Automated Identification of Complementarity Determining Regions (CDRs) Reveals Peculiar Characteristics of CDRs and B Cell Epitopes

Yanay Ofran, Avner Schlessinger and Burkhard Rost

J Immunol 2008; 181:6230-6235; doi: 10.4049/jimmunol.181.9.6230
http://www.jimmunol.org/content/181/9/6230

References This article cites 34 articles, 4 of which you can access for free at: http://www.jimmunol.org/content/181/9/6230.full#ref-list-1

Why The JI? Submit online.
• Rapid Reviews! 30 days* from submission to initial decision
• No Triage! Every submission reviewed by practicing scientists
• Fast 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

The Journal of Immunology is published twice each month by The American Association of Immunologists, Inc., 1451 Rockville Pike, Suite 650, Rockville, MD 20852

Copyright © 2008 by The American Association of Immunologists All rights reserved.

Print ISSN: 0022-1767 Online ISSN: 1550-6606.
Automated Identification of Complementarity Determining Regions (CDRs) Reveals Peculiar Characteristics of CDRs and B Cell Epitopes

Yanay Ofran, Avner Schlessinger, and Burkhard Rost

Exact identification of complementarity determining regions (CDRs) is crucial for understanding and manipulating antigenic interactions. One way to do this is by marking residues on the antibody that interact with B cell epitopes on the antigen. This, of course, requires identification of B cell epitopes, which could be done by marking residues on the antigen that bind to CDRs, thus requiring identification of CDRs. To circumvent this vicious circle, existing tools for identifying CDRs are based on sequence analysis or general biophysical principles. Often, these tools, which are based on partial data, fail to agree on the boundaries of the CDRs. Herein we present an automated procedure for identifying CDRs and B cell epitopes using consensus structural regions that interact with the antigens in all known antibody-protein complexes. Consequently, we provide the first comprehensive analysis of all CDR-epitope complexes of known three-dimensional structure. The CDRs we identify only partially overlap with the regions suggested by existing methods. We found that the general physicochemical properties of both CDRs and B cell epitopes are rather peculiar. In particular, only four amino acids account for most of the sequence of CDRs, and several types of amino acids almost never appear in them. The secondary structure content and the conservation of B cell epitopes are found to be different than previously thought. These characteristics of CDRs and epitopes may be instrumental in choosing which residues to mutate in experimental search for epitopes. They may also assist in computational design of antibodies and in predicting B cell epitopes.


Interactions between Abs and other proteins involve several mechanisms of molecular recognition and protein-protein interaction (1–5). The specific immunological response to non-self molecular challenges is conducted by Abs that bind specifically to the antigenic molecule. The specificity of an Ab to its target is determined by the complementarity determining regions (CDRs), a set of relatively small structural elements, three on each of the two chains. The overall structure of each of the Ab chains is a skeleton of β-sheets with six loops that constitute the CDRs. The main differences between the sequences of Abs of a given individual are in the CDRs. Apart from their ability to recognize and bind to Ags through the CDRs, Abs can use other patches on their surface to form interactions with proteins and other molecules. However, these interactions should be considered “nonantigenic” as they do not involve the CDRs and hence do not require the specific identification of an antigenic epitope. While some studies use antigenic interaction as a model for understanding the general phenomenon of molecular recognition (1–5), it is commonly assumed that the mechanisms that govern general protein-protein interactions are different than those that govern antigenic interactions (5, 6). Thus, to study antigenic interactions, including the specific recognition of antigenic sites, one needs to sieve the nonantigenic contacts from the antigenic ones. This could not be done without precise identification of the CDRs.

The term CDR is sometimes used interchangeably with hypervariable regions or segments assembled to form complete variable (V)-region genes. However, some studies have used it to describe the residues that participate in binding. We use the term to refer to those elements on the Ab that actively participate in the antigenic interaction. We acknowledge that this may not be the way that many people in the field use it. Hereafter, we use the term CDR to refer to the residues in the antigenic interface, namely the residues on the Ab that contact the epitopes.

The first systematic attempts to identify the CDRs have been based on analysis of the amino acid sequence of Abs. The underlying assumption of this approach was that CDRs are the most variable residues in the Ab sequence (7, 8), and therefore they could be identified through the analysis and comparison of the sequences of various Abs. The Kabat variability plot (7) offers a consensus sequence motif that could be aligned with sequences of new Abs to identify their CDRs. CDRs often have very unusual sequences that typically lack sequence motifs, consensus, profiles, or other similar elements (9) and, therefore, they are hard to align. Thus, any method based only on sequence information is prone to misaligning, and therefore misidentifying, loopy CDRs. Chothia and colleagues, therefore, based their CDR identification on
structural information (10, 11). Analyzing a small number of experimentally solved three-dimensional structures of Abs, they defined the hypervariable loops in terms of their secondary structure and the sequence position where each of them starts and ends. However, as more experimentally solved structures became available, the analysis was run again and the boundaries of the CDRs were changed (10, 12). These new definitions were still based mostly on manual analysis of structures and are expected to be changed with the ever-growing availability of experimentally determined three-dimensional structures of Abs. The different definitions of CDRs are also based on different definitions of secondary structures, thereby increasing the inconsistency in defining hypervariable loops. To overcome this problem, the contact definition of CDRs (13) uses complexes of Ag and Abs to identify the residues that bind the Ag. A pitfall of this approach may be the fact that Abs can create nonantigenic interaction with proteins. Thus, if the interaction between the Ab and a protein is not part of a specific antigenic response, the residues that bind the protein may not be part of the CDR. Additional analysis of the advantages and disadvantages of sequence- and structure-based methods is available elsewhere (www.bioinf.org.uk/abs/).

Several studies have used the increasing number of protein-Abs in the Protein Data Bank (PDB) (14) to characterize antigenic interactions (1, 5, 6, 12, 13, 15–17). Based on the current definitions of CDRs, these studies have attempted to conclude the general properties of CDRs and the general characteristics of antigenic interactions. Two recent studies have suggested that the amino acid composition and the length of CDRs determine the type of Ag it can bind (6, 15).

Identifying the CDRs on the Ab side of the interaction is, naturally, only half of the story of the antigenic interaction. The properties of epitopes, namely the structural elements on the Ag to which the CDR bind, could be studied only once the CDRs are identified. Several studies have taken this approach to characterize the epitopes on the antigenic sites (5, 16, 17). The results of these studies were rather inconsistent. Among the possible reasons for these inconsistencies are the small datasets used by some studies and the significant differences in the methodologies. Most importantly, the definitions of the CDRs often differ greatly; that is, if two studies investigate the same PDB complex and use the same methodology, they might disagree on which of the interactions are antigenic (9, 16).

Chothia et al. (10) suggested that some of the residues within CDRs may not create contacts with the Ag but are important to maintain the “canonical backbone conformation” of the Ab. MacCallum and coworkers observed that the hypervariable loops of CDRs only adopt a limited number of backbone conformations that are determined by a few key residues (13). Analyzing several antibody-Ag complexes, they defined different types of Ags (haptens, peptides, proteins) and the typical positions of each CDR to which they bind.

In essence, every attempt to define CDRs is based on comparing the common denominators of different Abs. The amount of structural data for Abs available today does not allow for careful, manual analysis. To benefit from all available Ab information we developed a procedure for the identification of CDRs using all available protein-Ab complexes. As described in Fig. 1, we structurally align all available Abs (separating heavy and light chains) to each other. Then, we mark the positions that are in contact with the Ag in each Ab structure. Next, we search for structurally aligned residues that create contacts with the Ag in most known complexes. By thus defining CDRs, the procedure also identifies the epitopes on the antigenic protein in the complex, allowing for large-scale analysis of B cell epitopes.

Materials and Methods

Extraction of 3D structures and Ab-Ag contacts

The overall scheme for identifying the CDRs and the epitopes is described in Fig. 1. To identify all structures in the PDB that contain at least one Ab-Ag complex, we searched with BLAST (18) for a consensus sequence of an Ab against the PDB. The rationale for using BLAST rather than PSI-BLAST was to avoid capturing molecules such as T cell receptors that, despite their similarity to Abs, participate in cell-mediated immune response and therefore represent a different type of antigenic interaction. We then added to the database PDB structures that contain an Ig fold from the Structural Classification of Proteins (SCOP) database (19), and using PDB entries containing keywords (e.g., “antibody” and “Ag”) identifying them as Ab-Ag complexes. Then, we manually discarded all complexes with T cell receptors or MHC molecules, as these are formed during cell-mediated immune response and are not relevant to the task at hand. In each complex, we labeled residues as interacting if any of their respective atoms were within 6 Å of one another (20). We thus created a list of interactions between Abs and Ag.

Structural alignment

We located the CDRs in the known protein-Ab complexes through the following knowledge-based approach. We began by creating multiple structure alignment of Ab structures using SKA (21, 22). Since the light and heavy chains have different CDRs, we conducted two different multiple structure alignments, one for all heavy chains and one for all light chains. Additionally, since our dataset included several redundant sequences, we ran the structural alignment program on a sequence-unique subset of all protein-Ab complexes. As Ab sequences are highly similar to each other, the criteria for the redundancy of the complex set was determined by the Ag sequences; sequence redundancy was reduced at HSSP-values of 0 (corresponding to <33% pairwise sequence identity for long alignments) (23, 24). Then, we identified structurally aligned positions that interact with a protein in >10% of the complexes of the alignment. The choice of the cutoff was based on analysis of the number of contacting positions that were structurally aligned in the non-CDR part of the Abs. These highly populated positions were defined by us as the boundaries of the CDRs.

After we identified the CDRs in the aligned Abs we transferred the location of those CDRs to the Ab chains of the family that they represent by structural pairwise alignment using combinatorial extension (25) (Fig. 1). Finally, we defined all the residues on the protein surface that are in contact with the residues on the Ab CDRs as antigenic residues.

Content statistics

Overall, our analysis was based on 140 Abs from protein-Ab complex structures with a current total of 10,180 interactions. From this set we generated a sequence-unique set of Ags, that is, a list of Ags such that Ag in the dataset has a level of sequence similarity that would enable coarse-grained homology modeling of another Ag in the dataset. All the data are available at the Epitome database (26).

Secondary structure, solvent accessibility, and protein interfaces

Ag residue secondary structure state and solvent accessibility were computed using DSSP (Dictionary of Secondary Structure of Proteins) (27) (one-letter code; G, H, and I correspond to helical structures, E and B to strands, and T, S, and L to other). The protein interfaces were extracted from PDB as described elsewhere (20).

Results

Using the alignments of the Ab chains we found, both on the heavy and on the light chains, three sequence stretches that contact the Ag in all structures in the alignments. Individual positions in these stretches were in contact with the Ag in up to 98% of the aligned Abs. This figure was higher for positions closer to the center of the stretch and lower toward the boundaries. We noticed that there was no position outside of these stretches that created contacts with the Ag in >7% of the Abs. Therefore, to set a stringent definition of boundaries of the CDRs, we defined the first and last position of a CDR as the position where <10% of the aligned Abs created contacts with the Ag.

Our automatically identified CDRs are typically only in partial agreement with other definitions of CDRs. Looking at all available protein-Ab complexes we were able to characterize...
the amino acid composition of CDRs. All experimental datasets, particularly those derived from PDB, are biased in more than one way: PDB is biased toward proteins that could be expressed, purified, crystallized, phased, and so forth. It is also biased by the interest of the researchers and funding agencies and by the simplicity and cost of certain proteins and systems. In the case of Abs, this bias may include biases toward certain species (it is easier to obtain rodent Abs than human ones). These biases should be kept in mind when looking at the result of any bioinformatics analysis. However, by pooling all the data together, reducing sequence homology, and focusing only on the CDRs and epitopes, we were hoping to identify signals that are stronger than any specific bias.

To the best of our knowledge, this is the first study that offers a comprehensive description of CDRs based on all available structural data. Earlier studies have suggested that the amino acid composition of CDRs is peculiar (6). We found that these peculiarities are more striking than previously thought; that is, four amino acids account for 55% of the CDR residues: tyrosine (24%), serine (12%), asparagine (10%), and tryptophan (9%). Ten other amino acids account for <10% of the CDR (C, M, Q, K, A, P, V, E, H, L). Fig. 2 shows under- and overrepresentation of all types of residues in CDRs (with respect to their frequencies in proteins in general). Almost all residues are underrepresented in all CDRs. There are only two residues that are present in all known CDRs: tryptophan and tyrosine. Interestingly, glycine, histidine, and aspartic acid show different preferences in CDRs on the L chain and in CDRs on the H chain. This is most obvious in the case of aspartic acid, which is substantially underrepresented in all the CDRs on the L chain but is overrepresented in all the CDRs on the H chain. This pattern is very different than that of glutamic acid (universally underrepresented) or asparagine (mostly overrepresented). Phenylalanine is underrepresented in CDRs 1 and 2 on both types of chains but is overrepresented in CDR 3 on both chains. The reverse is true for arginine.

The meticulous identification of CDRs led, naturally, to a better identification of the antigenic epitopes on the antigenic side of the interaction. Thus, we were able to analyze the general characteristics of B cell epitopes. Fig. 3 shows under- and overrepresentation of all types of residues in epitopes found in a nonredundant set of antigenic proteins in PDB. The results are compared with under- and overrepresentation of residues in all protein-protein interfaces. Epitopes are not a simple subset of all protein-protein interfaces; that is, other than histidine, no residue shows similar behavior in both types of interfaces. Aliphatic-hydrophobic residues (A, I, L, V) are strongly underrepresented. However, glycine, which is aliphatic but not hydrophobic, and some

---

**Figure 1.** Structure-based identification of antigenic contacts. A, We automatically extracted all complex structures containing immunoglobulins and other proteins from the PDB. The contacts between the immunoglobulins and the antigenic proteins may be antigenic (i.e., involving the CDRs) or nonantigenic. B, After manually removing complexes with T cell receptors we performed multiple structure alignments of Ab chains from a unique set of complexes. C, We then mapped the interacting residues of the corresponding epitope residues. A residue in one of the Abs that was in contact with the Ag in the original PDB complex is marked in this illustration with a filled circle. D, We searched for aligned positions that systematically create contacts with the Ag in our dataset (light blue filled circles) and disregarded positions that contact the Ag only sporadically (red filled circles). Looking at a sequence representation of the structural alignment, we found sequence stretches that are in contact with the Ag in all known complexes. We defined those as the CDRs. We marked the boundaries of the CDRs as the positions that have at least 10% of their residues interacting with an Ab chain (see Results). E, Lastly, we marked the locations of the CDRs on all Abs in PDB.
residues that are hydrophobic but not aliphatic (C, M, P, W) are overrepresented. Radical differences between the composition of protein-protein interfaces and epitopes are observed in several residues: proline is highly overrepresented, probably due to its tendency to break secondary structure elements (note that Ags are highly enriched with loop regions). Tyrosine is overrepresented in protein-protein interfaces but is underrepresented in epitopes, while the reverse holds for cysteine.

Fig. 4 shows the secondary structure composition (i.e., what fraction of the residues are in each secondary structure state) of the epitope residues. We compare the secondary structure composition of epitopes to that of: 1) the general population of protein-protein interfaces, 2) PDB in general, 3) residues that are buried in the core of known structures, and 4) residues that are exposed to the solvent. It is clear that epitopes are very different from any of these sets of residues. It has been claimed before that the intrinsic flexibility of surface loops plays an important role in the ability of Abs to recognize and bind them (17). Indeed, we see that residues that are neither in helices nor in strands constitute the largest group of secondary structure states in epitopes. However, this enrichment (52%) is only marginally higher than their enrichment in surface residues (49%).

The most dramatic difference in this comparison is in the underrepresentation of helices and, although to a lesser extent, the overrepresentation of strands in epitopes. Overall, it is possible to generalize and say that with respect to the exposed residues, CDRs prefer to bind strands and shun helices.

On average, residues involved in protein-protein interactions are slightly more conserved than all other surface residues (28, 29). However, as shown in Fig. 5, we found that this is not the case for antigenic residues. Surface antigenic residues are less conserved than other surface residues. This may be due to the fact that the immune system avoids determinants that resemble self-proteins, and this is more likely to bind to the nonconserved elements of the Ag. In contrast, sometimes identifying conserved epitopes could be an advantage if the epitope is conserved among different pathogens but have no homolog in the host. For example, Jun al is a protein found in mountain cedar pollen (30). When humans are infected by this protein an allergic reaction occurs and Abs bind the protein. Interestingly, two-thirds of the epitopes of this protein are conserved throughout the family and were also found to be functional in its homolog, Cry j. In this example, the immune system is efficient: by binding a conserved epitope it can be triggered when any of the proteins of this family penetrates the organism (30).
a large-scale, continuous analysis of all known CDRs and all known epitopes. A database in which all of these data are curated is described elsewhere (26).

To our knowledge, this is the first method for objective and fully automated identification of CDRs. It relies on protein structure and not on sequence considerations. The results presented herein are the most comprehensive large-scale quantification of the peculiarities of both CDRs (e.g., the fact that only four amino acids account for most of the CDRs) and B cell epitopes (e.g., their peculiar secondary structures), thus providing a firm ground for the distinction between CDRs/epitopes and other molecular interfaces. This knowledge has vast potential applications. It could be useful when one attempts to identify the epitopes used by a specific Ab. It provides the basis for selecting residues to mutate in mutation analysis of epitopes. The distribution of amino acids in germline sequences encoding the CDRs may be different than that of other genomic sequences. This may account for part of the peculiarities we report herein. However, the amino acids we identify as rare can be found in the germline, but in reading frames that are typically selected against (34). Therefore, to fully understand the unusual composition of CDRs, one needs to understand the selective pressure against them. Attempts have been made to force use of alternative amino acids in CDRs, and immune responses appear to be altered (34).

Furthermore, this dataset could serve as the basis for a computational method for predicting putative epitopes in a given protein, a task that has been shown to be one of the hardest challenges of molecular bioinformatics (35). Finally, the distinctions we made herein may serve as the basis for tools that will improve the design of Abs and epitopes.

Acknowledgments
Thanks to Jinfeng Liu and Guy Yachdav (Columbia University) for computer assistance, and to Guy Nimrod (Tel Aviv University) for helpful comments on the manuscript.

Disclosures
The authors have no financial conflicts of interest.

References


