Antibody and Antigen Contact Residues Define Epitope and Paratope Size and Structure

James W. Stave and Klaus Lindpaintner

*J Immunol* published online 24 June 2013
http://www.jimmunol.org/content/early/2013/06/22/jimmunol.1203198

Supplementary Material
http://www.jimmunol.org/content/suppl/2013/06/24/jimmunol.1203198.DC1

Why *The JI*?

- **Rapid Reviews!** 30 days* from submission to initial decision
- **No Triage!** Every submission reviewed by practicing scientists
- **Speedy Publication!** 4 weeks from acceptance to publication

*average

Subscription
Information about subscribing to *The Journal of Immunology* is online at:
http://jimmunol.org/subscription

Permissions
Submit copyright permission requests at:
http://www.aai.org/About/Publications/JI/copyright.html

Email Alerts
Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/alerts
Antibody and Antigen Contact Residues Define Epitope and Paratope Size and Structure

James W. Stave and Klaus Lindpaintner1

A total of 111 Ag–Ab x-ray crystal structures of large protein Ag epitopes and paratopes were analyzed to inform the process of eliciting or selecting functional and therapeutic Abs. These analyses illustrate that Ab contact residues (CR) are distributed in three prominent CR regions (CRR) on L and H chains that overlap but do not coincide with Ab CDR. The number of Ag and Ab CRs per structure are overlapping and centered around 18 and 19, respectively. The CR span (CRS), a novel measure introduced in this article, is defined as the minimum contiguous amino acid sequence containing all CRs of an Ag or Ab and represents the size of a complete structural epitope or paratope, inclusive of CR and the minimum set of supporting residues required for proper conformation. The most frequent size of epitope CRS is 50–79 aa, which is similar in size to L (60–69) and H chain (70–79) CRS. The size distribution of epitope CRS analyzed in this study ranges from ∼20 to 400 aa, similar to the distribution of independent protein domain sizes reported in the literature. Together, the number of CRs and the size of the CRS demonstrate that, on average, complete structural epitopes and paratopes are equal in size to each other and similar in size to intact protein domains. Thus, independent protein domains inclusive of biologically relevant sites represent the fundamental structural unit bound by, and useful for eliciting or selecting, functional and therapeutic Abs.

M onoclonal Abs (mAbs) are important tools, diagnostic reagents, and therapeutic drugs. High-value commercial diagnostic, functional, and therapeutic mAbs must recognize full-length native proteins folded in native conformations as they exist in patients or patient samples. Given their potential value, strategies for developing mAbs to native protein Ags is of intense interest; however, the underlying biology dictating performance makes the task of developing such Abs both complicated and costly (1). For an Ab to exert a desired biological effect, it is usually the case that it must bind to a specific, predefined site. In such cases, B cell epitope prediction algorithms are of limited value because the Ab must bind to the predefined site regardless of its inherent immunogenicity. Of paramount importance is binding at the target site to the three-dimensional conformation the protein exists in under the conditions where the Ab must function.

The most common method for developing Abs is to inject purified protein into animals and isolate mAbs that bind to native proteins. A difficulty with this approach is the preparation of protein of suitable purity and functional conformation. For decades, researchers have attempted to use peptides as surrogates of full-length native proteins with little success (2). Anti-peptide Abs react well to their cognate peptides and may be useful for detecting linear epitopes in techniques with little success (2). Anti-peptide Abs react well to their cognate peptides and may be useful for detecting linear epitopes in techniques with little success (2).

Purified, full-length native and recombinant proteins are useful for eliciting Abs that bind native proteins, but are often difficult to prepare in quantities and qualities sufficient for immunization and selection of mAbs. In addition, immunization with full-length proteins results in a polyclonal response where the majority of B cells will react randomly to epitopes or sites of the protein (5, 6). In cases where the predefined site of interest is nonimmunodominant, or even nonimmunogenic, it is very difficult to isolate mAbs with the required performance specifications. Thus, there is a critical need for a means of directing Ab development to predefined functional epitopes of native protein Ags regardless of their inherent immunogenicity.

In the case of therapeutics, once candidate mAbs have been isolated with the requisite binding characteristics, the CDR can be taken from the isolated mAb (7) and inserted into engineered Ab constructs having desired effector function and characterized immunogenicity and toxicity profiles. It has been reported for six protein Ags that only a fraction of the CDR residues participate in binding (8), and that Ab contact residues (CRs) in nine crystal structures do not precisely correspond with residues modified by somatic hypermutation in Vκ (9). It is critical to identify specific residues involved in contacting Ag and understand how they are organized within CDR to engineer Ab reactivity. Thus, the native three-dimensional structure of both Ag epitope and Ab paratope are fundamental to the success of developing and engineering high-performance functional and therapeutic mAbs.

With these considerations in mind, we undertook the analysis of a large number of protein–Ab x-ray crystal structures to better understand both epitope and paratope size, structure, and interaction. A number of researchers have reported identifying CR of Abs and Ags by analysis of publicly available crystal structures (10–12). A large body of work in this field is aimed at identifying CR as a means of developing algorithms for predicting B cell epitopes (10, 13–17), whereas others have focused on analyzing or predicting CR in CDR of Abs (8, 18–20). The majority of the existing studies define two residues to be in contact when they are within a specified distance of each other. Many researchers use 4 Å as the definition of...
CR (10, 14–17), and the Molecular Modeling Database (21), at the National Center for Biotechnology Information, defines protein interactions using a 4-Å cutoff. One group has published several works using 6 Å (13, 20, 22) and reported that their conclusions were unaffected by cutoffs ranging from 4 to 6 Å (20).

Haste Andersen et al. (14) identified epitope CR using a 4-Å cutoff in 76 crystal structures composed of Ags > 25 aa (29 of the structures were variants of lysozyme). Using this approach, these researchers were able to determine the number of CRs per epitope, the maximum continuous sequence of CR within each epitope, and the distribution of continuous CR per segment. These results highlighted the discontinuous nature of protein Ag epitopes.

Padlan et al. (8) determined Ab CR in six crystal structures with protein Ags and compared the sequence positions with CDR defined by nucleic acid sequence variability (23). MacCallum et al. (12) determined Ab CR sequence positions for 7 protein–Ag structures (5 of which are lysozyme), and Almagro (19) determined numbers and sequence positions of $V_L$ and $V_H$ CR for 19 protein Ags. Besides analyzing only a limited number of protein–Ag structures, these studies did not report on CR sequence positions outside established CDRs. Recently, Kunik and coworkers (20) identified Ab CR in 69 Ag–Ab structures with proteins $\geq$ 5 aa using a 6-Å cutoff and reported that ~20% of CRs fall outside CDR. It is not clear how many of these structures contained large protein Ags, and the authors did not report sequence positions of the identified CR.

The existing studies define CR in different ways, and the number of unique structures analyzed is small. In addition, the number of Ag amino acids in the crystals are often not reported or are too few to draw conclusions about the nature of Ab binding to large protein Ags. Furthermore, some studies report only on Ag CR, whereas others focus solely on Ab.

In this article we apply a uniform analysis to both Ag and Ab proteins in a large number of x-ray crystal structures containing relatively large numbers of Ag amino acids ($\geq$ 85). The number of sequence positions of all Ab CRs and map them against CDR revealing CR Ab, as well as H and L chains. In addition, we identify the sequence positions of Ab CR, and report on CR sequence positions outside established CDRs. Recently, Kunik and coworkers (20) identified Ab CR in 69 Ag–Ab structures with proteins $\geq$ 5 aa using a 6-Å cutoff and reported that ~20% of CRs fall outside CDR. It is not clear how many of these structures contained large protein Ags, and the authors did not report sequence positions of the identified CR.

The existing studies define CR in different ways, and the number of unique structures analyzed is small. In addition, the number of Ag amino acids in the crystals are often not reported or are too few to draw conclusions about the nature of Ab binding to large protein Ags. Furthermore, some studies report only on Ag CR, whereas others focus solely on Ab.

In this article we apply a uniform analysis to both Ag and Ab proteins in a large number of x-ray crystal structures containing relatively large numbers of Ag amino acids ($\geq$ 85). The number of sequence positions of all Ab CRs and map them against CDR revealing CR Ab, as well as H and L chains. About half of all CRs belong to Ag (52%) and 65% of these lie within the H chain. Overall, there are ~7 CRs per L chain, 13 per H chain, and 18 per Ag; however, there is significant variation between structures. The largest number of CRs in a single structure, including both Ag and Ab, is 101 (1BGX), whereas the smallest is 20 (2EXW).

The frequency distribution of numbers of CR in all 111 Ag–Ab structures is shown in Fig. 1. CRs are distributed around ~18–19 for both Ag and Ab, and the distributions exhibit significant overlap. Whereas Fig. 1 illustrates that the number of CRs across all 111 structures is distributed in a similar and overlapping manner for both Ag and Ab, there is only a modest correlation between the number of Ag and Ab CRs within individual structures (Pearson correlation coefficient, $r^2 = 0.5394$).

The ratio of Ab to Ag CR for each of the 111 structures is shown in Fig. 2 and is represented as the log(AB CR/AG CR). Negative values result when there are more CR in the Ag (36 structures), and positive values result when there are greater numbers of Ab CR (59 structures). Sixteen structures have equal numbers of Ag and Ab CRs. The structure having the highest ratio in the data set is 2J88 (21 Ab CRs, 10 Ag CRs), and the lowest is 1ADQ (9 Ab CRs, 16 Ag CRs).

### Results

**Ag and Ab CR**

The list of 110 unique Ab–Ag x-ray crystal structures analyzed in this work is presented in Table I. A summary of the amino acids analyzed in this study is presented in Table II. The data set is composed of 78,666 aa, of which 5.3% are identified as CR. Ab amino acids represent 61% of the total and are divided almost equally between heavy (IgH) and light (IgL) chains. About half of all CRs belong to Ab (52%) and 65% of these lie within the H chain. Overall, there are ~7 CRs per L chain, 13 per H chain, and 18 per Ag; however, there is significant variation between structures. The largest number of CRs in a single structure, including both Ag and Ab, is 101 (1BGX), whereas the smallest is 20 (2EXW).

The frequency distribution of numbers of CR in all 111 Ag–Ab structures is shown in Fig. 1. CRs are distributed around ~18–19 for both Ag and Ab, and the distributions exhibit significant overlap. Whereas Fig. 1 illustrates that the number of CRs across all 111 structures is distributed in a similar and overlapping manner for both Ag and Ab, there is only a modest correlation between the number of Ag and Ab CRs within individual structures (Pearson correlation coefficient, $r^2 = 0.5394$).

The ratio of Ab to Ag CR for each of the 111 structures is shown in Fig. 2 and is represented as the log(AB CR/AG CR). Negative values result when there are more CR in the Ag (36 structures), and positive values result when there are greater numbers of Ab CR (59 structures). Sixteen structures have equal numbers of Ag and Ab CRs. The structure having the highest ratio in the data set is 2J88 (21 Ab CRs, 10 Ag CRs), and the lowest is 1ADQ (9 Ab CRs, 16 Ag CRs).

### Table I. Ag–Ab crystal structures analyzed

<table>
<thead>
<tr>
<th>PDB Structures Analyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ADQ</td>
</tr>
<tr>
<td>1AFV</td>
</tr>
<tr>
<td>1BGX</td>
</tr>
<tr>
<td>1CZS</td>
</tr>
<tr>
<td>1F6D</td>
</tr>
<tr>
<td>1EGI</td>
</tr>
<tr>
<td>1EZV</td>
</tr>
<tr>
<td>1FIJ</td>
</tr>
<tr>
<td>1FNS</td>
</tr>
<tr>
<td>1FSK</td>
</tr>
</tbody>
</table>
Ag CRS

In addition to counting the number of CRs within each structure, the amino acid sequence position of each CR was also recorded for both Ag and Ab. From the sequence positions, the CRSs were determined for all 111 structures. Fig. 3 shows the distribution of the size of the CRS across all 111 Ags using 3.5, 4.0, and 4.5 Å as the definition of CR. Although there appears to be a trend toward longer distances between CRs resulting in longer CRS, the effect is small using the distances examined in this study. The large majority of Ag structures (86.5%) have CRS ranging from 20 to 229 aa. Only 2 of the 111 structures analyzed have all of the Ag CRs contained within a CRS of <20 aa (2J88 and 3BSZ). The most common class of CRS in this data set is 50–79 residues. The average and median CRS of all 111 Ag structures is 125 and 95 aa, respectively.

There is a significantly higher average number of CR for large Ag structures >230 aa (19.8, n = 51), compared with smaller Ags <230 aa (16.6, n = 60, p < 0.01), and there is a weak but significant correlation between Ag size and CRS (Pearson correlation coefficient, $r^2 = 0.273$). The relationship is readily seen in the distribution of average CRS depicted in Supplemental Fig. 1. The epitope CRS increases proportionally to the size of Ag in the crystal. Also, the average CRS of Ags >230 aa (179) is significantly greater than Ags <230 aa (79).

A small but significant percent of the structures (11.7%) have Ag CRS >229 aa. The 442-aa CRS in structure 3TT3 is the longest in the data set, whereas the 12-aa CRS in structure 2J88 is the shortest. The 12-aa CRS containing the 2J88 epitope contains 10 Ag CRs. In this respect, the 2J88 epitope is the most like a small, contiguous peptide of all 111 structures. Although 2J88 is the only structure where the number of CR ≈ CRS, the majority of the structures do indeed have epitope CRs that are clustered together in short, contiguous stretches ranging from 2 to 12 aa, with the most frequent longest contiguous sequence of CR being 5 (Fig. 4). These findings are in good agreement with previous work (14). The IFSK structure is an example of an epitope containing a long, contiguous sequence of 12 CRs (i.e., a linear peptide); however, the complete epitope contains a total of 17 CR spread over a CRS of 56 residues. Thus, Ag epitopes are discontinuous individual and peptide CRs contained within protein regions ranging from ~20 to 400 aa.

The size and distribution of native mAb neutralization epitopes is illustrated in Fig. 5. The number of CRs for 9 HIV-1 gp120 neutralizing Abs range from 13 to 39 and are contained within long CRS from 194 to 384 aa. Many of these mAbs contact the same amino acid positions and the SUM line is a “heat map” indicating the number of different neutralizing mAbs contacting residues at each position. HIV-1 gp120 residues 281, 365, 366, 367, 368, 419, 421, and 473 are contacted by five or six of the mAbs. Together, these 8 CRs represent the core of an epitope recognized by multiple broadly neutralizing mAbs spread over a linear amino acid sequence of 193 residues.

Although each CR epitope is unique, the number of CRs and pattern of sequence positions may be useful for grouping neutralization epitopes into classes. Structures 2NY1 and 3JWD have relatively few CRs located at almost identical positions and contained within a 319-aa CRS. Similarly, structures 3NGB, 3SE9, 3SE8, and...
NIH45-46 all have large numbers of CRs spread over very long CRS and located at very similar sequence positions. All of the HIV-1 Abs examined in this study bind and neutralize structurally intact infectious virus illustrating the general phenomenon that functional Abs contact Ag residues in highly conformationally dependant structures consisting of long CRS. The identification of CRs recognized by multiple neutralizing mAbs illustrates the utility of applying a uniform analysis to multiple structures and lends credence to the use of CR defined as a simple distance between amino acids as a means for identifying residues of viral pathogens that may be critical determinants of vaccines.

Ab CR

The distribution of the number of CRs present within the H and L chains of each Ab structure is shown in Fig. 6. Overall, 65% of the Ab CRs identified here are contained within the H chain. The frequency of CRs are distributed around 5–7 per L chain and 11–13 per H chain. The minimum Ig structure retaining both IgH and IgL variable domains (Fv) is ∼230 aa (∼110 aa V L and ∼120 aa V H).

Ab CR sequence positions

The sequence positions of the Ab CR are depicted in Fig. 9. All PDB Ab sequences were renumbered according to the Kabat convention using Abysis. Abysis did not number structures 3SE8.

The ratio of IgH to IgL CRs for each of the 111 structures is depicted in Fig. 7. Three of the 111 structures have no CRs within the L chain (1Z3G, 29L, and 3GBN). The overwhelming majority of the structures (92) have more IgH CR than IgL; however, 11 Ab structures have more IgL CRs than IgH. The structure having the highest ratio in the data set is 3EOA (13 IgH and 1 IgL), and the lowest is 3HB3 (5 IgH and 10 IgL).

A significant observation from these studies is the degree of variability in the number and ratio of CRs between individual structures within an Ag–Ab complex. Although all the data taken together suggest a model of Ag–Ab interaction involving equal numbers of Ab and Ag CR, this is the case in only 16 of the 111 structures analyzed. Similarly, the composite data indicate that 65% of Ab CRs reside within H chains, but the IgH to IgL CR ratio within an individual Ab is highly variable ranging from no IgL CRs to twice as many L chain CRs as H chain. It is commonly reported that the majority of Ab CRs reside within IgH, and the analyses reported in this article support these observations; however, there are 11 examples in this data set of Abs having more CRs in IgL than IgH.

Although the large number of structures analyzed in this study does indeed enable construction of useful models of Ag–Ab interaction, the significant variability within and between structures limits the ability to predict individual CRs based on these composite data.

Ab CRS

As with Ag (Fig. 3), Ab CRs, comprising the Ab paratope, are distributed along the IgH and IgL chains in a noncontiguous manner. The CRs was determined for all 111 IgL and IgH structures, and the frequency was plotted in Fig. 8. The IgL CRs are typically contained within a CRS of 60–69 aa, whereas the IgH CRS is typically 70–79 residues. Abs make contact with Ag at three discrete regions along each IgH and IgL chain known as CDR, and the predominant CRS depicted in Fig. 8 represents Ig chains having CR in both the first and third CDRs. Eleven IgH structures have CRS ≥90 aa (e.g., 3VG9). In all 11 of these structures, there is a CR in amino acid sequence position 1 or 2, which accounts for the extraordinary length of these CRSs. Similarly, there are 3 IgL structures where amino acid sequence position 1 or 2 is a CR resulting in CRS of 90–99 aa (e.g., 3SE8).

Ab CR sequence positions

The sequence positions of the Ab CR are depicted in Fig. 9. All PDB Ab sequences were renumbered according to the Kabat convention using Abysis. Abysis did not number structures 3SE8.
and 3T2N, and these were omitted from further analysis. The IgH sequence positions that are most frequently CRs are 100, 52, 31, 97, and 98. The most frequent IgL CR sequence positions are 32, 92, 91, 30, and 50. IgL positions 32, 92, and 91 are in agreement with previous findings (9). In this study, we add as predominant CR the IgL positions 30 and 50 and the IgH positions above.

Further inspection of the data reveals three prominent regions of CR for both IgH and IgL chains. These CRRs, defined by distance between Ag and Ab residues, are organized in a manner similar to established CDR defined by Ab gene sequence variability (23). Although the great majority of the total number of Ab CRs identified in this study lie within these prominent regions, a small but significant percentage of CRs have sequence positions that fall outside of these regions.

A more detailed analysis of Ab CRR is presented in Fig. 10. The upper panel is an analysis of IgL CRR and the lower panel is the same analysis for IgH. No CRs were identified beyond IgL sequence position 96 and IgH sequence position 103. The heat maps highlight sequence positions where multiple structures have CR. Sixty-three IgL and 59 IgH amino acid sequence positions, representing ~54% of all positions within an Ab V region fragment (Fv = 230 aa), are CRs in at least 1 of the 111 structures analyzed. Many of these sequence positions, however, are CRs in only a small number of structures.

CRRs are defined in this article as contiguous sequences where >10% of the structures have CR. With these boundaries, 91.6% of all IgL and 93.7% of all IgH CRs lie within the indicated CRR. IgH CR sequence positions 1–2 and 73–79 lie outside of defined CDR/CRR. Eleven structures have CRs within IgH region 73–79 (e.g., 1S78) and 11 within region 1–2 (see also Fig. 9). Region 3 contains more CR than regions 1 and 2 for both IgL and IgH. The third region of IgH is the largest, containing 24.6% of all Ab CR. There are many Ab structures in this data set that do not have CR in each of the six CRRs. There are 7 IgH (e.g., 2QQN) and 64 IgL structures that do not make contact in all 3 CRRs. The sequence positions of all CRs in relation to CDR and CRR for all 111 structures are shown in Supplemental Figs. 2 (IgL) and 3 (IgH).

In individual Abs, CRs generally do not occur as long, contiguous sequences, but are scattered in a discontinuous manner within and across regions. It is interesting to note that in IgH and IgL CDR2/CRR2, there are large numbers of structures having CRs at positions 50 and 52, but only a small number at position 51. Similarly, IgH position 29 is included within CRR1 even though only 1 Ab of 111 contacts Ag at that position because the positions on either side are significantly above the threshold. These observations illustrate the discontinuous nature of CR within CRR and CDR of individual Abs. The fact that IgH positions 29 and 51 and IgL position 51 are in-
cluded within CDR/CRR and yet are only infrequently CRs suggests the possibility of a specific role for these positions apart from directly binding Ag, perhaps in properly positioning actual CRs.

**Comparison of CRR with CDR**

A comparison of the boundaries, number of sequence positions, and CR in each CRR with the corresponding CDR is presented in Fig. 11. Although CRR and CDR significantly overlap, the boundaries are different. The differences between CDR and CRR boundaries are greater for IgH than for IgL. There are more total sequence positions within CDR than CRR. IgH region 3 contains the largest number of CRs for both CDR and CRR, followed by IgH region 2. IgL region 2 contains the fewest CRs for both CDR and CRR. Using the Kabat definitions (discussed in Ref. 25), 13.0% of all Ab CRs lie outside of established CDR boundaries and 7.0% outside of CRR. Recently, Kunik et al. (20) reported ∼20% of the Ab residues that actually bind the Ag fall outside the CDR. Fig. 12 illustrates the overlap between CDR and CRR. The greatest difference between the two is the inclusion of IgH residues 26–30 in CRR1 compared with CDR1. Another major difference between the two is inclusion of IgL position 49 in CRR2 compared with CDR2. The CRRs defined in this article contain a larger number of CR and are more uniform in number of sequence positions in a region, and yet comprise fewer total sequence positions compared with CDR; that is, they are more densely populated with CR than CDR.

The analyses presented in this article demonstrate a significant overlap between CRR and established CDR defined using completely orthogonal methodologies. Indeed, the CRR boundaries reported in this article are even more closely aligned with refinements of CDR reported by others (26). CRRs are defined by measuring physical distance between atoms within crystal structures, whereas CDRs are defined by Ig gene sequence variability. The overlap further confirms the validity of both methodologies. Although the two approaches point to similar structural architecture, there are important differences as well (Fig. 12, Supplemental Figs. 2 and 3). Ab engineers relying solely on definitions of CDR may not give adequate consideration to sequence positions that are actually important CR. Two examples are IgL sequence position 49 and IgH position 30. These positions lie outside established CDR and yet frequently occur as CR. Also, IgL positions 24, 25, 26, 90, and 97 and IgH positions 61 and 63 do not contact Ag in this data set but are included within CDR. These observations point out the advantage of using multiple methodologies to characterize Ab binding residues and shed new light on the organization of regions of Abs harboring important CR.

**Discussion**

These analyses were undertaken to better understand the nature of Ab binding to large protein Ags, and in so doing provide insight to strategies for developing high-performance, high-value functional and therapeutic Abs. Abs are used in a variety of ways, and the structure of an Ag is influenced by the assay or application in which the Ab is used (4). In the Western blot, the Ag is typically boiled in detergent and reducing agent, and the Ab must be capable of recognizing the denatured form of the Ag. In sandwich immunoassay methods using minimally processed patient serum or plasma samples, the Ab must be able to bind to the properly folded native Ag structure. Therapeutic Abs must bind to native protein structures as they exist in the patient. Functional Abs must bind to

---

**FIGURE 10.** Analysis of IgL (upper) and IgH (lower) CRRs. Heat maps for IgL and IgH were constructed by counting the number of occurrences of each position in all structures and applying a color scale (red > yellow > green) to the distribution. Sequence positions were numbered using Abysis and the Kabat convention. CRRs are highlighted in gray.

**FIGURE 11.** Comparison of established CDR (Kabat) and CRR. The total number of CRs in the data set for each group is in parentheses.
native protein structures at specific sites in a way that influences protein function.

In many applications, it is of greater interest to predict the sites on an Ag that will result in functional modulation or therapeutic effect than it is to predict sites that will elicit the production of Abs in an immunized animal. Indeed, the site of interest may not be immuno-dominant or even immunogenic, and yet an Ab to the epitope is highly desired. B cell epitope prediction algorithms are of little value in such instances, and thus strategies for developing Abs to specific sites of native proteins regardless of their inherent immunogenicity are critical.

An important aspect of these analyses is that the majority of the Abs are described in functional and therapeutic terms providing information about this important class of molecules. Another important aspect of this work is that all of the Ags consist of ≃85 aa. Structures with large Ags were selected in an attempt to reveal the full extent of Ab interaction with large, natively folded proteins. In this data set, there is a significantly higher average number of CRs for large Ag structures compared with smaller Ags, and the epitope CRS increases proportionally to the size of Ag in the crystal. Given the earlier considerations and the fact that 92% of the 111 Ags analyzed in this study are less than full-length, it seems likely that the earlier considerations and the fact that 92% of the 111 Ags analyzed in this study are less than full-length, it seems likely that the values reported in this article for CRS, and to a lesser extent CR, may be underestimates of actual values, and previously reported estimates of CR based on even smaller protein fragments are lower still.

In this study, we identify Ag and Ab CRs in publicly available x-ray crystal structures defined as amino acids within a distance of 4 Å. Other researchers have performed similar analyses using this methodology (10, 14, 15); however, prior studies have been more limited in scope and have not focused exclusively on structures comprising large protein Ags. In addition, the length of the linear sequence containing all CR, the CRS, has not been previously reported. The large number of structures analyzed in this article highlights the extent of variability associated with the number of CRs and their relative ratios, the length of CRS, and the organization of CR within prominent Ab CRRs. The large number of structures also highlights specific differences between CDR and CRR. The identification of viral Ag sequence positions contacted by multiple neutralizing Abs, and the organization of Ab CR into CRR similar to established CDR lend credence to distance between CR as a useful means of characterizing Ag–Ab interactions.

The observation that the distributions of the numbers of Ag and Ab CR significantly overlap and are both centered around ≃18 or 19 CRs per structure suggests that, in general, Abs contact regions on large protein Ags that are similar in size to their own physical dimensions. Thus, one characteristic of an Ag epitope is that the two-dimensional surface area containing all CRs is less than or equal to the physical dimensions of the binding region of an Ab. Davies et al. (11) estimated these dimensions from studies of lysozyme–Ab crystal structures as an oval with the approximate dimensions of 20 × 30 Å. These dimensions represent the “footprint” of Ab paratopes and large protein Ag epitopes.

A unique aspect of this work is the characterization of both the Ag and Ab CRS. The analysis of many Ag–Ab complexes indicates that CRs are contained within relatively long CRSs that come into close spatial proximity because of complex folding inherent within native protein structures. This is true of both Ag and Ab proteins. In the case of Abs, the CRSs are highly uniform in length (60–69 for IgL and 70–79 for IgH), and CRs are organized into prominent CRRs. Ag CR organization and CRS are more variable; however, the most common CRS class, 50–79 residues, is very similar in size to IgL and IgH CRS. Given that the function of all Abs is to bind, it seems reasonable that Ab CRs are organized into uniform CRRs and distributed across CRSs that are highly homogeneous in length. In contrast, epitopes reside on Ags that have a variety of different functions, so it is perhaps not surprising that Ag CR and CRS are more variable.

The fact that the most frequent Ag, IgH, and IgL CRs are very similar in size suggests a common underlying structural organization. In addition, the similarity in size between the large majority of Ag CRSs (20–229 residues), as well as the minimum Ab structure retaining complete binding (Fv ≃230 aa), further suggests common structural elements.

The basic structural unit within a protein, Ag and Ab, is a domain. Domains fold independently, mediate specific biological function, and combine to form larger modular proteins. The smallest domains are ≃35 aa and the largest can be several hundred (27). The size of the most frequently occurring protein domain is ≃100 aa (28), very similar in size to the average (125 aa) and median (95 aa) CRs of the 111 Ag structures analyzed in this article. In addition, the size distribution of complete protein domains (28, 29) is strikingly similar to the distribution of Ag CRSs presented in Fig. 3. Together, these observations suggest that complete structural epitopes, represented by CRS, are the size of intact protein domains.

It is widely recognized that Ab framework regions are required for proper paratope structure and Ab binding, and complete functional paratopes are formed only when CDR and framework regions are structurally organized into VH (≃120 aa) and VL (≃110 aa) domains that together form a functional binding unit (Fv). In a similar fashion, functional protein Ags are made up of “frameworks” of domains (30), and functional epitope conformation is determined by the underlying domain structure. When viewed in composite, the observations that the number of Ag CRs is equal to the number of Ab CRs, Ag CRs are similar in size to IgH and IgL CRSs, and the size of the majority of Ag CRs is less than or equal to Fv suggest that Abs recognize structural epitopes that are similar in size to VL, VH, and Fv, that is, one to two protein domains.

Large protein epitopes are a collection of not only CRs, but many more residues that are necessary to present a native structure that may be as large as, and larger than, the physical footprint of an Ab. Attempting to elicit or select Abs that bind full-length, native proteins with peptides and polypeptides smaller than an entire epitope not only risks non-native Ag folding, but Abs elicited in this way, although capable of binding native structure, may still not bind to the full-length protein because of steric hindrance imposed by Ag residues within the footprint of the Ab but not included in the immunogen.

In the analysis of the functional, neutralizing epitopes of HIV-1 presented in this article, the most broadly neutralizing epitopes are composed of structures of ≃380 aa. Immunization with smaller polypeptides would likely result in Abs with less CRs and, arguably, less
affinity or perhaps less broadly neutralizing activity. In instances where Abs bind to CR on multiple domains, the size of an immunogen required to elicit such an Ab would be expected to be larger still, to encompass all CRs and all supporting framework amino acids required for maximum contact.

A great deal of effort is expended attempting to predict protein epitopes. Zhang and coworkers (31) recently reported an improved method of predicting conformational epitopes based on a “thick surface patch” model including an “adjacent residue distance” feature. This concept of an epitope includes the idea of CRs layered on top of supporting amino acids but fails to appreciate that the structure of the immediately underlying amino acids is, in turn, dependent on even more amino acids comprising the complete protein domain. Currently, B cell epitope prediction is largely focused on physicochemical attributes of a protein Ag recognized by Abs such as surface exposure, solvent accessibility, spatial configuration, and so on. It is unlikely, however, that such characteristics will fully account for issues of immunodominance and tolerance that influence Ab development and involve other components of the immune system such as T cell help.

When attempting to develop functional and therapeutic Abs, it is less important to identify the sites on a protein that may elicit large amounts of Ab and is more important to direct Ab development to sites where binding would modulate biological function, regardless of the site’s inherent immunogenicity. Given that there are a limited number of immunodominant sites on a protein, immunization with full-length proteins, although ensuring proper conformation and maximizing the potential interaction between Ag and Ab, may not elicit Abs directed to the biologically relevant site. Immunization with peptides and polypeptides representing less than a complete structural epitope has proved time and again to elicit Abs that may be directed to the site of interest but either do not bind to native protein or bind in a way that does not modulate function. To develop functional and therapeutic Abs, it is critical to focus Ab development to potentially weakly immunogenic, biologically relevant sites in a way that the resulting Abs bind to full-length native proteins and effect a desired function.

The analyses presented in this article demonstrate that functional and therapeutic Abs bind structures composed of linear strings of contiguous amino acids that are the size of protein domains. As defined by the CRS, a structural epitope is significantly larger than previous definitions of epitopes that are focused primarily on CR.

A structural epitope includes CRs as well as the minimum number of contiguous amino acids required to yield a compact native structure that is independent of other protein regions or even other molecules. Thus, immunization with independent protein domains representing complete structural epitopes is indicated for development of functional and therapeutic mAbs targeted to biologically relevant and potentially weakly immunogenic sites.

Acknowledgments
We thank Dr. Ross Chambers for insight and review of the manuscript and Dr. Fenglin Yin for review of the database and Ab sequence numbering.

Disclosures
The authors have no financial conflicts of interest.

References