# Shapes of Antibody Binding Sites: Qualitative and Quantitative Analyses Based on a Geomorphic Classification Scheme

Michelle Lee,[†] Peter Lloyd,[‡] Xiyun Zhang,[†] Julie M. Schallhorn,[†] Keiki Sugimoto,[†] Andrew G. Leach,[†] Guillermo Sapiro,*[,‡] and K. N. Houk*[,†]

Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, and the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, Minnesota 55455

guile@ece.umn.edu; houk@chem.ucla.edu

Received December 27, 2005
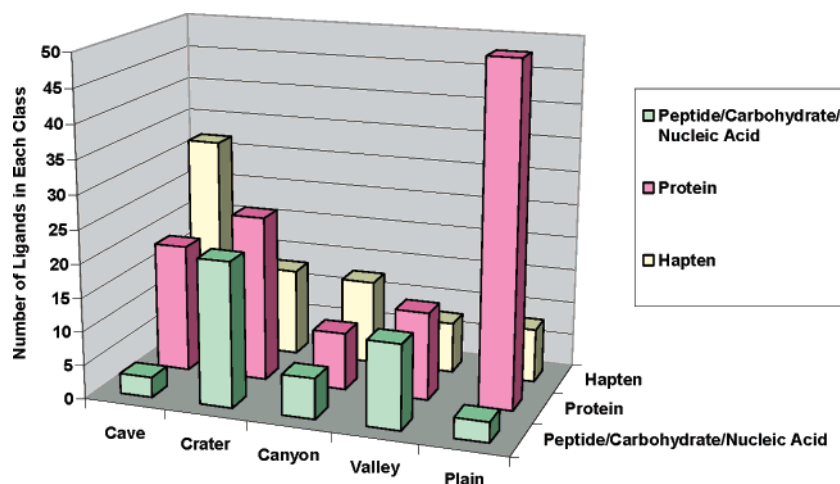
The topography of antibody binding sites has been classified into five types that evoke familiar geomorphic features of the Earth. The 229 antibody crystal structures from the Protein Data Bank were analyzed and classified into these classes. Relationships to previous topography classifications by Rees et al., who defined three classes, and Thornton et al., who defined four classes, are identified. An algorithm was developed to identify the antibody binding site class automatically based on the definition and the shape of the binding site. A three-dimensional convex hull was formed around the complementarity determining regions (CDRs) of the antibody. The convex hull was then "trimmed" to fit the binding site by using distance criteria and morphological techniques. Once the program identified the binding site shape, a statistical and distance based analysis was performed to classify automatically the antibody into one of the five geomorphic classes. The five antibody topography classes are as follows: cave (mostly hapten binders), crater (mostly protein and peptide/carbohydrate/nucleic acid binders), canyon, valley, and plain (mostly protein binders). Comparisons of the binding sites of empty and of complexed antibody binding sites gave an indication of how the shape of the binding site is influenced by binding of the antigen.

Introduction

Immunoglobulin G (Ig) antibodies are Y-shaped proteins composed of four polypeptide chains—two ~23 kD identical light chains and two 53—75 kD heavy chains.[1] The structure of an Ig antibody consists of two Fab regions (fragment antigen binding) and one Fc (fragment crystallizable). The two Fabs provide each antibody with two identical antigen binding sites consisting of the variable domains from both the heavy (VH) and the light chains (VL).[2] Each Fab is further subdivided into

(1) Voet, D.; Voet, J.; Pratt, W. *Fundamentals of Biochemistry*, 2nd ed.; John Wiley & Sons: New York, 2004.

(2) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Keith, R.; Walter, P. *Biology of the Cell*, 4th ed.; Garland Science: New York, 2002.

† University of California Los Angeles.
‡ University of Minnesota.

hypervariable (HV) and framework (FR) regions. The HV includes six hypervariable complementarity-determining regions (CDR H1−H3, CDR L1−L3) which are critical for antigen recognition. Through random genomic splicing of genes and mutations on the light and the heavy chains, the six CDRs on each Fab provide billions of different sequences perfected through selection for the binding of individual molecules.[3] Despite the enormous variability in these binding sites, certain patterns in the shape and size of binding sites are observed.

We have carried out theoretical studies on antibody catalysis[4] and realized that many catalytic antibodies have similar cavelike binding sites for haptens. This led us to perform a general analysis of the available antibody crystal structures. Rees et al.[5] and Thornton et al.[6] independently proposed classification schemes of antibody topography based on a handful of antibody crystal structures (31 by Webster, Henry, and Rees[5] and 45 by McCallum, Martin, and Thornton[6]). Rees et al. visually sorted the antibodies into three classes—cavity, groove, and plane. Comparisons with the structures bound by the antibodies show that the three classes preferentially bind hapten, peptide, and protein antigens, respectively.[5] Thornton et al. classified the antibodies more quantitatively by using Kuhn's fractal atomic density measurement as an estimate of the surface curvature. The method reduces the three-dimensional structure to a two-dimensional composite profile which is then quantified based on the relative degree of surface concavity and convexity.[6] Using the fractal measurement and cluster analysis, 45 antibodies structures were sorted into four categories. Upon visual inspection, the four classes were named on the basis of their general shape properties: concave, moderately concave, ridged, and plane.[6]

In our study, 229 antibody crystal structures in the Protein Data Bank were analyzed and classified.[7] The structures analyzed in this paper include all complete antibody crystal structures available in the Protein Data Bank as of April 2002, excluding incomplete or duplicate structures. If more than one structure of the same antibody was available, the one with the highest resolution was chosen. The PDB references of the crystal structures and for the analysis are given in the Supporting Information. Not all antibodies structures in the bank are complexed with an antigen. The antibodies without known antigens were excluded from the study. The classification scheme helps uncover the basic principles underlying antibody−antigen interactions and the mechanism of antigen recognition. The classification scheme presented here differs from previous ones in that there are five classes, and the antibody topographies are named after familiar geological features for easy communication and understanding of shapes. The antibody binding sites were initially analyzed in 3D at the UCLA Visualization Portal to recognize and establish key features unique to each of the five geomorphic classes. A computer algorithm was created to recognize the unique criteria for each of the five

geomorphic classes. Unlike Kuhn's fractal atomic density measurement used by Thornton which only measures amounts of surface concavity and convexity,[6] the computerized classification scheme developed here measures a number of geometric characteristics of the site and automatically differentiates between the shapes of antibody binding pockets and gives a numerical score between 0 (unlikely) and 1 (most likely) based on how closely the antibody binding site characteristics match the "ideal" characteristics of each of the five geomorphic classes.

Results and Discussion

The names of the classification scheme presented here correlate with the geological features of the earth. Unlike names such as cavity or ridged, the geomorphic names give a clearer image of the binding site morphologies and are defined in such a way that all chemists can readily identify the antibody cavity type.

The Cave. Figure 1 consists of pictures of a geological cave (Figure 1a) and various views of the Fab of antibody 48G7 (PDB code: 1GAF)[8] represented by VMD[9] (Figure 1b−d). The defining characteristics of the cave class are its small, approximately circular aperture of the pocket and the depth of the binding site. The walls of the cave are approximately perpendicular to the surface of the binding pocket; they extend straight into the antibody. These visually defining features were then used to create an automatic classification scheme as discussed later. The VMD pictures of the antibodies are colored based on the position of the atoms with respect to the center of volume of the Fab. The red color within the cave signifies the great depth of the cave. Figure 1d shows a typical antigen bound in the cave of 48G7.

The Crater. Figure 2 consists of pictures of a geological crater (Figure 2a) and various views of the Fab of antibody Se155−4 (PDB code: 1MFD)[10] (Figure 2b−d). This class is similar to the cave class in that the opening of the crater is approximately circular. The difference between the crater and the cave is that unlike the small circular opening of the cave, the aperture of the crater is rather large and the crater is not as deep as the cave. Instead of extending straight into the antibody, the walls of the crater slopes toward the center of the aperture forming a shallow bowl.

The Canyon. Figure 3 consists of pictures of a geological canyon (Figure 3a) and various views of the Fab of antibody Fv4155 (PDB code: 1BFV)[11] (Figure 3b−d). The characteristic of the canyon is that the binding site is a steep cut across the top of the Fab region. At least one end of the canyon is open to the external environment. Although it is not circular, the canyon is similar to the cave in that the opening is small and pocket is deep. The walls of the canyon are approximately perpendicular to the surface of the binding pocket and extend straight into the antibody.

The Valley. Figure 2 consists of pictures of a geological valley (Figure 4a) and various views of the Fab of antibody CB4−1 (PDB code: 1HH6)[12] (Figure 4b−d). This class is

(3) Janeway, C. A.; Travers, P.; Walport, M.; Schlomchik, M. *Immuno Biology*, 5th ed.; Garland Publishing: New York, 2001.

(4) Tantillo, D.; Leach, A. G.; Zhang, X.; Houk, K. N. *Catalytic Antibodies*; Keinan, E., Ed.; Weinheim, Germany: Wiley-VCH: New York, 2004; pp 72−117.

(5) Webster, D. M.; Henry, A. H.; Rees, A. R. *Curr. Opin. Struct. Biol.* **1994**, *4*, 123−129.

(6) MacCallum, R. M.; Martin, A. C. R.; Thornton, J. M. *J. Mol. Biol.* **1996**, *262*, 732−745.

(7) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
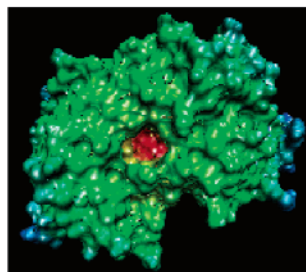
(8) Wedemayer, G. J.; Patten, P. A.; Wang, L. H.; Schultz, P. G.; Stevens, R. C. *Science* **1997**, *276*, 1665−1669.

(9) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.

(10) Bundle, D. R.; Baumann, H.; Brisson, J. R.; Gagne, S. M.; Zdanov, A.; Cygler, M. *Biochemistry* **1994**, *33*, 5183−5192.

(11) Trinh, C. H.; Hemmington, S. D.; Verhoeyen, M. E.; Phillips, S. E. *Structure* **1997**, *5*, 937−948.
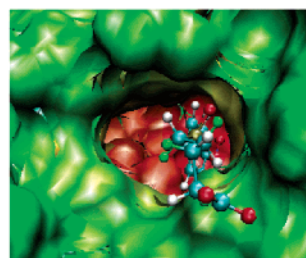
1a: Geological cave

1b: Antibody "cave" 48G7 (PDB code: 1GAF)[8]
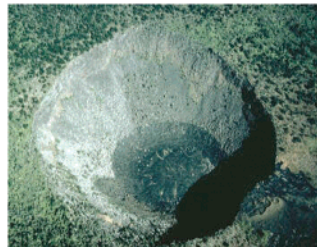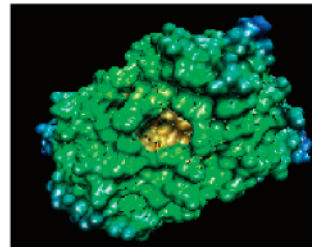
1c: Close up of antibody "cave"

1d: Antibody "cave" complexed with hapten 5-(p-nitrophenyl phosphonate)-pentanoic acid

FIGURE 1. Geological and antibody pictures of the class cave. The antibody pictures are generated with VMD. Geological cave photo (a) used with permission of the National Speleological Society (www.caves.org).

2a: Geological crater

2b: Antibody "crater" murine Se155-4 (PDB code: 1MFD)[10]

2c: Close up of antibody "crater"
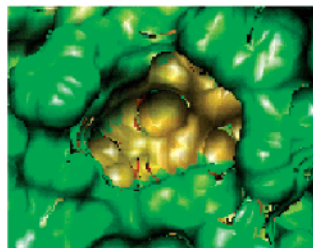
2d: Antibody "crater" complexed with trisaccharide: α-D-galactose(1-2)[α-D-abequose(1-3)]α-D-mannose (P1-OMe)
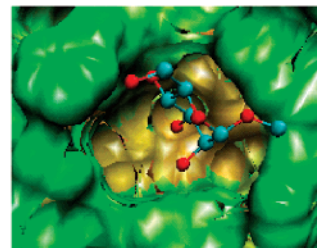
FIGURE 2. Geological and antibody pictures of the class crater. Geological crater photo (a) credit: U.S. Geological Survey, Department of the Interior/USGS, U.S. Geological Survey/photo by J. J. Griggs.

similar to the canyon in that it cut across the surface of the Fab region of antibody 4155 with at least one end of the pocket open to the aqueous environment that surrounds the antibody. However, the valley differs from the canyon in that the opening of the pocket is larger and shallower. The walls of the valley are not as steep as the walls of the canyon and they do not extend straight into the antibody, but rather slope toward the center of the binding pocket. The depth of the valley is smaller than the distance between the two valley walls.
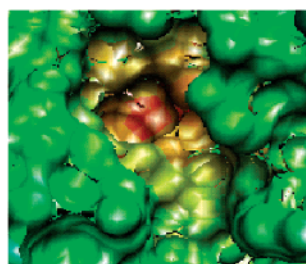
(12) Hoffmuller, U.; Knaute, T.; Hahn, M.; Hohne, W.; Schneider-Mergener, J.; Kramer, A. *EMBO J.* **2000**, *19*, 4866−4874.
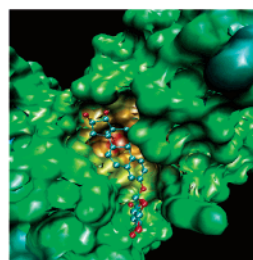
3a: Geological canyon

3b: Antibody "canyon" Fv4155 (PDB code: 1BFV)[11]

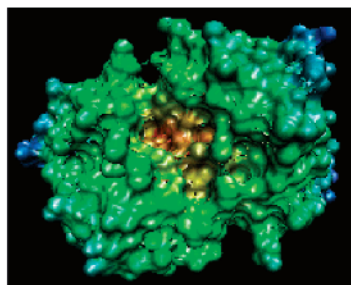3c: Close up of antibody "canyon"

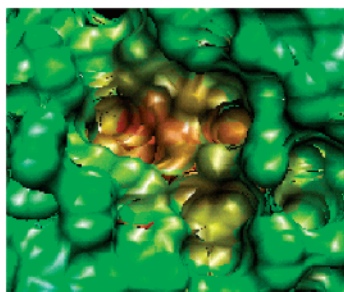3d: Antibody "canyon" complexed with estriol 3-(β-D-glucuronide)

FIGURE 3. Geological and antibody pictures of the class canyon. Canyon photo (a) used with permission of Robin Garrell.
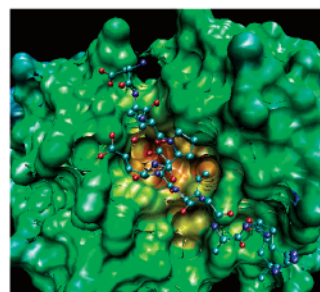
4a: Geological valley

4b: Antibody "valley" CB4-1 (PDB code:1HH6)[12]

4c: Close up of antibody "valley"

4d: Antibody "valley" complexed with an epitope-homologous peptide

FIGURE 4. Geological and antibody pictures of the class valley. Geological valley photo (a) used with permission of Jinhua Chen.

The Plain. Figure 5 consists of pictures of a geological plain (Figure 5a) and various views of the Fab of antibody NC10 (PDB code: 1NMB)[13] (Figure 5b−d). In this class, the antibody does not have a visually distinctive binding site but rather a bumpy surface. Like a r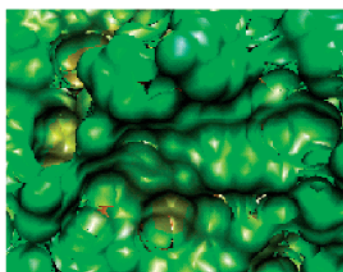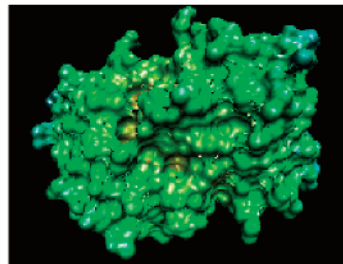eal plain, the surface is not perfectly smooth. The binding site contains bumps and ditches but no definable binding region.

(13) Malby, R. L.; Tulip, W. R.; Harley, V. R.; McKimm-Breschkin, J. L.; Laver, W. G.; Webster, R. G.; Colman, P. M. *Structure* **1994**, *2*, 733−746.

5a: Geological plain

5b: Antibody "plain" NC10 (PDB code:1NMB)[13]

5c: Close up of antibody "plain"

5d: Antibody "plain" complexed with a 5 residue linker and influenza virus neuraminidase
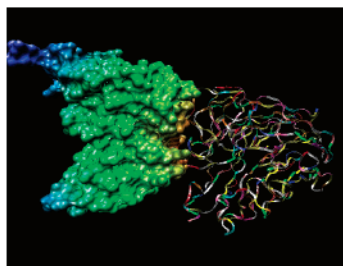

FIGURE 5. Geological and antibody pictures of the class plain. Plain photo (a) used with permission of Jinhua Chen.
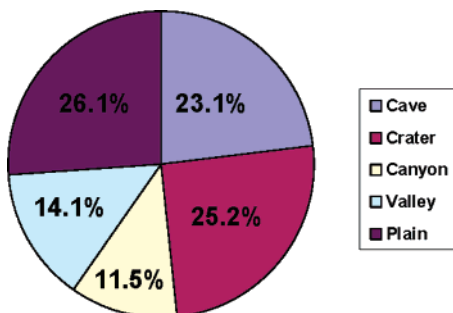

FIGURE 6. Percentage of antibodies per geomorphic class.

These distinguishing characteristics for each of the five antibody binding site classes were used to create an automatic algorithm for the classification of antibody binding sites, and are described in detail in the Algorithm Development section of this paper.

Out of the 229 antibody crystal structures currently available, following our automatic detection and classification procedure, 54 are cave type, 59 are crater type, 27 are canyon type, 33 are valley type, and 61 are plain type. There are 37 borderline structures; for these borderline structures, the algorithm gave the same numerical score to two classes since these antibodies have characteristics of both classes. In the statistical analysis, the borderline structures are divided in half and placed into both classes. For example, if a structure is valley/crater, 0.5 was given to the valley class and 0.5 to the crater class. Figure 6 shows graphically the percentage of antibodies that fall into a particular geomorphic class. Because the number of antibody crystal structures is a relatively small fraction of the colossal variety of potential antibodies, it is not known whether this distribution will be general.

TABLE 1. Comparison of Our Classification and Correlation with Ligand Type to Rees' and Thornton's Classifications and Correlations

ours	Rees'	Thornton's
cave−hapten	cavity−hapten	concave−hapten
crater−P/C/NA[a]		moderately concave−hapten
canyon	groove−P/C/NA	ridged−P/C/NA
valley		
plain−protein	plane−protein	plane−protein

[a] P/C/NA stands for protein/carbohydrate/nucleic acid.

The Supporting Information provides a complete list of antibodies used in this study. The antibodies are denoted by their Protein Data Bank codes followed by their corresponding binding site classification and antigen classification type.

The definitions of the classes defined here have some relationships to those that were defined earlier. A comparison of the previous and new classifications is shown in Table 1. The classification cave defined here corresponds qualitatively to Thornton's concave and Rees' cavity classes. Our crater corresponds to moderately concave (Thornton) and cavity (Rees). The new class, valley, was created to describe structures that resemble a canyon but are not as deep or narrow as the canyon. Together, the valley and the canyon class correspond to the class ridged (Thornton) and the class groove (Rees). Our plain corresponds to the class called plane by both Thornton and Rees. Our name refers to the plains common to the Midwest of the United States; these do more closely resemble the bumpy topography of these antibodies than the rigid smooth mathematical plane.

Correlation between Antigen Types and Antibody Classes. The relationship between the antigen type and antibody binding site topography was explored to determine whether a certain

TABLE 2. Number of Antigens Bound to Each Class of Antibody Classification[a]

antigen type	cave	crater	canyon	valley	plain	total
hapten	31	13	12.5	7.5	8	72
P/C/NA	3	20	5.5	12.5	3	44
protein	19	23.5	8.5	13	49	113

[a] 0.5 is due to borderline structures; P/C/NA stands for protein/carbohydrate/nucleic acid.

TABLE 3. Percentage (%) of Antigens Bound to Each Class of Antibody Classification

antigen type	cave	crater	canyon	valley	plain	total
hapten	43.1	18.1	17.4	10.4	11.1	100
P/C/NA[a]	6.8	45.5	12.5	28.4	6.8	100
protein	16.8	20.8	7.5	11.5	43.4	100

[a] P/C/NA stands for protein/carbohydrate/nucleic acid.

type of antigen preferentially binds to one type of antibody or elicits a type of antibody binding site topography.

The antigens in this study fall into three classes: hapten, peptide/carbohydrate/nucleic acid, and protein. The antigen classification is found in the literature for each of the antibody/antigen crystal structures. The hapten antigen is a small organic molecule, usually with less than 100 atoms. Peptides, carbohydrates, and nucleic acids are fragments of larger biological polymers and have similar shapes. A protein antigen is usually much larger than the binding site of the antibody.

Rees et al. first noted a relationship between binding site topography and the types of antigens bound for the 31 crystal structures available at the time.[5] They noted that antibodies, binding to protein antigens, are characterized by a flat binding site, while those binding to peptide and DNA antigens have a groove-like binding site and those binding to haptens have a cavity-like binding site.[5]

Thornton et al. used Kuhn's fractal atomic density measurement to detect the amounts of surface concavity and convexity on 45 structures.[6] They proposed four distinct classes by subdividing the cavity class proposed by Rees et al. into concave and moderately concave.[6] They found that haptens mainly bind to concave and moderately concave classes, peptides, nucleic acid, and carbohydrate bind more often to ridged antibodies, and proteins most frequently bind to planar antibodies.[6]

In this study, 229 antibody structures have been analyzed using a geomorphic classification scheme (cave, crater, canyon, valley, and plain) and a newly developed algorithm for automatic classification of antibody binding site. The antigen type was found to have some correlation with the shape of the binding site. Table 2 shows the number of antigens bound to each class of antibody classification. Table 3 is calculated from Table 2 and gives the percentage of antigens bound to each class of antibody classification. For example, for the peptide/carbohydrate/nucleic acid type of antigen, the percentage of cave (6.8%) is obtained by dividing the number of antigens (3) bound to a cave by the total number of this type of antigens (44).

The observation that hapten antigens bind mostly to the cave type antibodies suggests that small organic molecules prefer a narrow but deep binding site; the molecule is almost entirely lodged into the small and deep indenture. The preference of peptide/carbohydrate/nucleic acid antigens for crater type antibodies is consistent with the shape of the antigens. Proteins bind to the plain type antibodies because of the large, relatively flat surfaces. Our correlation is compared with those of Rees

TABLE 4. Number of Free and Bound Antibody Crystal Structures

antigen	free form	bound form
hapten	26	46
P/C/NA[a]	15	29
protein	48	65
total	89 (39%)	140 (61%)

[a]P/C/NA stands for protein/carbohydrate/nucleic acid.

and Thornton in Table 1. While we produced similar results for the "cave−hapten" and "plain−protein" correlation, the new classification indicates that protein/carbohydrate/nucleic acid type of antigens bind more often in crater antibodies.

Comparison between Free and Bound States. The 229 crystal structures we have investigated include both free and bound antibodies. As shown in Table 4, when the antigen is a hapten, 26 and 46 crystal structures are in the free and bound form, respectively; when the antigen is peptide/carbohydrate/nucleic acid, 15 and 29 crystal structures are in the free and bound form, respectively; when the antigen is protein, 48 and 65 crystal structures are in the free and bound form, respectively. Overall, the minority (39%) of the crystal structures is in the free form and the majority (61%) is in the bound form.

The generality of induced fit as a mechanism governing antibody−antigen interactions has frequently been discussed. We have investigated the conformational relationship of antibodies with and without antigen guests by comparing the topographic classes of the binding site of both the free and bound antibody structures. The results are shown in Table 5. There are only 10 entries in total, where both the crystal structures of the free antibody and bound antibody are available in our 229 data set. Interestingly, half of them (entries 1, 2, 4, 9, and 10) undergo class changes and half of them do not (entries 3, 5, 6, 7, and 8).

We first examined the antibodies that undergo class changes. Entry 1 is the esterolytic and amidolytic antibody 43C9.[14] Upon binding *p*-nitrophenol, the binding site changes from canyon to crater/cave. Entry 2 is the anti-digoxin monoclonal antibody 26−10.[15] Upon binding digoxin, the binding site changes from valley to cave. Entry 4 is an estradiol monoclonal antibody 10G6D6.[17] Upon binding 6-CMO-estradiol, the binding site changes from plain to valley. Entry 9 is a potent anti-human-TF monoclonal antibody 5G9,[21] and upon binding the tissue factor (TF), the binding site changes from cave to valley. Entry

(14) Thayer, M. M.; Olender, E. H.; Arvai, A. S.; Koike, C. K.; Canestrelli, I. L.; Stewart, J. D.; Benkovic, S. J.; Getzoff, E. D.; Roberts, V. A. *J. Mol. Biol.* **1999**, *291*, 329−345.
(15) Jeffrey, P. D.; Strong, R. K.; Sieker, L. C.; Chang, C. Y.; Campbell, R. L.; Petsko, G. A.; Haber, E.; Margolies, M. N.; Sheriff, S. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 10310−10314.
(16) Arevalo, J. H.; Stura, E. A.; Taussig, M. J.; Wilson, I. A. *J. Mol. Biol.* **1993**, *231*, 103−118.
(17) Monnet, C.; Bettsworth, F.; Stura, E. A.; Du, M. H.; Menez, R.; Derrien, L.; Zinn-Justin, S.; Gilquin, B.; Sibai, G.; Battail-Poirot, N.; Jolivet, M.; Menez, A.; Arnaud, M.; Ducancel, F.; Charbonnier, J. B. *J. Mol. Biol.* **2002**, *315*, 699−712.
(18) Rini, J. M.; Schulze-Gahmen, U.; Wilson, I. A. *Science* **1992**, *255*, 959−965.
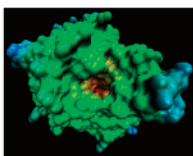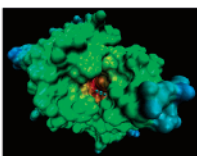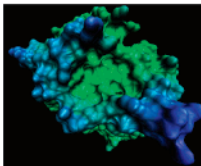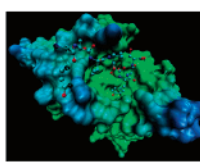(19) Herron, J. N.; He, X. M.; Ballard, D. W.; Blier, P. R.; Pace, P. E.; Bothwell, A. L.; Voss, E. W., Jr.; Edmundson, A. B. *Proteins* **1991**, *11*, 159−175.
(20) Kanyo, Z. F.; Pan, K. M.; Williamson, R. A.; Burton, D. R.; Prusiner, S. B.; Fletterick, R. J.; Cohen, F. E. *J. Mol. Biol.* **1999**, *293*, 855−863.
(21) Huang, M.; Syed, R.; Stura, E. A.; Stone, M. J.; Stefanko, R. S.; Ruf, W.; Edgington, T. S.; Wilson, I. A. *J. Mol. Biol.* **1998**, *275*, 873−894.
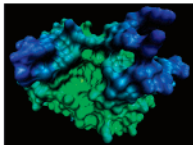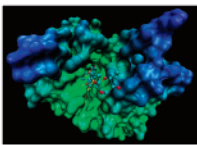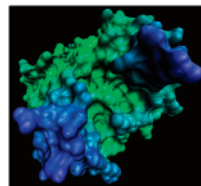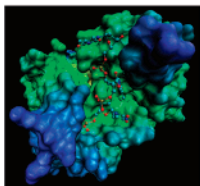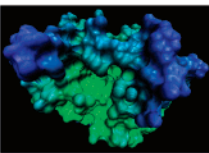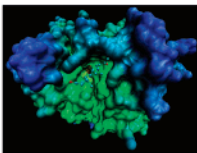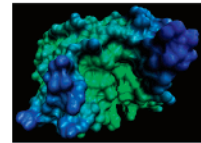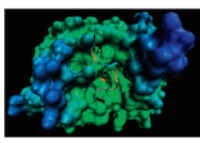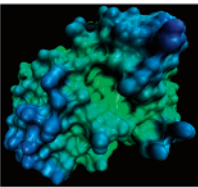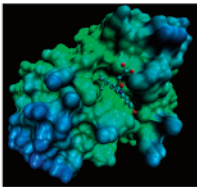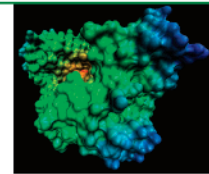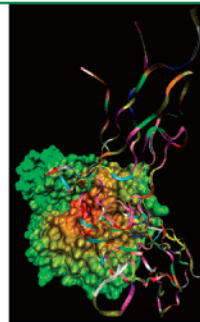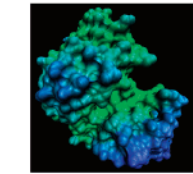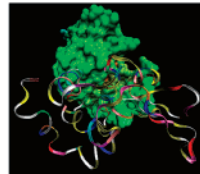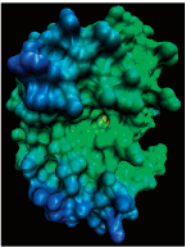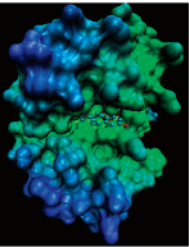
TABLE 5. Binding-Site Classes Compared between Free and Bound Antibody Crystal Structures

Entry	Antigen	Free form	Bound form	Entry	Antigen	Free form	Bound form
1[14]	Hapten	43C9 (Canyon)	43CA (Crater/Cave)	6[18]	Peptide	1HIL (Crater)	1HIN (Crater)
2[15]	Hapten	1IGI (Valley)	1IGJ (Cave)	7[19]	Nucleic acid	1NBV (Valley/Canyon)	1CBV (Valley)
3[16]	Hapten	1DBA (Plain)	1DBB (Crater/Plain)	8[20]	Protein	1CR9 (Cave)	1CU4 (Cave)
4[17]	Hapten	1JN6 (Plain)	1JNH (Valley)	9[21]	Protein	1FGN (Cave)	1AHW (Valley)
5[17]	Hapten	1JNL (Crater/Cave)	1JNN (Crater/Cave)	10[22]	Protein	1DQQ (Plain/Valley)	1DQJ (Cave)

10 is a monoclonal antilysozyme antibody HyHEL-63.[22] Upon binding the anti-hen egg white lysozyme (HEL), the binding site changes from plain/valley to cave.

These changes would be difficult to identify merely by inspection. The analysis by Thornton et al. has suggested that while conformational changes do occur when antibodies change from free state to bound state, they are not of great magnitude.[6] Our automatic classification of antibody binding site predicts that those changes do occur, and most often when a plain distorts toward a crater or valley to partially enclose the antigen.

We also examined the antibodies that experience few class changes. Entry 3 is the monoclonal anti-progesterone antibody DB3,[16] that is the only example of a binding site that is a plain both before and after binding. Entry 5 is an estradiol monoclonal antibody 17E12E5.[17] Both the free antibody and antibody-estradiol complex have a crater/cave binding site. Entry 6 is an antibody (17/9) with specificity for influenza virus hemagglutinin.[18] Both the free antibody and antibody−peptide immunogen complex have a crater binding site. Entry 7 is an autoantibody (BV04-01) with specificity for single-stranded DNA.[19] Both the free and antibody−deoxynucleotide complex have a valley binding site. Entry 8 is an anti-Syrian hamster prion protein (SHaPrP) monoclonal antibody 3F4.[20] Both the free and antibody-SHaPrP complex have a cave binding site.

Algorithm Development

Computational Methods: Automatic Classification of Antibody Binding Shapes. Using the criteria described above

(22) Li, Y.; Li, H.; Smith-Gill, S. J.; Mariuzza, R. A. *Biochemistry* **2000**, *39*, 6296−6309.
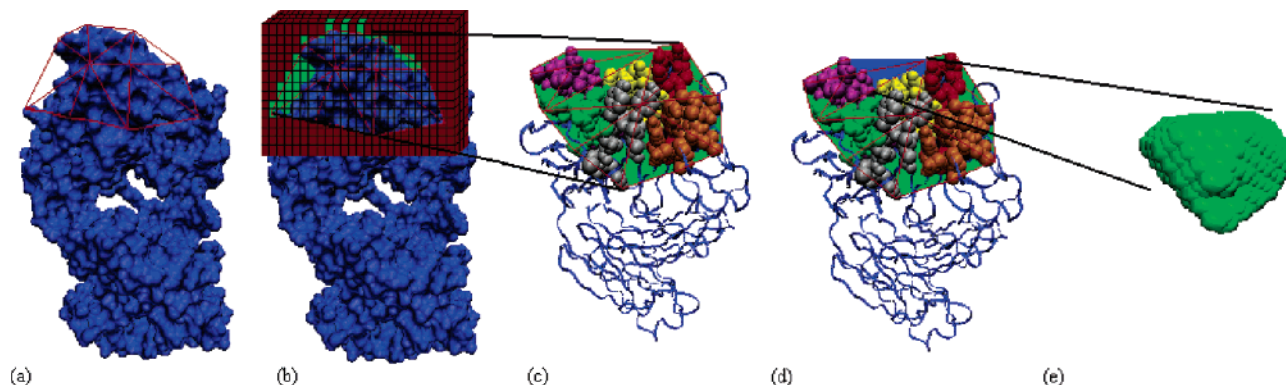
FIGURE 7. Automatic detection process algorithm used to identify the binding site region of the antibody. All of the steps (a–e) are described in more detail in the text: (a) the antibody with a 3D convex hull placed around the binding site; (b) a 3D grid placed around the convex hull, grid cubes lay inside (green region) or outside (red region) the convex hull; (c) the many disconnect objects/regions present on the binding site; (d) using morphological techniques from binary image analysis we isolate the largest (primary) object on the binding site; (e) the final object representing the three-dimensional shape of the binding site to be classified.

for each antibody class, an automatic classification scheme was developed. The automatic classification of the shape of antibody binding sites involves two steps: (I) automatic identification of antibody binding sites; (II) analysis and classification of binding site shape. In the following, we describe each one of these two steps in detail.

(I) Automatic Identification of Antibody Binding Sites. The process to automate the detection and classification of antibody binding sites begins by placing a convex hull around the binding site region on the antibody. The convex hull of a given three-dimensional shape is the smallest convex object that contains it. The convex hull is formed from the three-dimensional atomic coordinates of each antibody binding site given to us from the PDB file for the corresponding antibody. The hull is constructed using the Matlab command convhulln-(V), which is based on the Qhull algorithm.[23] Here, V is a vector of the form (x, y, z), where x, y, and z are the Cartesian atomic coordinates of those atoms making up the binding site (how these atoms are determined is described below). The convex hull is then represented as a set of indices that comprise the facets for the convex hull of the coordinate set V (see Figure 7a).

The next step in the binding site detection process is to construct a minimum sized three-dimensional grid to place around the binding site (see Figure 7b). The default size of each grid cube is 0.8 Å^3. This size was determined to clearly identify the binding site on the antibody structure and do so in a computationally efficient fashion. The size of each grid cube (or grid point) can be changed by the user if desired, both computational time and accuracy constraints should be considered. Through out this paper, the terms "cube" and "point" will be used interchangeably since a point in the algorithm can be thought of as the center of a cube. Each of the grid points may lie outside the convex hull or inside/underneath the convex hull. For a convex shape, the interior can be considered to be the set of points which are on the "positive side" of all the faces for the convex hull.[24] That is, when a test point is dotted (inner product) with the normal vectors of each plane, the dot product is positive.[24] On the other hand, if a point is on the outside of the convex hull, then the dot product of that point with at least one of the normal vectors of each plane will be zero or

TABLE 6. Our CDR Definitions (in some antibodies, chains L (light) and H (heavy) are replaced with chains A and B, respectively)

CDR	residues (Kabat numbering)
L1	L30–L33
L2	L50–L55
L3	L91–L96
H1	H30–H32
H2	H48–H56
H3	H94–H101

negative.[24] The algorithm used here identifies those grid points that are inside the convex hull, since these will fill in the different shape features that are present on the antibody binding site region. In addition, when this grid structure is being formed, the algorithm must avoid placing grid cubes/points where they overlap or collide with atoms in the antibody itself. Therefore, a grid point is added if and only if the distance between the prospective grid point and an atom center from the antibody obeys a simple rule, $\delta \leq (0.8 \text{ Å} + r)$, where r is the predetermined van der Waals radius of an atom in the antibody, and $\delta = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$, the ordinary Euclidean distance.[25] This process creates the object we call S_i, which is a set of (x, y, z) coordinates for the centers of all cubes in the object.

(Ia) Identification of Binding Site Atoms. The convex hull that is placed around the binding site region of the antibody is created from a set of atomic coordinates that compose the atoms on the binding site. We now discuss how these specific binding site atoms are identified. This identification starts by considering the complementarity determining regions (CDRs) of the binding site. Our CDR definitions are shown in Table 6.

Our CDR definition is slightly different from the one found in Thornton et al.'s work.[6] Thornton et al. note that some of the groupings of atoms in their original definition extend too far beyond the binding site region of the antibody, and as a result, the binding site shape can sometimes be misidentified.

(23) Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. T. *ACM Transactions on Mathematical Software* **1996**, *22*, 469–483.

(24) Garrity, M. "Matlab Central: 3D Objects and Random Point Generation", Retrieved December 29, 2004, from the Mathworks Central Archive: http://newsreader.mathworks.com/WebX?14@60.7NUMavZ9pRd.0@.eee175e.

(25) Winter, M. "Periodic Table: Scholar edition: links to elements: compounds (electronegatives and radii)", Retrieved July 28, 2004, from the University of Sheffield: www.webelements.com/webelements/scholar/elements/periodic-table/compounds.html.
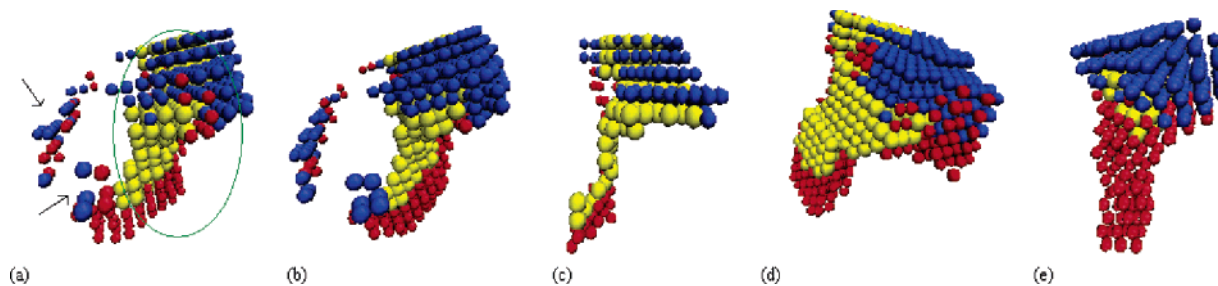
FIGURE 8. Processing steps taken to remove small objects and then rebuilding the primary binding site object circled in green on (a). The shapes have been qualitatively color coded according to the method described in the text, but are only accurate for (e): (a) the primary object with three groupings of smaller objects near to it; (b) we perform a morphological smoothing (closing then opening) with a cube shaped structuring element of 1 Å^3; (c) we then perform a morphological erosion with a sphere shaped structuring element of radius 2 Å; (d) rebuilding of the larger object by performing a dilation with a sphere-shaped structuring element of radius 3 Å; (e) cutting away of cubes that collide with atoms in the antibody itself and then performing morphological smoothing and finally color code the object appropriately.

In our case, we ignore those remote groups, and as a result, the amount of atoms that are under consideration in the binding site is decreased.

(Ib) Morphological Refinement of the 3D Binding Site Shape. Once the appropriate CDRs (as defined above) have been identified, the grid points which make up the shape of the 3D binding site (green region in Figure 7c) are refined. Oftentimes, the detected binding site contains many disconnected clusters of "green" cubes (inside the convex hull), since this area is usually not smooth, but quite "bumpy" with many high and low lying areas. The algorithm should detect and subsequently use only one uniform and connected shape on the binding site, multiple disconnected shapes could lead to inconsistencies in the eventual classification of these objects (see Figure 8a). The process used to eliminate these smaller less significant objects consists of three simple steps. First, the object that represents the binding site (S_i) can be very jagged looking, and might contain small gaps between cubes inside the object. This is because the process described above removes grid cubes that are touching an atom in the antibody, even if the majority of the cube is not intersecting the atom at all. To address this phenomenon, a classical morphological closing operation with a cube shaped structuring element of size 1 Å is performed, followed by a morphological opening.[26] The result is then a new shape $S_c := \text{open}\{\text{close}\{S_i\}\}$. Morphological closing/opening (or geometric smoothing) operations instead of, for example, a straight dilation operation is used because the smoothing operation will fill in the small gaps and "smooth" the overall shape of the object without significantly distorting it. The second step is to actually remove the small disconnected objects on the binding site and only keep the one primary object. This is accomplished through a series of additional morphological erosions and dilations with different sized structuring elements. The removal of the smaller objects begins by applying morphological erosion with a (discrete) sphere shaped structuring element of radius 2 Å. In this way the smaller objects are completely removed, but some portion of the largest object from S_c remains (see Figure 8c). This results in the shape $S_e := \text{erode}\{S_c\}$. The algorithm then rebuilds the larger object remaining from S_e by using a morphological dilation with a sphere shaped structuring element of radius 3 Å (see Figure 8d), and this results in shape $S_d := \text{dilate}\{S_e\}$. This results in an object that is not entirely accurate because some of the grid

cubes could be overlapping with atoms in the antibody, since the dilation might enlarge the binding site object beyond the size constraints of the binding region. To remove these overlapping grid cubes, the algorithm applies the same technique described above to check for collisions, and then applies another morphological smoothing operation to reduce the roughness of the shape. This results in the final shape which we call S_f. The reason that we decided to use a collection of morphological operations to remove the smaller objects instead of a more simple operation such as a closing/opening or a connected components detection technique was because we wanted to use varying sizes of structuring elements, which would not have been possible if we had used one of the more simplistic approaches. The sizes and shapes for all of the structuring elements used in this portion of the algorithm were developed though extensive testing and experimentation and found to produce very satisfactory results.

(II) Analysis and Classification of Binding Site Shape. The analysis of the shape starts by color coding the cubes which the object is made up of (see Figure 9b). Each cube is colored with one of three different colors: cubes which are on the surface of the object are colored either blue to mark them as hydro proximal (bordering the water environment in which the protein is submerged) or red to mark them as hydro distal (not bordering the water environment); cubes which are not on the surface are colored yellow. The color coding is represented by placing the 3D coordinates for the center of each cube for a particular color into individual matrixes that identify the color appropriately. The way that the surface points/cubes on the object are distinguished is by measuring the distance between the center of a cube in the object and the center of a cube that has been defined as being outside the convex hull. If the distance between these cubes is less than or equal to 1.5 Å, then that point on the object is marked as a surface point, otherwise the point is marked as an inner point and will be colored yellow. This parameter was determined through extensive experimentation with various binding site shapes, and has been found to produce very satisfactory results in separating inner points from points that compose the surfaces of these shapes. To then distinguish surface points which are either hydro proximal or hydro distal, we measure the smallest distance between the center of a cube on the surface and the centers of all of the cubes outside the convex hull. Then, all surface cubes which have $\delta \le (\text{smallest} + 3 \text{ Å})$ are marked as hydro proximal and cubes which have $\delta > (\text{smallest} + 3 \text{ Å})$ are marked as hydro

(26) Soille, P. *Morphological Image Analysis Principles and Applications*; Springer Publishing: Germany, 1999; pp 89−92, 163−164.
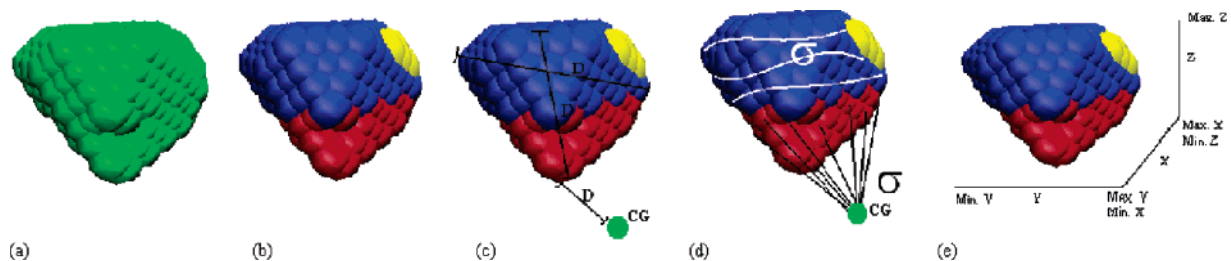
FIGURE 9. Different measurements that are taken on the binding site shape in order to classify it into one of the five categories. All images were generated using the VMD (Visual Molecular Dynamics) software, and although the shape is truly a collection of cubes, VMD represents the center of each cube as a point/sphere: (a) original binding site object; (b) color-coded object; (c) distances between various cubes; (d) standard deviations taken for the distances between all pairs of blue cubes, and all red cubes and the center of gravity; (e) maximum and minimum dimensions along each axis used to measure elongation for the shape.
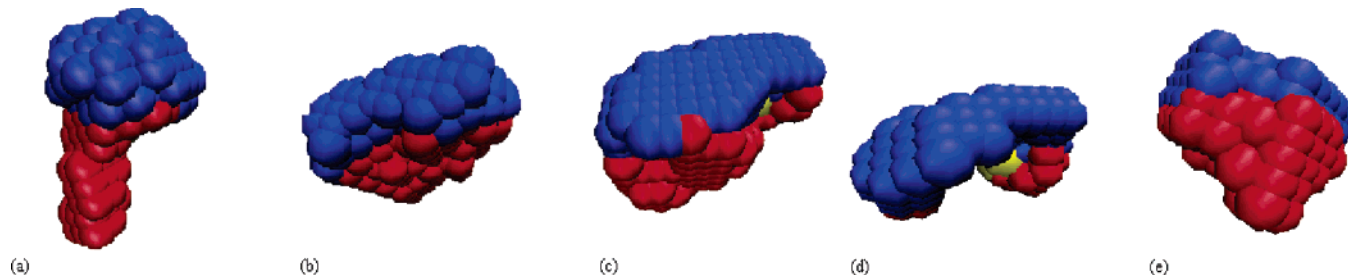
FIGURE 10. Examples of five classes detected and classified by our algorithm. (a) A cave produced from antibody 1HKL. (b) A canyon from antibody 1ADQ. (c) A typical valley produced from antibody 2IGF. (d) A plain from antibody 1DVF. (e) A crater from antibody 1FL5. These shapes were visualized by VMD software based on coordinates created by our algorithm.

distal. Here δ represents the Euclidean distance between a point outside the convex hull and one on the surface of the object, and smallest represents the minimum distance between a surface point and a point outside the convex hull. Again, these parameters were established through extensive testing and selected to produce good classification results.

The color coding is a first and necessary step in the analysis of the binding site shape, and its importance will become more apparent as the techniques used to further analyze and categorize this object are described. To formulate an effective technique to classify accurately the binding site into one of the five shape categories described before, key aspects of the binding site involved in a visual inspection and classification must be understood. Important primary aspects are the depth of the binding site region, its shape, and the size of this region. These visual detection parameters, and the results from our computations to classify the binding site shape automatically, are described below.

First, the volume of the shape is computed from volume $\approx N \times (0.8$ Å$)^3$, where N is the number of cubes composing the shape and 0.8 Å is the size of each cube. The volumes of various shapes, especially caves and plains, differ quite a bit and allow an easy and immediate way to separate these two shapes. Another simple measurement that can be computed is what percentage of the surface on the binding site object is red (hydro distal) versus blue (hydro proximal). In shapes that resemble caves the percentage of surface points/cubes which are blue is much lower than the percentage of surface points/cubes which are red. This is because caves typically have a small opening; thus, the amount of the surface exposed to the solvent environment is small in comparison, for example, to a valley or canyon. In these latter shapes, much of the surface is exposed to the solvent environment since these shapes are long and wide.

To better analyze the shape and size of the binding site object, distances between cubes were measured (see Figure 9c). The

algorithm computes the largest distance between a red and blue cube, along with the largest distance between any pair of blue cubes. The objective here is to understand how spread out the shape is, and the distance between a red and blue cube is typically larger for caves and canyons than for the other three shapes; these two objects usually penetrate deeply into the antibody and there is a large degree of separation between the top and bottom of the objects. On the other hand, the distance between pairs of far away blue cubes is used to differentiate between canyons and valleys, where canyon shapes are conventionally larger than valleys. The last distance-based measure that the algorithm computes is the smallest distance between a red cube and the center of gravity for the antibody. The center of gravity is computed as the point at coordinate $(X\mu, Y\mu, Z\mu)$, where $X\mu$, $Y\mu$, $Z\mu$ are the mean positions of the x, y, and z coordinates for each atom in the antibody. This measurement helps to capture the depth of the shape and it is particularly important for distinguishing caves, which go deep into the antibody and have a small distance to the center of gravity.

The elongation on the shape, which is important for distinguishing valleys and canyons, is measured by finding the largest and smallest values along each of the axes, subtracting the two pairs of axis values from each other, and comparing these values to other subtracted axis values. For example, the highest X axis value (X_h) is subtracted from the lowest X (X_l) axis value, and then this is subtracted from the difference between the highest (Y_h) and lowest Y (Y_l) axis values; $|(X_h - X_l) - (Y_h - Y_l)|$. These differences are smaller for objects that are more compact, and larger for objects that are more spread out and elongated. Figure 8e shows the maximum and minimum values along each axis, which are used to analyze the elongation of the object. The algorithm also performs some more formal measures of spread by computing the standard deviation for the distances between all pairs of blue cubes (see Figure 9d). In binding site shapes that resemble caves or craters, the standard deviation is

TABLE 7. Eleven Different Parameters Used for Categorizing the Binding Site

	cave	crater	canyon	valley	plain
volume[a] (Å^3)	200−600	400−535	≥600	300−500	≤300
% blue[b]	0.30−0.49	0.47−0.70	0.60−0.80	0.47−0.70	>0.70
% red	0.50−0.70	0.30−0.53	0.20−0.40	0.30−0.53	<0.30
σ CG[c]	≥2.0	1.10−1.50	k	<1.1	≤1.07
σ blues[d]	≤1.3	1.40−1.70	≥2.0	1.7−2.0	k
depth red[e]	≤24	24−27	< 20	>27	≥27
dist R and B[f]	>8.0	≤8.0	≥10	k	<6.0
dist blue[g]	<6.0	6.0−9.0	>9.0	>7.0	<4.0
diff yx[h]	15−17	16−19	>19	<13	≤8.0
diff yz[i]	12−18	< 12	>18	7.0−11	≤7.0
diff xz[j]	7.0−12	10−15	≥16	15−18	≤6.0

[a] Volume is computed by $N \times (0.8 \text{ Å})^3$, where N is the number of cubes composing our binding site object. It should be noted that the volume for plains is by far the smallest because of the flatness associated with this type of binding site, and canyons typically have a large volume because they are quite deep like a cave, but are also much wider/longer. [b] The % blue measurement is the percentage of cubes on the shape which are hydro proximal and can be used to separate similarly shaped objects such as caves from craters. Since it has been noticed that craters have slightly larger openings than caves, the % blue for the craters is somewhat larger than the % blue for the caves. In addition, plains usually have extremely high % blue values because these shapes are not very deep, and as a result, much of the surfaces on the plains are considered hydro proximal. [c] σ CG measures the standard deviation between a red cube and the center of gravity for the antibody. [d] σ blues measures the standard deviation for the distances between all pairs of blue cubes and was only used to distinguish between valleys and canyons. [e] Depth red is the smallest distance between a red cube and the center of gravity for the antibody. Although the numbers in this measurement are comparable to one another due to the similarities in antibody structures, the cave and canyon typically have smaller values compared to the other three shapes due to the great depth that these shapes often penetrate into the antibody. [f] Dist R and B is the largest distance between a red and blue cube. This is useful for separating valleys or canyons from craters; this measurement is smaller for craters than it is for a valley or a canyon, since valleys and canyons have large diagonal distances between blue and red cubes due to their stretched out form. [g] Dist blue is the largest distance between any pair of blue cubes. [h] Diff yx is the elongation measurement computed by $|(X_h - X_l) - (Y_h - Y_l)|$. [i] Diff yz is computed by $|(Y_h - Y_l) - (Z_h - Z_l)|$. [j] Diff xz is computed by $|(X_h - X_l) - (Z_h - Z_l)|$, where X_h, Y_h, Z_h are the largest values along each of the x, y, and z axis, respectively, and X_l, Y_l, Z_l are the smallest values along each of the x, y, and z axis, respectively. [k] Parameter was not used in determining the specified shape. All units on parameters except for the % blue and % red are measured in angstroms.

very small because many of the points are close together, and thus, the distances are similar to one another. This is also true for plains which are usually a very small shape on the binding site. The opposite is true for canyons and valleys that have more elongated shapes, with more variation in the distances between pairs of hydro proximal (blue) cubes. The algorithm also computes the standard deviation of the distances from the red cubes (hydro distal) to the center of gravity (see Figure 9d). In shapes such as caves, there are very large distances to the center of gravity at the top of the shape, but small distances to the center of gravity at the bottom. This is because objects that resemble caves are typically shaped like a shaft (see Figure 10a). As a result of this great variation in distances, this standard deviation is usually larger for caves than any other shape.

Table 7 summarizes the numbers and parameters used for each of the geometric calculations based on the predicted shape of the binding site. At the end of the algorithm, each of the five categories of shapes is given a score between 0 and 1, with 0 being the lowest. This score indicates how likely the binding site object is to fit into each one of these categories. The score is computed by giving one point to a specified shape if certain criteria are satisfied for that category. The final score is then NP/TP, where NP is the number of points given to a certain category, and TP is the total number of points available for that category. In Figure 10, examples of typical binding site shapes are shown color-coded based on the method described above.

Conclusion

The algorithm developed in this work has been tested on the antibody crystal structures available up to April 2002, and it will be applicable to any antibody crystal structures reported afterward. The rapid and automatic classification of antibody binding site, when used cautiously, should prove useful in choosing a germ-line antibody to bind a specific antigen, such as a hapten when immunizing catalytic antibodies, and in the design of antigens for a known antibody.

An automatic binding site topography classification system based on easily recognizable features of the Earth's surface has been developed to quantify the similarities between types of binding pockets. The antibodies are classified into five categories according to the binding site topography: cave, crater, canyon, valley, and plain. The antigens to which the antibodies bind are classified into three groups: hapten, peptide/carbohydrate/nucleic acid, and protein. Different classes of antigens have a preference for one class of binding sites. Hapten antigens have a preference for cave binding pockets, protein antigens are bound preferentially in plain binding sites, and carbohydrate/peptide/nucleic acid antigens are bound mostly in crater binding sites. The most general observation is that haptens prefer small and deep pockets while larger molecules such as peptide/nucleic acid/carbohydrates and proteins bind mostly to shallow and less defined binding sites. The conformational change of antibodies on binding site has also been examined by comparing the topographic classes of the binding site of both the free and bound antibody structures. The automatic algorithm is able to identify class changes on bindings that have not been noticed previously.

Supporting Information Available: List of antibodies analyzed and classified. This material is available free of charge via the Internet at http://pubs.acs.org.

JO052659Z