IMMUNOME RESEARCH

RESEARCH Open Access

Does difference exist between epitope and non-epitope residues?

Analysis of the physicochemical and structural properties on conformational epitopes from B-cell protein antigens.

Jing Sun[1], Tianlei Xu[1], Shuning Wang[2], Guoqing Li[2], Di Wu[2§], Zhiwei Cao[1,2§]

Abstract

Background

As an essential step of adaptive immune response, the recognition between antigen and antibody triggers a series of self-protection mechanisms. Therefore, the prediction of antibody-binding sites (B-cell epitope) for protein antigens is an important field in immunology research. The performance of current prediction methods is far from satisfying, especially for conformational epitope prediction. Here a multi-perspective analysis was carried on with a comprehensive B-cell conformational epitope dataset, which contains 161 immunoglobulin complex structures collected from PDB, corresponding to 166 unique computationally defined epitopes. These conformational epitopes were described with parameters from different perspectives, including characteristics of epitope itself, comparison to non-epitope surface areas, and interaction pattern with antibody.

Results

According to the analysis results, B-cell conformational epitopes were relatively constant both in the number of composing residues and the accessible surface area. Though composed of spatially clustering residues, there were sequentially linear segments exist in these epitopes. Besides, statistical differences were found between epitope and non-epitope surface residues with parameters in residual and structural levels. Compared to non-epitope surface residues, epitope ones were more accessible. Amino acid enrichment and preference for specific types of residue-pair set on epitope areas have also been observed. Several amino acid properties from AAindex have been proven to distinguish epitope residues from non-epitope surface ones. Additionally, epitope residues tended to be less conservative under the environmental pressure. Measured by topological parameters, epitope residues were surrounded with fewer residues but in a more compact way. The occurrences of residue-pair sets between epitope and paratope also showed some patterns.

Conclusions

Results indicate that, certain rules do exist in conformational epitopes in terms of size and sequential continuity. Statistical differences have been found between epitope and non-epitope surface residues in residual and structural levels. Such differences indicate the existence of distinctiveness for conformation epitopes. On the other hand, there was no accordant estimation for higher or lower values derived from any parameter for epitope residues compared with non-epitope surface residues. This observation further confirms the complicacy of characteristics for conformational epitope. Under such circumstance, it will be a more effective and accurate approach to combine several parameters to predict the conformation epitope. Finding conformational epitopes and analysing their properties is an important step to identify internal formation mechanism of conformational epitopes and this study will help future development of new prediction tools.

BACKGROUND

The adaptive immune response takes on the main protective responsibility for human body to eliminate antigens. Among various cells being involved in this response, B-cell has attracted widely interests for its central role in the antigen sterilization process. Antibodies are secreted by B-cell to recognize and bind the antigens specifically. The corresponding antibody recognizes a small antigen part which is known as the epitope. Epitope is composed of epitope residues defined in spatial approach. The contact of antigen and antibody erects on the structural complementary and residual affinity between epitope of the antigen and paratope of the antibody.

According to the sequential continuity, epitopes can be divided into linear and conformational epitope. The linear epitopes are a stretch of residues continuous in sequence, and the conformational ones are constituted by residues with sequential discontinuity but spatial vicinity [1]. Previous studies have shown that most of B-cell epitopes are conformational ones, which contributes to the complexity of B-cell epitope identification [2].

1 Department of Biomedical Engineering, College Life Science and Technology, Tongji University, Shanghai, 200092, China.
2 Shanghai Center for Bioinformation Technology, Qinzhou Rd 100, Building 1, 12F, Shanghai, 200235, China.
§Corresponding Author
Email addresses:
Email addresses:
JS: sunjing1010@gmail.com
SNW: wshuning.wang@gmail.com
GQL: linational86@gmail.com
DW: wudi@scbit.org
ZWC: zweao@tongji.edu.cn

Identifying epitopes is crucial for antibody design, disease diagnosis and immunological therapy [2]. Although many experimental methods have been tried to determine B-cell epitopes, the validated ones are still insufficient and limited [4]. Under such circumstance, various computational methods have been applied to predict epitope residues [5].

The initial attempt of B-cell epitope prediction dates back to 1981 when Hopp and Wodds brought forward the correlation between the charged hydrophilic amino acids and epitope residues [6]. Since then, various parameters have been proposed to be associated with protein antigenicity and assist in locating B-cell epitopes. In Westhof et al.'s work, the flexibility of backbone was used as an effective criterion for judging antigenicity [7]. Amino acid composition of antigenic regions has been calculated. The special preference for amino acid in epitope regions was used as antigenicity criterion and applied to predict the epitope region of bovine ribonucleas by Welling et al. [8]. In 1986, Novotny et al. demonstrated that the correlation between accessible surface area (ASA) and antigenicity was superior to previous findings [9]. At the same year, Parker et al. have derived new hydrophilicity scale and found correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites [10]. On the other hand, the structural character was also discussed to determine the epitope regions of a protein. Thornton et al. showed that antigenic sites normally protruded considerably from the protein surface and concluded that this property was an extraordinary characteristic of antigenicity [11]. Following, a semi-empirical method was developed to predict antigenic determinants on proteins according to both physicochemical properties of amino acid residues and their frequencies of occurrence by Kolaskar et al [12].

Based on the results from previous researches, several B-cell conformational epitope prediction methods have been proposed. The idea of these methods is to determine which residues are more likely to be epitope residues on the surface of protein structure. The CEP server [13] was developed as the first conformational epitope prediction server, where The concept of "accessibility of residues" was introduced to describe the structural characteristics of antigen conformational epitopes. A subsequent effort was chosen to combine propensity scales with the effects of conformational proximity and surface exposure in DiscoTope [2]. PEPITO [14] was developed based on DiscoTope with the use of a combination of amino acid propensity scores and half sphere exposure values at multiple distances. Comparing with peers, better performance was achieved by SEPPA [15]. More conformational factors have been taken into consideration in its algorithm, such as topological features and residue-triangle units' occurrences. Generally, the performances of conformational epitope prediction have been improved with more features introduced in prediction methods and the increasing number of the antigen-antibody complex crystal structures for analysing [16]. However, in a review work done by Xu et al. [17], comparison was performed to elucidate that the performances of current web-servers were significantly affected by their training datasets and the algorithms adopted: under a testing dataset of 110 experimentally determined conformational epitopes, AUC (Area Under the Curve) values were under 0.65 with all the servers. All these results indicate that despite of the continuous efforts, the precision of computational prediction is still less than satisfactory. And several questions related to conformational epitope are still open to be answered [4]. Does difference exist between epitope and non-epitope residues? Does B-cell conformational epitope describable?

In order to answer above questions, a latest comprehensive dataset was firstly collected to derive various epitope residues from available immunoglobulin complex structures. Then a series of analysis were examined on these data from three aspects: (1) parameters describing conformational B-cell epitope, such as epitope size and sequential continuity; (2) comparison between epitope and non-epitope surface residues from residual and geometrical levels, including 7 parameters and 544 indices from AAindex database; (3) residue interacting pattern between antigen and antibody. This work is aimed to systematically detect the intrinsic features of conformational B-cell epitope.

RESULTS

A variety of parameters have been applied to describe B-cell conformational epitope and make comparison between residues from epitope and non-epitope surface regions (Table 1). These parameters are classified into four perspectives, including the general, residual,

Perspectives to depict a B-cell conformational epitope		
	General characteristics	Epitope size
		Sequential continuity
	Residual characteristics	Residue accessibility
		Amino acid preference
		Residue-pair preference
		AAindex indices
		Evolutionary conservation
	Structural characteristics	Topological analysis
		Planarity analysis
	Interaction pattern with antibody	Epitope-paratope residue-pair preference

Table1. Schematic table of parameters used in analysis. Epitope residues have been investigated from four main perspectives in our work.

structural and interaction pattern characteristics. From these aspects, a comprehensive analysis has been performed.

General characteristics of B-cell conformational epitope

Size of conformational epitope

Antibody specifically recognizes antigen epitope, which is composed of spatially clustering residues. So the size of epitope is a general characteristic for conformational epitope.

Considering the diversity of protein antigen sizes, from 50 residues to more than 1000 residues, it is interesting to detect whether the size of conformational epitope is also varying with the variation of protein antigen size. A direct way to measure the epitope size is the number of residues included in epitope. In our dataset, the number of residues in an epitope mainly distributed from 15 to 30 residues (137 in 166 data), with an average number of 22.40±8.03 (μ ± σ). Considering that epitopes with same number of residues might vary considerably in areas, the accessible surface area (ASA) was introduced as another measure for the epitope size. ASA values have been calculated for residues in epitopes. These values of residues were summed up in each epitope, ranging from 208.29 to 5705.42Å2, with an average value of 846.59±278.87Å2.

Above two size parameters have also been calculated for the whole antigen. The number of residues ranges from 51 to 1267, with an average number of 209.04±154.40 (μ ± σ). The ASA ranges from 3673.57 to 52820.72Å2, with an average value of 10827.18±6616.09Å2. The distribution of these data has been plotted in Figure 1 and there was no clearly correlation between epitope and antigen size in both parameters. Besides, in order to evaluate the variation scope of epitope and protein antigen, the coefficient of variation ($C_v = \frac{\sigma}{\mu}$) was calculated for the number of residues (0.36 and 0.74 for epitopes and antigens respectively) and ASA values (0.31 and 0.61 for epitopes and antigens respectively).

Besides the number of residues in a conformational epitope, the diameter of epitope region was considered as another instrument for measuring epitope size. A series of distances among residues were inspected for each epitope in our dataset, including the largest distance between epitope residues, the average distance among residues, and the average distance between central residue and peripheral ones. Averaged among the 166 epitopes in our dataset, the largest residue distance is 26.39±6.87Å, and the average distance is 10.37±2.54Å. As to the average distance between the central residue and peripheral residues, it is 7.15±1.91Å. The coefficient of variation was also calculated for the distribution of these distances, and the results were 0.26, 0.25 and 0.27 for these distances.

Concluded from the number of residues, the residue distances and the residue ASA values, the size of epitopes was relatively constant compared with the size of antigen proteins which vary largely among our dataset. This partly attributes to the reality that the binding region of antibody is mainly constituted by the CDR regions, which has a comparatively similar structure and constant ASA areas. Thereafter, it is necessary to keep a constant size for the antigen in opposite side of binding interface.

Sequential continuity of residues in conformational B-cell epitope

Though the conformational epitope is mainly composed of discrete residues, it has been noticed that there still are some residues linear in primary sequence included in the conformational epitope. Are these linear continuous residues prevalent in conformational epitope? What is the percentage for these residues in conformational epitope?

Here, a concept of segment was introduced to describe the sequential continuity of residues in epitope [2]. A stretch of epitope residues linear in primary sequence was considered as a segment, and the number of residues in a segment was taken as the length of the segment (the length of segment composed of a single residue is one). As to the conformational B-cell epitopes in our dataset, the proportion of residues in different lengths of segments, the number of segments in an epitope, and the

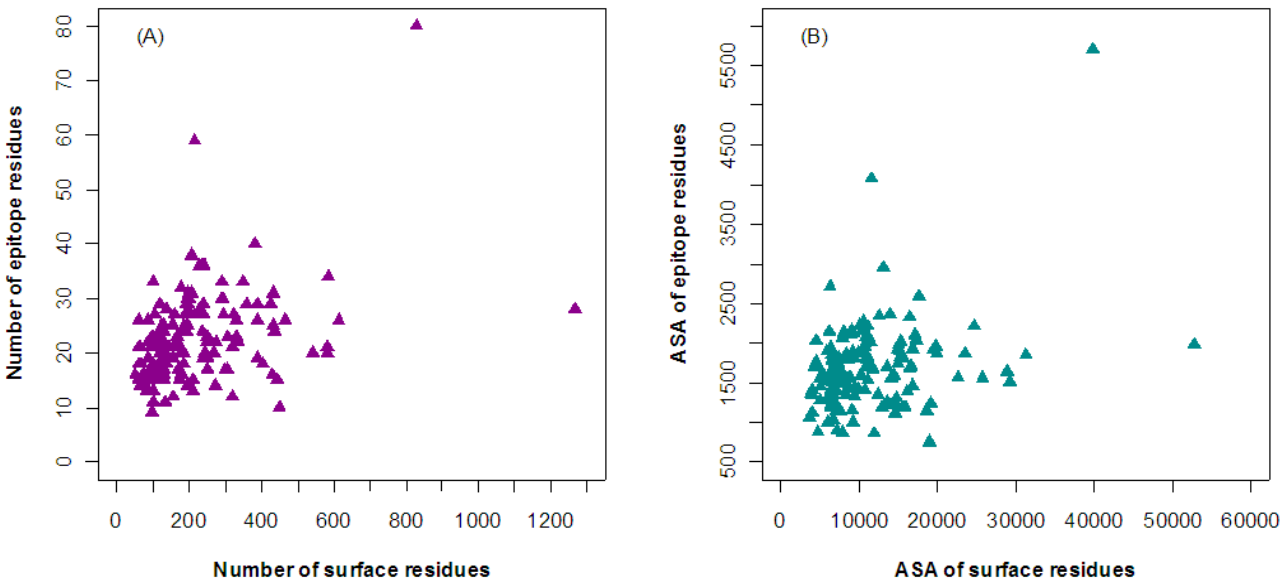

Figure 1. Size variation of epitope and surface areas. (A) the X-axis refers to the residue number of antigen surface, while Y-axis refers to corresponding epitope; (B) the X-axis indicate the sum of ASA values of surface residues, while Y-axis is the sum of epitope residues.

lengths of segments were analysed for all the 166 epitopes.
The sequential continuity analysis was concluded in two aspects:
1. The number of segments in one epitope. Except one epitope which contains only one segment, all the other conformational epitopes (165 in 166) contain several sequential segments with diverse segment lengths.
2. The length of segments. Although the lengths of about 80% segments in all epitopes were of less than 3 residues, there always existed a segment with a length more than 3 residues included in most conformational epitopes (165 in 166), and almost 85% (143 in 166) epitopes have a segment with length more than 5 residues. Although the sequentially discrete residues took up a large proportion in conformational epitope, it was a common phenomenon that there exists at least one linear segment in these epitopes. It might imply a significant role of these linear segments in the specific recognition of antibody.

Residual characteristics of conformational B-cell epitope

Residue accessibility

In the process of protein binding, interacting residues are expected to have relatively higher accessibility to facilitate the contact with interacting counterpart [18]. Whether the residues in epitope are more accessible than non-epitope surface residues? The ASA was first described for drawing the van der Waal's surface of a protein molecule by Lee & Richards in 1971 [19]. In Kulkarni-Kale U and his colleagues' work in 2005 [13], the ASA value was firstly used as a discriminator and the only one in their prediction server of CEP, the first conformational epitope prediction server. Here, relative ASA value was introduced to evaluate the accessibility of residues in the consideration that amino acid type with larger volume tends to present higher surface accessibility area. It was calculated as ASA value against the ASA index to eliminate the volume bias for different amino acids. The ASA index for twenty types of amino acids was the ASA value of residue X in tri-peptide ALA-X-ALA[20]. Results showed that the relative ASA values of epitope residues were generally higher than that of non-epitope surface residues, and such difference was statistically significant in 82 (49.40%) out of all 166 epitopes (*Mann-Whitney test, $p<0.05$*).
To observe the difference in more detail, the accessibility was further compared between epitope and non-epitope surface residues for 20 types of amino acids separately. Except for CYS, all other types of amino acids are more accessible in epitope than in non-epitope surfaces with statistical significance (*Mann-Whitney test, $p<0.05$*). Although the difference of accessibility for CYS residues in epitope and surface was not statistically significant, higher accessibility for CYS in epitope has been observed in our dataset. It was probably due to several appearances of CYS residues reside at the terminal parts of antigen protein chain, which are always non-epitope regions and with higher ASA values.

Preference of amino acid in epitope

The preference of amino acids in epitope and non-epitope surface was also calculated. The results were shown in Figure 2. Compared with the occurrence of amino acids in non-epitope surface regions, TRP, TYR, ARG and HIS seem to be more preferred in epitope regions, while CYS, ALA and VAL have less appearance frequency.
Amino acids on the surface adopt different conformation which might contribute extremely different ASA values. To eliminate this possible bias, the preference of amino acid was re-calculated with the consideration of ASA values. The results were similar, and there was no obvious difference between two kinds of preference observed.

On the other hand, amino acids were grouped into different classes according to the properties, including charge, polarity, residue volume, side chain group and so on (Table S1). The epitope preference for special

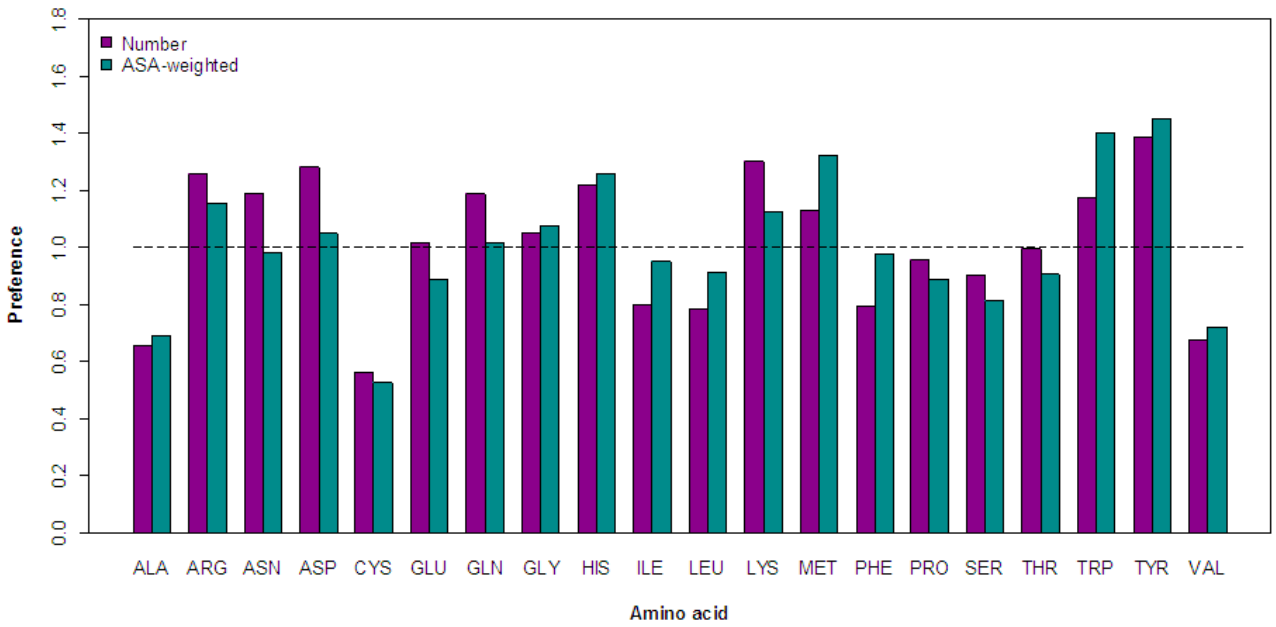

Figure 2. The preference of amino acid in epitope. Twenty types of amino acid are listed in X-axis; the Y-axis indicates their preference in epitope areas. The bars in magenta color are the preference based on the number of residues and the cyan bars refer to ASA-weighted preference.

characteristics was inspected among these classes. It could be found that charged amino acids showed more preference in epitope regions than the neutral ones, especially amino acids in the basic group, including ARG, HIS and LYS. Comparing other opposite characters, the polar, large and aromatic groups show higher preference than the hydrophobic, small and aliphatic groups. In general, the aromatic, charged and polar residues are generally preferred in epitopes due to their capability to form a multitude of interactions with antibodies. These results are consistent with previous studies which have been done on protein-protein interaction [21].

Preference of residue-neighbor set in epitope

The preference of residue-neighbor sets has been calculated. The results were displayed in heat map (Figure 3). In the figure, amino acids are sorted according to the epitope preference in both x-axis and y-axis. The color palette from green to red indicates a growing preference for residue-neighbor sets in epitope. The red color indicates the higher probability of appearance for residue-neighbor sets in epitope areas, while the green color mean less appearance. In general, it was more prevalent for the combination of residues among the epitope preferred amino acids in epitope region. Compared with non-epitope surface region, ASN-TYR, HIS-TYR and HIS-MET residue-neighbor sets were preferred in epitope region. Residue-neighbor sets were also grouped according to the property classification. Charged and aromatic residues have been observed to be preferred as the epitope residues, and the residue-neighbor sets involving the acidic, basic or aromatic residues were more frequent observed in epitope region.

AAindex indices

To further discuss the relationship between the properties of amino acid and the characteristics of conformational epitope, the indices in AAindex [22] were calculated for the residues in epitope regions and non-epitope surface. The results of AAindex scoring were compared to detect the properties which could be used to discriminate the epitope and non-epitope residues. According to the classification of AAindex, the indices have been grouped into

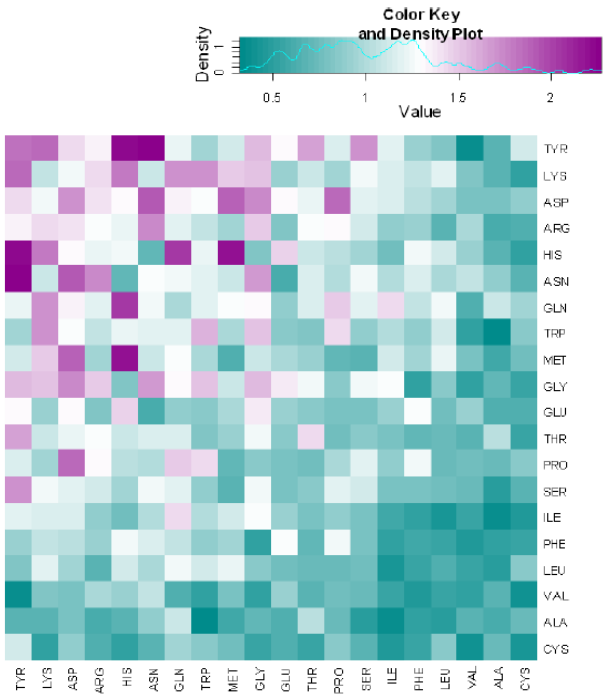

Figure 3. The preference of residue-neighbor sets in epitope. Residues are sorted according to individual amino acid epitope preference. The color palette from cyan to magenta indicates an increasing preference in epitope.

six groups to depict different kinds of properties: Group A (helix and turn conformation) and Group B (sheet conformation) for the secondary structure conformation, Group C for residue composition, Group P for physical-chemical character, Group H for hydrophobic character, and Group O for other indices. With these indices, the character of residues was evaluated quantitatively. The overall results of AAindex scoring distribution were shown in Figure 4. The scores of the indices for residues in epitope regions were compared with those for residues in non-epitope surface region. The x-coordinate refers to the number of antigens with higher scores for epitope residue than non-epitope surface residues, and the y-coordinate is the number of antigens in opposite case.

In Figure 4, the sets of indices (points in the plot) with higher x-coordinate or y-coordinate are supposed to

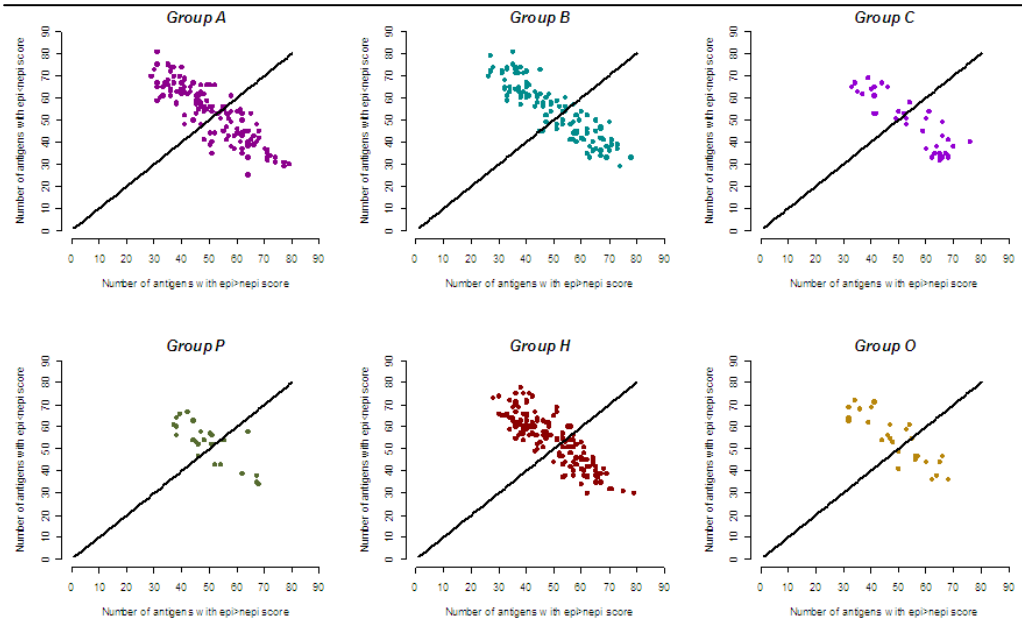

Figure 4. The distribution of scoring result with AAindex indices. There are six major groups of AAindex indices: alpha helix and turn (Group A), beta sheet (Group B), composition (Group C), physic-chemical (Group P), hydrophobic (Group H) and other indices (Group O). For every single point in the plot, it refers to a set of index. The X-axis indicates to the number of antigens which epitope residues score higher with this set of index; while the Y-axis indicates the number of antigens with non-epitope surface residues scoring higher. All the comparisons are statistically significant.

distinguish epitope and non-epitope surface residues effectively. In general, there was no one group of indices which could make an accordant discrimination between residues in epitope and non-epitope surface regions. As we can see from the figure, the indices in each group have made different evaluation for residues in epitope regions and non-epitope surfaces. There was no apparently trend in these groups.

Even so, there are some indices in each group which give relatively better result in discrimination of residues in epitope regions and non-epitope surface. According to the statistical test, 21 sets of indices show strong ability to distinguish the residues from epitope and non-epitope surface among all the 544 sets of indices in AAindex. As to these indices, there are six indices belonging to Group H and others, four to Group A and Group P respectively and one to Group B. The detail indices were listed in Table 2. We have divided these indices into two sets. The indices in set (A) are with significantly higher scores for residues in epitope than that for residues in non-epitope surface regions, and indices in set (B) are with significantly lower scores for residues in epitope than that for residues in non-epitope surface regions.

The correlation between these indices has been inspected. According to Equation (6), the correlation coefficient was calculated between any two sets of indices from above 21 sets. Lower correlation values (defined as correlation coefficient < 0.8 according to method part) have been observed for most sets of indices, except the *corr*QIAN880108-MUNV940102 (-0.84), *corr*ROSG850102-KRIW790101 (-0.94), *corr*CASG920101-ROSG850102 (0.95), *corr*RADA880104-KUHL950101 (-0.85) and *corr*CASG920101-KRIW790101 (-0.91). As to the majority of our selected 21 indices sets, it can be concluded that these sets of indices are independent to each other and non-redundant in the discrimination between epitope and non-epitope surface residues.

Conservation of residues on sequence

The sequence conservation is another characteristic for binding sites. Generally, the binding sites are considered to be more conserved for protein-protein or protein-ligand interaction as functional parts. However, for the binding sites of antibody-antigen complexes: epitope residues are supposed to be less conservative than non-epitope surface ones for the purpose of evading from the antibody recognition. With our dataset, the sequence conservation for binding sites of antibody-antigen complexes has been analyzed. Rate4site was used to calculate the rate for sequence conservation. The multi-sequent alignment files (MSA files) for the antigens were downloaded from HSSP database (http://swift.cmbi.ru.nl/gv/hssp/, dated 28th Apr, 2011) [23]. The conservation rates for residues from epitope and non-epitope surface regions were calculated respectively. Results of conservation rates analysis demonstrated that the epitope residues were less conservative than non-epitope surface residues. In 57 out of all 166 data (34.34%), the difference is significant (*Mann-Whitney test, $p<0.05$*).

Table 2. List of indices with significant result

(A) Indices sets: the scores of epitope residues are significantly higher than that of non-epitope surface ones.

Header	Description	Group	Number
CHAM830105	The number of atoms in the side chain labelled 3+1 (Charton-Charton, 1983)	P	76
FAUJ880101	Graph shape index (Fauchere et al., 1988)	P	80
GEIM800108	Aperiodic indices (Geisow-Roberts, 1980)	A	75
HUTJ700103	Entropy of formation (Hutchens, 1970)	P	75
KRIW790101	Side chain interaction parameter (Krigbaum-Komoriya, 1979)	H	78
MAXF760106	Normalized frequency of alpha region (Maxfield-Scheraga, 1976)	A	81
QIAN880108	Weights for alpha-helix at the window position of 1 (Qian-Sejnowski, 1988)	A	79
RADA880104	Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)	H	75
VASM830102	Relative population of conformational state C (Vasquez et al., 1983)	H	75
WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)	H	75
WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White, 1996)	/	75
KUHL950101	Hydrophilicity scale (Kuhn et al., 1995)	/	75
CASG920101	Hydrophobicity scale from native protein structures (Casari-Sippl, 1992)	/	81

(B) Indices sets: the scores of epitope residues are significantly lower than that of non-epitope surface ones.

Header	Description	Group	Number
FAUJ880111	Positive charge (Fauchere et al., 1988)	H	79
KANM800102	Average relative probability of beta-sheet (Kanehisa-Tsong, 1980)	B	76
RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)	A	78
RADA880106	Accessible surface area (Radzicka-Wolfenden, 1988)	P	79
ROSG850102	Mean fractional area loss (Rose et al., 1985)	H	75
MUNV940102	Free energy in alpha-helical region (Munoz-Serrano, 1994)	/	79
CEDJ970104	Composition of amino acids in intracellular proteins (percent) (Cedano et al., 1997)	/	77
COSI940101	Electron-ion interaction potential values (Cosic, 1994)	/	77

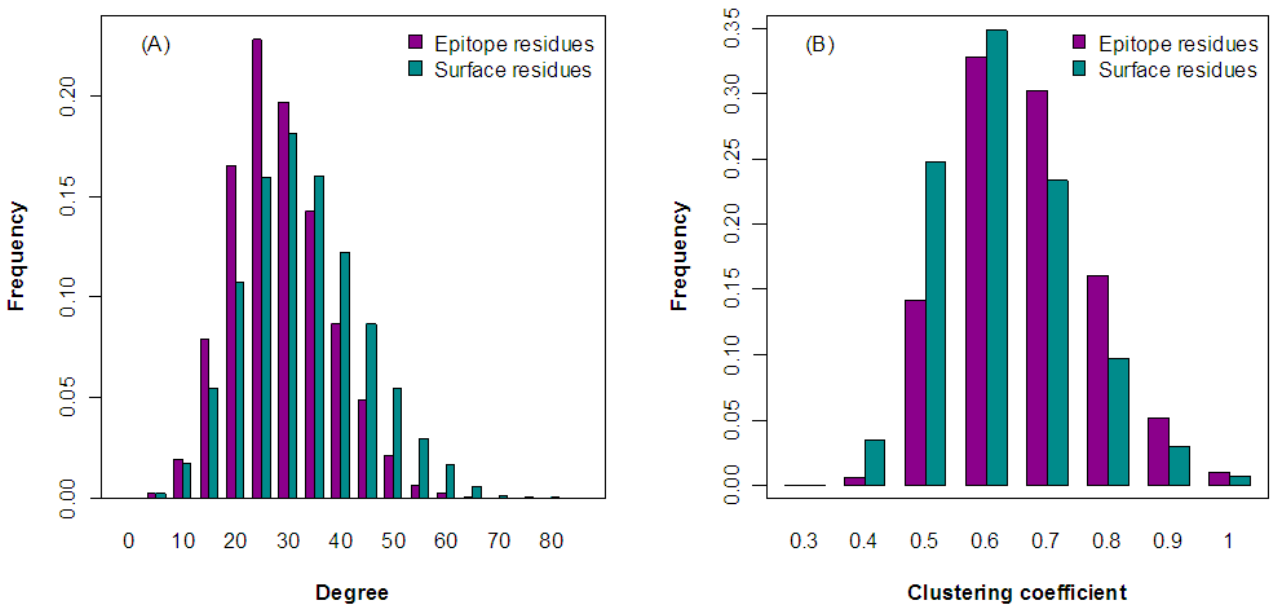

Figure 5. Distribution of two topological parameters. (A) Degree distribution of epitope and non-epitope surface residues; (B) Clustering coefficient distribution of epitope/non-epitope residues.

Structural character of B-cell conformational epitope

Topological analysis

The topological characteristics for residues at the surface and in epitope regions were also analysed in our research. The parameters of degree and cluster coefficient have been selected and calculated to depict the conformational topology for residues. As to the residues in 166 epitopes, the degrees range from 20 to 35 for more than 75% epitope residues. In comparison of epitope and non-epitope surface residues, the distributions of degree are similar. However, the distribution peak of degrees for epitope residues is lower than that of non-epitope surface residues (Figure 5(A)). Furthermore, in 75 out of 166 (45.18%) data, the values of epitope residues' degree were significantly lower than that of non-epitope residues (*Mann-Whitney test, $p<0.05$*). Such difference indicates that fewer connections were formed between the epitope residues. Oo the contrary, the number of connections is relatively higher between residues in non-epitope surface regions. In terms of clustering coefficient, higher values have been observed for epitope residues than that of non-epitope surface residues (Figure 5(B)). As to the residues in 74 out of 166 (44.58%) epitopes in our dataset, this difference was significant (*Mann-Whitney test, $p<0.05$*). From above two topological parameters, it can be concluded that epitope regions are composed of residues with relatively less neighbors but more compact topological distribution.

Surface planarity analysis

The shape complementary is necessary for protein-protein binding. For antigen-antibody complexes, the binding region of antibody is Y shaped structure while the antigen epitope region always keeps an obvious protruding conformation. Therefore, the planarity of surface may be considered as another parameter to depict the epitope region of antigens. Planarity index can describe the planarity of regional planarity conformation for binding sites of antibody and antigen from different views.

As to the planarity, the statistical scores for 70 epitopes out of 166 (42.17%) in our dataset were significantly lower than the planarity scores for non-epitope regions (*Mann-Whitney test, $p<0.05$*). Such difference means that the epitope regions would be more rugged than non-epitope surface regions. With these results, we can characterize the epitope regions of antigen as the local region with the rugged and protruding conformation.

Interaction between paratope and B-cell epitope

The statistics of residue-pair sets between antigen and antibody address the context dependent issue in

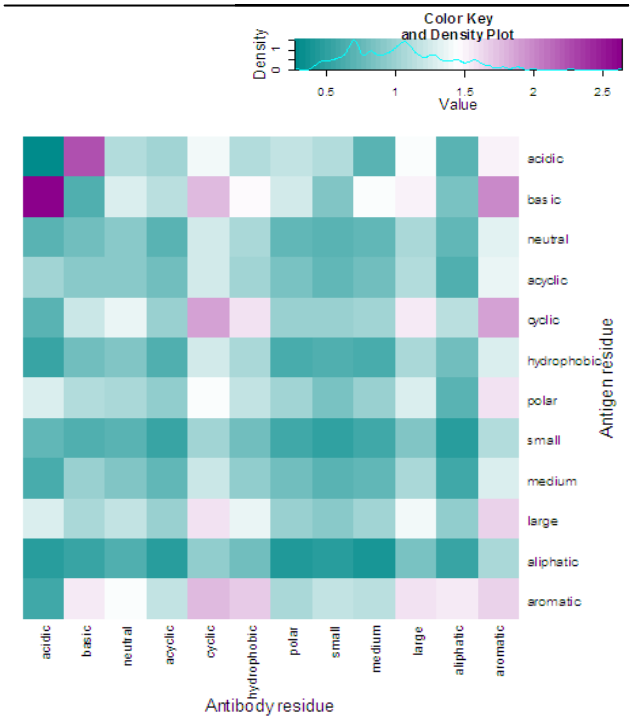

Figure 6. The occurrence of epitope-paratope residue-pair sets. Residues from antigen and antibody sides are grouped according to properties. The color palette from cyan to magenta indicates an increasing of occurrences.

antigen-antibody interaction [24]. The occurrence rate of residue-pair sets across the binding interface of antibody and antigen is shown in Figure 6. Residues have been grouped according to tie properties (Table S1). Acidic-basic residue-pair sets are highly enriched in binding interactions across interface. The composition of aromatic residues is relatively lower in epitope region, whereas aromatic residues involved epitope-paratope interaction are preferred.

DISCUSSION

Prediction of B-cell conformational epitope has long attracted the interests of researchers for its potential application. However, the complexity of this problem baffles the further progress, such as the improvement of prediction accuracy for conformational epitope. According to previous research work, this obstruction mainly derives from the dependence of spatial information for conformational epitope prediction. In previous study, restricted number of structures was used for feature extraction or prediction model training, which may be a reason for incomprehensive conclusion or inaccurate prediction. A representative dataset has been collected in our study. The crystal structures of the antibody-antigen complexes have high resolution and unique epitope, which can guarantee conducting a reliable and unbiased analysis.

In this research, the epitope residues were identified according to the change of solvent accessible surface area before and after the formation of complex. In order to avoid bias, another definition method was also adopted based on the Euclidean distance between atoms from antigens and antibodies. The residues at antigen surface residue was taken as an epitope residue if the minimal atom distance to the atoms at antibody side was within a threshold (4Å, 5Å and 6Å were used most in previous studies) [2,25] . Comparing the identified epitope residues, there was no significant difference between the results of these two methods.

The key question of this study is to estimate whether there is any rules for characterizing B-cell conformational epitope residues from non-epitope surface residues. Earlier work have already proposed many features in this field, which are discrete and unsystematic [26-28] Here, a comprehensive analysis from multiple perspectives was performed on above dataset.

First, B-cell conformational epitopes were examined in size and continuity. Data have shown that a B-cell conformational epitope is relatively constant in residue number and region radius, composing of linear segments and single residues. Second, in order to find characteristics which can distinguish epitope from non-epitope surface areas, various characters were compared between epitope residues and non-epitope surface residues in each antigen. Significant differences were observed in residual composition, sequence conservation and structural formation. However, none of the characteristics can solely predict the conformational epitope residues with a satisfying accuracy, which also implies the complexity and arduous of conformational epitope prediction. Conformational epitope residues are distinctive in many aspects. We believe that with the accumulation of immunoglobulin structures, some combinatorial pattern might be discovered to improve the accuracy. As to the combination methods, machine learning methods have been widely used in other previous researches and set good examples [29-30]. Besides, epitopes are highly context dependent and cannot exist without a corresponding antibody. It is a promising measure to take the paratope-epitope interaction pattern into consideration is for conformation epitope prediction in future Immunoinformatics research.

CONCLUSIONS

In summary, a large scale analysis has been done focusing on spatial description of B-cell conformational epitope. The results of this paper prove the existence of difference between epitope and non-epitope surface residues.

According to the analysis, B-cell conformational epitope is an area on antigen surface with relatively constant size and distance diameters. Although the concept of epitope is defined in 3-D approach, linear segments exist in most of the epitopes. Compared to non-epitope surface residues, epitope ones have larger ASA values. Also, epitope residues are enriched with polar and aromatic residues and show preference of specific residue-pair sets. Besides, epitope residues tend to be less conservative under the environmental pressure. Measured by the parameters describing residue topological features, epitope residues can be distinguished from non-epitope surface significantly for less neighbor residues but more compact neighborhood. Planarity index is another structurally parameter which infers epitope area as rugged region. Future works on B-cell conformational epitope analysis will not only benefit the mechanism comprehension, but also facilitate the antibody design and potential clinical application.

METHODS

Dataset

A comprehensive and non-bias dataset is required because of highly dependence of training in epitope prediction. Four hundred and two crystal structures of antigen–antibody complexes have been obtained from PDB [32], dated April 28th, 2011. Only those with refined resolution better than 3.0Å and the length of protein antigen more than 50 residues were retained. Considering that similar epitopes may be presented in different complexes by antigens sequentially resembles or even identical, the dataset redundancy is removed according to the similarity of antigen sequences and conformational epitopes. The likeness of conformational epitopes was measured according to their composing residues' spatial distribution. An in-house algorithm was developed based on this idea. Strict criteria were set to compare this similarity in our study: in a group of conformational epitopes with high similarity scores, only the complex with the best resolution is kept. Finally, 161 complex structures were re-

tained as our dataset, including 166 unique conformational epitopes (it happens occasionally that more than one epitopes are presented in one complex).

Epitope identification

In our research, method used in the identification of conformational epitope residues relies on the residue Accessible Surface Area (*ASA*). The *ASA* values of antigen residues in complexes (ASA_{bound}) and monomer ($ASA_{unbound}$) were computed using Naccess V2.1.1, with a probe radius of 1.4Å. The structure of monomer antigens were extracted from antigen-antibody complexes. According to *ASA* values, residues at the antigen surface and core were discriminated by a threshold of 1Å2 *ASA* in their monomer status. After binding to antibody, surface residues with *ASA* lose ($ASA_{delta} = ASA_{unbound} - ASA_{bound}$) more than 1Å2 were taken as the epitope residues [25]:

$$r_i \in \begin{cases} core\ residue, & ASA_{unbound} < 1Å \\ epitope\ residue, & ASA_{unbound} > 1Å\ \&\ ASA_{delta} \geq 1Å \\ non-epitope\ residue, & ASA_{unbound} > 1Å\ \&\ ASA_{delta} < 1Å \end{cases}$$

Residue-neighbor and residue-patch

In the process of immune binding, epitopes recognized by antibody are always composed of several residues in proximity. To reflect such cooperative relation for further analysis, the concept of residue-neighbor and residue-patch were defined for each epitope. A residue-neighbor was defined if the atoms from any two residues in an epitope have a minimal distance less than 4Å. As for the residue-patch, it was a group of residues with minimal atom distance less than 10Å to a central residue *ri*. Since the residues in the core of protein antigens are not supposed to form direct interactions with antibody, the inspection for the residue-neighbor and residue-patch was limited to surface residues.

Epitope preference of amino acid and residue-neighbor set types

The preference to be in an epitope was calculated for 20 types of amino acids respectively, as Equation (2):

$$preference_i = \frac{\frac{epi_i}{\sum_i epi_i}}{\frac{nepi_i}{\sum_i nepi_i}} \quad i = 1, 2, 3, \ldots 20$$

where $preference_i$ represents the epitope preference of *i* -type amino acid; epi_i is the number of *i*-type amino acid shown in the epitope, and $\sum_i epi_i$ is the number of all types amino acid composing the epitope. As to the denominator term $nepi_i$ and $\sum_i nepi_i$ are the numbers of *i*-type amino acid and all types amino acids in non-epitope surface areas.

The surface residues have different ASA exposure due to their volume and conformational differences in the structure. Take this factor into consideration, another epitope preference with ASA values as weight was calculated by Equation (3):

$$preference_i' = \frac{\frac{epi_i * ASA}{\sum_i epi_i * ASA}}{\frac{nepi_i * ASA}{\sum_i nepi_i * ASA}} \quad i = 1, 2, 3, \ldots 20$$

where $preference_i'$ is the ASA-weighted epitope preference of *i*-type residue.

With regard to the cooperativeness among epitope residues, the epitope preference of residue-neighbor set types was given by Equation (4):

$$preference_{ij} = \frac{\frac{epi_{ij}}{\sum_i \sum_j epi_{ij}}}{\frac{nepi_{ij}}{\sum_i \sum_j nepi_{ij}}} \quad i, j = 1, 2, 3, \ldots 20$$

Where $preference_{ij}$ denotes the epitope preference of *ij* -type residue-neighbor set, epi_{ij} is the number of *ij*-type residue-neighbor set in epitope, $\sum_i \sum_j epi_{ij}$ is the number of all types residue-neighbor sets in epitope, $nepi_{ij}$ and $\sum_i \sum_j nepi_{ij}$ are the numbers of *ij*-type residue-neighbor set and all types residue-neighbor sets in non-epitope surface areas.

AAindex indices mapping

A variety of amino acid properties have been applied in T -cell epitope prediction and made great progress. Are these properties equally effective when they are used in the B-cell epitope prediction? Indices from AAindex database [22] were introduced to evaluate this ability by using the properties of amino acid to discriminate epitope and non-epitope. As a database collecting indices of amino acid or amino acid pair from literatures, AAindex (ver.9.1) have 544 amino acid indices, including alpha and turn propensities, beta propensity, composition, hydrophobicity, physicochemical properties and other properties. These 544 indices were introduced to score surface residues in this work. The steps are described as follows:

1. Map the amino acid type–corresponding values from *indexi* onto antigen surface residues;
2. Identify residue-patch for each surface residue;
3. Sum and average the index scores among residue-patch for each residue;
4. Compare the average scores between epitope and non-epitope surface residues on antigen basis: significantly higher or lower result will be recorded;
5. Mark *indexi* distinguishable for epitope and non-epitope surface residues: more than 45% of our data (75 out of 166 data) give same significantly higher or lower results with *indexi*;
6. Repeat the steps 1-5 for all 544 indices.

In AAindex, two set of indices might be strongly correlated for describing similar or contrary attributes. To confirm that the final determined distinguishable indices are

independent for each other, the correlation coefficient was calculated for any two sets of indices:

$$corr_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \qquad i = 1, 2, 3, \ldots 20$$

where *x* and *y* represent different sets of indices. In AA-index, two sets of indices are supposed to be correlated when their correlation coefficient is higher than 0.8.

Topological parameters

Vertex and edge are the key elements in the graph theory, and are often used to describe the topological properties for complicated network. Here, the whole antigen surface was viewed as a network with surface residues as vertices and residue-patch relations as edges [33, 34]. Following this thought, each antigen surface formed a network. Then topological properties were analysed while residues were observed in their respective residue patch. Here, the degree and the clustering coefficient were selected to depict the residue network.

In the network, degree of a vertex is the number of edges incident to this vertex. For residue *ri*, the union of its patch residues is its degree, which is calculated as following:

$$degree_i = \sum_{j \neq i} e_{ij}$$

where *eij* indicates the existing edges of *ri*. Higher value of degree means more surface residues being included in the patch with residue *ri* as central residue.

Clustering coefficient is a measure of the extent that vertices in a graph tend to cluster together:

$$clusterCoef_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} e_{il} e_{im} e_{ml}}{\{(\sum_{l \neq i} e_{il})^2 - \sum_{l \neq i} e_{il}^2\}}$$

where *eil* indicates the edge between residue *ri* and *rl*, same with *eim* and *eml*. Equation (8) calculates the ratio of actually existing edge number among *ri*'s patch residues and the potential maximum number. This parameter measures the compactness of *ri*'s residue-patch.

Geometrical parameters

B-cell conformational epitope is a 3-D entity extracted from antigen protein structure. Geometrical characterization is of great significance and potential in the epitope prediction research. The planarity index for the residue-patches at the surface of antigens has been calculated in this work.

For a region of protein surface, a least square plane can be fit based on the coordinates of region's composing residues. By summing the distances of these residues to the least square plane, planarity index was used to evaluate the flatness of plane the residue-patch residing on. Lower value indicates flatter region. With our data, surface residue *ri*'s residue-patch was selected to fit a least square plane. Then the planarity index was calculated for *ri*.

Epitope-paratope interacting residue-pair and occurrence rate

The paratope is the part of an antibody which recognizes an antigen, the antigen-binding site of an antibody. Interacting residue-pair sets are analysed to describe the association between epitope and paratope. Epitope residue *re* and paratope residue *rp* make an interacting residue-pair if their atom distance is less than 4Å. All the epitope-paratope interacting residue-pairs are detected following this distance threshold. The frequencies of these interacting residue-pair sets are biased by different residue composition in epitope and paratope regions. Equation (9) is used to eliminate the bias:

$$pair_{ij} = \frac{epi - para_{ij}}{\sqrt{epi_i * para_j}} \qquad i, j = 1, 2, 3, \ldots 20$$

Where $epi - para_{ij}$ is the number of *ij*-type interacting residue-pair set, *i*-type belong to epitope residues and *j*-type belong to paratope residues, epi_i and $para_j$ indicate the composition of *i*-type residue in epitopes and *j*-type residue in paratopes.

Statistical inference

A series of comparisons between epitope residues and non-epitope surface residues have been carried on in our work. Considering the fact of un-paired and possibly non-homogeneous data, Mann-Whitney *U* test was used for the statistical inference. This test is a non-parametric test for assessing whether two independent samples of observations have equally large values. One tail $p<0.05$ has set for statistical significance.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

JS carried out the program design, performed the statistical analysis and drafted the manuscript. DW carried out the data collection and the statistical analysis. TLX, SNW and GQL performed the statistical analysis. ZWC drafted the manuscript. All authors read the final manuscript.

ACKNOWLEDGEMENTS

This work was supported in part by grants from Ministry of Science and Technology China(2010CB833601, 2008BAI64B02), National Natural Science Foundation of China (30900832, 30976611), Program for New Century Excellent Talents in University(NCET-08-0399)。

REFERENCES

1. Greenbaum JA, Andersen PH, Blythe M, Bui HH, Cachau RE, Crowe J, Davies M, Kolaskar AS, Lund O, Morrison S, Mumey B, Ofran Y, Pellequer JL, Pinilla C, Ponomarenko JV, Raghava GP, van Regenmortel MH, Roggen EL, Sette A, Schlessinger A, Sollner J, Zand M, and Peters B, Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. J Mol Recognit, 2007. 20(2): p. 75-82.
2. Haste Andersen P, Nielsen M, and Lund O, Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci, 2006. 15(11): p. 2558-67.
3. Flower DR, Towards in silico prediction of immunogenic epitopes. Trends Immunol, 2003. 24(12): p. 667-74.

4. Blythe MJ and Flower DR, Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci, 2005. 14(1): p. 246-8.
5. Korber B, LaBute M, and Yusim K, Immunoinformatics comes of age. PLoS Comput Biol, 2006. 2(6): p. e71.
6. Hopp TP and Woods KR, Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A, 1981. 78(6): p. 3824-8.
7. Westhof E, Altschuh D, Moras D, Bloomer AC, Mondragon A, Klug A, and Van Regenmortel MH, Correlation between segmental mobility and the location of antigenic determinants in proteins. Nature, 1984. 311 (5982): p. 123-6.
8. Welling GW, Weijer WJ, van der Zee R, and Welling-Wester S, Prediction of sequential antigenic regions in proteins. FEBS Lett, 1985. 188 (2): p. 215-8.
9. Novotny J, Handschumacher M, Haber E, Bruccoleri RE, Carlson WB, Fanning DW, Smith JA, and Rose GD, Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). Proc Natl Acad Sci U S A, 1986. 83(2): p. 226-30.
10. Parker JM, Guo D, and Hodges RS, New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry, 1986. 25(19): p. 5425-32.
11. Thornton JM, Edwards MS, Taylor WR, and Barlow DJ, Location of 'continuous' antigenic determinants in the protruding regions of proteins. Embo J, 1986. 5(2): p. 409-13.
12. Kolaskar AS and Tongaonkar PC, A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett, 1990. 276(1-2): p. 172-4.
13. Kulkarni-Kale U, Bhosle S, and Kolaskar AS, CEP: a conformational epitope prediction server. Nucleic Acids Res, 2005. 33(Web Server issue): p. W168-71.
14. Sweredoski MJ and Baldi P, PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics, 2008. 24(12): p. 1459-60.
15. Sun J, Wu D, Xu T, Wang X, Xu X, Tao L, Li YX, and Cao ZW, SEPPA: a computational server for spatial epitope prediction of protein antigens. Nucleic Acids Res, 2009. 37(Web Server issue): p. W612-6.
16. Pellequer JL, Westhof E, and Van Regenmortel MH, Predicting location of continuous epitopes in proteins from their primary structures. Methods Enzymol, 1991. 203: p. 176-201.
17. Xu XL, Sun, J., Liu, Q., Wang, X.J., Xu, T.L., Zhu, R.X., Wu, D., Cao, Z.W., Evaluation of spatial epitope computational tools based on experimentally-confirmed dataset for protein antigens. Chinese Science Bulletin, 2010. 55(20): p. 6.
18. Rubinstein ND, Mayrose I, Halperin D, Yekutieli D, Gershoni JM, and Pupko T, Computational characterization of B-cell epitopes. Mol Immunol, 2008. 45(12): p. 3477-89.
19. Lee B and Richards FM, The interpretation of protein structures: estimation of static accessibility. J Mol Biol, 1971. 55(3): p. 379-400.
20. Chothia C, The nature of the accessible and buried surfaces in proteins. J Mol Biol, 1976. 105(1): p. 1-12.
21. Lo Conte L, Chothia C, and Janin J, The atomic structure of protein-protein recognition sites. J Mol Biol, 1999. 285(5): p. 2177-98.
22. Kawashima S and Kanehisa M, AAindex: amino acid index database. Nucleic Acids Res, 2000. 28(1): p. 374.
23. Dodge C, Schneider R, and Sander C, The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res, 1998. 26(1): p. 313-5.
24. Zhao L and Li J, Mining for the antibody-antigen interacting associations that predict the B cell epitopes. BMC Struct Biol. 10 Suppl 1: p. S6.
25. Ponomarenko JV and Bourne PE, Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct Biol, 2007. 7: p. 64.
26. Jones S and Thornton JM, Analysis of protein-protein interaction sites using surface patches. J Mol Biol, 1997. 272(1): p. 121-32.
27. Ghate AD, Bhagwat BU, Bhosle SG, Gadepalli SM, and Kulkarni-Kale UD, Characterization of antibody-binding sites on proteins: development of a knowledgebase and its applications in improving epitope prediction. Protein Pept Lett, 2007. 14(6): p. 531-5.
28. Chou PY and Fasman GD, Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry, 1974. 13(2): p. 211-22.
29. Sollner J and Mayer B, Machine learning approaches for prediction of linear B-cell epitopes on proteins. J Mol Recognit, 2006. 19(3): p. 200-8.
30. Saha S and Raghava GP, Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins, 2006. 65(1): p. 40-8.
31. Rubinstein ND, Mayrose I, and Pupko T, A machine-learning approach for predicting B-cell epitopes. Mol Immunol, 2009. 46(5): p. 840-7.
32. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, and Abola EE, Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr, 1998. 54(Pt 6 Pt 1): p. 1078-84.
33. Huang J, Kawashima S, and Kanehisa M, New amino acid indices based on residue network topology. Genome Inform, 2007. 18: p. 152-61.
34. Huang J, Honda W, and Kanehisa M, Predicting B cell epitope residues with network topology based amino acid indices. Genome Inform, 2007. 19: p. 40-9.