

Use of amino acid composition to predict epitope residues of individual antibodies

Shinji Soga¹, Daisuke Kuroda^{2,3}, Hiroki Shirai¹,
Masato Kobori¹ and Noriaki Hirayama^{4,5}

¹Molecular Medicine Research Laboratories, Drug Discovery Research, Astellas Pharma Inc., 21 Miyukigaoka, Tsukuba, Ibaraki 305-8585, Japan,

²Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan, ³Graduate School of Frontier Biosciences, Osaka University, 1-3 Yamadaoka, Suita, Osaka 565-0871, Japan and ⁴Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1143, Japan

⁵To whom correspondence should be addressed.
E-mail: hirayama@is.icc.u-tokai.ac.jp

Received September 6, 2009; revised January 7, 2010;
accepted February 16, 2010

Edited by Andreas Plueckthun

We identified specific amino acid propensities at the interfaces of antigen–antibody interactions in non-redundant qualified antigen–antibody complex structures from Protein Data Bank. Propensities were expressed by the frequency of each of the 20 × 20 standard amino acid pairs that appeared at the interfaces of the complexes and were named the antibody-specific epitope propensity (ASEP) index. Using this index, we developed a novel method of predicting epitope residues for individual antibodies by narrowing down candidate epitope residues which was predicted by the conventional method. The 74 benchmarked antigens were used in ASEP prediction. The efficiency of this method was assessed using the leave-one-out approach. On elimination of residues with ASEP indices in the lowest 10% of all measured, true positives were enriched for 49 antigens. On subsequent elimination of residues with ASEP indices in the lowest 50%, true positives were enriched for 40 of the 74 antigens assessed. The ASEP index is the first benchmark proposed to predict epitope residues for an individual antibody. Used in combination with mutation experiments, this index has the potential to markedly increase the success ratio of epitope analysis.

Keywords: amino acid propensity/antibody/ASEP/epitope prediction

Introduction

The coming years should see the development of a great number of antibody drugs. To ensure that an antibody possesses the targeted potency and efficacy, the drug must have sufficient affinity for the antigen and be able to recognize the appropriate epitope. Although epitope analysis of antibodies is thus an important step in the development of antibody drugs, such as in the rational design of antibodies for affinity maturation, current approaches to this analysis are far from satisfactory.

The two methods of epitope analysis now available are the peptide-based (Frank, 2002; Bublil *et al.*, 2007; Huang *et al.*, 2008) and the mutation-based approaches (Lu *et al.*, 2001; Hu *et al.*, 2008; Zou *et al.*, 2008). In the peptide-based approach, epitope prediction is based on the SPOT-synthesis technique (Frank, 2002) and requires peptides ~10 residues long, obtained by cleaving the complete antigen sequence. The epitope itself is composed of several epitope residues, and if a peptide competitively blocks the antigen–antibody interaction, the sequence is considered to contain epitope residues. Epitope prediction based on mimotopes (Bublil *et al.*, 2007; Huang *et al.*, 2008) is performed by selecting from a set of affinity-isolated peptides obtained by screening phage-display peptide libraries with the antibody of interest. These mimotopes are considered as functional epitope mimics. Discontinuous epitopes are observed much more frequently than continuous epitopes, however, and the peptide-based approach is less effective at discerning discontinuous than continuous epitopes. Further, given that peptide fragments are too flexible to construct a real tertiary structure for an antigen, this approach is not always effective, even for continuous linear epitopes.

The mutation-based approach requires the construction of many proteins by mutating residues into other amino acids. If the mutant subsequently loses affinity for the antibody, the position is proven to be an epitope residue. Although this approach is more responsive to discontinuous epitopes, however, expression and purification of the multiple mutant proteins is extremely costly and requires significant effort. In addition, for practical purposes, the positions to be mutated must be narrowed down, requiring even further preparation and analysis.

Many sophisticated methods of epitope prediction have been developed (Hopp and Woods, 1981; Pellequer *et al.*, 1991; Alix, 1999; Odorico and Pellequer, 2003; Haste Andersen *et al.*, 2006; Larsen *et al.*, 2006; Söllner and Mayer, 2006). Although these have sufficient efficacy for the prediction of epitopes for general antibodies, they are less effective when required to identify the epitope of a particular antibody of interest. One reason for this may be that they assume the likelihood of epitope residues for general antibodies, whereas an individual antibody is able to accurately and efficiently recognize its corresponding epitope. To achieve such highly specific recognition, the characteristics of the interaction between an epitope and an antigen-binding site of an antibody must be well formulated.

In contrast, prediction of a paratope, an antigen-binding site, is comparatively easy. Paratopes are basically comprised of the solvent-exposed residues of the six complementarity-determining regions (CDRs), mainly CDR-H3. Given the relative ease of identifying these antigen-binding sites, prediction of antibody-specific epitope residues based on the specific characteristics of the paratope of an individual antibody would facilitate mutation analysis as well as the rational design of antibodies.

Table 1. Non-redundant X-ray structures of 74 protein antigen–antibody complexes from the PDB

PDB code	Resolution	R_{free}	Description
1A2Y	1.50	0.251	Lysozyme
1AR1	2.70	0.261	Cytochrome c oxidase
1BGX	2.30	0.253	<i>Taq</i> DNA polymerase
1BJ1	2.40	0.266	Vascular endothelial growth factor
1EGJ	2.80	0.288	Cytokine receptor common β -chain precursor
1FE8	2.03	0.264	von Willebrand factor
1FNS	2.00	0.207	von Willebrand factor
1H0D	2.00	0.272	Angiogenin
1IQD	2.00	0.253	Human factor VIII
1JPS	1.85	0.224	Tissue factor
1JRH	2.80	0.314	Interferon- γ receptor α -chain
1KB5	2.50	—	KB5-C20 T-cell antigen receptor
1KB9	2.30	0.249	Ubiquinol-cytochrome c reductase iron-sulfur subunit
1LK3	1.91	0.240	Interleukin-10
1N8Z	2.52	0.284	Receptor protein-tyrosine kinase ERBB-2
1NFD	2.80	0.309	TCR
1NL0	2.20	0.270	Human factor IX
1NMB	2.20	—	Neuraminidase
1OAZ	2.78	0.276	Recombinant thioredoxin
1ORS	1.90	0.251	Potassium channel
1OSP	1.95	0.295	Outer surface protein A
1OTS	2.51	0.299	Voltage-gated ClC-type chloride channel EriC
1QFU	2.80	0.284	Hemagglutinin
1R0A	2.80	0.272	HIV-1 reverse transcriptase
1RJL	2.60	0.235	Outer surface protein B
1TXV	2.75	0.242	Integrin α -IIb /integrin β -3
1TZI	2.80	0.254	Vascular endothelial growth factor A
1UAC	1.70	0.248	Lysozyme
1V7M	2.51	0.316	Thrombopoietin
1W72	2.15	0.248	HLA class I histocompatibility antigen
1WEJ	1.80	0.256	Cytochrome c
1XIW	1.90	0.241	T-cell surface glycoprotein CD3 epsilon chain
1YJD	2.70	0.282	T-cell-specific surface glycoprotein CD28
1YQV	1.70	0.234	Lysozyme
1YY9	2.60	0.289	Epidermal growth factor receptor
1ZTX	2.50	0.282	West Nile Virus envelope protein DIII
2ADF	1.90	0.220	von Willebrand factor
2AEP	2.10	0.224	Neuraminidase
2B2X	2.20	0.272	Integrin α -1
2BDN	2.53	0.277	Small inducible cytokine A2
2CMR	2.00	0.258	Transmembrane glycoprotein (5-helix, residues 543–582 and 625–662)
2DD8	2.30	0.261	Spike glycoprotein
2FD6	1.90	0.276	Urokinase-type plasminogen activator/urokinase plasminogen activator surface receptor
2GHW	2.30	0.295	SARS spike protein
2H9G	2.32	0.282	Tumor necrosis factor receptor superfamily member 10B precursor
2HFG	2.61	0.252	Tumor necrosis factor receptor superfamily member 13C
2IH3	1.72	0.242	Potassium channel KcsA
2J4W	2.50	0.241	Apical membrane antigen 1
2JEL	2.50	0.280	Histidine-containing protein
2NY1	1.99	0.250	Envelope glycoprotein gp120
2NY7	2.30	0.255	Envelope glycoprotein gp120
2NYY	2.61	0.246	Botulinum neurotoxin type A
2UZI	2.00	0.271	GTPase HRas
2VXQ	1.90	0.226	Pollen allergen Phl p 2
2VXS	2.63	0.264	Interleukin-17A
2VXT	1.49	0.196	Interleukin-18
2W9E	2.90	0.269	Major prion protein
2ZJS	3.20	0.280	Preprotein translocase SecY subunit
2ZUQ	3.30	0.351	Disulfide bond formation protein B
3B9K	2.70	0.284	T-cell surface glycoprotein CD8 β -chain
3BSZ	3.38	0.312	Retinol-binding protein 4
3EFD	2.60	0.276	Potassium channel KcsA
3EO1	3.10	0.279	Transforming growth factor β -3
3EOA	2.80	0.264	Integrin α -L
3ETB	3.80	0.276	Protective antigen
3FKU	3.20	0.286	Hemagglutinin
3FMG	3.40	0.238	Outer capsid glycoprotein VP7
3G6D	3.20	0.263	Interleukin-13
3G6J	3.10	0.282	Complement C3
3GI9	2.48	0.296	Uncharacterized protein MJ0609
3GRW	2.10	0.228	Fibroblast growth factor receptor 3 (fragment)
3H42	2.30	0.209	Proprotein convertase subtilisin/kexin type 9
3HI6	2.30	0.226	Integrin LFA-1
3HQB	3.20	0.324	arginine/agmatine antiporter

Recent progress with X-ray crystallography has generated large numbers of protein structures complexed with other proteins, peptides and small compounds. We recently developed a method for predicting binding sites on protein surfaces for drug-like compounds on the basis of the specific amino acid composition observed at the ligand-binding sites of ligand–protein complexes, as determined by X-ray analysis (Soga *et al.*, 2007a,b). A profile describing the propensity of each of the 20 standard amino acids at the binding sites of the drug-like molecules was obtained for high-quality complex structures and used to define the propensity for ligand binding (PLB). The PLB index is used to predict binding sites based on the likelihood of ligand binding for each protein concavity as evaluated from its amino acid composition. We also introduced a novel index which characterizes the binding site based on the concurrency rate of amino acids (Soga *et al.*, 2008) and used this index to identify a specific concavity, known as a chemocavity, where a specific group of small molecules was extremely likely to bind. Our study showed that amino acid propensity and concurrency were essential characteristics in determining the interactions between a protein and a small molecule. These findings encouraged us to use these characteristics to analyze interactions between epitopes and antigen-binding sites of antibodies.

Here, we used a profile describing the propensity of each of 20×20 standard amino acid pairs at the interface of an antigen–antibody interaction to develop a novel method of predicting epitope residues on the surface of proteins of antigens for an individual antibody.

Materials and methods

Data set

In October 2009, we extracted 200 X-ray structures of protein antigen–antibody complexes from the Protein Data Bank (PDB; Berman *et al.*, 2000). For the present study, proteins were defined as polypeptides more than 21 residues long. Protein structure data were obtained from the non-redundant PDB (<http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpd.html>), maintained by the National Center for Biotechnology Information. A P -value of $10e-7$ was used to determine sequence similarity by BLAST (Altschul *et al.*, 1990). Complex structures containing homologous antigens defined above the threshold were classified into the same antigen–antibody structural groups. If multiple structures in the same structural group shared less than 50% epitope residues, they were further classified into different subgroups. To eliminate redundancy, the high-quality structures were selected using two criteria, the resolution limit of X-ray diffraction and the reliability index of R_{free} . The former guarantees the quality of the experimental data and the latter the quality of refinement of the structure. Complex structures with the highest resolution were selected as representative of each subgroup. If there were multiple structures with the same resolution limit, the structure with the smallest R_{free} was chosen as representative. Seventy-four non-redundant qualified antigen–antibody complex structures were selected (Table I).

As reference for our analyses, we also constructed homo- and heterodimer sets as follows. PDB entries containing only

two chains of protein sequences were classified as homo- or heterodimers. For the present study, sequences with at least one differing residue were defined as heterodimers. The complex structure with the smallest resolution was selected as representative. If several structures had the same resolution, we selected that with the smallest R_{free} as representative. We ultimately obtained 3601 and 649 non-redundant X-ray structures for homo- and heterodimers, respectively.

Interface residues

Surface atoms of a protein within $4.5 \text{ \AA}$ of those on the partner protein were defined as ‘interface atoms’, and residues which included at least one interface atom were defined as ‘interface residues’. As a reference, residues on the protein surface which were not interface residues were defined as ‘surface residues’. The surface atoms were identified using a probe sphere with a radius of $1.4 \text{ \AA}$ by the use of the software system MOE (Molecular Operating Environment, version 2007.0902, Chemical Computing Group, Quebec, Canada). If the probe came into contact with a non-hydrogen atom of the complex, the atom was regarded as a surface atom. The interface residues of antigens were defined as epitope residues, and those of antibodies as paratope residues.

Amino acid propensity

When the number of x -type amino acids at the epitope residues of all of the non-redundant antigen–antibody complex structures was $n_e(x)$, the occurrence rate of amino acid x of epitope residues, $f_e(x)$, was defined as:

$$f_e(x) = \frac{n_e(x)}{\sum_{y=1}^{20} n_e(y)} \quad (1)$$

When the number of x -type amino acids at the surface residues of all of the structures for the homo- and heterodimer sets was $n_s(x)$, the occurrence rate of amino acid x for the surface residues, $f_s(x)$, was defined as:

$$f_s(x) = \frac{n_s(x)}{\sum_{y=1}^{20} n_s(y)} \quad (2)$$

The log-odds ratio of $f_e(x)$ to $f_s(x)$ as amino acid propensity of the epitope was calculated as follows:

$$p_e(x) = \log \left(\frac{f_e(x)}{f_s(x)} \right) \quad (3)$$

Using the occurrence rates of amino acid x in the paratope, $f_p(x)$, homodimer interface, $f_{ho}(x)$, and heterodimer interface, $f_{he}(x)$, the amino acid propensities of the paratope, $p_p(x)$, homodimer interface, $p_{ho}(x)$, and heterodimer interface, $p_{he}(x)$, were calculated similarly.

Occurrence rate of pair-residue sets

The interface residue of a protein and the nearest residue of its partner protein were defined together as a ‘pair-residue set’, and the occurrence rates of pair-residue sets at the epitope–paratope interface, $f_{ep}(x, y)$, were calculated. Here, ‘ x ’ and ‘ y ’ denote an epitope residue and a corresponding paratope residue, respectively.

The occurrence rate of amino acids at an epitope, $f_e(x)$, can be expressed using $f_{ep}(x, y)$ as:

$$f_e(x) = \sum_{y=1}^{20} f_{ep}(x, y) \quad (4)$$

where amino acid x denotes the epitope residue and amino acid y denotes the paratope residue for $f_{ep}(x, y)$.

Similarly, we calculated the occurrence rate of pair-residue sets at the homo-, $f_{ho}(x, y)$, and heterodimer interfaces, $f_{he}(x, y)$. With regard to heterodimers, we conformed the occurrence rate of (x, y) to that of (y, x) by averaging it.

Antibody-specific epitope propensity

Here, we introduced a weighted occurrence rate, $f'_e(x)$, for individual antibodies depending on the amino acid composition of the paratope. The value of $f'_e(x)$ can be calculated by the following equation:

$$f'_e(x) = \sum_{y=1}^{20} [f_{ep}(x, y) \times n_p(y)] \quad (5)$$

where the weighting factor, $n_p(y)$, denotes the number of y -type amino acids in the paratope. Using this weighted occurrence rate of epitope amino acids, we calculated antibody-specific epitope propensity (ASEP), $p'_e(x)$, as follows:

$$p'_e(x) = \log \left[\frac{f'_e(x)}{f_s(x)} \right] \quad (6)$$

First step in epitope prediction

Our prediction method consists of two steps. Candidate epitope residues are first identified using a conventional epitope prediction method, and then narrowed down to predict final epitope residues. The first step was done with reference to the DiscoTope algorithm (Haste Andersen *et al.*, 2006), a method which incorporates amino acid propensity and the tertiary protein structure of antigens. However, we used the amino acid propensity index of $p_e(x)$ in place of DiscoTope's epitope propensity because the method with $p_e(x)$ was almost same or more efficient than the one with DiscoTope's epitope propensity (Supplementary materials, Supplementary data are available at PEDS online).

The details of the DiscoTope algorithm are as follows: we first calculated the sequential average of the epitope propensity values with a window size of 9, and then summed up the averages for all residues with α -carbons within a 10-Å distance of each residue in the tertiary structure, a summation known as the structural proximity sum. We then calculated the prediction score by subtracting half of the contact number (number of α -carbons within 10 Å of the α -carbon of the residue) from the structural proximity sum for each residue. As the prediction score increases, the estimated likelihood of the candidate actually being an epitope increases. We then extracted epitope residue candidates with a prediction score above -7.7 , the default value recommended by DiscoTope.

Second step in epitope prediction

ASEP index for a residue ' i ' predicted in the first step is calculated by summing up the $p'_e(x)$ value of amino acids x

whose α -carbons are located within 10 Å of the α -carbons of the residue ' i '. The ASEP index was defined as follows and employed in the second step.

$$\text{ASEP}(i) = \sum_{x=1}^{20} p'_e(x) c_i(x) \quad (7)$$

Here, the $c_i(x)$ denotes the number of amino acid x . The higher the ASEP index of a residue, the higher is its epitope propensity. We predicted final epitope residues by eliminating candidate residues with ASEP indices in the bottom 10%, 20%, 30%, 40% or 50%.

Assessment of efficiency of the second step using the leave-one-out approach

We assessed the efficiency of the second step using the leave-one-out approach. Amino acid propensity of an epitope and ASEP data were prepared by leaving out each of the complexes in turn and using that antigen–antibody complex for validation. For each antigen in the data set, we examined the degree of enrichment with true positives. The positive predictive value (PPV) is the ratio of the number of true positive residues to that of predicted residues and is given as follows:

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

where TP denotes the number of true positives and FP the number of false positives. Using this formula, we compared

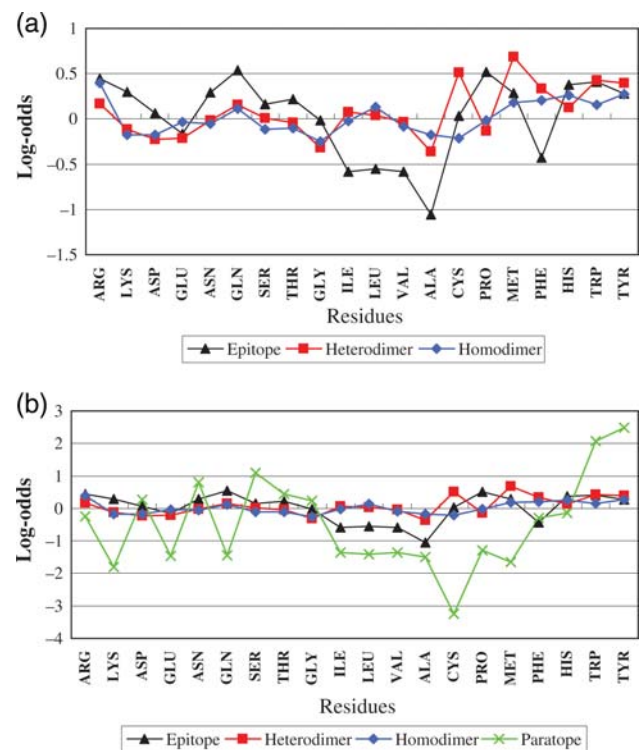


Fig. 1. (a) Amino acid propensities of the epitope (black), heterodimer interface (red) and homodimer interface (blue). (b) Amino acid propensities of the epitope (black), paratope (green), heterodimer interface (red), and homodimer interface (blue). The amino acid propensities of epitope and paratope were calculated from 74 antigen–antibody complexes.

the PPV of the second-step prediction (PPV_{step2}) with that of the first-step prediction (PPV_{step1}). For validation, we examined the enrichment effect in the second step by the following formula:

$$\Delta PPV_{\text{step2-step1}} = PPV_{\text{step2}} - PPV_{\text{step1}} \quad (9)$$

The positive value of $\Delta PPV_{\text{step2-step1}}$ indicates the substantial enrichment in the second step. In addition, to demonstrate the statistical significance of the results, we tested the null hypothesis that the number of complexes with $\Delta PPV_{\text{step2-step1}} > 0$ produced by the random ranking method was the same as that produced by the ASEP index. Briefly, random ranking was performed with 10 000 iterations using the Fisher–Yates randomizing shuffle algorithm. For each iteration, we counted the number of complexes with $\Delta PPV_{\text{step2-step1}} > 0$ and generated a null distribution. From this distribution, we then computed P -value as a statistical significance.

Results and discussion

Amino acid propensity of epitopes

In our previous work, we determined amino acid PLB by comparing the amino acid composition of small ligand-binding sites and the general protein surface. These previous findings enabled the development of methods for predicting ligand-binding sites (Soga *et al.*, 2007a,b), a process which requires the comparison of binding site residues not with overall residues, but rather with surface residues only. Applying this idea to the present study, we calculated the amino acid propensities of epitopes and homo- and heterodimer interfaces (Fig. 1a and Table 4S, Supplementary data are available at PEDS online). Propensities of epitopes and paratopes were calculated for 74 antigen–antibody complexes. Propensities of homo- and heterodimers were closely similar: both preferred residues such as Phe, His, Trp, Tyr, Met, Leu, Gln and Arg, and avoided Lys, Asp, Glu, Asn, Thr, Gly, Val, Ala and Pro. Of particular note, the two sulfur-containing amino acids, Cys and Met, were more preferred by heterodimers than homodimers.

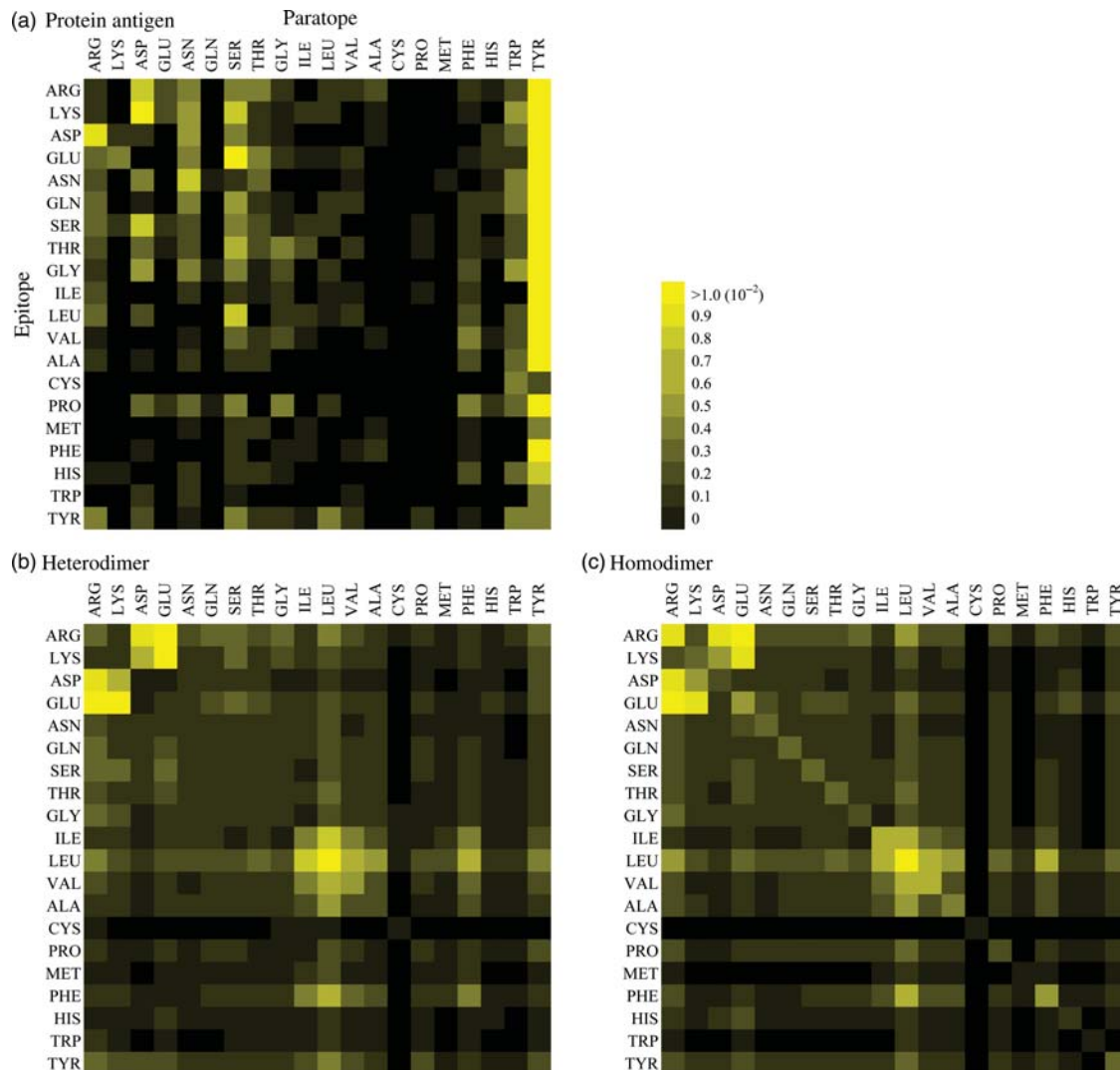


Fig. 2. Occurrence rates of 20 × 20 amino acid combinations for pair-residue sets. (a) Antigen–antibody interface with the left side indicating epitope amino acids and the upper side antibody amino acids, namely paratope amino acids. Occurrence was determined from 74 antigen–antibody complexes; (b) heterodimer interface with the left and upper amino acids indicating the interface residues of the heterodimer; (c) homodimer interface with the left and upper amino acids indicating the interface residues of the homodimer.

The propensities of amino acids identified here differed from those obtained by Jones and Thornton (1997). One reason may be that Jones and Thornton compared the interface residue with all residues in a protein, whereas we compared the interface residue with the surface residues only. Additional causes may be differences in the data sets, and in details of the algorithms used.

The amino acid propensities of epitopes differed significantly from those of homo- and heterodimers. Small hydrophilic residues, such as Ser, Thr, Asn and Gln, were preferred, whereas small hydrophobic residues, such as Ala, Val, Leu and Ile, were not. Also interesting is that Trp, Tyr and His were preferred, whereas Phe was not. Arg, Lys and Asp were preferred, whereas Glu was not. These propensities are considered intrinsic characteristics of epitopes.

We also compared the amino acid propensities of paratopes with those of epitopes and homo- and heterodimer interfaces (Fig. 1b). Paratopes had markedly different amino acid propensities, preferring Trp, Tyr, Asn and Ser.

Occurrence rate of amino acid pairs in pair-residue sets

Occurrence rates of the 20 × 20 amino acid pairs in the pair-residue sets are shown in Fig. 2. The occurrence rate of epitope–paratope interactions was determined from 74 antigen–antibody complexes. The profile of heterodimers (Fig. 2b and Table 6S, Supplementary data are available at PEDS online) was similar to that of homodimers (Fig. 2c and Table 7S, Supplementary data are available at PEDS online). Interestingly, aromatic residues were preferred as interface residues in epitopes, heterodimers and homodimers (Fig. 1), whereas aromatic–aromatic interactions were less preferred in pair-residue sets. It is also noteworthy that interactions between small hydrophobic residues were highly frequent in the pair-residue sets of homodimer and heterodimer, even though small hydrophobic residues were less favorable as interface residues (Fig. 1). These results indicate that preferences for amino acid pairs did not depend on the frequency at which the relevant amino acids appeared at the interfaces. Interactions between negative- and positive-charged residues were frequently observed in the pair-residue sets. In homodimer interfaces in particular, Arg–Arg interactions were frequently observed.

The profile of the occurrence rates of epitope–paratope interactions (Fig. 2a and Table 5S, Supplementary data are available at PEDS online) differed markedly from those of other protein–protein interfaces. A strong preference for Tyr in paratopes was preserved in the pair-residue sets. Interestingly, however, interactions between Tyr in paratopes and His, Trp or Tyr in epitopes were less frequent, despite the status of these aromatic residues as preferred epitope residues. The interaction between Tyr in the paratope and Phe, however, was more frequently observed, although Phe was not preferred as an epitope residue. In addition, Asp was more frequently involved than Glu in interactions with Arg and Lys in epitope–paratope pairs. This preference was not observed in homo- and heterodimer interactions. These results indicate that epitope–paratope interactions require specific amino acid pairs.

Prediction of paratope-specific epitope residues

We predicted epitope residues for 74 antigens in two steps using the leave-one-out approach. ASEP indices were calculated for 4477 candidate residues in 74 antigens.

Results are given in Table II. The first column shows the lowest percentage which is eliminated and the next three columns show the number of antigens with $\Delta PPV_{step2-step1}$

Table II. Validation results of predictions

Bottom %	$\Delta PPV_{step2-step1}$			P-value
	>0	<0	0	
50	40	27	7	0.1238
40	43	28	3	0.0630
30	43	28	3	0.1285
20	50	20	4	0.0005*
10	49	22	3	0.0045*

The first column shows the lowest percentage which was eliminated, and the next three columns show the number of cases with $\Delta PPV_{step2-step1}$ greater than, less than or equal to zero, respectively. The last column shows P-values obtained with random permutation. *Statistically significant at $P < 0.05$.

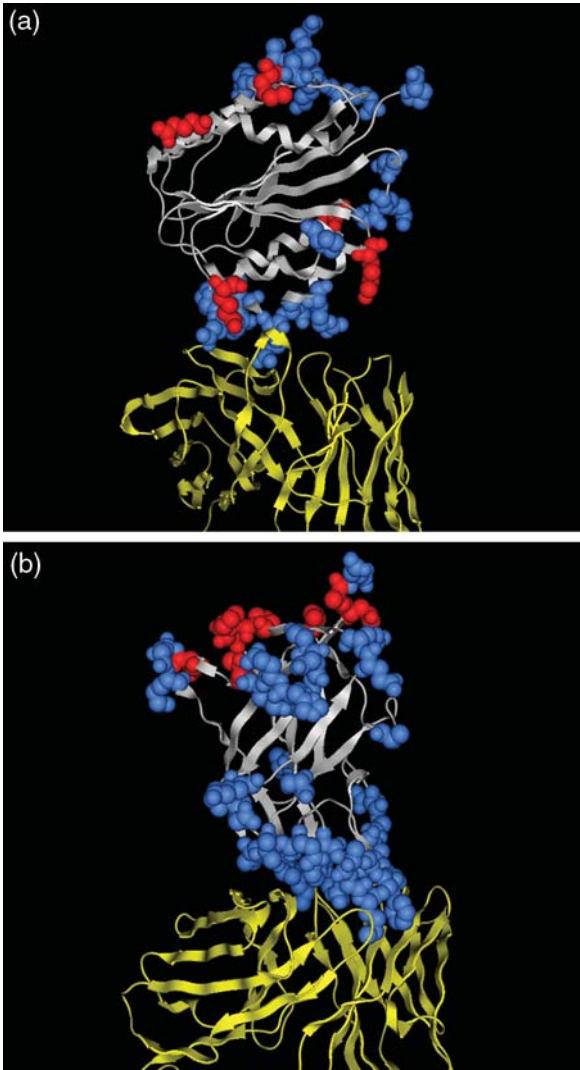


Fig. 3. Predicted and true epitope residues for (a) PDB code: 1FNS and (b) PDB code: 1IQD. Antigen and antibody backbones are indicated by white and yellow ribbons, respectively. Candidate epitope residues predicted in the first step are shown using space-filling models. Residues with ASEP indices in the bottom 20% are colored red, and the remaining residues are colored cyan. For both cases, several false positives were successfully eliminated.

greater than, less than or equal to zero, respectively. Success rates were 66% and 54% following elimination of the bottom 10% and 50%, respectively, and the bottom 10% ($P = 0.0045$) and 20% ($P = 0.0005$) were found to be statistically significant. Several other cases also supported the conclusion that our method was more effective than the random method. These results clearly indicate that the ASEP index effectively eliminated false positives and enriched PPV.

Two typical examples of prediction

The following two examples illustrate the efficacy of the ASEP index in distinguishing true epitope residues from a pool of candidate epitope residues if the bottom 20% of candidates is eliminated. Unsuccessful examples are described in supplementary materials (Supplementary data are available at *PEDS* online).

The first example is the complex between a mutant von Willebrand factor A1 domain and antibody NMC4 (PDB code: 1FNS). In the first step, 24 residues were predicted as candidate residues. Predicted and true epitope residues are compared in Fig. 3a and Table III, which also shows the ASEP index, the true epitope residue (answer) and the contact number for each candidate residue. In Fig. 3a, the backbone of the antigen is colored white and the antibody yellow. Residues predicted in the first step are depicted by a space filling model. The residues predicted using the ASEP index (top 80%) and those eliminated by the ASEP index (bottom 20%) are shown in blue and red, respectively. These results clearly demonstrate the ability of the ASEP index in predicting epitope.

The second example is a complex between the human factor VIII C2 domain and antibody BO2C11 (PDB code:

Table III. Residues predicted in the first step (PDB code:1FNS)

Residue number	Residue	ASEP index	Answer	Contact number
571	Arg	5.17		13
573	Arg	4.11		12
628	Gln	3.86	True epitope residue	9
572	Lys	3.83		13
629	Arg	3.51	True epitope residue	9
632	Arg	3.40	True epitope residue	12
630	Met	3.26		16
608	Lys	3.23		12
627	Pro	3.03	True epitope residue	12
576	Glu	2.73		11
508	Tyr	2.58		15
636	Arg	2.35	True epitope residue	15
633	Asn	2.28	True epitope residue	14
610	Asp	2.24		12
607	Ser	1.66		12
702	Pro	1.52		6
579	Arg	1.48		13
575	Ser	1.41		14
670	Pro	1.37		10
.....				
507	Met	1.27		13
685	Gln	1.26		12
643	Lys	1.18		12
660	Lys	1.10	True epitope residue	10
604	Gln	0.96		12

Residues predicted using the ASEP index (top 80%) are shown above the dotted line, whereas those eliminated using the index (bottom 20%) lie below. The 'answer' column defines the true epitope residues.

IIQD). In the first step, 42 residues were predicted. Residues predicted using the ASEP index (top 80%) and those eliminated using the index (bottom 20%) are colored in red and blue, respectively. Final results are shown in Fig. 3b and Table IV. These findings also clearly demonstrate the ability of the ASEP index to successfully enrich candidate residues by eliminating false-positive residues.

Conclusion

We developed a novel two-step method of predicting epitope residues using information on the amino acid composition of the paratope of an individual antibody. In the first step, candidate epitope residues were predicted by the conventional prediction method. In the second step, these candidate residues were narrowed down using the ASEP index, with assessments carried out using the leave-one-out approach. Success rates of 66% and 54% were achieved after

Table IV. Residues predicted in the first step (PDB code:IIQD)

Residue number	Residue	ASEP index	Answer	Contact number
2252	Leu	-1.57	True epitope residue	7
2251	Leu	-1.70	True epitope residue	6
2270	Gln	-1.78		10
2200	Phe	-2.14	True epitope residue	8
2269	His	-2.14		10
2199	Met	-2.40	True epitope residue	6
2277	Asn	-2.45		7
2227	Lys	-2.48		12
2214	Gly	-2.61		9
2267	Asp	-2.67		13
2250	Ser	-3.00	True epitope residue	13
2329	Gln	-3.27		7
2201	Ala	-3.27		12
2278	Gly	-3.41		8
2193	Ser	-3.57		13
2198	Asn	-3.63	True epitope residue	10
2197	Thr	-3.68	True epitope residue	15
2225	Asn	-4.13		16
2196	Phe	-4.28	True epitope residue	17
2223	Val	-4.36	True epitope residue	14
2253	Thr	-4.43	True epitope residue	12
2266	Gln	-4.51		18
2222	Gln	-4.60	True epitope residue	17
2268	Gly	-4.61		16
2236	Lys	-4.65		13
2235	Gln	-4.77		14
2213	Gln	-4.96		12
2215	Arg	-5.09	True epitope residue	13
2279	Lys	-5.56		10
2226	Pro	-5.61		15
2315	His	-5.75	True epitope residue	13
2271	Trp	-5.95		16
2188	Ala	-6.06		11
2291	Thr	-6.47		15
.....				
2272	Thr	-6.76		15
2276	Gln	-7.14		13
2328	Ala	-7.14		13
2174	Cys	-7.63		14
2327	Glu	-7.96		14
2298	Asp	-7.97		13
2299	Pro	-8.85		13
2300	Pro	-9.40		15

Residues predicted using the ASEP index (top 80%) are shown above the dotted line, whereas those eliminated using the index (bottom 20%) lie below. The 'answer' column defines the true epitope residues.

elimination of the bottom 10% and 50% of residues, respectively. These results suggest that prediction by the present method in combination with suitable mutation experiments will greatly increase the success rate of epitope analysis. Prediction of paratopes is comparatively easy because these are generally composed of the solvent-exposed residues of the six CDRs, mainly CDR-H3. We recently developed methods for the structural classification of CDR-H3 and CDR-L3, the two most important CDRs for antigen binding (Chothia *et al.*, 1989; Al-Lazikani *et al.*, 1997; Kuroda *et al.*, 2008, 2009). Given the efficacy of these methods for antibody modeling, we expect that the combined use of the ASEP index with these recent advances in antibody modeling will improve epitope analysis, allow rational antibody design using the antibody sequence, and eventually accelerate antibody drug development.

Supplementary data

Supplementary data are available at *PEDS* online

Acknowledgements

We thank Professor Haruki Nakamura of Osaka University for his helpful discussions.

Funding

This work was partly supported by a Grant-in-Aid for Scientific Research (A) (19200024) from MEXT (Ministry of Education, Culture, Sports, Science and Technology) for N.H. and by Research Fellowships from the Japan Society for the Promotion of Science for Young Scientists for D.K.

References

- Alix,A.J. (1999) *Vaccine*, **18**, 311–314.
- Al-Lazikani,B., Lesk,A.M. and Chothia,C (1997) *J. Mol. Biol.*, **273**, 927–948.
- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J (1990) *J. Mol. Biol.*, **215**, 403–410.
- Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E (2000) *Nucleic Acids Res.*, **28**, 235–242.
- Bubli,E.M., Freund,N.T., Mayrose,I., Penn,O., Roitburd-Berman,A., Rubinstein,N.D., Pupko,T. and Gershoni,J.M (2007) *Proteins*, **68**, 294–304.
- Chothia,C., Lesk,A.M., Tramontano,A., Levitt,M., Smith-Gill,S.J., Air,G., Sheriff,S., Padlan,E.A., Davies,D., Tulip,W.R., Colman,P.M., Spinelli,S., Alzari,P.M. and Poljak,R.J (1989) *Nature*, **342**, 877–883.
- Frank,R. (2002) *J. Immunol. Methods*, **267**, 13–26.
- Haste Andersen,P., Nielsen,M. and Lund,O (2006) *Protein Sci.*, **15**, 2558–2567.
- Hopp,T.P. and Woods,K.R. (1981) *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.
- Hu,S., Zhu,Z., Li,L., Chang,L., Li,W., Cheng,L., Teng,M. and Liu,J (2008) *Proteins*, **70**, 938–949.
- Huang,Y.X., Bao,Y.L., Guo,S.Y., Wang,Y., Zhou,C.G. and Li,Y.X (2008) *BMC Bioinformatics*, **16**, 538.
- Jones,S. and Thornton,J.M. (1997) *J. Mol. Biol.*, **272**, 121–132.
- Kuroda,D., Shirai,H., Kobori,M. and Nakamura,H (2008) *Proteins*, **73**, 608–620.
- Kuroda,D., Shirai,H., Kobori,M. and Nakamura,H (2009) *Proteins*, **75**, 139–146.
- Larsen,J.E., Lund,O. and Nielsen,M (2006) *Immunome. Res.*, **2**, 2.
- Lu,C., Ferzly,M., Takagi,J. and Springer,T.A (2001) *J. Immunol.*, **166**, 5629–5637.
- Odorico,M. and Pellequer,J.L. (2003) *J. Mol. Recognit.*, **16**, 20–22.
- Pellequer,J.L., Westhof,E. and Van Regenmortel,M.H (1991) *Methods Enzymol.*, **203**, 176–201.
- Soga,S., Shirai,H., Kobori,M. and Hirayama,N (2007a) *J. Chem. Inf. Model.*, **47**, 400–406.
- Soga,S., Shirai,H., Kobori,M. and Hirayama,N (2007b) *J. Chem. Inf. Model.*, **47**, 2287–2292.
- Soga,S., Shirai,H., Kobori,M. and Hirayama,N (2008) *J. Chem. Inf. Model.*, **48**, 1679–1685.
- Söllner,J. and Mayer,B. (2006) *J. Mol. Recognit.*, **19**, 200–208.
- Zou,P., Liu,W., Wu,F. and Chen,Y.H (2008) *FEMS Immunol. Med. Microbiol.*, **53**, 79–84.