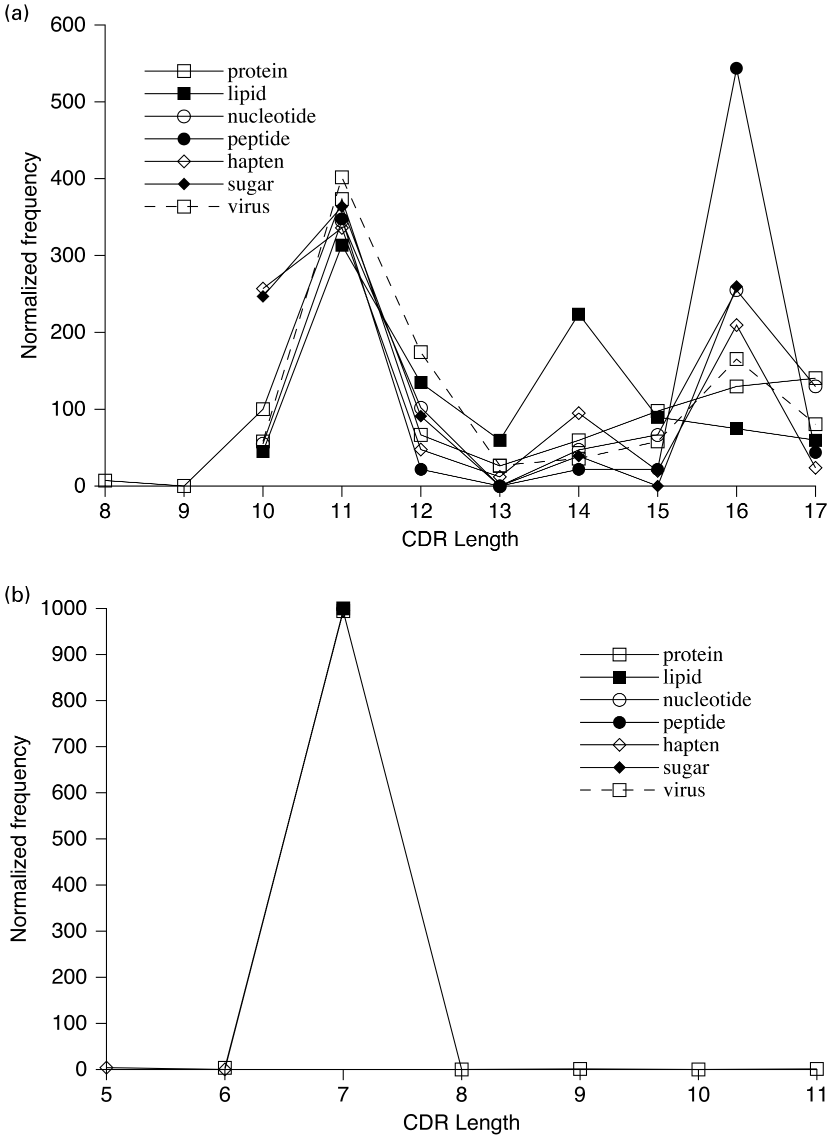

Figure 2 (*legend on page 342*)

CDR-H1) that affect the topography of the combining site.

Having previously shown a relationship between the topography of the combining site as assessed using the fractal measure of concavity/convexity, we set out to ask the question of whether the distribution of CDR lengths, and therefore the implied topography of the combining site, can be correlated directly with the general class of antigen.

Other work from the Almagro group analyzed the sequence composition of CDR1 and CDR2 regions from 2000 antibodies.[13] They showed that some positions fit an inverse power-law distribution, while others fit an exponential distribution, and proposed that the first class is critical to maintaining the conformation while the second class is involved exclusively in recognition. They did not make any general statistical comparisons with loops from other proteins or between loops binding to different classes of antigen. Since it is well known that the combining sites of antibodies binding to DNA are rich in arginine side-chains, the question can be asked as to whether, like the lengths of the CDRs, the amino acid distributions in the CDRs can be correlated with the nature of the antigen.

Results and Discussion

The Kabat loop definitions were used for analysis of CDR lengths. The contact definitions described by MacCallum *et al.* were used for analysis of sequence composition.[9] The contents of the data set extracted from the Kabat database are summarized in Table 2. Overall, approximately three-quarters of antibody sequences come from

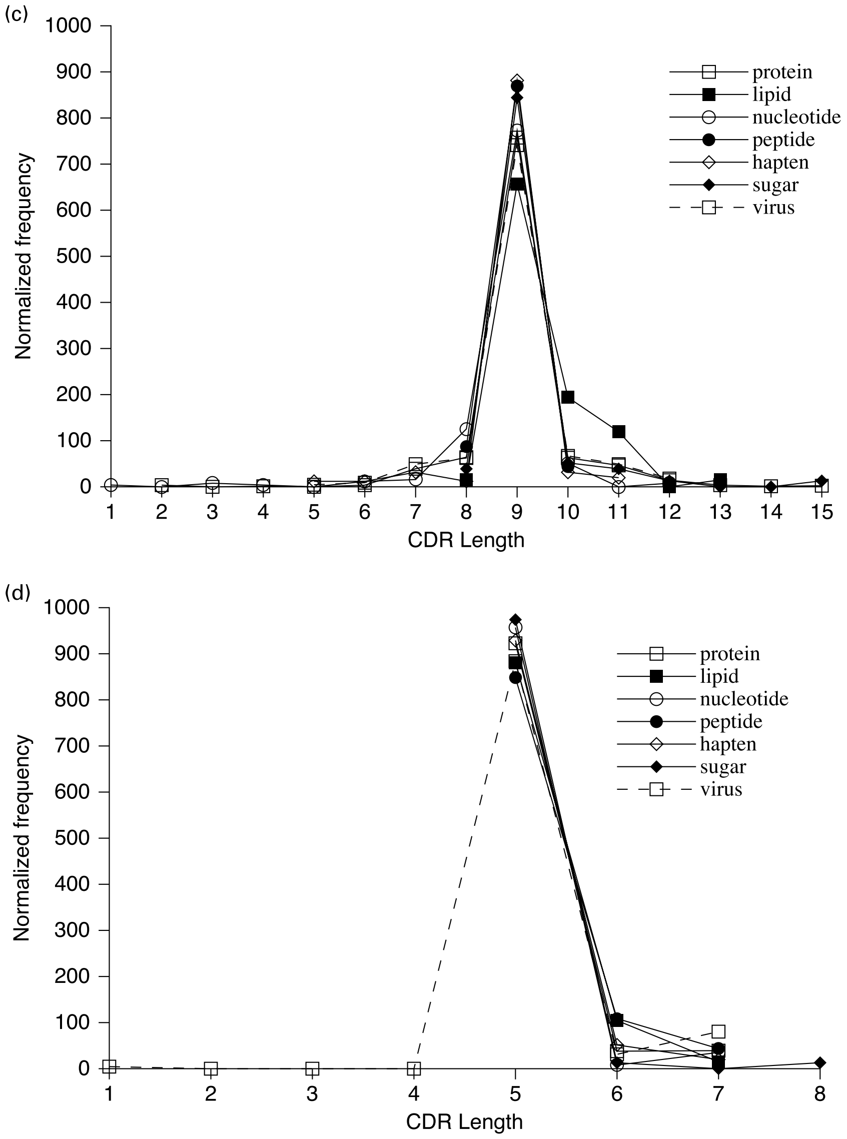

Figure 2 (*legend on page 342*)

mouse while a quarter are human and only around 3% come from other species. These data are summarized in Table 3.

CDR length distributions

We have shown previously that the gross topography of the antigen combining site is correlated with the general class of the antigen.[9] Given that the chief factor influencing the topography is the lengths of the six CDR loops, we posed the question of whether the lengths of the CDRs are correlated with the nature of the antigen.

Initially, we examined the mean lengths of the CDRs in the different antigen classes as shown in Table 4. It has been noted that the distribution of lengths of the CDRs, in particular CDR-H3, varies between species. Wu *et al.* showed that the lengths of CDR-H3 followed an approximately Poisson distribution in both human and mouse sequences, with average lengths of 11.6 and 8.7 residues, respectively.[14] They suggested that this may be due to the existence of relatively longer D-minigenes in human compared with mouse. The more recent dataset used here agrees with their data, with mouse antibodies showing CDR-H3 loops ranging from 1–19 residues and human antibodies ranging from 2–28 residues.

The data in Table 4 therefore show these species (which make up approximately 97% of the available data) separately as well as the overall means. Clear trends can be seen, especially in the length of CDR-H3: longer CDR-H3 loops tend to favor large antigens, while shorter loops favor small antigens; i.e. a long CDR-H3 implies a flat (or protruding) antigen combining site, while short loops favor pockets or grooves. CDR-H2 shows some variation: in mouse, it is longer for large antigens except viruses and shorter for small antigens except sugars; in human, it is long for the large

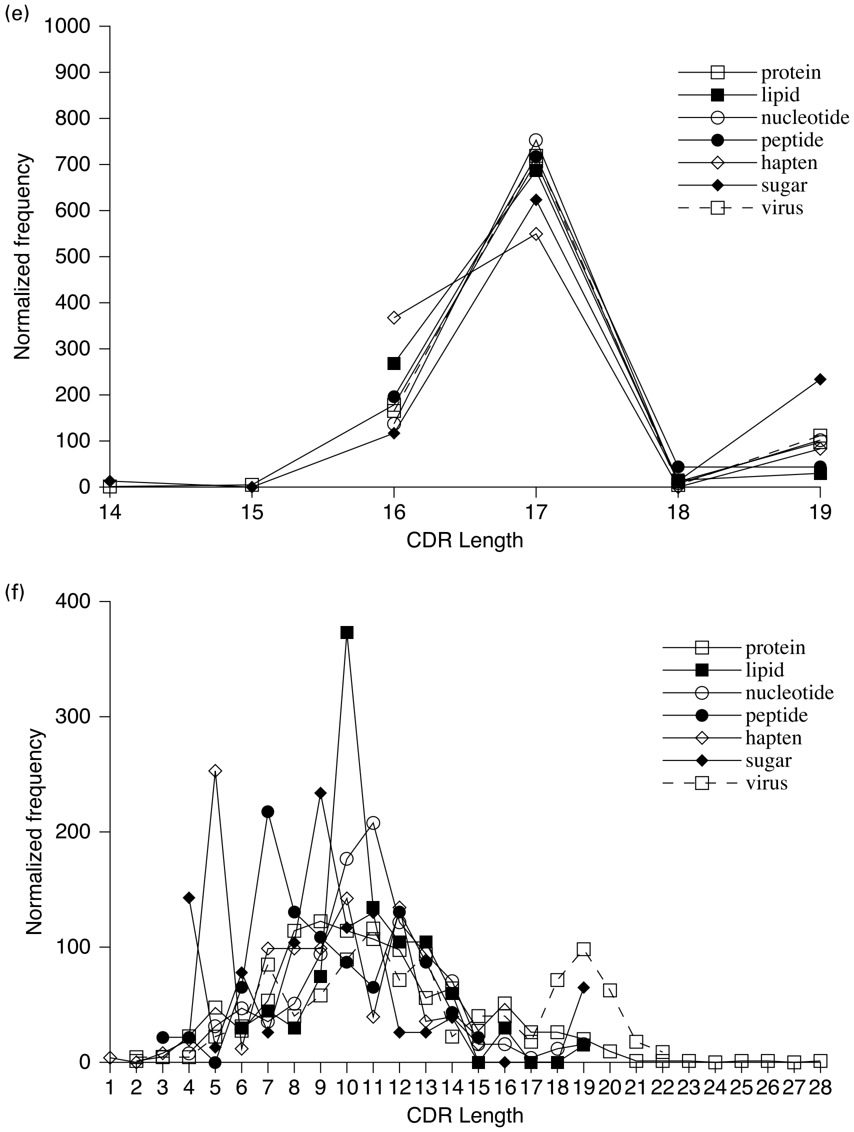

Figure 2. Length distributions for antibody CDRs binding the seven classes of antigen. (a) CDR-L1, (b) CDR-L2, (c) CDR-L3, (d) CDR-H1, (e) CDR-H2, and (f) CDR-H3.

antigens and shorter for smaller antigens except sugars and peptides. In mouse, CDR-L1 is longer for the more elongated antigens (peptides and nucleotides); in human, this trend is not clear, but this may be a result of the small amounts of data available. For the rest of this analysis, we consider the combined dataset, and consider the influence of species variations separately.

Figure 2 shows the length distributions for the six CDRs in all species. These graphs are normalized such that the total number of each antigen class is 1000. Some clear trends are immediately obvious, and there are some particularly striking features, as discussed below.

CDR-L1. Long loops (16 and 17 residues) are two to three times more abundant in peptide binders than in other classes.

CDR-L3. Longer loops are seen almost exclusively in viral binders.

CDR-H1. Longer loops are seen mostly in viral binders, though some are seen in protein and sugar binders.

CDR-H2. Hapten binders show more long and short loops, medium-length loops being less common. The shortest loops are seen only in sugar and protein binders.

CDR-H3. Hapten binders have no longer loops; peptide, hapten and sugar binders have a high occurrence of short loops, while virus binders have a high occurrence of long loops.

We went on to consider the statistical significance of these differences in distributions, using a χ^2 test to compare the distributions in all species. CDR-L2 and CDR-H1 do not vary in length enough to make further analysis worthwhile. The results are shown in Table 5.

The statistics of length distributions can be summarized as follows.

CDR-L1. Haptens and sugars are both different from all others; peptides are different from everything except nucleotides; proteins are different from everything except viruses; nucleotides are different from everything except peptides; viruses are different from everything except proteins.

Table 5. Significances calculated from a χ^2 test to compare the length distributions in each pair of antibody classes

	Significance (p)					
	Nucleotide	Sugar	Peptide	Protein	Virus	Lipid
A. *CDR-L1*						
Hapten	2.3×10^{-12} ++++	3.2×10^{-1}	6.1×10^{-7} ++++	3.3×10^{-16} ++++	2.6×10^{-11} ++++	1.1×10^{-7} ++++
Nucleotide	–	1.2×10^{-6} +++	4.0×10^{-4} +	2.9×10^{-6} +++	$1.1\text{x}10^{-2}$	1.4×10^{-5}
Sugar	–	–	1.3×10^{-4} +	5.0×10^{-6} +++	3.8×10^{-5} ++	2.3×10^{-6} +++
Peptide	–	–	–	5.1×10^{-13} ++++	7.5×10^{-6} +++	3.0×10^{-6} +++
Protein	–	–	–	–	1.8×10^{-6} +++	3.7×10^{-4} +
Virus	–	–	–	–	–	2.3×10^{-4} +
B. *CDR-L3*						
Hapten	1.8×10^{-5} ++	9.7×10^{-2}	8.7×10^{-1}	3.8×10^{-5} ++	3.3×10^{-1} +	1×10^{-10} ++++
Nucleotide	–	5.7×10^{-3}	3.7×10^{-1}	1.2×10^{-4} +	1.1×10^{-3}	3.3×10^{-10} ++++
Sugar	–	–	2.3×10^{-1}	1.1×10^{-1}	8.0×10^{-2}	2.1×10^{-3} +
Peptide	–	–	–	1.7×10^{-1}	1.4×10^{-1}	2.8×10^{-4} +
Protein	–	–	–	–	9.3×10^{-1}	9.8×10^{-6} +++
Virus	–	–	–	–	–	3.9×10^{-4} +
C. *CDR-H2*						
Hapten	1.7×10^{-8} ++++	5.2×10^{-6} +++	7.1×10^{-2}	8.0×10^{-9} ++++	4.6×10^{-6} +++	1.2×10^{-1}
Nucleotide	–	9.9×10^{-3}	5.6×10^{-1}	2.1×10^{-1}	6.5×10^{-1}	1.6×10^{-2}
Sugar	–	–	7.7×10^{-2}	1.6×10^{-3}	2.1×10^{-2}	1.3×10^{-3}
Peptide	–	–	–	8.9×10^{-1}	7.8×10^{-1}	4.8×10^{-1}
Protein	–	–	–	–	7.9×10^{-1}	8.9×10^{-2}
Virus	–	–	–	–	–	6.0×10^{-2}
D. *CDR-H3*						
Hapten	1.4×10^{-18} ++++	1.6×10^{-3}	1.3×10^{-3}	0 ++++	0 ++++	3.8×10^{-8} ++++
Nucleotide	–	5.7×10^{-6} +++	1.0×10^{-3}	2.2×10^{-7} ++++	2.7×10^{-12} ++++	5.0×10^{-2}
Sugar	–	–	4.6×10^{-2}	2.0×10^{-3}	1.3×10^{-8} ++++	1.1×10^{-4} +
Peptide	–	–	–	2.5×10^{-2}	6.5×10^{-5} ++	1.8×10^{-4} +
Protein	–	–	–	–	2.1×10^{-15} ++++	1.8×10^{-7} ++++
Virus	–	–	–	–	–	1.2×10^{-7} ++++

The p-values shown as zero are less than $< 5.4 \times 10^{-20}$, the limit of precision in our calculations. The degree of significance is indicated (+, $p < 0.1\%$; ++, $p < 0.01\%$; +++, $p < 0.001\%$; ++++, $p < 0.0001\%$).

Table 6. Summary of most commonly occurring lengths for the six CDRs in antibodies binding to different antigen classes

	CDR-L1	CDR-L3	CDR-H2	CDR-H3
Hapten	**10,11,14,16**	9	**16,17,19**	**5,7,8,9,10,12**
Lipid	11,12,14,15	9,10,11	16,17	9,10,11,12,13
Nucleotide	**11,12,16**	8,9	17	**9,10,11,12,13**
Peptide	**11,16**	8,9	17	7,8,12
Protein	**11,12**	9	17	**8,9,10,11**
Sugar	**11,12,16**	9	17,19	**4,8,19,11,19**
Virus	**11,12,16**	**7,8,9,10,11**	17,19	**7,11,18,19,20,21**

These data were obtained by visual selection from Figure 2. Length selections of high significance are highlighted in bold.

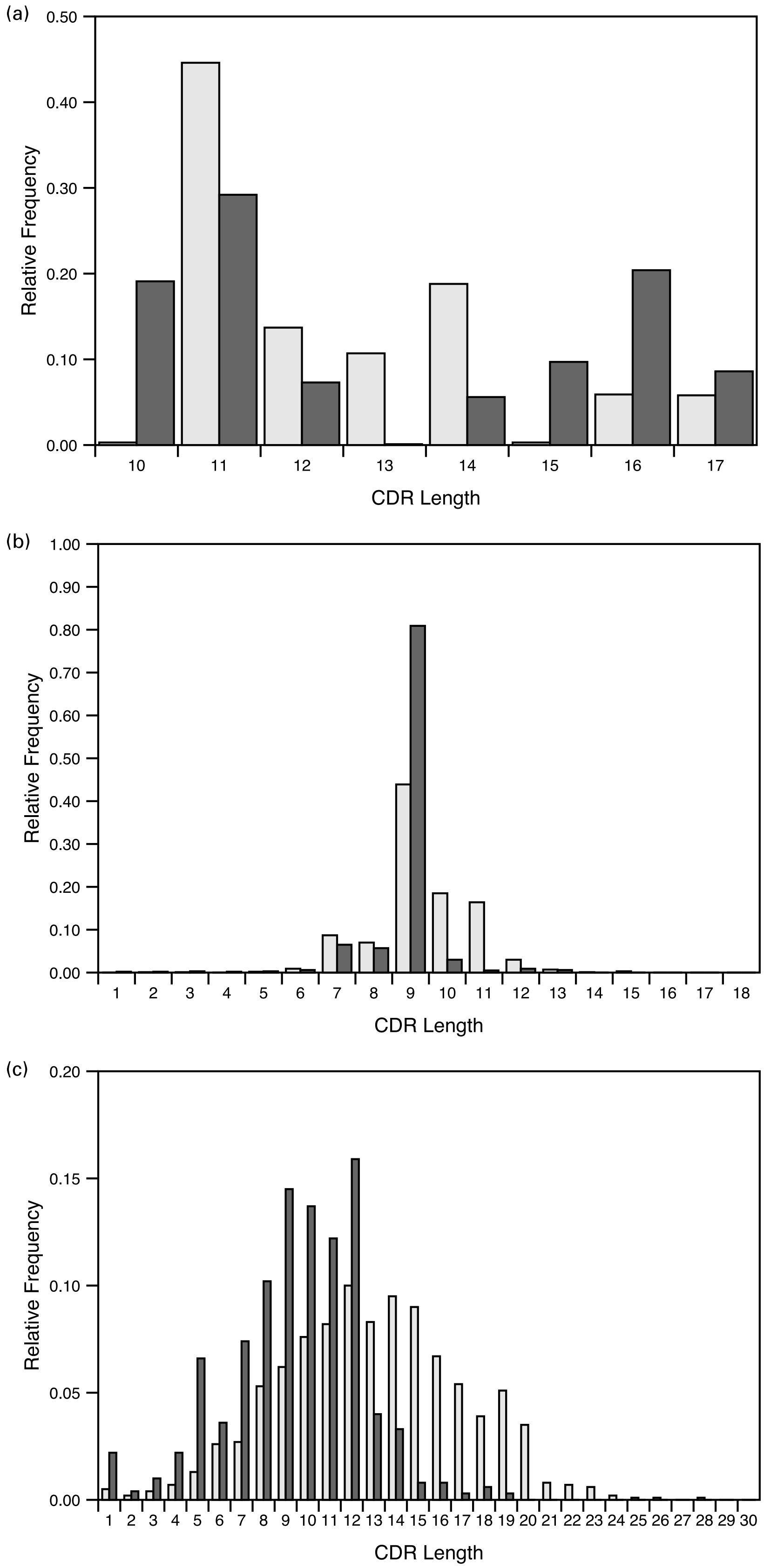

Figure 3. Distributions of CDR lengths observed in the dataset for known antigens in human and mouse antibodies for (a) CDR-L1, (b) CDR-L3 and (c) CDR-H3. Light shaded bars (left of each pair) are human and dark shaded bars (right of each pair) are mouse.

Table 7. Significances calculated from a χ^2 test to compare length distributions in CDR-H3 for each pair of antibody classes within human and mouse antibodies

	Significance (p)					
	Nucleotide	Sugar	Peptide	Protein	Virus	Lipid
A. *Human*						
Hapten	5.4×10^{-2}[a]	1.0[a]	3.0×10^{-1}[a]	6.5×10^{-2}[a]	1.5×10^{-1}[a]	5.1×10^{-2}[a]
Nucleotide	–	2.5×10^{-2}	9.3×10^{-1}[a]	7.6×10^{-1}	3.6×10^{-5} ++	2.5×10^{-2}
Sugar	–	–	7.9×10^{-2}[a]	4.1×10^{-2}	5.6×10^{-3}	3.0×10^{-1}
Peptide	–	–	–	8.3×10^{-1}	1.0×10^{-2}	1.9×10^{-1}[a]
Protein	–	–	–	–	2.2×10^{-12} ++++	1.8×10^{-1}
Virus	–	–	–	–	–	1.7×10^{-4} +
B. *Mouse*						
Hapten	3.1×10^{-15} ++++	4.5×10^{-4} +	8.7×10^{-4} +	3.3×10^{-17} ++++	1.0×10^{-9} ++++	4.4×10^{-6} +++
Nucleotide	–	2.4×10^{-7} ++++	3.9×10^{-4} +	7.0×10^{-6} +++	6.4×10^{-2}	9.8×10^{-2}
Sugar	–	–	2.2×10^{-2}	1.6×10^{-3}	1.5×10^{-6} +++	1.1×10^{-5} ++
Peptide	–	–	–	9.8×10^{-2}	5.6×0^{-2}	2.4×10^{-4} +
Protein	–	–	–	–	8.2×10^{-5} ++	7.3×10^{-6} +++
Virus	–	–	–	–	–	4.8×10^{-3}

The degree of significance is indicated (+, $p < 0.1\%$; ++, $p < 0.01\%$; +++, $p < 0.001\%$; ++++, $p < 0.0001\%$).

[a] Indicates calculations for which insufficient data were available to obtain expected values greater than 5 even after grouping of loop lengths.

CDR-L3. There are no significant differences except: Viruses are different from everything except nucleotides; proteins are different from haptens and nucleotides; haptens are different from nucleotides.

CDR-H2. There are no significant differences except: haptens are different from everything except sugars; proteins and sugars are different.

CDR-H3. Haptens are different from everything else; nucleotides are different from everything except sugars; sugars are different from everything except nucleotides and peptides; proteins are different from everything except peptides; viruses are different from everything except peptides; peptides are different from haptens and nucleotides.

Table 6 summarizes suggested preferred lengths for the CDRs of antibodies binding to each antigen class obtained from a visual analysis of the data in Figure 2. Length selections of high significance are highlighted in bold. Note that in some cases two length distributions may be significantly different, yet the most commonly occurring lengths (shown in the Table) are the same. These suggested lengths may be used to bias phage library generation,[15–17] in selection from libraries[18] or in modifying specificity[19] with relative length usage being derived from the graphs.

We have thus shown that there are clear, statistically significant, differences in the distributions of CDR lengths in antibodies binding to different classes of antigens. Differences in distributions of lengths are, in many cases, highly significant.

Variation in CDR length between species

As stated above, length distributions in the CDRs, particularly CDR-H3, vary between species. Analysis of the Kabat data shows that CDR-L1 and CDR-L3 show rather different length distributions in human and house. Some differences are seen in the other CDRs, but these do not appear relevant to the correlations with antigen type discussed here. For example, 18 residue CDR-H2 loops are seen only in human antibodies, but Figure 2(e) does not show significant variation in the usage of 18 residue CDR-H2 loops between antigen classes. The distributions for human and mouse CDR length usage for CDR-L1, L3 and H3 in the antibodies analyzed here are shown in Figure 3.

CDR-L1. The most notable differences between the length distributions in human and mouse are (i) the low usage of mid-length CDR-L1 (13 or 14 residues) in mouse and (ii) the high usage of longer CDR-L1 (15–17 residues) in mouse.

Examining the graph of length usage in different antigen classes (Figure 2(a)) shows that 13 or 14 residue CDRs are over-used in anti-lipid antibodies. Table 3 shows that the majority of these are, in fact, mouse antibodies. This feature of anti-lipid antibodies is dominated by the minority human antibodies.

Figure 2(a) shows that long CDR-L1s (15–17 residues) are used preferentially by hapten, nucleotide and sugar binders, and, in particular, by

Table 8. Raw distributions of amino acid frequencies in the contact definitions of the CDRs

		Antibody loops							
Residue	SRep loops	Hapten	Lipid	Nucleotide	Peptide	Protein	Sugar	Virus	Total
A	7590	692	221	856	132	3206	264	868	6239
C	1808	13	7	33	3	92	11	34	193
D	8584	497	160	572	97	2356	166	664	4512
E	6303	137	62	190	35	925	103	273	1725
F	3702	307	105	396	91	1249	105	330	2583
G	12,199	1176	319	1015	204	2669	268	982	6633
H	2789	394	79	384	73	1197	115	313	2555
I	3974	707	216	792	129	2592	293	679	5408
K	6678	262	65	337	63	870	80	258	1935
L	7088	967	260	1056	178	3261	353	960	7035
M	1838	224	81	304	56	1193	113	357	2328
N	6698	878	252	961	144	2992	297	786	6310
P	8896	380	114	400	77	1509	132	445	3057
Q	3801	386	104	376	62	1395	124	403	2850
R	4961	603	179	850	115	2193	178	620	4738
S	8242	1236	474	1439	249	5283	489	1358	10,528
T	6922	823	205	825	175	2662	227	749	5666
V	5224	475	110	429	98	1347	112	429	3000
W	1457	861	258	887	135	2839	276	728	5984
Y	3553	1598	473	1833	314	6246	504	1722	12,690
Total	112,307	12,616	3744	13,935	2430	46,076	4210	12,958	95,969

peptide binders. The datasets for all these are all at least 75% mouse antibodies, so it is possible that the significance of these results is over-estimated because of the preference for mouse antibodies to use longer CDR-L1 loops. Conversely, however, 43.6% of the mouse antibodies fall into one of these four antigen classes (as opposed to 14.4% of the human antibodies), so the over-representation of long CDR-L1s in the mouse dataset may be because the mouse antibody dataset is biased towards long CDR-L1s through its high number of antibodies binding to these classes of antigens.

CDR-L3. The key observation is that more than 35% of human antibodies have a 10–12 residue CDR-L3 compared with <5% of mouse antibodies. These longer CDR-L3 loops are characteristic of lipid-binding antibodies. As with CDR-L1, the majority of lipid-binding antibodies are mouse antibodies. This dominant feature of lipid-binding antibodies arises from the minority human antibodies.

CDR-H3. In our dataset, human CDR-H3 loops range in length from two to 30 residues, while mouse CDR-H3 loops range from two to 19 residues. Figure 2(f) shows that all antibodies with CDR-H3 longer than 19 residues bind protein, or more frequently, virus. These very long CDR-H3 loops consist of 13 human protein binders, 19 human virus binders and two bovine antibodies binding protein and virus, respectively.

Long CDR-H3 loops are characteristic of virus binders (particularly 18–20 residues) and sugar binders (19 residues). However, these characteristics are restricted to human antibodies; they are not observed in mouse antibodies (only one mouse anti-virus and zero anti-sugar antibodies have CDR-H3 longer than 17 residues). While the use of longer CDR-H3 loops in mouse anti-virus antibodies is less characteristic than it is in human antibodies, the distribution still favors longer loops (Table 4).

Since the length distributions are so different for CDR-H3, the statistical analysis for this loop was repeated for human and mouse antibodies separately. The results are shown in Table 7. Human data were somewhat restricted, meaning that valid statistical data could be obtained only for some of the comparisons. Fewer significant differences are seen in the human data simply because the data sets are too small. However, the trends observed and the conclusions are essentially no different from the combined dataset.

Sequence composition

Initially, we set out to ask whether the sequence composition of the antigen combining site is significantly different from that of typical loops in proteins. Using the CATH SReps†[20] loop regions of at least four residues in length were identified using the SSTRUC (D. K. Smith & J. M. Thornton, unpublished results) implementation of the DSSP algorithm[21] and the sequence composition of all loops in the database was calculated. CATH SReps are sequence family representatives selected from the Protein Databank.[22,23] The sequence composition of these loops provided our expected values for the distributions in antibody combining sites. Raw distribution data are shown in Table 8.

The χ^2 test was then applied to compare the sequence composition of all antibody combining sites, and each of our seven classes of antigen in turn, with the expected values from the CATH

† http://www.biochem.ucl.ac.uk/bsm/cath/

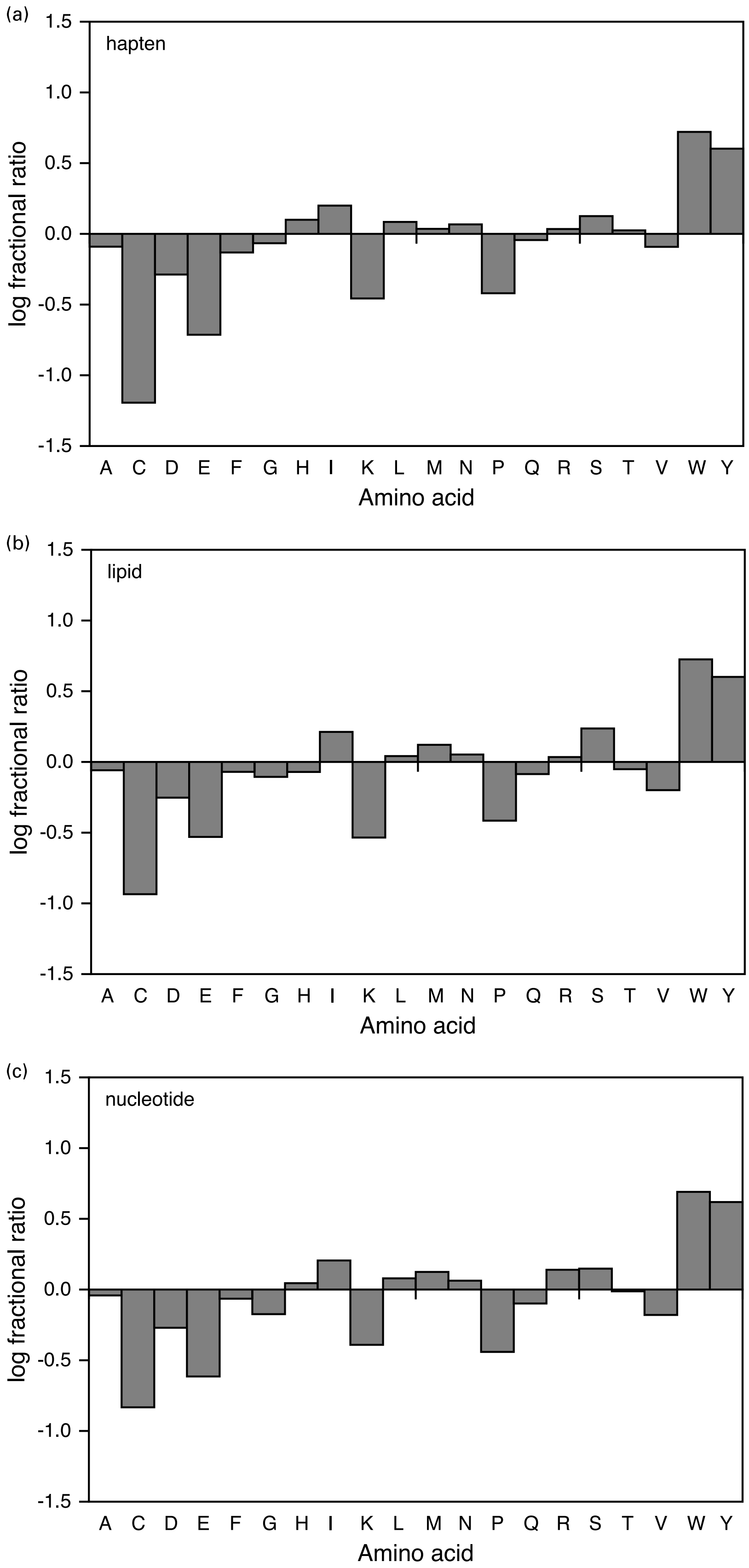

Figure 4 (*legend on page 349*)

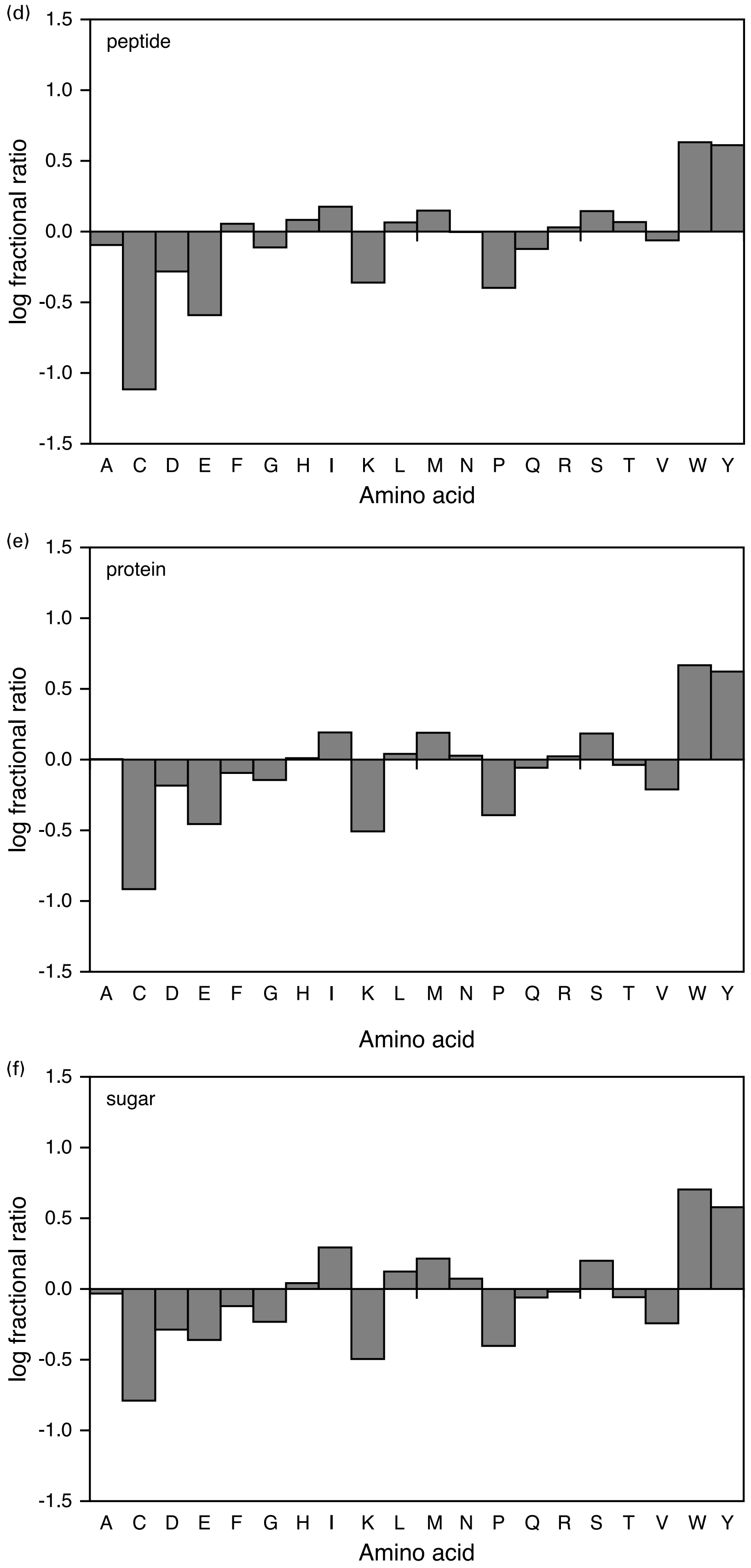

Figure 4 (*legend opposite*)

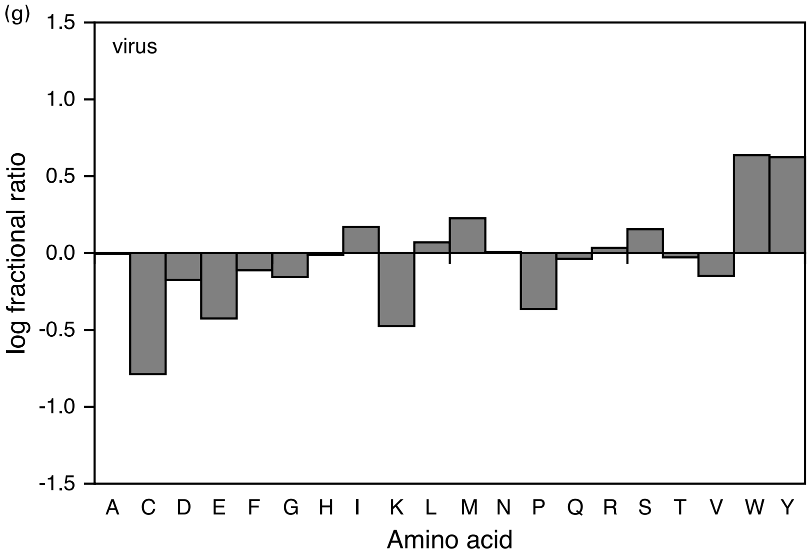

Figure 4. Sequence composition of the combining sites for antibodies binding the seven classes of antigens: (a) hapten, (b) lipid, (c) nucleotide, (d) peptide, (e) protein, (f) sugar, and (g) virus. The graphs show the fractional distribution of amino acid residues in the contact definition of the CDRs divided by the fractional distribution in CATH SReps, expressed as a $\log_{10}$ value.

SReps. The null hypothesis is that the sequence composition of the antibody combining sites is the same as generic protein loops. Thus, expected values were obtained simply by scaling the observations from the CATH SReps to the number of observations in each antibody dataset. χ^2 values ranged from 1433.34 (peptides) to 30,700.52 (proteins) with the χ^2 value for the combined antibody sequences being 63,001.22. In all cases, p-values, calculated using 19 degrees of freedom, were zero to the level of accuracy used in our calculations (i.e. much less than 5.42×10^{-20}, a value reached with a χ^2 of 139). In all cases, the distributions of amino acid frequencies in antibody combining sites were totally different from that of normal loops.

To display the results graphically, the fractional distributions were calculated and divided by the fractional distribution for generic protein loops (CATH SReps). The results, displayed as log values, are shown in Figure 4.

In order to assess the significance of the abundance of each individual amino acid second-level χ^2 tests were applied. When this type of analysis is applied *post hoc*, one effectively has no degrees of freedom and a method for achieving reliable statistics has been described briefly.[24] The details of the method, which has proved to be of wide applicability, are shown in the Appendix. With 20 amino acid types, we need to have a χ^2 value of 17.668 to have a true significance of better than $p < 0.0005$, which was selected as a cutoff for this work. Table 9 summarizes the results of this comparison.

The high abundance of tyrosine (Y) in the antibody combining sites is very striking in all classes of antibodies, as is that of tryptophan (W). Both these residues have a mixture of hydrophilic, hydrophobic and aromatic character. This mixed nature may have a clear advantage, in being able to make a range of types of interactions and therefore being generally "sticky". In fact the abundance of tryptophan is not particularly high; its high relative abundance is more a reflection of the fact that it occurs so extremely rarely in the SRep loops.

Isoleucine (I), and serine (S) show a higher abundance in all types of antibodies than in generic loops. In contrast, glutamate (E), aspartate (D), lysine (K), cysteine (C) and proline (P) have a noticeably lower abundance in the CDRs of all types of antibodies when compared with generic protein loops.

Certain amino acids are significantly over-represented or under-represented only in individual classes of antibodies. Alanine (A) is under-represented significantly only in hapten binders. Phenylalanine (F) is under-represented significantly only in hapten, protein and virus binders. Glycine (G) and valine (V) are under-represented significantly in all classes except peptide binders. Histidine (H) is over-represented significantly only in hapten binders. Leucine (L) is over-represented significantly in all classes except lipid and peptide binders. Methionine (M) is over-represented significantly in all classes except hapten, lipid and peptide binders. Asparagine (N) is over-represented significantly only in hapten and nucleotide binders. Glutamine (Q) is under-represented significantly only in nucleotide and protein binders. Arginine (R) is over-represented significantly only in nucleotide binders. Threonine (T) is under-represented significantly only in protein binders.

The reasons for many of these trends are not obvious. However, in some cases explanations are clear. For example, the very strong over-representation of arginine in nucleotide binders is clearly a reflection of interactions with the negative charge of the phosphate backbone. As stated above, methionine (M) is over-represented significantly in all classes except hapten, lipid and peptide binders. Its over-representation in protein binders is particularly striking and may reflect a preference for the use of ten residue CDR-L1s falling in

Table 9. Summary of sequence composition significance

	Significance (χ^2)							
Amino Acid	All	Hapten	Nucleotide	Lipid	Peptide	Protein	Sugar	Virus
A	15.15	30.26 − −	7.93	4.05	6.32	0.19	1.48	0.07
C	1199 10 − − − −	177.93 − − − −	163.35 − − − −	47.09 − − −	33.35 − −	585.03 − − − −	47.56 − − −	146.15 − − − −
D	1134.92 − − − −	226.44 − − − −	228.83 − − − −	55.62 − − −	42.39 − − −	428.83 − − − −	75.42 − − −	107.58 − − − −
E	2539 46 − − − −	460.56 − − − −	448.76 − − − −	104.42 − − − −	75.36 − − −	1115.89 − − − −	75.18 − − −	283.72 − − − −
F	117.83 − − − −	28.50 − −	8.82	2.75	1.48	59.08 − − −	8.22	22.09 −
G	798.94 − − − −	27.57 − −	164.87 − − − −	18.90 −	13.62	408.05 − − − −	78.36 − − −	128.64 − − − −
H	8.93	20.79 +	4.10	2.10	2.65	0.67	1.04	0.24
I	1138.27 + + + +	152.10 + + + +	180.64 + + + +	52.65 + + +	21.52 +	514.99 + + + +	139.25 + + + +	106.02 + + + +
K	2545.87 − − − −	317.68 − − − −	292.16 − − − −	111.60 − − − −	45.96 − − −	1329.63 − − − −	115.90 − − − −	340.90 − − − −
L	136.61 + + + +	36.63 + + +	35.15 + +	2.38	3.96	28.29 + +	28.68 + +	24.72 +
M	345.82 + + + +	1.49	25.16 +	6.35	6.62	231.76 + + + +	28.23 + +	99.05 + + +
N	47.87 + + +	20.96 +	20.11 +	3.69	0.01	12.11	8.40	0.22
P	2784.38 − − − −	383.83 − − − −	449.45 − − − −	112.39 − − − −	69.29 − − −	1321.61 − − − −	121.73 − − − −	329.35 − − − −
Q	56.92 − − −	3.93	19.51 −	4.07	4.98	24.67 −	2.40	2.88
R	48.13 + + +	3.75	88.89 + + +	1.12	0.55	6.19	0.34	3.96
S	1634.68 + + + +	103.89 + + + +	168.78 + + + +	144.47 + + + +	28.00 + +	967.42 + + + +	104.91 + + + +	174.22 + + + +
T	15.18	2.65	0.70	2.88	4.25	19.77 −	4.07	3.09
V	506.13 − − − −	21.31 −	74.38 − − −	23.63 −	2.00	324.35 − − − −	35.89 − − −	50.08 − − −
W	17,751.31 + + + +	2970.98 + + + +	2755.78 + + + +	902.98 + + + +	339.63 + + + +	8129.83 + + + +	897.33 + + + +	1864.73 + + + +
Y	30,175.73 + + + +	3601.12 + + + +	4391.01 + + + +	1061.30 + + + +	731.40 + + + +	15,192.18 + + + +	1032.36 + + + +	4199.30 + + + +
All	63,001.22	8592.36	9528.38	2664.45	1433.34	30,700.52	2806.74	7887.03

In all cases, the null hypothesis is that the sequence composition is the same as in general protein loops (taken from the CATH SReps). The total χ^2 values are shown at the bottom (All) together with individual χ^2 values for each amino acid type. While significance for the total χ^2 value is calculated with 19 degrees of freedom, a double test must be applied for individual χ^2 values. Significant over-representation or under-representation is indicated at $p < 0.0005$ (+/−), $p < 0.000005$ (++/−−), $p < 0.00000005$ (+++/−−−) and for χ^2 values >100 (++++/−−−−). p is zero ($<5.4\text{x}10^{-20}$) with a χ^2 value of >85 at the level of precision used in our calculations.

canonical class 1 (class 10A in the numbering scheme of Martin & Thornton[25]), which have methionine or leucine at position L33.

Comparison of antigen classes

Having shown that for all classes of antigens, antigen combining sites have a sequence composition different from that of loops in general, we set out to compare the combining sites of antibodies that bind to different classes of antigens with one another. This may give clues about cross-reactivity preferences.

Again, χ^2 tests were used to compare distributions. In this case, the residue distributions of each of the seven classes were compared with each of the other classes in turn. Total χ^2 values were calculated as described in the Appendix, and the results are shown in Table 10. In some cases, it was necessary to group amino acid types, since expected values for cysteine were below five. This occurred in the comparison of hapten with lipid, hapten with peptide, lipid with peptide, and protein with peptide. In all cases, cysteine was grouped with methionine, since both contain sulphur.

In most cases, residue distributions are significantly different between classes at least at the $p < 0.1\%$ level. However, there are a number of noticeable exceptions. For example, given that most viral antigens are proteins, we would not expect the sequence composition to be very different between viral and protein binders, and

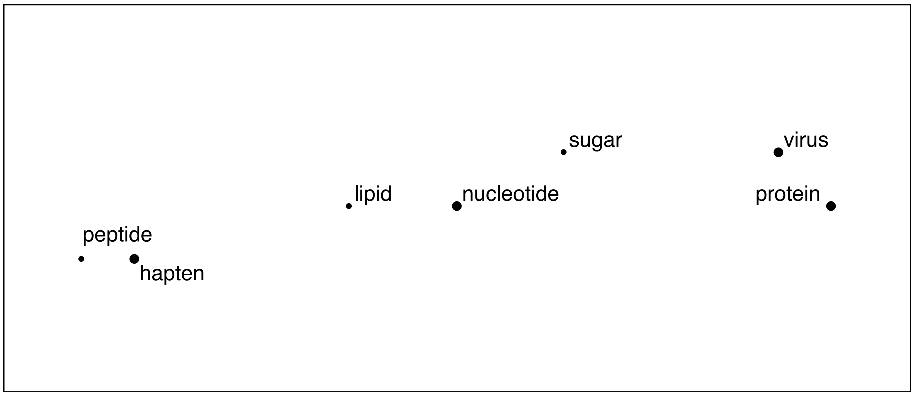

Figure 5. Principal components projection of the fractional composition vectors for antibodies binding the seven classes of antigens. Large spots are used for classes with more than 200 examples; small spots for classes with fewer than 100 examples.

this is indeed the case, with the distributions being different only at $p = 0.029$. While this ($p < 5.0\%$) level of significance may be regarded as significant for many areas of research, it is only marginally significant compared with the levels of significance seen for other comparisons.

Differences in sequence composition of low significance were found between peptides and haptens; between peptides and nucleotides; between lipids and peptides; between lipids and proteins; and between lipids and sugars. In general, lipids show the least significant differences with other classes and this may be important for cross-reactivity.

Composition space

The composition of the seven different classes of antigen-binding antibodies was also visualized by projection of the 20-dimensional fractional composition space into two dimensions using principal components projections with the program Xgobi.[26] The results are shown in Figure 5, which should be considered in the light of the statistical information. Some of the data are available only in small quantities and therefore apparent similarities or differences in composition visible in this Figure may not be statistically significant. For example, the similarities in hapten and peptide composition and in the protein and virus composition, visible in this plot, are supported by the statistics (in neither case do the statistics suggest that these are significantly different). However, the plot suggests that peptide and nucleotide composition are very different, while cluster analysis groups nucleotides with virus and protein rather than with peptides (data not shown). However, neither of these results is supported by the statistical analysis.

Conclusions

We have shown clear correlations between both the lengths of the CDRs and the sequence composition of the residues forming the contact definition of the CDRs with the type of antigen. These are merely statistical preferences: protein-binding antibodies with very short CDR-H3 loops do, of course, occur. For example, Gloop2,[27] which binds the loop region of lysozyme has a CDR-H3 of just

Table 10. Significance of difference in sequence composition from comparing pairs of antigen classes

	Significance (p)					
	Hapten	Nucleotide	Lipid	Peptide	Protein	Sugar
Nucleotide	2.8×10^{-14} ++++	–	–	–	–	–
Lipid	3.7×10^{-7} ++++	2.1×10^{-4} +	–	–	–	–
Peptide	6.7×10^{-3}	7.9×10^{-3}	1.1×10^{-3}	–	–	–
Protein	0 ++++	0 ++++	4.1×10^{-2}	1.1×10^{-6} +++	–	–
Sugar	0 ++++	7.6×10^{-10} ++++	2.2×10^{-3}	1.0×10^{-7} ++++	3.7×10^{-7} ++++	–
Virus	0 ++++	6.1×10^{-15} ++++	1.0×10^{-4} ++	1.3×10^{-4} +	2.9×10^{-2}	3.7×10^{-7} ++++

The p-values calculated from the χ^2 values using 19 degrees of freedom (18 in the cases where cysteine and methionine had to be merged; see the text) are shown together with a set of + signs to indicate the degree of significance (+, $p < 0.1\%$; ++, $p < 0.01\%$; +++, $p < 0.001\%$; ++++, $p < 0.0001\%$).

four residues. While there are many such exceptions, general rules about the nature of the antigen as correlated with the lengths of the CDRs are clear.

The results of this analysis lead us to suggest that previously demonstrated preferences for certain combinations of canonical classes[11,12] binding to given types of antigen are driven by the nature of antigen recognition as well as genetic and structural factors.

In addition, the results give very clear guidance for the design of phage display libraries,[15–17] or the enhancement of affinity (*in vitro* maturation) through biased random mutagenesis.[28,29] For example, to create anti-viral antibodies, one should bias the library in favor of antibodies with long CDR-L3, CDR-H2 and CDR-H3 having high contents of serine, isoleucine, and methionine, and low contents of phenylalanine, cysteine, glycine and valine content.

Interestingly, the nucleotide binders are relatively similar in sequence composition to the lipid binders (Figure 5) and the composition is different only at the $p < 0.1\%$ level (Table 10). These show significant CDR length difference only for CDR-L3 (Table 5). A number of anti-DNA antibodies are known to cross-react with lipids, such as cardiolipin. While Figure 5 does not suggest that nucleotides and peptides are particularly similar in residue composition, the statistics show that the composition of this pair is not significantly different. Given also similarity in shape, it is possible that both lipids and peptides are able to act as immunogens to illicit an immune response generating cross-reactive antibodies with weak binding to DNA, later maturing to create anti-DNA antibodies observed in auto-immune disease such as systemic lupus erythamatosus.[30]

We have begun some initial work to use the results of this analysis predictively. Neural networks trained with the sequence composition and length data from the four largest antigen categories (protein, hapten, nucleotide and virus) show an 85% improvement over a random prediction (46% correct compared with a random prediction of 25% correct). We intend to continue this work using different representations of the data, types of network, training parameters, etc.

In summary, we have shown that the general shape and sequence composition of the antibody combining site do influence the class of antigen to which the antibody binds. These data show promise in prediction of antigen class and can be used to guide phage library design.

Materials and Methods

Data were collected from the April 2000 release of the Kabat sequence database, this being the most recent version for which the complete dataset is available by FTP. All analysis was performed using scripts written in Perl accessing the Kabat data *via* the KabatMan software.[31] KabatMan makes links between associated light and heavy chains in the Kabat data allowing complete antibodies to be studied. The April 2000 dataset contains 2140 complete antibodies for which the antigen is identified in 1875 examples.

Antigens were classified into one of seven classes: protein, peptide, hapten, lipid, nucleotide, sugar, and virus. Most viral antigens are proteins, but previous observations had suggested that the antigen combining sites of anti-viral antibodies are somewhat different from normal protein-binding antibodies. The classification process was semi-automated by using a Perl script that presented the name of each antigen, allowing the user to classify the associated antibody. A portion of the text that caused the user to select that classification could then be highlighted by the user and all subsequent antibodies containing that text in their antigen descriptions were placed in the same class with no manual intervention. An initial pass through the antigen names enabled some 90% to be classified. For the remaining ca 200 antigens, the original literature was checked to classify the antigen type.

In rare cases, more than one antibody in Kabat has the same name, although all sequences have a unique KADBID accession code. For example, there are two antibodies with the name 2B2. One is a complete anti-2-phenyl oxazolone antibody composed of light chain 006442[32] and heavy chain 001350,[33] while the other has only a light chain of undefined specificity (KADBID: 029804[34]). We found that no complete antibodies had the same name as another complete antibody, so filtering on completeness (necessary in any case for our analysis) proved sufficient to overcome this problem. Sequence data for the CDRs were then obtained using Kabatman for each of the antigen classes and associated with each antibody name.

To collect data for analysis of sequence composition, lists of antibody names were compiled for each of the antibody classes. The sequences of the contact-definition CDRs of these (complete) antibodies were extracted using KabatMan. The contact-definition CDRs[9] restrict the CDRs to those residues most likely to be involved in interactions with antigen: CDR-L1, L30–L36; CDR-L2, L46–L55; CDR-L3, L89–L97; CDR-H1, H30–H35 (Chothia numbering), H30–H35B (Kabat numbering); CDR-H2, H47–H58; and CDR-H3, H93–H101. Any X or ? character was stripped from the sequences and any antibodies containing no information for any one of the six CDRs were rejected. The total composition of the six CDRs was pooled i.e. we made no attempt to examine the composition of each loop independently. The total and fractional composition could then be calculated simply by counting the number of occurrences of each amino acid in the resulting data files.

In the case of CDR length analysis, the Kabat definitions of the CDRs were used: CDR-L1, L24–L34; CDR-L2, L50–L56; CDR-L3, L89–L97; CDR-H1, H30–H35 (Chothia numbering), H30–H35B (Kabat numbering); CDR-H2, H50–H65; CDR-H3, H95–H102. Any antibodies with zero-length CDRs were rejected, but no special handling of X or ? characters was performed. The assumption was made that these are used in the datafiles to indicate the correct length of the CDRs even if the sequence is not known. In performing the χ^2 tests to compare loop length distributions between classes, some grouping of loop lengths was necessary to obtain expected counts >5.0. A complete list of the groupings used is available on the web†

† http://www.bioinf.org.uk/papers/

Acknowledgements

A.V.J.C. was funded by Nuffield Undergraduate Research Bursary URB/00044/A; Darren Nesbeth performed some initial evaluation of amino acid composition as an undergraduate project student under the direction of A.C.R.M. and Janet M. Thornton. Richard Mott is thanked for his very valuable help with the statistics.

References

1. Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917.
2. Riechmann, L., Clark, M., Waldmann, H. & Winter, G. (1988). Reshaping human antibodies for therapy. *Nature (London)*, **332**, 323–327.
3. Wu, T. T. & Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light, chains and their implications for antibody complementarity. *J. Exp. Med.* **132**, 211–250.
4. Davies, D. R., Padlan, E. A. & Sheriff, S. (1990). Antibody–antigen complexes. *Annu. Rev. Biochem.* **59**, 439–473.
5. Padlan, E. A. (1977). The structural basis for the specificity of antibody–antigen reactions and structural mechanisms for the diversification of antigen binding specificities. *Quart. Rev. Biophys.* **10**, 35–65.
6. Braden, B. C. & Poljak, R. J. (1995). Structural features of the reactions between antibodies and protein antigens. *FASEB J.* **9**, 9–16.
7. Wilson, I. A. & Stanfield, R. L. (1993). Antibody–antigen interactions. *Curr. Opin. Struct. Biol.* **3**, 113–118.
8. Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1987). The structural basis of antigen–antibody recognition. *Annu. Rev. Biophys. Bioeng.*, 139–159.
9. MacCallum, R. M., Martin, A. C. R. & Thornton, J. M. (1996). Antibody–antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* **262**, 732–745.
10. Kuhn, L. A., Siani, M. A., Pique, M. E., Fisher, C. L., Getzoff, E. D. & Tainer, J. A. (1992). The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *J. Mol. Biol.* **228**, 13–22.
11. Lara-Ochoa, F., Almagro, J. C., Vargas-Madrazo, E. & Conrad, M. (1996). Antibody–antigen recognition: a canonical structure paradigm. *J. Mol. Evol.* **43**, 678–684.
12. Vargas-Madrazo, E., Lara-Ochoa, F. & Almagro, J. C. (1995). Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. *J. Mol. Biol.* **254**, 497–504.
13. Lara-Ochoa, F., Vargas-Madrazo, E. & Almagro, J. C. (1995). Distributions of the use frequencies of amino acids in the hypervariable regions of immunoglobulins. *J. Mol. Evol.* **41**, 98–103.
14. Wu, T. T., Johnson, G. & Kabat, E. A. (1993). Length distribution of CDRH3 in antibodies. *Proteins: Struct. Funct. Genet.* **16**, 1–7.
15. Engberg, J., Jensen, L. B., Yenidunya, A. F., Brandt, K. & Riise, E. (2001). Phage-display libraries of murine antibody Fab fragments. In *Antibody Engineering* (Kontermann, R. & Dübel, S., eds), pp. 65–92, Springer Verlag, Heidelberg.
16. Winter, G., Griffiths, A. D., Hawkins, R. E. & Hoogenboom, H. R. (1994). Making antibodies by phage display technology. *Annu. Rev. Immunol.* **12**, 433–455.
17. Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G. *et al.* (2000). Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol.* **296**, 57–86.
18. Jespers, L. S., Roberts, A., Mahler, S. M., Winter, G. & Hoogenboom, H. R. (1994). Guiding the selection of human antibodies from phage display repertoires to a single epitope of an antigen. *Biotechnology,* **12**, 899–903.
19. Miyazaki, C., Iba, Y., Yamada, Y., Takahashi, H., Sawada, J. & Kurosawa, Y. (1999). Changes in the specificity of antibodies by site-specific mutagenesis followed by random mutagenesis. *Protein Eng.* **12**, 407–415.
20. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
21. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure. *Biopolymers*, **22**, 2577–2637.
22. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
23. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R. *et al.* (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
24. Martin, A. C. R., Toda, K., Stirk, H. J. & Thornton, J. M. (1995). Long loops in proteins. *Protein Eng.* **8**, 1093–1101.
25. Martin, A. C. R. & Thornton, J. M. (1996). Structural families of loops in homologous proteins: automatic classification, modelling and application to antibodies. *J. Mol. Biol.* **263**, 800–815.
26. Swayne, D. F., Cook, D. & Buja, A. (1992). XGobi: interactive dynamic graphics in the X window system with a link to S. In *ASA Proceedings of the 1991 American Statistical Association Meetings*, pp. 1–8, American Statistical Association, VA.
27. Darsley, M. J. & Rees, A. R. (1985). Nucleotide sequences of five anti-lysozyme monoclonal antibodies. *EMBO J.* **4**, 393–398.
28. Chowdhury, P. S. & Pastan, I. (1999). Improving antibody affinity by mimicking somatic hypermutation in vitro. *Nature: Biotechnol.* **17**, 568–572.
29. Daugherty, P. S., Chen, G., Iverson, B. L. & Georgiou, G. (1999). Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc. Natl Acad. Sci. USA*, **97**, 2029–2034.
30. Kalsi, J. K., Martin, A. C. R. & Isenberg, D. A. (1995). Structure-function relation of anti-DNA antibodies. *Lupus*, **4**, 245–248.
31. Martin, A. C. R. (1996). Accessing the Kabat antibody sequence database by computer. *Proteins: Struct. Funct. Genet.* **25**, 130–133.
32. Kaartinen, M. & Makela, O. (1987). Functional analogues of the VKOx1 gene in different strains of mice: Evolutionary conservation but diversity based on V–J joining. *J. Immunol.* **138**, 1613–1617.

33. Kaartinen, M., Solin, M.-L. & Makela, O. (1989). "Allelic" forms of immunoglobulin V genes in different strains of mice. *EMBO J.* **8**, 1743–1748.
34. Williams, S. C., Frippiat, J.-P., Tomlinson, I. M., Ignatovich, O., le Franc, M.-P. & Winter, G. (1996). Sequence and evolution of the human germline V-repertoire. *J. Mol. Biol.* **264**, 220–232.
35. Saul, F. A., Amzel, L. M. & Poljak, R. J. (1978). The preliminary refinement and structural analysis of the Fab fragment from human immunoglobulin New at 2.0 Å resolution. *J. Biol. Chem.* **253**, 585–597.
36. Lamminmaki, U. & Kankare, J. A. (2001). Crystal structure of a recombinant anti-estradiol Fab fragment in complex with 17-β-estradiol. *J. Biol. Chem.* **276**, 36687–36694.
37. Ban, N., Day, J., Wang, X., Ferrone, S. & McPherson, A. (1996). Crystal structure of an anti-anti-idiotype shows it to be self-complementary. *J. Mol. Biol.* **225**, 617–627.

Appendix: Statistics

Comparing two distributions is achieved simply by using the standard χ^2 test. However, if one then wishes to look at two distributions that one has shown to be significantly different and ask whether a particular parameter in the distribution (for example the frequency of arginine amino acids) differs significantly, one has a problem as the post-hoc observation effectively has no degrees of freedom. It is thus necessary to calculate the significance as if one were using one degree of freedom, but then to ask what the probability is of seeing that level of significance by chance alone.

In the normal χ^2 test, if T_1 and T_2 are the total numbers of amino acid residues in antibody classes 1 and 2, respectively, then $O_{1,R}$ will be the observed number of residues of type R in class 1. Then, if classes 1 and 2 are from the same population, the estimate of the expected number of residues of type R in class 1 is:

$$E_{1,R} = T_1(O_{1,R} + O_{2,R})/(T_1 + T_2)$$

and the χ^2 test for whether the residue R is the same in both types is based on the distribution of the statistic:

$$X_R = ((O_{1,R} - E_{1,R})^2/E_{1,R}) + ((O_{2,R} - E_{2,R})^2/E_{2,R})$$

The total χ^2 test to compare the distributions is found by summing this value for all n amino acid types with $(n-1)$ degrees of freedom.

In order to evaluate the significance of differences in individual data points we proceed as follows: If there were no significant differences between classes 1 and 2, then the data will appear to be $(n-1)$ random samples drawn from a χ^2 distribution with one degree of freedom. Consequently, we want to know the distribution of the largest of these $(n-1)$ χ^2 values. Let it have p-value P. The probability of getting $(n-1)$ p-values $\leq P$ by chance alone is:

$$Q = (1-P)^{n-1}$$

We want this probability to be high, for example 0.9995 (i.e. a 0.05% significance level), so we solve for P:

$$P = 1 - Q^{1/(n-1)}$$

As there are $n = 20$ residue types, this means:

$$P = 1 - 0.9995^{1/19}$$

$$P = 2.63 \times 10^{-5}$$

i.e. a p-value of 2.63×10^{-5} (or a χ^2 for one degree of freedom of greater than 17.668) is necessary to achieve a definite level of significance better than 0.05%. In other words, if there were no differences between types 1 and 2, we could still expect that the p-value for a residue examined *post hoc* would be as small as 2.63×10^{-5}, purely by chance, 0.05% of the time.

Edited by I. Wilson

(Received 19 April 2002; received in revised form 10 October 2002; accepted 23 October 2002)