Origins of specificity and affinity in antibody–protein interactions

Hung-Pin Peng^{a,b,c}, Kuo Hao Lee^a, Jih-Wei Jian^{a,b,c}, and An-Suei Yang^{a,1}

^aGenomics Research Center, Academia Sinica, Taipei 115, Taiwan; ^bInstitute of Biomedical Informatics, National Yang-Ming University, Taipei 11221, Taiwan; and ^cBioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan

Edited by Barry Honig, Howard Hughes Medical Institute, Columbia University, New York, NY, and approved May 19, 2014 (received for review February 6, 2014)

Natural antibodies are frequently elicited to recognize diverse protein surfaces, where the sequence features of the epitopes are frequently indistinguishable from those of non-epitope protein surfaces. It is not clearly understood how the paratopes are able to recognize sequence-wise featureless epitopes and how a natural antibody repertoire with limited variants can recognize seemingly unlimited protein antigens foreign to the host immune system. In this work, computational methods were used to predict the functional paratopes with the 3D antibody variable domain structure as input. The predicted functional paratopes were reasonably validated by the hot spot residues known from experimental alanine scanning measurements. The functional paratope (hot spot) predictions on a set of 111 antibody–antigen complex structures indicate that aromatic, mostly tyrosyl, side chains constitute the major part of the predicted functional paratopes, with short-chain hydrophilic residues forming the minor portion of the predicted functional paratopes. These aromatic side chains interact mostly with the epitope main chain atoms and side-chain carbons. The functional paratopes are surrounded by favorable polar atomistic contacts in the structural paratope–epitope interfaces; more than 80% these polar contacts are electrostatically favorable and about 40% of these polar contacts form direct hydrogen bonds across the interfaces. These results indicate that a limited repertoire of antibodies bearing paratopes with diverse structural contours enriched with aromatic side chains among short-chain hydrophilic residues can recognize all sorts of protein surfaces, because the determinants for antibody recognition are common physicochemical features ubiquitously distributed over all protein surfaces.

protein antigenic site | interface hot spot | paratope prediction | epitope prediction | functional epitope

It is incompletely understood as to how functional antibodies can almost always be elicited against unlimited possibilities of protein antigens from a limited repertoire of antibodies. Antibodies provide protection against foreign protein antigens by recognizing the antigen proteins with exquisite specificity and remarkable affinity, but the principles underlying the antibody affinity and specificity remain elusive. Consequently, current antibody discoveries are by and large limited by the uncontrollable animal immune systems (1) or by the recombinant antibody libraries with relatively infinitesimal coverage of the vast combinatorial sequence space in antibody–antigen interaction interfaces (2). In developing the efficacy of a therapeutic antibody, optimizing the affinity and specificity of the antibody–antigen interaction mostly relies on selecting and screening from a large pool of random candidates. As antibodies are becoming the most prominent class of protein therapeutics (3), a better understanding of the principles governing antibody affinity and specificity will facilitate in understanding humoral immunity and in developing novel antibody-based therapeutics.

Antibody paratopes are enriched with aromatic residues. Tyrosyl side chains are overpopulated on the paratopes, noticeable on solving the first structures of antibody–antigen complexes (4). Surveys thereafter showed that Tyr and Trp frequently occur

in the putative binding regions of antibodies as determined from structural and sequence data (5). Recent analyses of more than 100 high-resolution antibody–antigen complexes in the Protein Data Bank (PDB) confirm a similar conclusion that aromatic residues (Tyr, Trp, and Phe) are substantially enriched in antibody paratopes (6, 7). The fundamental role of the tyrosyl side chains in antibody–antigen recognition has been demonstrated (8), with the functional antibodies selected and screened from the minimalist designs of antibody complementarity determining region (CDR) libraries with only a small subset of amino acid types (Tyr, Ala, Asp, and Ser) (9) or with binary code (Tyr and Ser) (10).

Interactions involving aromatic side chains on the CDRs of antibodies with epitope residues on protein antigens have been demonstrated to contribute energetically to antibody–antigen recognition. Alanine scanning of the antibody paratope residues of the FvD1.3–hen egg white lysozyme (HEL) and FvD1.3–FvE5.2 (anti-idiotypic antibody) complexes and shotgun alanine scanning assessing the energetic contributions of paratope residues to VEGF and human epidermal growth factor receptor 2 (HER2) binding indicated that around half of the hot spots ($\Delta\Delta G \geq 1$ kcal/mol) are aromatic residues (20/40) (11, 12). Double-mutant cycle experiments dissecting the residue-pair coupling energies between the epitope and paratope for the two antibody–antigen complexes also indicated the predominant energetic contribution of the aromatic side chains (9/11) in the antibody–antigen interactions (13). These energetic assessments suggest that aromatic side chains contribute a substantial portion of the affinity of the antibody–antigen complexes in general. These results are consistent with the survey indicating that aromatic residues, in

Significance

Natural antibodies perform their biological function by recognizing all sorts of foreign proteins—seemingly unlimited structural and sequence diversities in antigens can be recognized by a limited repertoire of antibodies, for which the sequence and structure are relatively homogeneous. We found that the energetically critical epitope portions are largely composed of backbone atoms, side-chain carbons, and hydrogen bond donors/acceptors. These key components are ubiquitous on protein surfaces and can be recognized by the enriched aromatic side chains and, to a lesser extent, short-chain hydrophilic residues on the antibody paratopes; antibodies, with relatively limited sequence and structural diversities in the antigen binding sites, can recognize unlimited protein antigens through recognizing the common physicochemical features on all protein surfaces.

Author contributions: A.-S.Y. designed research; H.-P.P., K.H.L., and A.-S.Y. performed research; H.-P.P., K.H.L., and J.-W.J. contributed new reagents/analytic tools; H.-P.P. and A.-S.Y. analyzed data; and H.-P.P., J.-W.J., and A.-S.Y. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹To whom correspondence should be addressed. E-mail: yangas@gate.sinica.edu.tw.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1401131111/-DCSupplemental.

particular Tyr and Trp, account for a large portion of the hot spot residues in protein–protein interactions (14, 15).

Aromatic side chains interact favorably with diverse functional groups in natural amino acids, underlying the affinity and specificity of the antibody–antigen recognition through a cumulative collection of relatively weak noncovalent interactions. The aromatic side chains interact with other aromatic side chains through face-to-edge or parallel π -stacking, with positively charged side chains through the cation– π interaction, with backbone and side-chain hydrogen bond donors through hydrogen bonding to aromatic π -systems (X–H– π interaction, X = N, O, S), with alkyl carbons through the C–H– π interaction, with sulfur-containing side chains through sulfur–arene interactions, and with negative charged side chains through anion– π interactions (16, 17). Although each of the above mentioned interactions is relatively weak, on the order of a few kilocalories per mole in model systems (16, 17), the cumulative sum of the interactions involving the aromatic side chains can reasonably account for the binding energy of 12 kcal/mol for a typical antibody–antigen interaction of $K_d \sim 1$ nM at room temperature aqueous environment. Moreover, the specific preferences of the spatial geometries for the interacting functional groups involving aromatic side chains (see refs. 16–18 and references therein) also underlie the specificity of antibody–antigen recognitions. Direct hydrogen bonds bridging across the antibody–antigen interaction interface are expected to contribute to both the binding affinity and specificity, but the removal of an interface hydrogen bond is frequently inconsequential to the binding specificity and affinity due to compensating water-mediated interactions (13). These results suggest that the 3D distribution of the paratope aromatic side chains by and large determines the affinity and specificity of the antibody–antigen interaction.

Known epitopes on antigens, on the other hand, are not easily distinguishable from the solvent accessible surfaces of protein structures. A recent review of the public domain conformational epitope prediction algorithms, for which the performances were compared with an independent test set for benchmarking, shows that the conformational epitope prediction problem remain challenging, with an area under the curve (AUC) ranging from 0.567 to 0.638 and accuracy from 15.5% to 25.6% (19). This moderate success rate was attributed to incomplete understanding of the essence of the epitopes (6). It has been well accepted that the solvent accessible and protruding surface regions are more likely to be conformational epitopes (20–23) and that the epitopes encompass substantially more loop residues than α -helix and β -strand residues (23, 24). By contrast, the conclusions from various studies on the amino acid composition of conformational epitopes are not consistent (6), in large part due to the fact that the epitope amino acid composition is not particularly distinguishable from the nonantigenic solvent accessible surface area (6, 7, 22, 23, 25). The contradiction has been discussed recently (25), indicating that the physicochemical complementarity between the paratopes and the epitopes are strikingly incomparable, with overwhelmingly emphasized tyrosyl side chains in all CDR loops (25).

The goal of this work is to understand how a natural repertoire of antibodies, for which the sequence and structure are relatively limited in variation, can recognize protein antigens with seemingly unlimited structural and sequence diversities. An extensive examination on the monoclonal antibodies elicited with a set of model antigens has concluded that a protein antigen surface consists of overlapping conformational epitopes forming a continuum; that is, no inherent property of the protein molecule could restrict antigenic site locations on the protein surface (26). It would be conceivable that antibodies recognize a common feature shared by all protein surface sites, and as such, a relatively limited population of antibodies could recognize limitless protein antigen surfaces. That is, protein surfaces are not as di-

verse as one would expect by looking at the protein sequences. However, this common feature on protein surface recognizable by antibodies is not known. Although aromatic side chains are known, in principle, to be able to interact favorably with wide varieties of functional groups in natural amino acids (see above), atomic details of the paratope aromatic side chains interacting favorably with diverse epitope surfaces have not been systematically analyzed. In addition, residues with short hydrophilic side chains (Ser, Thr, Asp, and Asn) are known to be enriched alongside the aromatic side chains in the paratopes (5, 24, 27), but the roles of these short hydrophilic side chains in antibody–antigen interactions have not been systematically investigated. More importantly, it has been well accepted that only hot spot residues in an antigen combining site of an antibody, i.e., the residues in the functional paratopes, are indispensable for the antibody–antigen interaction; side chains contacting the antigen (i.e., structural paratope residues) outside the functional paratope can frequently be truncated to C β carbon without affecting the antibody–antigen interaction (11, 13, 15, 28). To search for the relevant protein surface features recognizable by antibodies, it is desirable to first elucidate the principles governing the interactions for the functional paratopes with the corresponding functional epitopes. Such studies require a large number of well-defined functional paratopes and functional epitopes, but only a small number have been determined with the labor-intensive alanine scanning experiment (11–13, 29). To circumvent the scarcity of the experimental data, we use computational methods to predict the functional paratopes/epitopes in antibody–protein complex structures so that the key interactions involving the hot spots in the antibody–protein interactions can be elucidated, at least to the reliable extent depending on the functional paratope prediction accuracy.

In this work, we applied computational methods to predict functional paratopes on antibody variable domains and analyzed the key atomistic contact pairs in the functional paratope–epitope interfaces. Although the structural paratope–epitope interfaces can be defined from the known complex structures, the functional binding interfaces involving hot spot residues are unknown experimentally and need to be defined with computational predictions. One set of predictions was carried out with our previously published computational method [In-silico Molecular Biology Lab–protein–protein interaction (ISMBLab-PPI)], where the protein–protein interaction confidence level (PPI_CL) for protein surface atoms to participate in protein–protein interaction is strongly correlated ($r^2 = 0.99$) with the averaged burial level of the atoms in the PPI interfaces (30). Another set of predictions was carried out with a recently published random forest algorithm, prediction of antibody contacts (proABC) (31), which was trained specifically with antibody–antigen complex structures in PDB with additional information from antibody germ-line family sequences, CDR residue positions, multiple antibody sequence alignments, CDR lengths and canonical structures, and antigen volume. Both sets of predicted functional paratope–epitope interfaces consistently led to the conclusion that antibodies, with relatively limited sequence and structural diversities in the antigen binding sites, can recognize unlimited protein antigens through recognizing the common and ubiquitous physicochemical features on all protein surfaces. The implication is that a limited repertoire of antibodies bearing paratopes with diverse structural contours enriched with aromatic side chains among short-chain hydrophilic residues can recognize all sorts of protein surfaces.

Results and Discussion

Paratopes Share Common Features Recognizable in PPI Interfaces: Epitope Locations Are Not Restricted by Any Inherent Property of the Protein Molecule. A subset of the principles governing non-antibody PPIs is also applicable in antibody–protein interactions. The principles have been embedded in the ISMBLab-PPI pre-

dictors, for which the machine learning algorithms and the benchmarks for protein–protein interaction site predictions have been published (30); a brief summary of the prediction method is included in the *SI Materials and Methods*. The protein antigen recognition sites (paratopes) on the antibody structures from the S111 dataset, which is the representative set of 111 antibody–protein complex structures listed in a previously published work (32) (*SI Materials and Methods*), can be predicted with significant average accuracy using the ISMBLab-PPI predictors. Quantitatively, Fig. 1*A* shows that the PPI_CLs for paratope surface atoms are strongly correlated (Pearson's correlation coefficient = 0.95, $r^2 = 0.9$) with the percentage change of the solvent accessible surface areas (dSAs; Eq. S1) of the atoms due to antigen binding. More prediction results are summarized in Fig. S1, which shows that antigen binding sites can be predicted on antibody surfaces to an extent based on the general protein–protein interaction principles embedded in the PPI prediction algorithms, and the protein atom types from the aromatic side chains (Tyr, Trp, Phe, and to some extent His) are predicted with the highest accuracy (Fig. S1 *A–D*), indicating that the paratopes share the common PPI features mostly involving the aromatic amino acid types.

The same set of PPI principles is not applicable in epitope prediction. The correlation between PPI_CL and dSASA for the surface atoms in the epitopes is statistically insignificant, as shown in Fig. 1*B* ($r^2 = 0.12$). This result, along with prediction results showing low prediction accuracies on the basis of amino acid types and protein atom types in epitopes in Fig. S1 *A–D*, indicates that epitopes are fundamentally different from paratopes in terms of determinants in protein–protein interactions. However, how could a paratope frequently recognizable as a potential PPI interface combine with the corresponding protein antigen surface unrecognizable as a PPI site? One likely hypothesis is that the antigenic surfaces recognizable by antibodies are composed of common and ubiquitous features shared by all protein surfaces, rendering the machine learning algorithms in the ISMBLab-PPI predictors incapable of contrasting the likelihood for antibody binding on a protein antigen surface with indistinguishable features. This hypothesis is consistent with the extensively experimentally validated conclusion that a protein antigen surface consists of overlapping conformational epitopes

forming a continuum with no inherent property of the protein molecule restricting antigenic site locations on the protein surface (26). This hypothesis can be validated, or challenged, by examining the interface subareas energetically responsible for the antibody–protein interactions, to see if the functional epitopes are indeed composed of common and ubiquitous features shared by all protein surfaces.

Validation of Computationally Predicted Hot Spot Residues in Antibody CDRs. To test the hypothesis above, it is essential to first define the functional paratopes on antibody surfaces in the S111 dataset with computational methods and valid the computational predictions with experimental data. Residues frequently recognizable in antigen binding sites on antibodies have been speculated to have energetic roles in stabilizing the antibody–antigen interaction interfaces, as demonstrated in the strong correlation between residues with increasing Antibody i-Patch score and those with stronger antibody–antigen interaction energy calculated with FoldX (33). In this section, we will test the correlation of the predicted functional paratopes in antibody–protein interactions with experimental measurements. The alanine scanning experimental data for the following three antibody–protein complexes were used for validating the computational predictions of hot spot residues: (i) anti-HEL antibody FvD1.3, which recognizes both HEL (PDB code: 1VFB) and FvE5.2 (PDB code: 1DVF), with largely overlapped paratopes (11); (ii) the two-in-one antibody recognizing both VEGF and HER2 through overlapped paratopes in the CDRs (PDB code: 3BE1 and 3BDY) (12); and (iii) anti-HEL antibody (HyHEL-63; PDB code: 1DQJ) (29). These datasets are suitable for the hypothesis validation because of the extensive alanine scanning measurements over a large number of CDR residues in each of the antibodies and the availability of the antibody–antigen complex structures.

Fig. 2 *A–F* compares the actual hot spot residues (alanine scanning $\Delta\Delta G > 1$ kcal/mol) with predicted hot spot residues defined by the threshold $\text{PPI_CL} \geq 0.45$ and threshold proABC_CP (contact probability) $\geq 80\%$. Conventionally, $\Delta\Delta G > 1$ kcal/mol is frequently used as the threshold for defining hot spot residues because this value is marginally greater than the uncertainty of the experimental alanine scanning $\Delta\Delta G$ measurement. The alanine scanning data are summarized in Fig. 2 *A–C* for the anti-HEL/FvE5.2 antibody (FvD1.3), the anti-VEGF/HER2 antibody (bH1), and the HyHEL-63 antibody, respectively. The predicted paratope hot spot residues defined with the optimal threshold $\text{PPI_CL} \geq 0.45$ for the CDR residues in each of the antibodies are highlighted in Fig. 2 *D–F* for the corresponding antibody. The predicted hot spot residues defined with the optimal threshold $\text{proABC_CP} \geq 80\%$ are also shown in the figures for comparison. Overlapped predictions are colored in orange.

The results in Fig. 2 indicate that the predicted functional paratopes defined by $\text{PPI_CL} \geq 0.45$ with the ISMBLab-PPI predictor correlate, to the optimal extent, with the experimental functional paratopes defined by alanine scanning measurements with $\Delta\Delta G > 1$ kcal/mol. Because the input for the ISMBLab-PPI prediction requires only the 3D structure of the antibody variable domains, it is conceivable that not all of the residues predicted to be able to contribute energetically to antigen recognition on an antibody would be simultaneously used in binding to its corresponding antigen. These false positives undermine the prediction accuracy: precision, $[\text{TP}/(\text{TP} + \text{FP})] = 0.68$ for the overall predictions shown in Fig. 2. Moreover, only 41% of experimental hot spot residues are considered as theoretical hot spots: sensitivity, $[\text{TP}/(\text{TP} + \text{FN})] = 0.41$ for the overall predictions shown in Fig. 2. Lowering PPI_CL threshold reduces the number of false negatives, but this also results in more false positives. As such, the threshold $\text{PPI_CL} \geq 0.45$ is selected so that the Matthews correlation coefficient (MCC) reaches a maxima (0.43) for the

Fig. 1. Prediction benchmarks of antibody/antigen binding sites on antibody/antigen surfaces. (A) Correlations of PPI_CL of antibody surface atoms to atomic burial in antibody–antigen interfaces. Atom-based PPI_CL range (shown in the x axis of the panel) for antibody surface atoms is correlated to the averaged burial level (measured by dSASA) of the subgroup of atoms in the antibody–antigen complexes predicted within the confidence level range. The correlation is shown by the square symbols, corresponding to the y axis on the right of the panel. The distribution of the atom-based predictions as shown by the diamond symbols, corresponding to the y axis on the left of the panel, is plotted against the PPI_CL range in the x axis. The data were derived from the independent test with the ISMBLab-PPI predictors on the antibodies in the S111 dataset. (B) The same as in A for the antigens in the S111 dataset. Additional benchmarks are shown in Fig. S1. The prediction algorithm and parameters follow the optimal settings without modification as previously described (30).

Fig. 2. Comparison of the functional paratopes with the predicted hot spot residues on the antibody structures. (A) The residues of the functional paratopes on antibody FvD1.3 (PDB code: 1VFB) are highlighted with carbon atoms colored in yellow, green, and cyan. The residues with carbon atoms colored in cyan and green are the hot spot residues ($\Delta\Delta G > 1$ kcal/mol) interacting with lysozyme. The residues with carbon atoms colored in yellow and green are the hot spot residues ($\Delta\Delta G > 1$ kcal/mol) interacting with FvE5.2. (B) The residues of the functional paratopes on antibody bH1 (PDB code: 3BDY) are highlighted in color. The residues with carbon atoms colored in cyan and green are the hot spot residues ($\Delta\Delta G > 1$ kcal/mol) interacting with VEGF. The residues with carbon atoms colored in yellow and green are the hot spot residues ($\Delta\Delta G > 1$ kcal/mol) interacting with HER2. (C) The residues of the functional paratope of the anti-lysozyme antibody HyHEL-63 (PDB code: 1DQJ) ($\Delta\Delta G > 1$ kcal/mol) are highlighted in cyan. (D–F) The carbon atoms of the residues in the potential functional paratope (PFP)(PPI) ($PPI_CL \geq 0.45$) are colored in orange and pink. The carbon atoms of the residues in the PFP(proABC) ($proABC_CP \geq 80\%$) are colored in orange and magenta. Oxygen atoms are colored in red, and nitrogen atoms are colored in blue in A–F. The prediction results are compared with the actual hot spots in the table below the structures, where TP (true positive), FP (false positive), TN (true negative), FN (false negative), PRE (precision), ACC (accuracy), SEN (sensitivity), MCC (Matthews correlation coefficient), SPC (specificity), and F1 (F-score) are defined in Eqs. S4–S9.

prediction of the hot spot residues defined by the threshold of $\Delta\Delta G > 1$ kcal/mol in the three sets of alanine scanning data shown in Fig. 2.

Fig. 2 D–F compares the predicted hot spot residues with the two computational methods. The optimal $proABC_CP$ threshold (80%) is selected to maximize the MCC of the $proABC$ predictions while avoiding the predicted functional paratopes to encompass the complete structural paratopes ($proABC_CP < 70\%$). The $proABC$ predictions are superior in optimal MCC (0.54 vs. 0.43; Fig. 2), due to higher sensitivity (0.67 vs. 0.41). However, the precision of the ISMBLab-PPI predictions is better (0.68 vs. 0.62). Overall, $proABC$ predicted 45 hot spots and ISMBLab-PPI predicted 25 hot spots, with 18 predicted hot spots shared by both methods. All these shared predictions are aromatic residues and 13 of the shared predictions are true positives. These two sets of results are the most accurate predictions from the known algorithms available in the public domain (30, 31, 33, 34).

The interpretations of the prediction results by the two prediction methods are fundamentally different. The ISMBLab-PPI algorithm uses 3D probability density maps of interacting protein atom types derived from known protein structures as inputs for prediction training with nonantibody PPI complexes (30). The trained ISMBLab-PPI predictors are then used without modification in the antibody–antigen interaction predictions. By contrast, the $proABC$ predictor uses a database of the antibody–antigen complex structures in PDB augmented with known antibody sequences for training and prediction. Although the ISMBLab-PPI predictors retain no specific memory of antibody–antigen interactions and predict antibody–protein interactions with general principles governing PPIs, the information from the $proABC$'s antibody database dominates the prediction results of $proABC$, and the specific memories related to the known antibody–antigen complexes cannot be easily separated from the predictions shown in this section and the section below. As such, ISMBLab-PPI predictors enable comparisons of nonantibody protein–protein interactions with antibody–antigen interactions (as shown in Fig. 5), but the $proABC$ predictions are limited to antibody–antigen interaction predictions.

Both computational methods will be used in parallel for the discussion below. The prediction methods are not perfect; on average, about 41% of the actual functional paratope on an antibody CDR surface can be identified with the threshold $PPI_CL \geq 0.45$ (67% with $proABC_CP \geq 80\%$), and a lower limit of at least 68% (62% with $proABC_CP \geq 80\%$) of the predicted functional paratope is composed of actual hot spot residues on the CDR surface. Conversely, the alanine scanning experiments are not problem free in dissecting the energetic contributions of the constituents in the complex interfaces due to the limitation that the experimental data reflect not only the energetic effects of the side-chain truncation but also the compensating responses of the local protein and solvent environment around the mutation site. Nevertheless, the computational methods are consistent in predicting aromatic residues as the main portion of the functional paratopes, enabling at least semiquantitative analysis of the potential hot spot residues in all of the antibody–protein interfaces known thus far.

Validation of Predicted Functional Paratopes with Antibody–Protein Complex Structures. Paratope hot spot residues are mostly buried in the antibody–antigen interaction interface (15). The predicted hot spot paratope residues (with $PPI_CL \geq 0.45$ or $proABC_CP \geq 80\%$; see previous section) have been further validated by the threshold of $dSASA \geq 0.2$, which is determined with the corresponding $dSASA$ value for $PPI_CL = 0.45$ from the PPI_CL - $dSASA$ correlation plot shown in Fig. 1A. Fig. 3 shows the validation results: representative antibody–antigen complex structures from the S111 dataset are used as testing cases; two-class classification prediction benchmarks for each of the test antibody structures are calculated with the prediction results of the CDR residues. The distributions of the precision and sensitivity for the test set antibody structures are shown in Fig. 3 A and B, respectively. The detailed prediction and benchmark results with ISMBLab-PPI are shown in Table S1. Graphic displays of the prediction results are shown in the web server <http://ismlab.genomics.sinica.edu.tw/paratope>. As shown in Fig. 3 A and B, on average, 31% of the actual buried paratope in an antibody–antigen interface can be predicted with the threshold $PPI_CL \geq 0.45$ (Fig. 3B), and on average, 72% of a predicted functional paratope ($PPI_CL \geq 0.45$) is composed of buried residues in the antibody–antigen interface (Fig. 3A). Fig. 3 compares the ISMBLab-PPI prediction results with those of $proABC$. The predictions of $proABC$ are substantially more accurate in both precision (Fig. 3A) and sensitivity (Fig. 3B). However, it is difficult to isolate the contribution of the memory

residues (Ala, Leu, Ile, Val, Met, Pro, and Cys) are all depleted in the paratopes in comparison with the background probabilities. Taken together, (i) the core paratopes defined as PFP(proABC)s and PFP(PPI)s are mainly composed of aromatic residues, especially Tyr, and these PFP residues contribute a substantial portion of antigen binding energy because on average more than two thirds of these residues are hot spot residues on the CDRs, determined with the prediction benchmarks shown in Fig. 2; (ii) the peripheral paratopes outside PFPs are predominantly populated with short-chain hydrophilic residues, in particular Ser, and to a lesser extent, Asp, Asn, Gly, and Thr; and (iii) long-chain hydrophilic residues and all hydrophobic residues are not preferred in the paratopes.

The amino acid type preferences in the epitopes defined with increasingly stringent criteria are all essentially indistinguishable from the background probabilities. Fig. 4*B* shows that there are slight tendencies for enrichment of hydrophilic/aromatic residues and for depletion of hydrophobic residues in the paratopes with increasingly stringent criteria. However, these tendencies are substantially insignificant in contrast to the amino acid preferences for the paratopes shown in Fig. 4*A*. As such, the enriched aromatic residues in the PFPs have no specific amino acid preference counterpart in the corresponding epitopes. How does a PFP enriched with aromatic residues, most of which are hot spot residues, bind with specificity and affinity to its corresponding core epitope for which the amino acid type preference is indistinguishable from that of other protein solvent accessible surfaces? If amino acid type in the epitope is not to be recognized by

the paratope hot spot residues, what physicochemical property of the epitope is responsible for the specificity and affinity in the antibody–antigen interaction?

Protein Atom Type Preferences of the Paratope–Epitope Interfaces.

Protein atom type preferences on the core paratope–epitope interfaces provide some answer to the questions above. Fig. 4*C* shows the protein atom type distribution of the paratope surface atoms that are in contact with the corresponding epitope surface atoms, for which the distribution of the protein atom types are depicted in Fig. 4*D*. The protein atom types are summarized in Table 1. The side-chain aromatic carbons CR1E, CY, CY2, CR1W, CW, and C5W [53% of the PFP(PPI) atoms] and side-chain hydrogen bond donor/acceptors OH1, NH1S, and NC2 [13% of the PFP(PPI) atoms] are increasingly enriched with increasingly stringent paratope criteria. As expected, the main portion of the enrichment is attributed to the enrichment of the tyrosyl side chains, contributing to the buildup of the CR1E, CY2, and OH1 atom types on the core paratopes. Fig. 4*C* also indicates that the main chain atoms [NH1, CB, CH1E, and OB; 16% of the PFP(PPI) atoms] and side-chain aliphatic carbons [CH0, CH2E, and CH3E; 14% of the PFP(PPI) atoms] are increasingly depleted in the paratope surfaces with increasingly stringent criteria.

By contrast, Fig. 4*D* shows that 63% of the PFE(PPI) atoms are the backbone atoms (NH1, CB, CH1E, and OB) or side-chain aliphatic carbons (CH0, CH2E, and CH3E). The rest is mainly composed of 6% aromatic carbons (CR1E) and 17% hydrogen bond donors/acceptors (OH1, OC, OS, NH1S, NC2,

Table 1. Atom type definition and atom radii

Atom type group	Polar atom group	Protein atom type	Atom radius (Å)	Description
B	N	NH1	1.65	Backbone NH
B		CB	1.76	Backbone C
B		CH1E	1.87	Backbone CA (exclude Gly)
B		OB	1.40	Backbone O
C		CH2G	1.87	Gly CA
C		CH0	1.76	Arg CZ, Asn CG, Asp CG, Gln CD, Glu CD
C		CH1S	1.87	Side chain CH1: Ile CB Leu CG, Thr CB, Val CB
C		CH2P	1.87	Pro CB, CG, and CD
C		CH2E	1.87	Tetrahedral CH2 (except CH2P and CH2G) All CB
C		CH3E	1.87	Tetrahedral CH3
A		CR1E	1.76	Aromatic CH (except CR1W, CRHH, CR1H)
A		CF	1.76	Phe CG
A		CY2	1.76	Tyr CZ
A		CY	1.76	Tyr CG
A		CR1W	1.76	Trp CZ2 and CH2
A		CW	1.76	Trp CD2 and CE2
A	O	C5W	1.76	Trp CG
C		CRHH	1.76	His CE1
C		CR1H	1.76	His CD2
C		C5	1.76	His CG
S		SC	1.85	Cys S
S		SM	1.85	Met S
P		OH1	1.40	Alcohol OH (Ser OG, Thr OG1, and Tyr OH)
P		OS	1.40	Side chain O: Asn OD1 and Gln OE1
P		OC	1.40	Carboxyl O (Asp OD1, OD2 and Glu OE1, OE2)
P		ProN	1.65	Pro N
P		NH1S	1.65	Side chain NH: Arg NE, His ND1, NE2, and Trp NE1
P		NH2	1.65	Asn ND2 and Gln NE2
P		NC2	1.65	Arg NH1 and NH2
P		NH3	1.50	Lys NZ

Twenty natural amino acid types are divided into 30 protein atom types, which are classified into five atom type groups. A, aromatic carbon atom types; B, backbone atom types; C, aliphatic carbon atom types; P, polar atom types; S, sulfur atom types.

and NH₂). The results indicate that the aromatic side chains in the core paratopes could interact mostly with the backbone atoms (NH₁, CB, CH₁E, and OB) and side-chain carbons of the core epitopes (CH₀, CH₂E, CH₃E, and CR₁E).

Together, the analyses above based on protein atom type preferences in the paratope–epitope interfaces provide further interaction information, especially for the interactions involving backbone atoms and side-chain carbons, which cannot be explicitly scrutinized with the analyses solely based on amino acid type preferences. The results show that the core paratopes (PFPs) are enriched with aromatic atom type, especially from the Tyr side chain, and the corresponding core epitopes (PFEs) are mostly backbone atoms and side-chain aliphatic carbon atoms and to a lesser extent aromatic carbons. The backbone atoms are common to all 20 natural amino acid types, and 19 natural amino acid types have the side-chain aliphatic or aromatic carbons. Hence, the atom type preference distributions provide a possible explanation for the lack of amino acid type preferences on the epitopes. In addition, on average, 79% of the average protein surface atoms are backbone atoms or side-chain carbons, providing easily assessable common targets for antibody paratopes enriched with aromatic side chains; indeed, 69% of the PFE (PPI) atoms remain backbone atoms or side-chain carbons.

Pairwise Atomistic Contacts in the Paratope–Epitope Interfaces. The analyses of pairwise atomistic contact preferences in the paratope–epitope interfaces confirm that the aromatic side chains in the PFP interact mostly with the backbone atoms and side-chain carbons of the PFE. The 2D distributions of pairwise atomistic contacts in the paratope–epitope interfaces are shown in [Tables S2–S4](#) for SP(0)–SE(0), SP(0.2)–SE(0.2), and PFP(PPI)–PFE(PPI) interfaces respectively. The SP(0)–SE(0) and SP(0.2)–SE(0.2) interfaces are different by only 2% additional pairwise atomistic contacts. However, 48% pairwise atomistic contacts in the SP(0.2)–SE(0.2) interfaces remain in the PFP(PPI)–PFE(PPI) interfaces [60% in the PFP(proABC)–PFE(proABC) interface].

Analysis of the pairwise atomistic contacts in [Tables S2–S4](#) shows that the core antibody–protein interaction interfaces are mainly composed of atomistic contacts involving paratope aromatic carbons. Fig. 5A summarizes the distribution of the atomistic contact pairs shown in [Tables S2–S4](#). The atomistic contact pairs can be sorted into five groups as shown in Fig. 5: B:B+P, contact pairs with paratope backbone atoms and epitope backbone and polar side-chain atoms; P:B+P, contact pairs with paratope polar side-chain atoms and epitope backbone and polar side-chain atoms; C:C+A, contact pairs with paratope aliphatic carbons and epitope aliphatic/aromatic carbons; A:all, contact pairs with paratope aromatic carbons and all epitope atoms; else, the polar-to-nonpolar contact pairs not included in the above groups. The atom type groups are summarized in Table 1. A fraction of the B:B+P and P:B+P contact pairs contribute direct hydrogen bonding across the interface; the C:C+A contact pairs contribute to hydrophobic interactions; the A:all contact pairs contribute to multiple types of interactions involving paratope aromatic side chains (Introduction); and the else contact pairs are not expected to contribute favorable binding energy. As summarized in Fig. 5A, contacts involving paratope aromatic carbons (A:all in Fig. 5A) contribute predominantly to the favorable PFP–PFE interfaces—67% of the energetically relevant PFP(PPI)–PFE(PPI) interface atomistic contact pairs (B:B+P, P:B+P, C:C+A, and A:all) involving paratope aromatic carbons [53% in PFP(proABC)–PFE(proABC) interface]—suggesting that interactions involving paratope aromatic side chains contribute to the majority of the antigen-binding energy. Hydrophobic contacts (C:C+A) contribute only 14–17% to the core interface atomistic contact pairs, suggesting that hydrophobic interaction is not a major driving force in antibody–protein recog-

nition. Overall, the general trends for the distributions of the atomistic contact pair types in PFP(PPI)–PFE(PPI) and PFP(proABC)–PFE(proABC) interfaces are similar as shown in Fig. 5A. The fraction of the atomistic contacts involving the paratope aromatic side chains does not stand out as much in the SP(0)–SE(0) and SP(0.2)–SE(0.2) interfaces. In particular, the group marked as else, which contain polar–nonpolar atomistic contacts, is the most prominent group in the SP(0)–SE(0) and SP(0.2)–SE(0.2) interfaces but could contribute little, if not negatively, to the antibody–antigen binding energy; these polar–nonpolar atomistic contacts are markedly depleted from the PFP–PFE interfaces (Fig. 5A).

The paratope aromatic carbons in the PFP–PFE interfaces interact mostly with backbone atoms and side-chain carbons on the PFE. As shown in Fig. 5B, the atomistic contact pairs involving the paratope aromatic side-chain carbons are sorted into five groups: A:B, contact pairs with paratope aromatic carbons and epitope backbone atoms; A:C, contact pairs with paratope aromatic carbons and epitope aliphatic carbons; A:A, contact pairs with paratope aromatic carbons and epitope aromatic carbons; A:S, contact pairs with paratope aromatic carbons and epitope sulfur atoms in Met and Cys; A:P, contact pairs with paratope aromatic carbons and epitope polar side-chain atoms. Three quarters [77% of the PFP(PPI)–PFE(PPI) interfaces and 74% of the PFP(proABC)–PFE(proABC) interfaces] of the atomistic pairwise contacts involving paratope aromatic carbons interact with backbone atoms and aliphatic carbon atoms on the PFE (Fig. 5B). The aromatic side chains interact with the protein backbone NH group (18, 35) and other hydrogen bond donors through hydrogen bonding to aromatic π -systems (X–H– π interaction, X = N, O, S), with the peptide bond through the π -electron interaction and with aliphatic carbons of the side chain through the C–H– π interaction (16, 17). These interactions contribute to the majority of interactions involving hot spots in the PFP–PFE interfaces.

The result above explains in part the notions that all surfaces of a protein may be antigenic and that a limited antibody repertoire is capable of recognizing seemingly unlimited number of protein structures. Seventy-nine percent of an average protein surface is composed of backbone atoms and side-chain carbons, indicating that, in principle, antigenic sites recognizable by antibodies are ubiquitous on all protein surfaces. The ubiquity explains the difficulty in conformational epitope prediction as shown in Fig. 1B by the ISMBLab-PPI predictors. This implication echoes the conclusions by Benjamin et al. (26), suggesting that most, if not all, of the surface of a protein maybe antigenic with multiple overlapping antigenic sites, and these antigenic sites require the native conformation integrity of the protein for their antigenicity.

Fig. 5A and B compares the differences between antibody–protein interactions and nonantibody PPIs. [Tables S5–S7](#) show the 2D distributions of pairwise atomistic contacts in the PPI interfaces defined by dSASA > 0, dSASA $\geq$ 0.2, and dSASA $\geq$ 0.2 & PPI_CL $\geq$ 0.45, respectively. As expected, an average non-antibody PPI interface has more pairwise atomistic contacts (about twice as many) in comparison with an average antibody–protein interface, and the hydrophobic contacts contribute predominantly to the PPI interfaces, in agreement with the previous study with the same PPI dataset (30).

The antibody–protein complex interfaces and the PPI interfaces share common features involving aromatic residues. Fig. 5B shows that the interaction features involving aromatic side-chain carbons contacting with backbone atoms and side-chain carbons are commonly shared by the two types of interfaces. These common interactions involving aromatic side chains in the interfaces also explain that the PPI_CL calculation algorithm (ISMBLab-PPI), which is modeled with the PPI dataset (30), is equally applicable in antibody paratope prediction (Fig. 1A).

Fig. 5. Distributions of pairwise atomistic contacts in antibody-protein interactions and in nonantibody protein-protein interaction interfaces. (A) Pairwise atomistic contacts are sorted in five groups as shown in the y axis (more details are discussed in the main text). The atom type groups are defined in the first column of Table 1. (Left) Results for antibody-protein interfaces from the S111 dataset, calculated with PFP(proABC)–PFE(proABC) interfaces. (Center) Results for antibody-protein interfaces from the S111 dataset, calculated with PFP(PPI)–PFE(PPI) interfaces. (Right) Results for PPI interfaces from the S430 dataset. The x axis shows the number of pairwise atomistic contacts per interface. The histograms are cumulative: the blue part of the histogram in the Left and Center shows the distributions for the PFP(proABC)–PFE(proABC) and PFP(PPI)–PFE(PPI) interfaces, respectively, in antibody-protein interactions; the blue+red histogram shows the distributions for the SP(0.2)–SE(0.2) interfaces; the blue+red+green histogram shows the distributions for the SP(0)–SE(0) interfaces. (Right) The same presentation method is used for the PPI interfaces defined by $PPI_CL \geq 0.45$ and $dSASA \geq 0.2$, $dSASA \geq 0.2$, and $dSASA > 0$, respectively. (B) Pairwise atomistic contacts involving paratope aromatic atoms are sorted into five groups as shown in the y axis. The x axis shows the number of atomistic contact pairs per interface. Other specifications are the same as in A. (C) (Left) The y axis shows the fraction of antibody-protein complexes in the S111 dataset with the number of A:B + A:C atomistic contact pairs (x axis) per interface; the purple, blue, red, and green curves are calculated with PFP(proABC)–PFE(proABC), PFP(PPI)–PFE(PPI), SP(0.2)–SE(0.2), and SP(0)–SE(0) interfaces, respectively. (Right) The y axis shows the fraction of protein-protein complexes in the S430 dataset with the number of A:B + A:C atomistic contact pairs (x axis) per interface; the blue, red, and green curves are calculated with $PPI_CL \geq 0.45$ and $dSASA \geq 0.2$, $dSASA \geq 0.2$, and $dSASA > 0$ interfaces, respectively. (D) Polar pairwise atomistic contacts are sorted into six groups as shown in the y axis (more details are discussed in the main text). The polar atom groups are defined in the second column of Table 1. The x axis shows the number of contacts per interface. Other specifications are the same as in A. (E) Direct H-bonded polar pairwise atomistic contacts per interface for the six polar contact groups are shown. Other specifications are the same as in A.

In general, nonantibody PPIs are driven energetically by hydrophobic packing and aromatic side chains binding to backbone atoms and side-chain carbons, whereas antibody-protein interactions are driven by the latter of the two types of interactions in PPIs, likely due to the fact that hydrophobic residues are much less accessible on protein surfaces for antibodies to bind. In any type of protein-protein interface, 72–77% of surface atoms in contact with aromatic side-chain carbons are backbone atoms or side-chain carbons, supporting the notion that this type of interaction plays an important energetic role in all types of protein-protein recognitions.

Fig. 5C supports the critical role of the interaction involving paratope aromatic side chains binding to backbone atoms and side-chain carbons in the corresponding epitope; almost all antibody-protein complexes in the S111 dataset have this type of interaction in the structural paratope-epitope interfaces (110/111 complex structures in S111) or the functional paratope-epitope interfaces defined by $PPI_CL \geq 0.45$ (95.5% of the complex structures in S111). By contrast, about 10% of the nonantibody PPI interfaces do not have this type of interaction, suggesting that hydrophobic contacts can compensate for the binding energy in the absence of the interactions involving aromatic side chains.

Contacts involving polar atoms in the interfaces are largely depleted in the PFP–PFE interfaces (B:B+P and P:B+P in Fig. 5A): only 33% of the SP(0)–SE(0) interface contacts involving paratope main chain atoms and 30% of the SP(0)–SE(0) interface contacts involving paratope side-chain hydrogen bond donor/acceptor remain in the PFP(PPI)–PFE(PPI) interfaces. The polar-to-polar atomistic contact pairs are sorted into six groups as shown in Fig. 5D: N:O+OH1+OCS, contact pairs with paratope backbone N and epitope H-bond acceptors; O:N+OH1+NS, contact pairs with paratope backbone O and epitope H-bond donors; OH1:all, contact pairs with paratope side-chain hydroxyl group and epitope H-bond donors/acceptors; OCS:N+OH1+NS, contact pairs with paratope side-chain H-bond acceptors and epitope H-bond donors; NS:O+OH1+OCS, contact pairs with paratope side-chain H-bond donors and epitope acceptors; else, donor-donor and acceptor-acceptor pairs that are not expected to form an H-bond. The polar atom groups are defined in Table 1. Fig. 5D shows that most (88%, the non-else groups together) of the polar-polar atomistic contacts in the SP(0)–SE(0) interfaces are matched in polarity for forming the H-bond. Fig. 5E shows that about 40% the polar-polar contacts in the SP(0)–SE(0) interfaces form direct H-bonding across the interface, as defined by the H-bond criteria $D \dots A < 3.5 \text{ \AA}$ and angle for $D-H \dots A > 150^\circ$. Fig. 5D and E also shows that the PFP(proABC)–PFE(proABC) interfaces have more polar contacts and direct H-bonds comparing with the PFP(PPI)–PFE(PPI) interfaces. Together, these results indicate that the hydrogen bonds across the paratope-epitope interface surround the core PFP–PFE interface, and the acceptor-donor polarity across the interface is mostly matched for H-bond formation.

Fig. 5D and E also compares the polar contacts and H-bond patterns between the antibody-protein interfaces and the non-antibody PPI interfaces. Both polar contacts (Fig. 5D) and H-bonds (Fig. 5E) in an average PPI interface are about twice as many as in an average antibody-protein interface, in agreement with the ratio of the overall contact areas of the two types of interfaces (Fig. 5A). Nevertheless, OH1 H-bond donors/acceptors in PPI from Thr, Ser, and Tyr are not used as frequently as in antibody-protein interfaces, indicating that the H-bonds involving OH1 are not particularly favorable for protein recognitions in general.

Together, the humoral responses where vast variations of protein antigens can be recognized with functional antibodies bearing a limited repertoire of paratopes that are enriched with aromatic side chains and short-chain hydrophilic residues can be

reconciled by the notion that the aromatic paratope side chains are able to recognize diverse protein antigens through accessible protein surfaces mostly composed of backbone atoms and side-chain carbons. In addition, the hydrophilic functional groups on the paratopes (mostly from Tyr and from short-chain hydrophilic residues Ser, Thr, Asn, and Asp) surrounding the core aromatic paratope side chains could contribute substantially to the specificity through direct hydrogen bonding across the paratope–epitope interface. Short-chain hydrophilic side chains are particularly suitable for this binding role because of the smaller side chain conformational entropy penalty by fixing the side chains in the interfaces. These notions are consistent with the studies showing that Tyr, Trp, and Ser are highly overrepresented in germ-line contacting residues (36), such that antibody repertoires derived from recombination of germ-line antibody sequences are highly functional in antigen recognition; somatic hypermutations reduce the population of these residue types where these residues are not involved in antibody–antigen interactions. Water-mediated hydrogen bonding could contribute also to the affinity and specificity, but the magnitude could be relatively insignificant. The rest of the atomistic contacts contribute insignificantly, if not negatively, to both affinity and specificity in antibody–protein antigen interactions.

Promptness for Protein Surface Residues to Interact with the Tyrosyl Side Chain. The studies above suggest an approach to evaluate the propensity for protein surfaces to be recognized by antibody paratopes enriched mostly with tyrosyl side chains: the propensity for each protein surface residue can be evaluated by the scoring matrix as shown in Table S8, where the tyrosyl side-chain contact frequency for each of the 30 protein atom types from each of the 20 amino acid types are derived from the S111 dataset; the propensity of a query residue to interact with a tyrosyl side chain is the simple linear combination of the frequencies in the look-up table for the solvent-exposed atoms (SASA > 0) of the query residue averaged over the number of atoms in the residue (SI Materials and Methods). The propensity is a simple measurement of the frequency for the residue to be in contact with a tyrosyl side chain. This analysis is best tested with anti-HEL antibodies, for which the epitopes on HEL have been well studied (13).

Fig. 6A shows the HEL structure with each of the residues color coded based on the propensity to interact with a tyrosyl side chain. The tyrosyl side chains from the antibodies known to bind to HEL are shown as in the composite structure in this figure. As shown in Fig. 6A, the high propensity areas are situated in the loop regions and the ends of the secondary structure elements, where the main chain atoms and side-chain carbons are mostly exposed to the protein surface. These high-propensity areas are also the areas where the anti-HEL antibodies bind to the HEL with tyrosyl side chains. The results explain the preference of the antibodies to recognize the protein antigens through interactions with the solvent-exposed loop regions on the antigen surfaces (20, 21).

Fig. 6B compares the tyrosyl side-chain interacting propensity distribution for the antigen surfaces with that of the nonantigen surfaces for the antigens in the S111 dataset. These two distributions are indistinguishable (the *t* test *P* value is close to 1), suggesting that many of the nonepitope areas defined by the antibody–antigen complexes in the S111 dataset are in principle equally recognizable by antibodies; the protein antigenic sites defined by the complexes in the S111 dataset are only the tip of an iceberg; Benjamin et al. (26) indicated that most of the surface of a protein may be antigenic. This result explains the difficulties in predicting the antibody binding sites on proteins as shown in Fig. 1B. The finding in Fig. 6B also suggests that conformational epitope predictors trained with the known antibody–antigen complexes tend to predict conformational epitopes that would not be necessarily incorrect but would be considered as

Fig. 6. Propensity for tyrosine side-chain interaction on protein surface residues. (A) The residues of the lysozyme are color coded according to the tyrosyl side-chain interacting propensity (from blue to white to red representing low, medium, and high interacting propensity for tyrosyl side chain). The stick models of tyrosyl side chains are from three known anti-HEL antibodies: yellow, HyHEL-5 (PDB code: 1YQV); green, HyHEL-10 (PDB code: 3HFM); cyan, FvD1.3 (PDB code: 1VFB). The quantitative propensities are shown in Fig. S2, which shows the residue-based tyrosyl side-chain interacting propensity of HEL, plotted against the residue number of the antigen protein HEL. The tyrosyl side-chain interacting propensities are calculated with the scoring matrix shown in Table S8 (main text and SI Materials and Methods). (B) Tyrosyl side-chain interacting propensity score distributions calculated with epitope atoms (blue) and non-epitope atoms (red) in the protein antigens of the S111 dataset are compared. The *t* test *P* value for the two distributions is close to 1.

false positives because these predicted epitopes have not been observed in the database, explaining the low prediction accuracy for the current conformational B-cell epitope predictions (19).

Conclusion. The hypothesis that the antigenic surfaces recognizable by known antibodies are composed of common and ubiquitous features shared by all protein surfaces is supported by the finding that functional paratopes are enriched with aromatic side chains mostly binding to epitope backbone atoms and side-chain carbons, which make up four fifths of surface atoms on an average protein structure. The significant implication is that a relatively limited repertoire of spatial distribution of aromatic side chains among short-chain hydrophilic residues in antibody CDRs can recognize the common features, including backbone atoms, side-chain carbons, and surface-exposed H-bond donors/acceptors, shared by all protein surfaces. Two fundamentally different functional epitope/paratope prediction algorithms are used to define the key energetic determinant in antibody–protein interactions in 111 representative antibody–protein complex structures. The conclusions derived from both methods are consistent: antibody functional paratopes are enriched with aromatic residues, especially Tyr side chains. These functional paratope residues are surrounded by short-chain hydrophilic side chains (Asp, Asn, Ser, Thr, and Gly) in the structural paratopes. These paratope aromatic side chains recognize the corresponding epitopes by interacting with the backbone atoms and the side-chain carbons, mainly through the interaction of the tyrosyl side chains on the functional paratopes. These interactions contribute the majority of the binding energy for the antibody–protein complexes as hot spots in the interfaces. A strikingly large fraction of the nonantibody protein–protein core interactions are also composed of the interactions between aromatic side chains and backbone atoms/side-chain carbons, suggesting a driving force commonly shared in all types of protein–protein recognitions. The short hydrophilic side chains on the structural paratopes form favorable short-range electrostatic interactions with the corresponding polar functional groups on the epitope, and slightly less than half of the favorable polar contacts form direct hydrogen bonding across the paratope–epitope interfaces. The direct H-bonds across the epitope–paratope interfaces could contribute to the specificity of the antibody–antigen recognition; short-chain hydrophilic side chains are particularly suitable for this interaction because of the smaller side-chain conformational entropy penalty in the interface.

The major contribution of the main-chain atoms and side-chain carbons on the antigen surfaces to the antibody recognitions explains that the antibodies are able to recognize diverse protein antigen structures with relatively limited configurations of the functional paratopes and that the sequence preferences of the conformational epitopes are hardly distinguishable from those of the average protein surfaces. This conclusion also explains the difficulties in B-cell epitope predictions. The results provide an explanation for the observations that most of the surface of a protein may be antigenic with multiple overlapping antigenic sites and that solvent accessible and protruding loop regions on the protein antigens are more likely to bind to antibodies because of the expose of the main-chain atoms and side-chain carbons in these regions. Artificial antibody repertoires aiming at recognizing diverse protein antigens could consider designing combinations of aromatic residues with short-chain hydrophilic residues on diverse CDR structural contours matching large varieties of antigen surfa-

ces, to mimic the natural strategy in recognizing seemingly unlimited variations of protein antigens with limited repertoire of antibodies.

Materials and Methods

Computation algorithm for PPI_CL was used without modification from a previous work (30). A brief description of the computational methodology for the ISMBLab-PPI predictors is included in the [SI Materials and Methods](#). Other computational details, including dSASA calculation, background probability of protein surface amino acid types and atom types, secondary structure assignment, hidden Markov models for CDR identification, atomistic contact pair calculation, hydrogen bonding criteria, tyrosyl side-chain interaction scoring matrix, datasets, and prediction capacity benchmarks are all documented in [SI Materials and Methods](#).

ACKNOWLEDGMENTS. We thank Dr. Peter Kwong (National Institutes of Health) and Dr. Lawrence Shapiro (Columbia University) for helpful discussions. This work was supported by National Science Council Grants 100DP006-3 and 99-2311-B-001-014-MY3 and Genomics Research Center at Academia Sinica Grant AS-100-TP2-B01.

1. Michnick SW, Sidhu SS (2008) Submitting antibodies to binding arbitration. *Nat Chem Biol* 4(6):326–329.
2. Ponsel D, Neugebauer J, Ladetzki-Baehs K, Tissot K (2011) High affinity, developability and functional size: The holy grail of combinatorial antibody library generation. *Molecules* 16(5):3675–3700.
3. Sliwkowski MX, Mellman I (2013) Antibody therapeutics in cancer. *Science* 341(6151):1192–1198.
4. Davies DR, Padlan EA, Sheriff S (1990) Antibody-antigen complexes. *Annu Rev Biochem* 59:439–473.
5. Mian IS, Bradwell AR, Olson AJ (1991) Structure, function and properties of antibody binding sites. *J Mol Biol* 217(1):133–151.
6. Kringelum JV, Nielsen M, Padkjær SB, Lund O (2013) Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol Immunol* 53(1–2):24–34.
7. Ramaraj T, Angel T, Dratz EA, Jesaitis AJ, Murney B (2012) Antigen-antibody interface properties: Composition, residue interactions, and features of 53 non-redundant structures. *Biochim Biophys Acta* 1824(3):520–532.
8. Koide S, Sidhu SS (2009) The importance of being tyrosine: Lessons in molecular recognition from minimalist synthetic binding proteins. *ACS Chem Biol* 4(5):325–334.
9. Fellouse FA, Wiesmann C, Sidhu SS (2004) Synthetic antibodies from a four-amino-acid code: A dominant role for tyrosine in antigen recognition. *Proc Natl Acad Sci USA* 101(34):12467–12472.
10. Fellouse FA, et al. (2007) High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J Mol Biol* 373(4):924–940.
11. Dall'Acqua W, Goldman ER, Eisenstein E, Mariuzza RA (1996) A mutational analysis of the binding of two different proteins to the same antibody. *Biochemistry* 35(30):9667–9676.
12. Bostrom J, et al. (2009) Variants of the antibody herceptin that interact with HER2 and VEGF at the antigen binding site. *Science* 323(5921):1610–1614.
13. Sundberg EJ, Mariuzza RA (2002) Molecular recognition in antibody-antigen complexes. *Adv Protein Chem* 61:119–160.
14. Moreira IS, Fernandes PA, Ramos MJ (2007) Hot spots—A review of the protein-protein interface determinant amino-acid residues. *Proteins* 68(4):803–812.
15. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* 280(1):1–9.
16. Salonen LM, Ellermann M, Diederich F (2011) Aromatic rings in chemical and biological recognition: Energetics and structures. *Angew Chem Int Ed Engl* 50(21):4808–4842.
17. Meyer EA, Castellano RK, Diederich F (2003) Interactions with aromatic rings in chemical and biological recognition. *Angew Chem Int Ed Engl* 42(11):1210–1250.
18. Chakrabarti P, Bhattacharyya R (2007) Geometry of nonbonded interactions involving planar groups in proteins. *Prog Biophys Mol Biol* 95(1–3):83–137.
19. Yao B, Zheng D, Liang S, Zhang C (2013) Conformational B-cell epitope prediction on antigen protein structures: A review of current algorithms and comparison with common binding site prediction methods. *PLoS ONE* 8(4):e62249.
20. Thornton JM, Edwards MS, Taylor WR, Barlow DJ (1986) Location of 'continuous' antigenic determinants in the protruding regions of proteins. *EMBO J* 5(2):409–413.
21. Novotný J, et al. (1986) Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc Natl Acad Sci USA* 83(2):226–230.
22. Sun J, et al. (2011) Does difference exist between epitope and non-epitope residues? Analysis of the physicochemical and structural properties on conformational epitopes from B-cell protein antigens. *Immunome Res* 7(3):1–11.
23. Rubinstein ND, et al. (2008) Computational characterization of B-cell epitopes. *Mol Immunol* 45(12):3477–3489.
24. Ofra Y, Schlessinger A, Rost B (2008) Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes. *J Immunol* 181(9):6230–6235.
25. Kunik V, Ofra 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(10):599–609.
26. Benjamin DC, et al. (1984) The antigenic structure of proteins: A reappraisal. *Annu Rev Immunol* 2:67–101.
27. Yu CM, et al. (2012) Rationalization and design of the complementarity determining region sequences in an antibody-antigen recognition interface. *PLoS ONE* 7(3):e33340.
28. Dall'Acqua W, et al. (1998) A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry* 37(22):7981–7991.
29. Li Y, Urrutia M, Smith-Gill SJ, Mariuzza RA (2003) Dissection of binding interactions in the complex between the anti-lysozyme antibody HyHEL-63 and its antigen. *Biochemistry* 42(1):11–22.
30. Chen CT, et al. (2012) Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces. *PLoS ONE* 7(6):e37706.
31. Olimpieri PP, Chaillyan A, Tramontano A, Marcatili P (2013) Prediction of site-specific interactions in antibody-antigen complexes: The proABC method and server. *Bioinformatics* 29(18):2285–2291.
32. Stave JW, Lindpaintner K (2013) Antibody and antigen contact residues define epitope and paratope size and structure. *J Immunol* 191(3):1428–1435.
33. Krawczyk K, Baker T, Shi J, Deane CM (2013) Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking. *Protein Eng Des Sel* 26(10):621–629.
34. Kunik V, Peters B, Ofra Y (2012) Structural consensus among antibodies defines the antigen binding site. *PLoS Comput Biol* 8(2):e1002388.
35. Tóth G, Watts CR, Murphy RF, Lovas S (2001) Significance of aromatic-backbone amide interactions in protein structure. *Proteins* 43(4):373–381.
36. Burkovitz A, Sela-Culang I, Ofra Y (2014) Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. *FEBS J* 281(1):306–319.