THE PROTEIN SOCIETY

The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops

Yuko Tsuchiya[1]* and Kenji Mizuguchi[1]

[1]National Institutes of Biomedical Innovation, Health and Nutrition, 7-6-8 Saito-Asagi, Ibaraki, Osaka, 567-0085, Japan

Received 14 October 2015; Accepted 30 December 2015
DOI: 10.1002/pro.2874
Published online 8 January 2016 proteinscience.org

Abstract: Of the complementarity-determining regions (CDRs) of antibodies, H3 loops, with varying amino acid sequences and loop lengths, adopt particularly diverse loop conformations. The diversity of H3 conformations produces an array of antigen recognition patterns involving all the CDRs, in which the residue positions actually in contact with the antigen vary considerably. Therefore, for a deeper understanding of antigen recognition, it is necessary to relate the sequence and structural properties of each residue position in each CDR loop to its ability to bind antigens. In this study, we proposed a new method for characterizing the structural features of the CDR loops and obtained the antigen-binding ability of each residue position in each CDR loop. This analysis led to a simple set of rules for identifying probable antigen-binding residues. We also found that the diversity of H3 loop lengths and conformations affects the antigen-binding tendencies of all the CDR loops.

Keywords: diversity of CDR-H3; diverse conformations of long H3 loops; antigen recognition by antibodies; antigen-binding tendency; hydrogen bond networks

Introduction

The complementary determining regions (CDRs) of antibodies play a key role in antigen recognition. In general, the CDR loops in the heavy chain are more frequently involved in antigen binding than those in the light chain. Of the three heavy chain loops, H3 is considered to be the most important to antigen recognition.[1,2] The contribution of each of the six CDR loops to antigen recognition is different from each other, and even within a single CDR loop, each residue position plays a different role in antigen binding.[3,4] It is necessary, therefore, to characterize the sequence and structural properties of each position in a CDR loop for estimating the utilization of each position in antigen binding and for understanding antigen recognition in more detail.

As mentioned above, the H3 loop is the main contributor to antigen recognition among the six CDR loops, because of its sequence diversity and location favorable to antigen binding.[1,5] The sequence diversity produces diverse conformations, particularly in long H3 loops, and the conformational variety may be required for maintaining antigen specificity and H3's predominant role in antigen binding. For example, the Protein Data Bank (PDB) includes several crystal structures of anti-HIV1 antibodies in complex with envelope glycoproteins. The antibodies with long H3 loops (>=14 residue long) appear to utilize their H3 loops to achieve

*Present address: Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan

Abbreviations: HB, hydrogen bond; PDB, Protein Data Bank; SeqCns, sequence conserved; SeqNoc, sequence non-conserved; StrDef, structurally definable; StrNod, structurally non-definable

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Japan Society for the Promotion of Science (Grants-in-Aid for Scientific Research); Grant number: 25430186 and 25293079; Grant sponsor: Japan Agency for Medical Research and Development ("The adjuvant database project") to K.M.

*Correspondence to: Dr. Yuko Tsuchiya, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan. E-mail: tsuchiya@protein.osaka-u.ac.jp or Dr. Kenji Mizuguchi, 7-6-8 Saito-Asagi, Ibaraki, Osaka 567-0085, Japan. E-mail: kenji@nibiohn.go.jp

Published by Wiley-Blackwell. © 2016 The Protein Society

high specificity and affinity (as observed in PDBID 1g9m). On the other hand, the antibodies with short H3 loops show different antigen-recognition patterns due to completely different structures of their CDRs (as in PDBID 2vxt). Thus, antibodies with different H3 loop lengths show different antigen-binding properties.

Table I. *Sequence- and Structure-Based Characteristics in Each CDR Position*

A) H1 loop (major length = 13 in length 12–15)

Number[a]	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	9	10	N-4	**N-3**	N-2	**N-1**	**N**
AB rate[b]	0.0	0.0	0.02	0.05	0.11	0.18	0.04	0.39	0.25	0.56	0.67	0.40	0.56	0.0	0.08
SC rate[c]	0.85	0.93	1.0	1.0	0.95	0.92	0.96	0.82	—	—	0.67	0.75	0.55	0.85	0.81
AA type[d]	KA	A	S	G	FY	TS	F	ST	—	—		Y		MI	HSN
Hy[e]	0.0	2.4	3.1	5.1	8.0	10.4	9.2	12.4	14.7	16.1	13.4	10.6	9.0	5.4	3.0
Dy[f]	17.4	16.4	18.7	19.1	17.0	16.6	15.0	14.6	13.9	10.7	11.5	9.2	7.7	8.6	6.9
HB[g]	**F**	**F**	**F**						—	—			E	**E**	**E**
Antibody[h]	171	171	171	171	171	171	171	170	16	9	171	171	171	171	171
Del (Ins)[i]								12	12–13	12–14					
Rmsd[j]	0.52	1.00	1.73	2.58	1.40	2.27	1.94	1.90	—	—	1.79	1.60	0.88	0.84	0.67
Chothia[k]	23	24	25										33	34	35
CDR[l]				CI	CI	CI	CI	CI	CI	CI	CIK	CIK	IK	K	K

B) H2 loop (major length = 10 in length 9–12)

Number	1	**2**	3	4	5	6	7	N-4	**N-3**	N-2	**N-1**	N
AB rate	0.39	0.05	0.70	0.33	0.64	1.0	1.0	0.60	0.20	0.63	0.19	0.50
SC rate	0.38	0.92	0.53	0.66	0.49	—	—	0.61	0.92	0.42	0.81	0.50
AA type		I				—	—		G		T	
Hy	5.6	8.0	11.2	11.7	14.4	18.9	17.3	14.8	12.7	12.5	9.2	7.8
Dy	8.8	10.9	10.2	12.6	13.9	11.1	9.8	14.1	15.0	13.2	12.9	11.8
HB	**A**	**E**	**A**						A	**A**		**A**
Antibody	171	171	171	171	134	5	5	171	171	171	171	171
Del (Ins)					9	9 – 10	9 – 10					
Rmsd	1.40	1.70	1.52	2.24	1.39	—	—	3.08	2.08	3.02	3.43	2.57
Chothia	50	51									57	58
CDR	K	IK	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK	IK	K(+7)[m]

C) H3 loop (major length = 12 in length 5–30)

Number	**1**	**2**	3	4	5	6	*7 ~ N-7 (17)*	N-6	N-5	N-4	N-3	**N-2**	**N-1**	**N**
AB rate	0.01	0.11	0.36	0.44	0.68	0.69	0.67[n]	0.62	0.57	0.60	0.17	0.05	0.09	0.05
SC rate	0.94	0.91	0.44	0.42	0.45	0.50	—	0.52	0.47	0.59	0.56	0.88	0.87	0.81
AA type	A	R					—					FM	D	Y
Hy	0.8	4.3	6.8	9.1	11.9	13.8	18.0[n]	13.2	11.0	8.7	5.9	3.4	3.0	1.0
Dy	6.7	7.4	5.0	6.2	5.4	6.0	7.2[n]	5.0	4.1	3.2	4.0	4.2	7.6	9.1
HB	**E**	**A**	E				—					F		**A**
Antibody	171	171	171	162	156	117	72	89	141	159	165	170	171	171
Del (Ins)				5–7	5–9	5–11	(14–30)	5–12	5–10	5–8	5–6	5		
Rmsd	0.49	1.03	1.23	2.18	3.19	4.36	—	5.09	5.05	3.40	2.39	1.90	1.30	0.78
Chothia	93	94	95											102
CDR	I	I	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK

D) L1 loop (major length = 11 in length 9–17)

Number	**1**	**2**	**3**	**4**	**5**	**6**	*7 ~ N-4 (7)*	**N-3**	N-2	**N-1**	N
AB rate	0.0	0.01	0.01	0.12	0.15	0.06	0.41	0.30	0.67	0.01	0.08
SC rate	0.88	0.96	0.97	0.85	0.81	0.84	—	0.78	0.69	0.90	0.71
AA type	R	A	S	Q	SD	VI	—	SNT		LV	A
Hy	0.0	2.4	3.2	6.9	8.9	8.9	12.9	10.0	8.8	5.2	3.0
Dy	17.4	16.3	18.5	18.1	16.9	13.6	13.2	12.3	9.3	9.9	8.4
HB	**F**	**F**	F			F	—		E	**E**	**E**
Antibody	171	171	171	171	171	171	168	154	171	171	171
Del (Ins)							(10–17)	9–10			
Rmsd	0.79	0.80	1.11	1.25	1.83	1.96	—	1.33	2.09	0.64	0.85
Chothia	24	25								33	34
CDR	CK	CK	CK	CIK	CIK	CIK	CIK	CIK	CIK	CK	CK

Table I. *Continued*

E) L2 loop (length=8)								
Number	**1**	2	**3**	**4**	5	**6**	7	**8**
AB rate	0.27	0.35	0.04	0.07	0.30	0.10	0.12	0.15
SC rate	0.89	0.50	0.82	0.94	0.62	0.98	0.56	0.90
AA type	Y		A	S		LR		S
Hy	5.6	8.5	6.4	6.4	7.2	4.4	3.9	3.9
Dy	9.9	10.6	13.4	15.8	14.8	15.6	13.9	16.7
HB	**A**	A	**E**	A	**A**		**F**	
Antibody	171	171	171	171	171	171	171	171
Del (Ins)								
Rmsd	0.90	0.84	1.76	1.14	1.95	0.73	0.76	1.46
Chothia	49				53	54	55	56
CDR		CIK	CIK	CIK	CK	CK	CK	CK

F) L3 loop (major length = 9 in length 5–13)											
Number	**1**	**2**	3	4	5	6	*7 ~ N-4 (3)*	N-3	**N-2**	N-1	**N**
AB rate	0.02	0.01	0.64	0.68	0.54	0.40	0.24	0.53	0.08	0.39	0.04
SC rate	0.87	0.86	0.55	0.47	0.54	—	—	0.45	0.79	0.52	0.95
AA type	Q	Q					—		P		T
Hy	1.0	4.3	7.0	9.4	8.7	10.5	9.7	7.9	5.3	3.8	0.8
Dy	7.1	8.2	6.1	9.7	9.8	12.9	12.8	9.9	9.7	6.9	8.3
HB	**E**	**A**	E				—				**A**
Antibody	171	171	171	164	164	25	14	150	164	171	171
Del (Ins)				5	5	5–9	(11–13)	5–8	5		
Rmsd	0.85	1.05	1.58	1.49	2.51	—	—	3.47	1.34	0.98	1.22
Chothia	89	90									97
CDR	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK	CIK

[a] Position numbering. The N means the length of a given loop. The positions in boldface or italic are StrDef_SeqCns- or StrNod-positions, respectively, where the number of StrNod-positions is indicated in parenthesis. The others are StrDef_SeqNoc positions.
[b] The averaged antigen-binding (AB) rate in a given position in the antibodies indicated in the eighth row.
[c] The averaged sequence-conservation (SC) rate in a given position in the antibodies indicated in the eighth row. The SC rates in the positions to which only a few antibodies belong, were not shown, such as those in positions 9 and 10 in H1, positions 6 and 7 in H2, position 6 in L3 and StrNod-positions.
[d] Amino acids that have an observed frequency >0.2 among the three most abundant amino acids in a given position, shown in order of the frequency.
[e] The averaged Hy value in a given position in the antibodies indicated in the eighth row.
[f] The averaged Dy value in a given position in the antibodies indicated in the eighth row.
[g] Conserved intraloop (A), interloop (E) and framework (F) hydrogen bonds that were observed over 75% of the 171 antibodies, respectively, in a given position, where especially boldfaced A, E and F show that over 90% of the antibodies have the hydrogen bonds.
[h] The number of antibodies, in which the residue belonging to a given position exists.
[i] The loop lengths are shown. In StrDef-positions, the antibodies that contain the loop with the length shown in this row have no residues in a given position, while in StrNod-positions, the antibodies including the loop with the length shown in parenthesis have residues in a given position.
[j] The rmsd among Cα atoms of the residues belonging to a given position in the antibodies indicated in the eighth row.
[k] The Chothia's numbering is shown, when a given position is defined as β-framework region based on Chothia's definition.
[l] The CDR positions defined by Chothia (C), IMGT (I) and Kabat (K). For example, the position with "CI" means that the position is defined as being in CDR by Chothia and IMGT.
[m] In Kabat definition, the subsequent seven positions are also defined as being in CDR.
[n] The averaged AB rate and Hy and Dy values in all StrNod-positions in the antibodies indicated in the eighth row.

A better understanding of the effect of diverse H3 loop conformations on antigen binding will be of use to antibody design and affinity maturation, and it will require a precise description of the loop conformations. The backbone conformations of the CDR loops have been examined and the CDR loops, with the exception of H3, have been classified into a small number of "canonical structures" based on their length and sequence features.[1,6–8] For H3 loops, several studies have revealed sequence–structure relationships, particularly in the stem region of the loops, and classified them into two groups, bulged or kinked, and non-bulged or extended.[9–12] Nonstem regions (particularly in long H3 loops) are crucial as main antigen-binding sites but their structures have not been fully characterized because of their diversity[13] and thus, a novel method will be required for describing nonstem conformations from new perspectives.

In this study, by using sequence and structural information obtained from a nonredundant set of

171 antibody–antigen complex structures, we aimed: (1) to characterize the antigen-binding propensity of each position in the six CDR loops, (2) to understand the effect of H3 loop lengths on the antigen recognition properties of all the CDR loops, and (3) to relate diverse conformations of long H3 loops to antigen recognition. We proposed a new method for describing structural features of each position in each CDR loop. The summarized structural features determined by the new method, along with sequence properties, were assigned to each position, and this analysis led to simple rules for distinguishing probable antigen-binding from non antigen-binding positions. Moreover, we found that H3 loop lengths affect the antigen-binding patterns of all the CDR loops and that diverse conformations of long H3 loops are largely pre-formed and may increase the specificity for the target antigen.

Results and Discussion

Characterization of antigen-binding propensity of each CDR position

Structurally definable and nondefinable residue positions. To identify the antigen-binding ability of a CDR position, we wished to name the residue positions systematically (e.g., position 1 of H1 and position 3 of H2). Since the CDR loops vary considerably in length and conformation even within a single loop type, it was necessary to distinguish "named" and "unnamed" positions. Table I shows a total of 68 named positions (1, 2, 3... and N, N-1, N-2... from either end of the loop) and their associated sequence and structural properties. We call these positions structurally definable (StrDef), because they correspond to well-aligned columns in a structure-based alignment (see Methods for details). The columns judged to be unaligned were excluded. We call the excluded positions (a total of 27 positions) structurally nondefinable (StrNod) and indicate them in italic in the top row of Table I.

While the definition of the StrDef- and StrNod-positions depended on a specific alignment program (Mustang[14]) and somewhat subjective decisions, we argue that these assignments are largely unambiguous and that different definitions would not alter the main conclusions of this study in any significant manner. The reason is that except for H3, most residue positions in the CDR loops adopt similar structures and can be aligned well by any method. For H3, the length of which ranged between five and 30 in our dataset, we tried to extract common features by tolerating some structural diversity. Based on the structural alignments, we identified the first six and the last seven positions to be structurally definable, which corresponded to the root mean square deviation (RMSD) of 0.5–5.1 Å in each position (the 10th row of Table I). Because of this tolerance, our definition differs from the previous studies, where structural conservation has been determined precisely based on the backbone torsion angles.[1,6–8] Note also that we adopted the CDR definition by Dunbrack,[13] because we felt it more suitable to relate structure features to antigen binding than other definitions but we also indicated the widely used Chothia, Kabat and IMGT definitions in Table I.[1,15–17]

Sequence and structural properties of each CDR position. After naming the positions of the CDR loops, our main interest was to relate sequence and structural properties to the antigen-binding rate, defined as the fraction of antibodies in our dataset, in which the residue in a given position is involved in antigen binding (the second row of Table I). As the main sequence property, we defined the sequence-conservation rate as a sum of the observed frequencies of the three most abundant amino acids in each position (as shown in the third row of Table I). Of the three most abundant amino acids, the amino acids with the observed frequency of over 20% are shown in the fourth row of Table I.

To capture structural features of the CDR loops, we introduced a new coordinate system to describe quantitatively the spatial location of each residue in the CDR loops relative to the geometric center of the CDR regions. The idea was based on our observation that antigen binding tends to take place around the geometric center of the CDRs. A similar notion was found in a previous study,[18] showing that the center of the antigen-binding site on an antibody has a high possibility of antigen binding. The coordinate system specifies a "standard view," in which the width, height or depth of an antibody structure is represented by the distance along the *x*-, *y*,- or *z*-axis, respectively (see Methods and Supporting Information Figs. S1A and S1B for details). We characterized each residue in each CDR loop with two properties: the height value, Hy (the *y*-coordinate of the Cα atom), and the distance Dy (of the Cα atom) from the *y*-axis. The averaged Hy and Dy values of the residues in a given position in the 171 antibodies are summarized in the fifth and sixth rows of Table I, respectively.

We calculated hydrogen bonds (HBs) by HBPLUS[19] between residues within a CDR loop (intraloop HB, labeled A in the seventh row of Table I, where over 75% of the 171 antibodies have intraloop HBs in a given position, and in boldface for over 90% of the antibodies.), those belonging to different CDR loops (interloop HB, labeled E), and those between a CDR loop and a framework (non CDR) region (framework HB, labeled F).

Sequence conserved positions are infrequently used in antigen binding. The antigen-binding and sequence-conservation rates in StrDef-positions

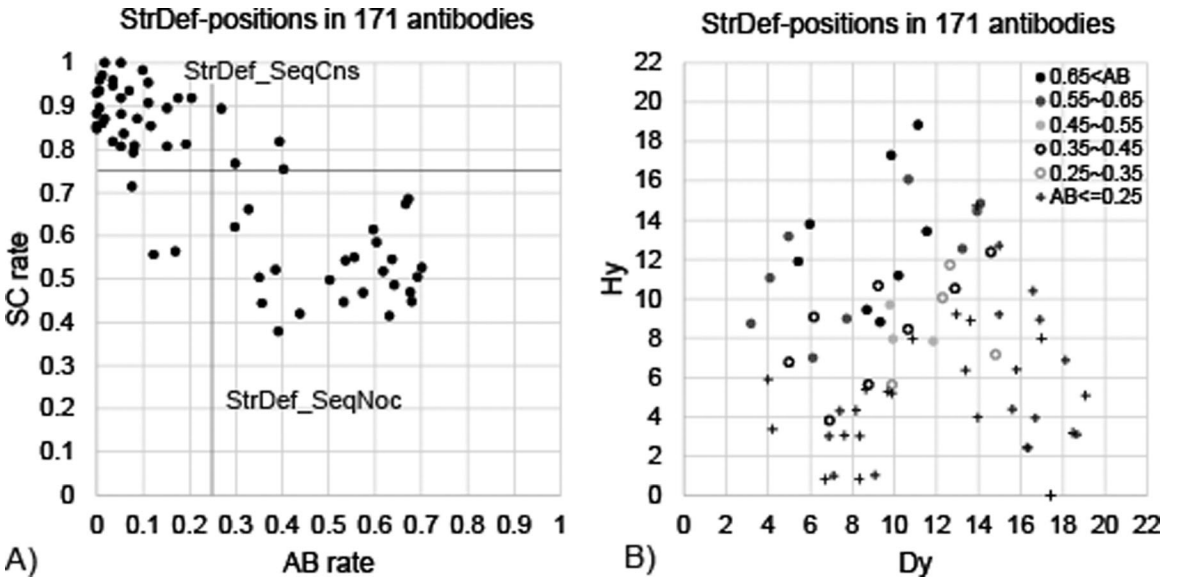

Figure 1. Antigen-binding properties in StrDef-positions. A) The relationship between antigen-binding (AB) and sequence-conservation (SC) rates in StrDef-positions. The correlation coefficient of the relationship is −0.82. B) The relationship between averaged Hy and Dy values in each StrDef-position in the dataset, being separated into six ranges of AB rate, AB rate>0.65 (black filled circle), 0.55< AB rate<= 0.65 (gray filled circle), 0.45<AB rate<=0.55 (light-gray filled circle), 0.35<AB rate<=0.45 (black unfilled circle), 0.25<AB rate<=0.35 (gray unfilled circle), and AB rate<=0.25 (plus).

are highly correlated with each other [a correlation coefficient of −0.82: Fig. 1(A)]. The lines drawn at 0.25 and 0.75 on the respective axes of this plot can define (with only six exceptions) two major groups; one contains the positions that are sequence conserved (the sequence-conservation rate > 0.75; labeled StrDef_SeqCns) and infrequently used in antigen binding (the antigen-binding rate < 0.25), and the other contains the positions that are sequence nonconserved (labeled StrDef_SeqNoc) and more frequently utilized for antigen binding. Thus, our first observation is that structurally definable, sequence conserved (StrDef_SeqCns) positions, the 36 cells with the numbers in bold in the top row of Table I, are infrequently used in antigen binding. We also examined sequence-conservation rates based on the two most abundant amino acids, and obtained a slightly lower correlation coefficient with the antigen-binding rates (−0.80). Therefore, we decided to define the sequence-conservation rate based on the three most abundant amino acids.

StrDef_SeqCns positions are mainly involved in maintaining the structures of antibody variable fragments, by connecting alternatingly the stem regions of the CDR loops. Previously, Chothia and Lesk defined structurally important positions as the β-sheet framework and observed these positions to be non antigen-binding in general.[1,6] Our definition of the CDR included 26 β-sheet framework positions (Table I) and we classified them into StrDef_SeqCns (19 positions) and StrDef_SeqNoc (seven positions). While most of these β-sheet framework positions are involved in maintaining antibody structures, the latter positions (StrDef_SeqNoc) were utilized also for antigen binding, as described above. MacCallum *et al*. have made an observation similar to ours.[18]

Highness and centrality in the CDRs determine the residue's tendency to engage in antigen binding. The relationship between Hy and Dy properties in StrDef-positions [Fig. 1(B)] indicates that the positions frequently used for antigen binding tend to have a very large Hy value or large Hy and small Dy values simultaneously. Our second observation, thus, is that antigen binding typically takes place in the "very high" (observed mainly in H2) or "high and centrally located" (in H3) StrDef-positions, most of which are StrDef_SeqNoc-positions in the heavy chains.

StrNod-positions exist only in H3, L1, and L3 and are located around the middle of the CDR loops (Table I). The H3 loops that consist of more than or equal to 14 residues contain StrNod-positions. Their Hy and Dy values are broadly distributed as shown in Supporting Information Figure S2A, where most of the residues have larger Hy values than the StrDef-positions [Fig. 1(B)]. Of the 329 residues belonging to the StrNod-positions, 221 (67%) are involved in antigen binding, which is significantly higher than the corresponding value for the StrDef_SeqNoc-positions (50%, 579 out of 1,160; $p < 3.0e\text{-}8$ by Fisher's exact text). This observation indicates that the StrNod-positions are even more frequently utilized for antigen binding than the StrDef_SeqNoc-positions in particularly long H3 loops, as shown in Supporting Information Figures S3A and S3B.

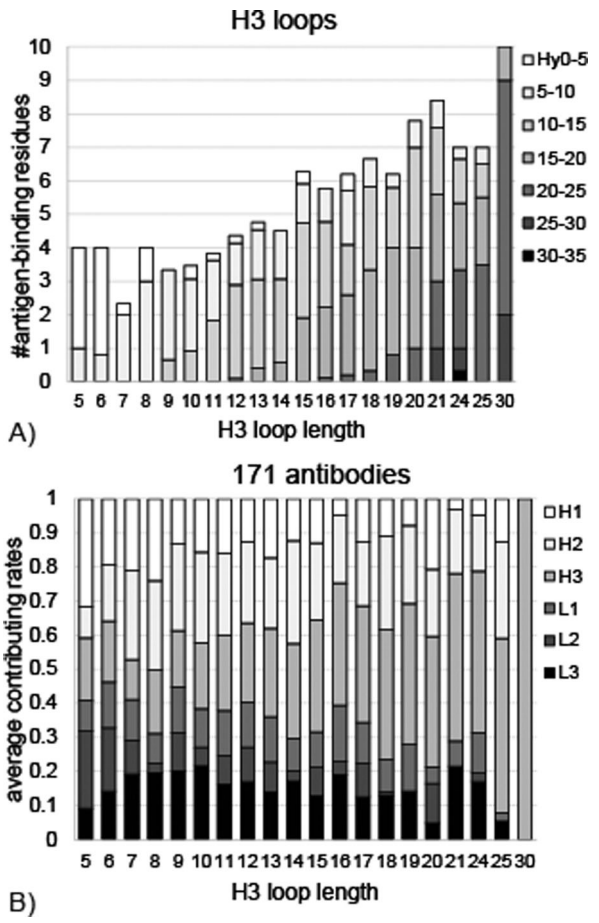

Figure 2. The effects of H3 loop lengths on antigen binding. A) The average number of antigen-binding residues in H3 loops as a function of the H3 loop length, being separated into seven ranges of Hy values, from 0 to 35 Å by 5 Å, whose colors are changed gradually from white to black. B) The averaged contributing rates to antigen binding in each of six CDR loops in antibodies with the same H3 loop length, whose colors are changed from white (H1) to black (L3) gradually. The contributing rates were calculated based on the number of antigen-binding residues.

Figure 2(A) shows the above observations from a different perspective, where the proportions of Hy values of the antigen-binding residues in H3 loops are shown as a function of the loop length. It indicates that in H3 with the length shorter than 14 residues, where all the positions in H3 are definable (StrDef-positions), the positions with low Hy values contribute to antigen binding. In those short H3 loops, antigen binding occurs primarily in the positions with Hy values of 5.0–15.0 Å, most of which are StrDef_SeqNoc (the range of Hy values: 5.9–13.8 Å, the mean value: 10.0 Å), although in the shortest H3 loops (five or six residues), the positions with Hy values smaller than 5.0 Å, which included only StrDef_SeqCns-positions (the range of Hy values: 0.8–4.3 Å, the mean value: 2.5 Å) are dominated (see also Supporting Information Figs. S3A and S3B). On the other hand, in H3 with the length longer than or equal to 14 residues, the positions with high Hy values (>10 Å), most of which are StrNod-positions (18.0 Å on average), are main contributors to antigen binding, where however, the Hy values or the types (StrDef_SeqNoc or StrNod) of positions utilized in antigen binding vary based on the loop length. Thus, our third observation is that the H3 loops utilize different positions for antigen binding depending on the loop length and that a cut-off of 14 residues differentiates the short and long H3 loops.

The StrNod-positions in L1 and L3 (the central seven and three positions, respectively) also showed antigen-binding tendencies different from those in the StrDef-positions. The relationship between Hy and Dy values in all the residues belonging to these positions (Supporting Information Figs. S2B and S2C) shows that a tendency in L1 is different from that in the StrDef-positions [Fig. 1(B)]. This is because the average Hy value in StrNod-positions in L1 is higher than those in StrDef-positions due to the formation of an α-helix or β-sheet in the nonstem regions; the proportion of antigen-binding residues in the StrNod-positions (0.41) is higher than the average antigen-binding rate in the adjacent StrDef-positions (0.05 in position 6 and 0.30 in position N-3). On the other hand, in L3, the proportion of antigen-binding residues in the StrNod-positions is lower (0.24) than the average antigen-binding rate in the adjacent StrDef-positions (0.40 in position 6 and 0.53 in position N-3). These StrNod-positions in L3 are bent to the outside of the CDR, which decreases the Hy values compared to the adjacent positions and thus, less likely to be involved in antigen binding. These observations suggest that the StrNod-positions in L1 or L3 have a moderate or low tendency to be involved in antigen binding, respectively.

A simple method for distinguishing probable antigen-binding positions. Based on the above observations, we constructed a simple method for distinguishing antigen-binding from non antigen-binding positions, which requires only the information about antibody sequences. We suppose that (1) StrDef_SeqCns-positions are not involved in antigen binding, (2) StrDef_SeqNoc-positions are involved in antigen binding, and (3) StrNod-positions in H3 and L1 are antigen-binding sites, while those in L3 are non antigen-binding sites. As summarized in Supporting Information Table S1, the method distinguished the residues in all the CDR loops correctly with the moderate accuracy of 0.72 (the Matthews correlation coefficient MCC = 0.44, the harmonic mean of precision and recall F measure=0.62, see the legend to Supporting Information Table S1). It also showed that the accuracy for the six CDR loops differ from each other, and that it is particularly low for L2 loops due to their low antigen-binding rates. Our discrimination accuracy is higher than that in a previous study with a similar rule-based method (Paratome, MCC = 0.23 and F measure = 0.48),[20,21] where the definition of antigen-binding residues is

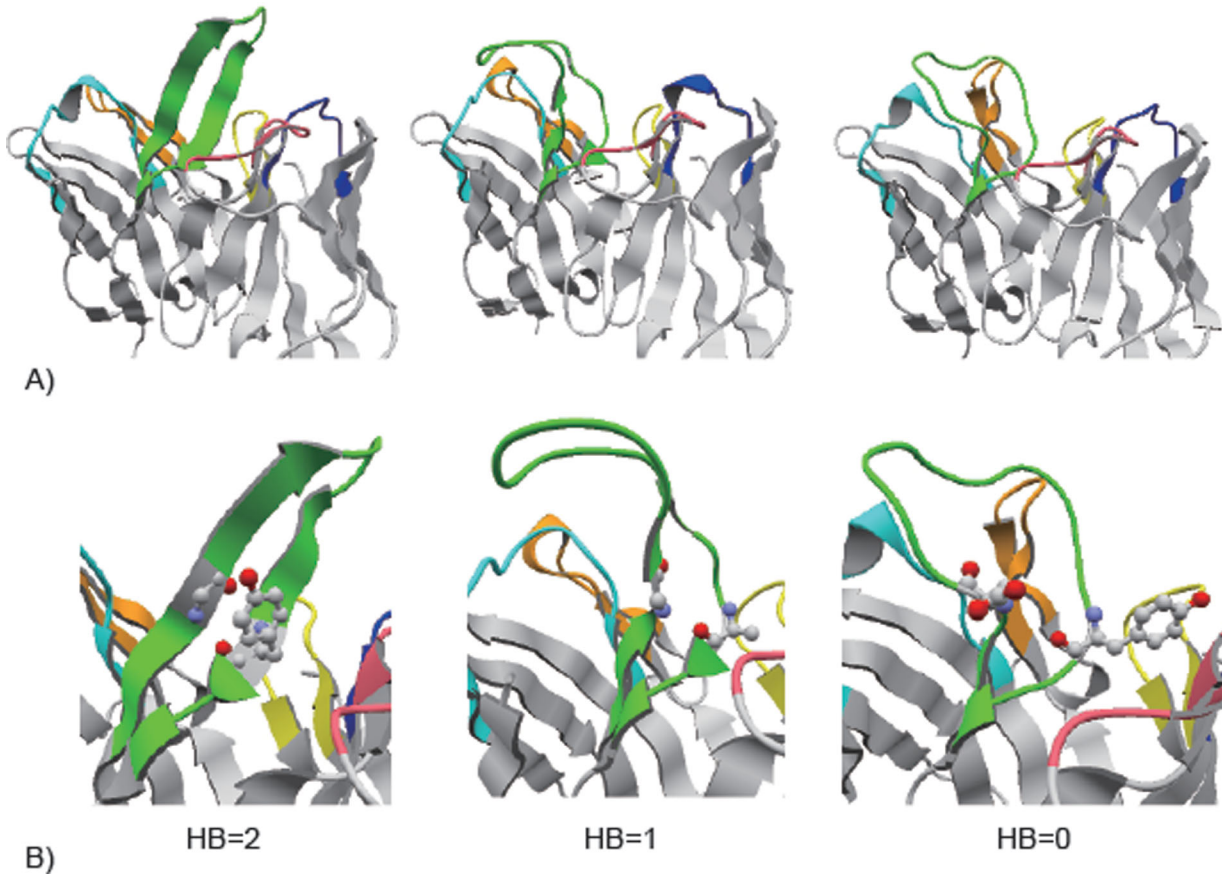

Figure 3. Diverse conformations of long H3 loops. A) Antibody structures with diverse H3 loop conformations. The structures of antibody variable regions are shown. The CDR loops are colored cyan, orange, green, blue, pink, and yellow for H1, H2, H3, L1, L2, and L3, respectively. All the figures were drawn by using the interactive molecular viewer, jV.[25] From the left, the structures of antibodies in PDBID 3vg9, 4m62, and 5e8e are shown, whose H3 loops consist of 17, 20, and 18 residues and have straight, bend and broad conformations in the nonstem region, respectively. B) Hydrogen bonds between main-chain atoms at positions 4 and N-3 in H3 loops in the antibodies shown in 3A. The positions 4 and N-3 are shown in ball and stick model. The numbers of the hydrogen bonds are also indicated.

different from ours (at least one atom within 6 Å from any antigen atoms, compared to our 4Å distance threshold), while it is lower than the prediction accuracy of antigen-binding residues in full length antibodies by a random forest-based method (MCC = 0.52).[22] These observations suggest that probable antigen-binding positions can be identified by using simple sequence and structural features.

The effect of H3 loop lengths on the antigen-binding properties of all the CDR loops

In H3 loops, the positions utilized in antigen binding differ depending on the loop lengths, as described above. The H3 loop lengths also affect the contributions of the other CDR loops to antigen binding (our fourth observation). In antibodies with short H3 loops, most of the six CDR loops participate in antigen binding, while in those with long H3 loops, H3 loops are the main contributor to antigen binding [Fig. 2(B)]. This observation is consistent with the shapes of the surfaces of the CDRs (Supporting Information Fig. S3C); in antibodies with short H3 loops, the CDR surfaces are concave, where all the six CDR loops can contact the antigen. On the other hand, the CDR surfaces in antibodies with long H3 loops are convex, where mainly the H3 loops bind to the antigens and the degree of convexity of the CDR surface depends on the H3 loop conformation. These observations suggest that in antibodies with long H3 loops, H3 loops, particularly their StrDef_SeqNoc- or StrNod-positions, are the main contributors to antigen binding among all the CDR loops.

Figure 2(B) also shows that H2 loop is often involved in antigen binding, which is also consistent with the observations in Figure 1(B), where the StrDef_SeqNoc-positions in H2 have very large Hy values.

The effect of adopting diverse conformations of long H3 loops on antigen recognition

The predominance of nonstraight conformations in long H3 loops. The distribution of Hy and Dy values is very broad in StrNod-positions in

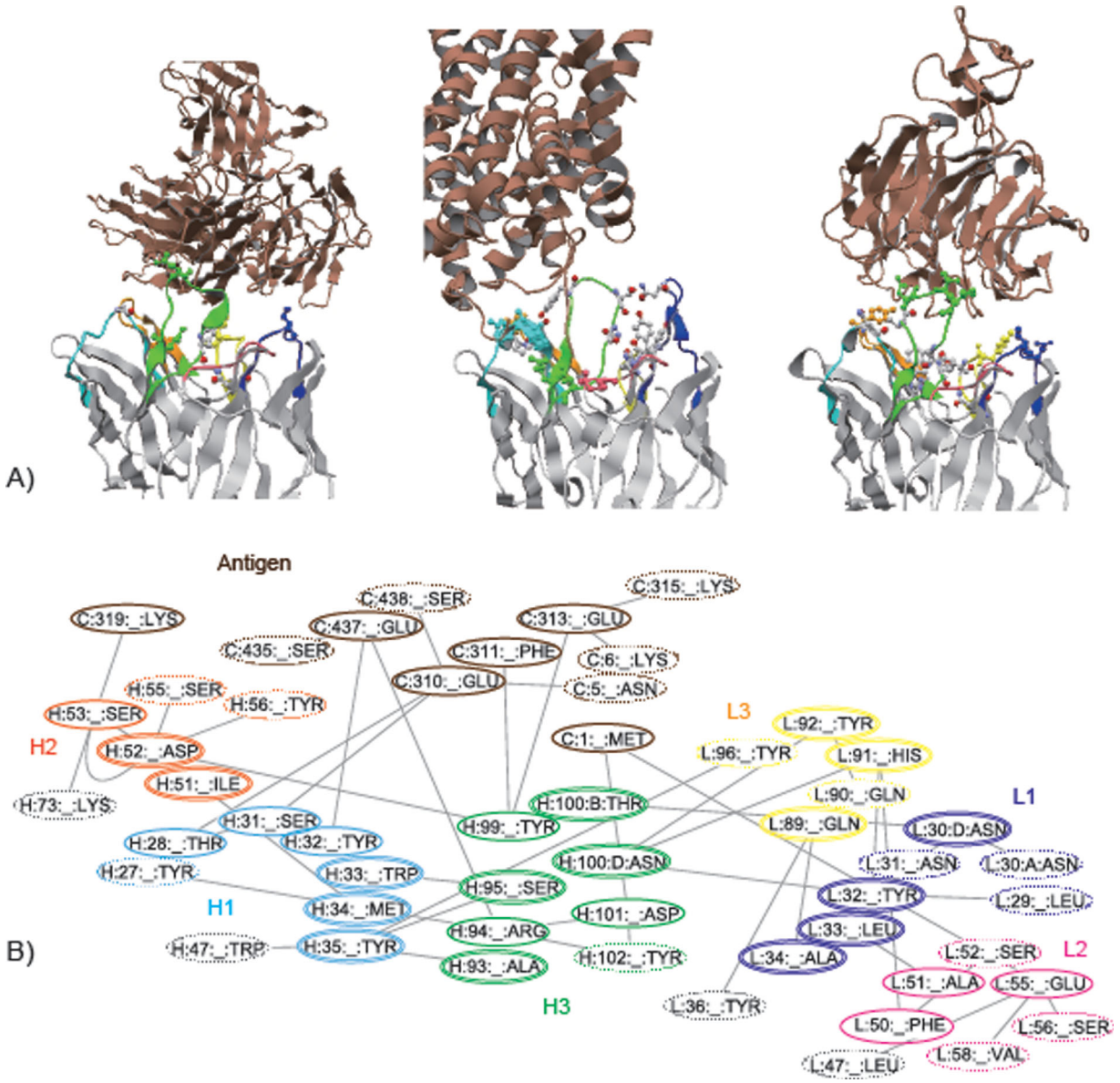

Figure 4. Hydrogen bond networks around antibody–antigen interaction sites. A) The structures of three antibody–antigen complexes (from the left, 2qqk, 3gi9, and 3sob). The antibody–antigen and nonconserved interloop HBs are shown in ball and stick models, where the former is colored in the same manner as in Figure 3(A), and the latter is colored cpk. The antigen molecules are colored brown. B) The hydrogen bond network in 3gi9. The hydrogen bond network that comprises the residues involved in antibody–antigen and interloop HBs, was calculated by RINerator[26] and drawn by Cytoscape 3.3.0.[27,28] The residues enclosed by single or double line are directly involved in antibody–antigen or interloop HBs, respectively. The other related residues are enclosed by dotted line. The non-CDR residues are enclosed by light-gray line.

H3 loops (Supporting Information Fig. S2A). These broad distributions come from the diversity of loop conformations in the nonstem regions as shown in Figure 3(A), where H3 loops with varying lengths and conformations are shown. The distribution of the largest Hy value of each H3 loop in the dataset shows that long H3 loops with the same length have different Hy values, indicating that these H3 loops adopt different conformations (Supporting Information Fig. S4A). We examined the conformations of long H3 loops visually, and found that a ladder-like "straight" β-sheet conformation, as in 3vg9 of Figure 3(A), was a minority. In many long H3 loops, we observed conformations that broaden in the nonstem regions, as in 5e8e (the "broad" conformation), or those with no broadening and are bent towards H1 or L3 loops (the "bent" conformation, as in 4m62, of which the H3 loop bends to H1). As summarized in Supporting Information Table S2, only 12 of the 72 long H3 loops have straight conformations. Thus, our fifth observation is that long H3 loops prefer to form nonstraight conformations.

The observed conformational varieties are not induced by antigen binding. We considered how the long H3 loops form nonstraight conformations.

To check whether antigen binding induced conformational changes in H3 loops, we examined 31 antibodies in the dataset, which have also been determined in antigen unbound forms. The H3 loops of the 31 antibodies have various lengths (shorter than 14 residues in 20 antibodies and longer than or equal to 14 residues in 11 antibodies, in four of which the nonstem regions of the H3 loops are disordered, therefore, we only used the remaining seven antibodies). Most (twenty four) of these 27 structures have conformations in H3 loops in identical categories, such as straight or nonstraight (bent or broad) as described above, between the antigen-bound and unbound forms (Supporting Information Fig. S4B). It indicates that antigen binding has little influence on the formation of H3 loop conformations. A previous study has reported only a few specific examples of conformational changes caused by antigen binding (particularly in H3 loops),[2] consistent with our observation. Note that the affinity data for ten of the 32 antibodies are available in SAbDab,[23] which range from 2.48e-12 to 4.70e-6. It suggests that the above observation is not biased toward high affinity antibodies.

The formation of a nonstraight conformation by breaking and forming intra and interloop HBs. Since the broad, bent or straight conformations are inherent features of the H3 loops, we sought to identify key determinants of these conformations. The CDR loops start with a β-ladder structure with regular intraloop hydrogen bonding patterns, but if the HB in a specific position is broken, the ladder cannot extend and it may result in forming various loop conformations, particularly in long loops, such as bent or broad [Fig. 3(A)]. We found that HB breaks between main-chain atoms in positions 4 and N-3 are an important factor to prevent an extension of the β-ladder [our sixth observation; Fig. 3(B)]. The numbers of these hydrogen bonds highly correlate with the observed long H3 loop conformations; all the long H3 loops with two main-chain hydrogen bonds between positions 4 and N-3 form a straight conformation [as in 3vg9 in Fig. 3(B)], while H3 loops with one or no hydrogen bonds tend to form bent (as in 4m62) or broad (5e8e) conformations, respectively, as summarized in Supporting Information Table S2.

We also found that long H3 loops tend to form nonconserved intra and interloop HBs in nonstem regions (our seventh observation; Supporting Information Figs. S4C and S4D, respectively). These nonconserved intraloop HBs differ from intraloop HBs observed in regular ladder-like structures and appear to play a key role in maintaining nonstraight conformations in nonstem regions. [In a straight conformation such as that in 3vg9 of Figure 3(A) and Supporting Information Figure S4E, regular ladder-like intraloop HBs, formed between main-chain atoms, stabilize the H3 loop conformation.] On the other hand, the formation of interloop HBs result in increasing the structural diversity of long H3 loops. Thirteen pairs of the 27 antigen-bound and unbound antibody structures show identical nonstem interloop HBs in H3 loops, five of which contain long H3 loops (Supporting Information Fig. S4F). Seven of the remaining pairs have no nonstem interloop HBs in both bound and unbound forms. The remaining seven pairs show interloop HBs different between the bound and unbound forms. These observations suggest that these nonconserved interloop HBs are formed before antigen binding.

Moreover, we observed that these nonconserved interloop HBs in long H3 loops are located close to the residues involved in antibody–antigen HBs, and this observation applies to almost all antibodies with nonconserved interloop HBs (our eighth observation). As shown in Figure 4(A), the interloop HBs exist at the edges of the antibody–antigen HB networks in many cases so as to facilitate the formation of antibody–antigen HBs. In addition, these interloop HBs form a larger HB network including the antibody–antigen HBs [Fig. 4(B)] and may contribute to increasing the antibody's affinity for the antigen.

Conclusions

In this study, we observed several important features for antigen recognition by antibodies as summarized in Supporting Information Table S3. We found simple rules for identifying probable antigen-binding positions in each CDR loop. The structural analysis of long H3 loops showed that the H3 loop length affects their loop conformations and thereby determines the H3 residues involved in antigen binding. The H3 loop length also influences the CDR surface shapes as a whole, defining to what extent the other CDR loops contribute to antigen binding. These observations suggest that in designing higher-affinity antibodies, we should consider the H3 loop length and (if possible) conformation in determining which positions to introduce mutations. The majority of long H3 loops adopt nonstraight conformations that are likely to be pre-formed before antigen binding and increase the antigen specificities.

Methods

Dataset

A nonredundant search of antibody–antigen complex structures was performed using CD-hit[24] in SAbDab,[23] with a cut-off of 95% sequence identity to cluster the antibody and protein antigen structures having a resolution of better than 2.8 Å (the search performed on December 7th, 2015). Of the 180 antibody–antigen complex structures obtained, the

antibodies in three complexes contained disordered CDR loop structures, and six bound to nonantigen proteins. Therefore, we excluded these nine complexes and used the remaining 171 antibody–antigen complex structures in our analyses.

Note that this set included 12 antibodies with H3 loops having extended conformations in their stem regions (where the longest extended H3 loop consisted of 16 residues). In this study, we did not mention the extended H3 loops, because we mainly focused on the diverse nonstem conformations of long H3 loops.

A structure alignment of CDR loops

A structure-based alignment of each of the six CDR loops was generated by using Mustang.[14] Each alignment included the CDR loops of a particular type (e.g., L1) and 10 residues before and after the loop region. We used the shortest loop in the dataset as a reference and examined additional positions in the alignment (shown as insertions for the shortest loop) one by one from both ends. If we confirmed sufficient structural similarity over the majority of the longer antibodies, we defined them as additional positions to the shortest loop (Supporting Information Fig. S5). All the positions represented by the shortest loop and additional positions thus extended were defined as structurally definable (StrDef). Positions that were judged to be nonextendable due to dissimilar structures were defined as structurally nondefinable (StrNod). When the alignment of an antibody was ambiguous, we manually compared the structure to those of clearly aligned antibodies to determine whether the positions are definable. Most of the CDR positions except for H3 were well-aligned and adopted similar structures, hence, they were defined to be structurally definable. In H3, we identified only the first six and the last seven positions to be structurally definable. There exist 17, seven and three nondefinable positions in H3, L1, and L3, respectively, as shown in Table I.

A new coordinate system

We developed a new coordinate system that characterizes CDR loop conformations and shows a standard view of an antibody structure. Three axes in the new coordinate system were determined by considering the following three lines: the line connecting between the Cα atoms in the first residues (position 1) in H1 and L1 loops, that between H2 and L2 loops, and that between H3 and L3 loops. This is because the structures of the stem regions of the CDR loops (including position 1) were well conserved in the structural alignment of the 171 antibodies. The line between H1 and L1 (H1-L1) is almost overlapped with the H3-L3 line. The H2-L2 line is located about 6 Å above the H1-L1 line, having around 60°, as shown in Supporting Information Figure S1A. We used the H1-L1 line as the first axis (x-axis). As the second axis (y-axis), the line perpendicular to the H1-L1 line and passing through the H2-L2 line was considered, which is identical to the shortest line between the H1-L1 and H2-L2 lines. Then, the exterior product between the x- and y-axes was used as the third axis (z-axis). Supporting Information Figure S1B shows a standard view of an antibody structure in this coordinate system, where the origin is put at around the center-bottom of the region formed by the CDR loops, and the distance along the x-, y-, or z-axis indicates the width, height, or depth of the antibody structure.

Acknowledgments

We thank N. Sakiyama (Mitsui Knowledge Industry CO., LTD) and H. Shirai (Astellas Pharma Inc.) for the construction of the dataset.

References

1. Al-Lazikani B, Lesk AM, Chothia C (1997) Standard conformations for the canonical structures of immunoglobulins. J Mol Biol 273:927–948.
2. Sela-Culang I, Alon S, Ofran Y (2012) A systematic comparison of free and bound antibodies reveals binding-related conformational changes. J Immunol 189:4890–4899.
3. Padlan EA, Abergel C, Tipper JP (1995) Identification of specificity-determining residues in antibodies. Faseb J 9:133–139.
4. Kunik V, Ofran Y (2013) The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. Protein Eng Des Sel 26:599–609.
5. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, Colman PM, Spinelli S, Alzari PM, Poljak RJ. (1989) Conformations of immunoglobulin hypervariable regions. Nature 342:877–883.
6. Chothia C, Lesk AM (1987) Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 196:901–917.
7. Sibanda BL, Blundell TL, Thornton JM (1989) Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. J Mol Biol 206:759–777.
8. Martin AC, Thornton JM (1996) Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. J Mol Biol 263:800–815.
9. Shirai H, Kidera A, Nakamura H (1996) Structural classification of CDR-H3 in antibodies. FEBS Lett 399:1–8.
10. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM (1998) Conformations of the third hypervariable region in the VH domain of immunoglobulins. J Mol Biol 275:269–294.
11. Shirai H, Kidera A, Nakamura H (1999) H3-rules: identification of CDR-H3 structures in antibodies. FEBS Lett 455:188–197.
12. Kuroda D, Shirai H, Kobori M, Nakamura H (2008) Structural classification of CDR-H3 revisited: a lesson in antibody modeling. Proteins 73:608–620.
13. North B, Lehmann A, Dunbrack RL Jr. (2011) A new clustering of antibody CDR loop conformations. J Mol Biol 406:228–256.

14. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM (2006) MUSTANG: a multiple structural alignment algorithm. Proteins 64:559–574.
15. Wu TT, Kabat EA (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.
16. Kabat EA, Wu TT, Perry H, Gottesman K, Foeller C (1991) Sequences of Proteins of Immunological Interest, 5th ed. NIH Publication, pp 91–3242.
17. Lefranc MP, Pommie C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27:55–77.
18. MacCallum RM, Martin AC, Thornton JM (1996) Antibody-antigen interactions: contact analysis and binding site topography. J Mol Biol 262:732–745.
19. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793.
20. Kunik V, Ashkenazi S, Ofran Y (2012) Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. Nucleic Acids Res 40:W521–W524.
21. Kunik V, Peters B, Ofran Y (2012) Structural consensus among antibodies defines the antigen binding site. PLoS Comput Biol 8:e1002388
22. Olimpieri PP, Chailyan A, Tramontano A, Marcatili P (2013) Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server. Bioinformatics 29:2285–2291.
23. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM (2014) SAbDab: the structural antibody database. Nucleic Acids Res 42:D1140–D1146.
24. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.
25. Kinoshita K, Nakamura H (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. Bioinformatics 20:1329–1330.
26. Doncheva NT, Klein K, Domingues FS, Albrecht M (2011) Analyzing and visualizing residue networks of protein structures. Trends Biochem Sci 36:179–182.
27. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504.
28. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2:2366–2382.