Analysis of Conformational B-Cell Epitopes in the Antibody-Antigen Complex Using the Depth Function and the Convex Hull

Wei Zheng[1], Jishou Ruan[1,2], Gang Hu[1], Kui Wang[1], Michelle Hanlon[3], Jianzhao Gao[1] *

1 School of Mathematical Sciences and LPMC, Nankai University, Tianjin, People's Republic of China, **2** State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin, People's Republic of China, **3** Department of Physical Sciences, Grant MacEwan University, Alberta, Canada

* gaojz@nankai.edu.cn

Abstract

The prediction of conformational b-cell epitopes plays an important role in immunoinformatics. Several computational methods are proposed on the basis of discrimination determined by the solvent-accessible surface between epitopes and non-epitopes, but the performance of existing methods is far from satisfying. In this paper, depth functions and the k-th surface convex hull are used to analyze epitopes and exposed non-epitopes. On each layer of the protein, we compute relative solvent accessibility and four different types of depth functions, i.e., Chakravarty depth, DPX, half-sphere exposure and half space depth, to analyze the location of epitopes on different layers of the proteins. We found that conformational b-cell epitopes are rich in charged residues Asp, Glu, Lys, Arg, His; aliphatic residues Gly, Pro; non-charged residues Asn, Gln; and aromatic residue Tyr. Conformational b-cell epitopes are rich in coils. Conservation of epitopes is not significantly lower than that of exposed non-epitopes. The average depths (obtained by four methods) for epitopes are significantly lower than that of non-epitopes on the surface using the Wilcoxon rank sum test. Epitopes are more likely to be located in the outer layer of the convex hull of a protein. On the benchmark dataset, the cumulate 10th convex hull covers 84.6% of exposed residues on the protein surface area, and nearly 95% of epitope sites. These findings may be helpful in building a predictor for epitopes.

Introduction

Epitopes are binding areas on antigens. The prediction of b-cell epitopes is critical for the development of vaccines and immunotherapeutic drugs [1]. B-cell epitopes are categorized into linear and conformational epitopes. A linear b-cell epitope is a contiguous amino acid segment in an antigen. A conformational b-cell epitope is located in close proximity in the protein 3-dimensional structure but discontinuous in the protein sequence. The majority of b-cell epitopes are conformational [2].

It is time-consuming and expensive to use experimental techniques to identify b-cell epitopes [3], especially on a genomic scale. Many computational methods have been developed for b-cell epitope prediction [4]. The prediction of linear b-cell epitopes from antigen sequences can be traced back to the 1980s. The first generation of prediction methods is the propensity model. These models utilized a single propensity of the amino acid [2, 5–10], or combined multiple physicochemical propensities to predict epitopes [10–14]. More complicated models were then used to predict epitopes including the neural network [15], hidden Markov model [16], Naïve Bayes model [17] and support vector machine [18–25].

Some other methods use protein structure to predict epitopes, building the model using simple scoring-based approaches. CEP [26] utilizes solvent accessibility to score amino acid surfaces. DiscoTope [27] utilizes surface/solvent accessibility, contact numbers, and amino acid propensity scores. SEPPA [28] combines propensity scores that were based on solvent accessibility and the packing density of amino acids. PEPITO [29] fuses amino acid propensity scores and solvent accessibility, quantified by using half-sphere exposure in a linear regression. EPSVR [30] takes epitope propensity scores, contact numbers, secondary structure composition, conservation, side chain energy surfaces and planarity scores as inputs, and a support vector regression was built. Zhang et al. [31] utilizes the random forest model, while Liu and Hu [32] uses logistic regression with B-factors and the relative accessible surface area as inputs. Bepar is an association patterns model [33]. EPMeta is a consensus model [30]. Epitopia [34, 35] utilizes the Naïve Bayes model, fusing physicochemical and structural-geometrical properties from a surface patch.

Although many methods are proposed, the performance of b-cell epitope predictors is moderate. With the increase of antigen-antibody crystal structures, it is possible to analyze these complex structures. A more detailed description of the b-cell epitope area becomes important. In this paper, we applied four types of depth functions; half-sphere exposure (HSE) [36], Chakravarty depth [37], DPX [38–40], and half-space depth (HSD) [41] to analyze the location of epitopes. Compared with solvent accessibility, depth functions distinguish between atoms just below the protein surface and those in the core of the protein [37–40]. The goal of this paper is to investigate these depth functions and the convex hull to distinguish conformational epitopes from non-epitopes. This information may provide useful clues for b-cell epitope prediction.

Materials and Methods

Dataset

The dataset for this paper was first used as benchmark dataset in Ansari and Raghava [42]. It contains 161 protein chains from 144 antigen-antibody complex structures. Sequence redundancy was removed by BLASTCLUST, at 40% cutoff, leaving 57 antigen chains remaining. In this paper, all exposed residues (relative solvent accessibility RSA>0) are considered. The dataset of 57 antigens contains 915 conformational epitopes and 9632 exposed non-epitopes (in S1 Table). In the following section, the term non-epitopes will refer to exposed non-epitopes, i.e. non-epitopes on the antigen protein surface.

Computed Features

RSA. Solvent accessible surface area (ASA) is calculated by NACCESS [43]. Relative solvent accessibility (RSA) is defined as the ratio of the ASA of a residue, observed in its three-dimensional structure, to that observed in an extended tri-peptide conformation. We found that the RSAs of all epitopes are positive, though some values are only slightly larger than zero. For example, the epitope site Val206 of paracoccus denitrificans two-subunit cytochrome C

oxidase complex (PDB ID:1AR1:B) has an RSA value of 0.008. To avoid losing any epitope sites, a residue is considered to be an exposed residue if the RSA is greater than 0.

Conservation. Conservation measure, our use of which was motivated by Valdar's 2002 work[44], is defined as follows,

$$\text{Conservation} = 1 - \left(-\sum\nolimits_{k=1..20} \left\{ p_k * \log_2(p_k) \right\} / \log_2(20) \right) \tag{1}$$

where p_k is the value from the Weighted Observation Percentage (WOP) matrix generated by PSI-BLAST [45], divided by 100. If all WOP values of a given residue equal zero, i.e. $p_1, p_2, \ldots, p_{20}$ is represented by 20 zeroes, then *conservation* is one.

Half-Sphere Exposure. Half-Sphere Exposure (HSE) is a 2D measure, introduced by Hamelryck [36]. HSE consists of the number of C_α atoms in two half-spheres around a residue's C_α atom. One of the half-spheres corresponds to the side chain's neighborhood, the other half-sphere is in the opposite direction. There are two ways to compute HSE, depending on whether information is available about both the C_α and C_β positions (HSEB) or only about the C_α positions (HSEA). HSE can be divided into HSEAU, HSEAD, HSEBU and HSEBD, depending on whether the half-sphere selected is an up half-sphere (U) or a down half-sphere (D). In this paper, we calculated HSEAU, HSEAD, HSEBU and HSEBD, with radius 13Å.

Chakravarty Depth. Chakravarty [37] defined the depth of an atom in a protein as the distance between the atom and the nearest surface water molecule. The residue depth is the average of the constituent atom depths. Residue depth is calculated by the program *depth-1.0* [46].

DPX. Atom depth (DPX), first introduced by Pintar [38–40], is defined as the distance between a non-hydrogen atom and its closest solvent-accessible protein neighbor. DPX is a good geometrical descriptor of the protein interior. Residue DPX is the average of the constituent atom DPX values. We calculated residue DPX using the software DPX [38].

Half Space Depth (HSD). Tukey [41] introduced half space depth (HSD) to order the high dimensional data. It is defined as:

$$\text{HSD}(x; P) = \inf \{ P(H) : H \text{ is closed half space}, \ x \in H, \ x \text{ in } R^d \} \tag{2}$$

where x is a point in d-dimensional space with probability measure P. HSD is defined as the minimum probability mass carried by any closed half space containing x. For a protein, the probability P(H) is estimated by the empirical distribution. The i-th residue HSD is then defined by:

$$\text{HSD}(\text{Res}_i) = \inf \left\{ \#(\text{plane}_{\text{Resi}}) / N \right\} \tag{3}$$

where $(\text{plane}_{\text{Resi}})$ is the number of residues in the half space which is divided by the plane through that i-th residue, and N is the total number of residues in the protein. The use of residue HSD is motivated by Shen [47, 48].

Convex Hull. In mathematics, the convex hull of a set X of points in Euclidean space is the smallest convex set that contains X. For instance, when X is a bounded subset of the plane, the convex hull may be visualized as the shape formed by a rubber band stretched around X. For a protein, we consider all residues in a protein to be a point set X. In this paper, MIConvexHull (http://miconvexhull.codeplex.com/) is used to calculate the convex hull of a point set. [49].

K-th Convex Hull. Let X be all the atoms of the exposed residues (RSA>0) in a protein. Atoms which are on the convex hull surface are defined as the first level of the convex hull, denoted as $CH_1(X)$. The remaining atoms of the protein comprise the set $X-CH_1(X)$. We then compute the convex hull of $X-CH_1(X)$, such that the atoms on the convex hull surface of $X-CH_1(X)$ are defined as the secondary level of the convex hull of the protein, denoted as

$CH_2(X)$. Generally speaking, the k-th level of the convex hull of the protein is the convex hull of the set $X-\cup_{i=1..k-1}\{CH_i(X)\}$, denoted as $CH_k(X)$. Finally, one protein can represent a union of n convex hulls, i.e. $X = \cup_{i=1..n}\{CH_i(X)\}$. The level of the k-th convex hull of a residue is defined as the minimal level of convex hulls of atoms that are in the residue. For instance, Glycine contains C_α, C_β, N,O atoms, where $C_\alpha \in CH_2(X)$, $C_\beta \in CH_3(X)$, $N \in CH_3(X)$, and $O \in CH_5(X)$. The level of the convex hull for Gly is 2, as that is the minimal of $\{2, 3, 3, 5\}$. The k-th convex hull is a useful tool for classifying the residues. Take the exposed residues of a protein for example; all exposed residues can be divided into k convex hulls.

Cumulate k-th Convex Hull. For given exposed atoms of a protein X, we can calculate the k-th convex hull ($CH_k(X)$) for each residue. The cumulate k-th convex hull of X (denoted as $CH^k(X)$) is defined as the union of k convex hulls, i.e. $CH^k(X) = \cup_{i=1..k}\{CH_i(X)\}$. Obviously, $CH^m(X)$ is a strict subset of $CH^n(X)$, if $m<n$. If a residue is in $CH_i(X)$, then this residue is also in the cumulate i-th convex hull $CH^i(X)$. Clearly, we can choose a finite value K_0 for which $CH^k(X)$ will cover (contain) all the exposed epitopes on the protein surface. K_0 is defined as follows:

$$K_0 = \min\{k| \text{ all epitopes} \in CH^k\} \qquad (4)$$

K_0 would be different for different antigen proteins. If there are more than two chains in the antibody-antigen complex, the exposed atoms of all chains will be used to compute the convex hull and the cumulate convex hull. We developed software for calculating the Convex Hull of Protein Surface (CHOPS), which is available at <www.sourceforge.net/projects/chops>.

The cumulate k-th convex hull of protein 1NCA:N is shown in Fig 1, where k = 1, 2,. . ., 9. Atoms on the cumulate k-th convex hull are colored blue, and the remaining exposed atoms are colored red. The proportion of exposed atoms on the convex hull to exposed atoms is increased with increasing value of k.

Statistical Features on the K-Th Convex Hull. The coverage ratio of *epitopes* in the k-th convex hull ($CREPI_k$) of a protein is the number of epitopes in the k-th convex hull divided by the total number of *epitopes* in this protein. The coverage ratio of *epitopes* in the *cumulate* k-th convex hull ($CREPI^k$) is the number of epitopes in the cumulate k-th convex hull divided by the total number of *epitopes* in the protein. The coverage ratio of exposed residues in the k-th convex hull ($CREXP_k$) is the number of *exposed residues* in the k-th convex hull divided by the number of total *exposed residues* in the protein. The coverage ratio of exposed residues in the cumulate k-th convex hull ($CREXP^k$) is the number of *exposed residues* in the *cumulate* k-th convex hull divided by the number of total *exposed residues* in the protein. The proportion of epitopes in the k-th convex hull ($PROP_k$) is defined as the number of *epitopes* in the k-th convex hull divided by the number of *all residues* in the k-th convex hull. The proportion of epitopes in the cumulate k-th convex hull ($PROP^k$) is defined as the number of *epitopes* in the cumulate k-th convex hull divided by the number of *all residues* in the cumulate k-th convex hull.

For example, there are 10 epitopes and 200 exposed residues on an antigen protein. Two epitopes and 20 residues are in the first convex hull. Then, $CREPI_1$ is 2/10 = 20%, $CREXP_1$ is 20/200 = 10%, and $PROP_1$ is 2/20 = 10%.

Results and Discussion

Amino Acid Composition of Epitopes and Non-Epitopes

All exposed residues (RSA>0) are considered. In the following sections, the term non-epitopes will refer to non-epitopes with RSA values larger than zero. Amino acid composition is defined as the count of a type of amino acid divided by the length of the antigen protein. Fig 2 shows
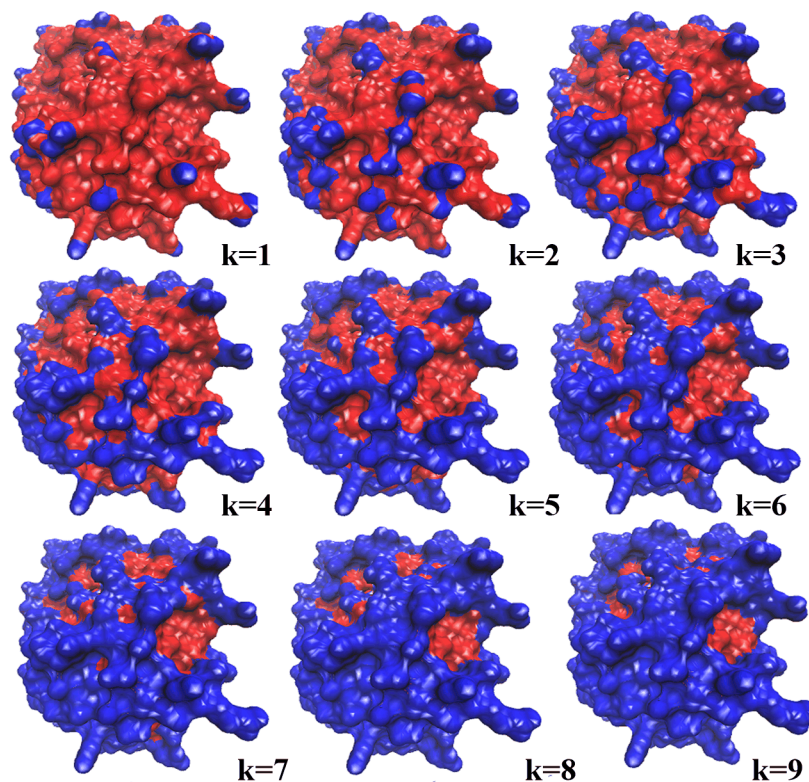
Fig 1. The cumulate k-th convex hull of the protein 1NCA:N. (k = 1, 2,. . .9). Atoms on the cumulate k-th convex hull are colored blue, remaining exposed atoms are colored red.
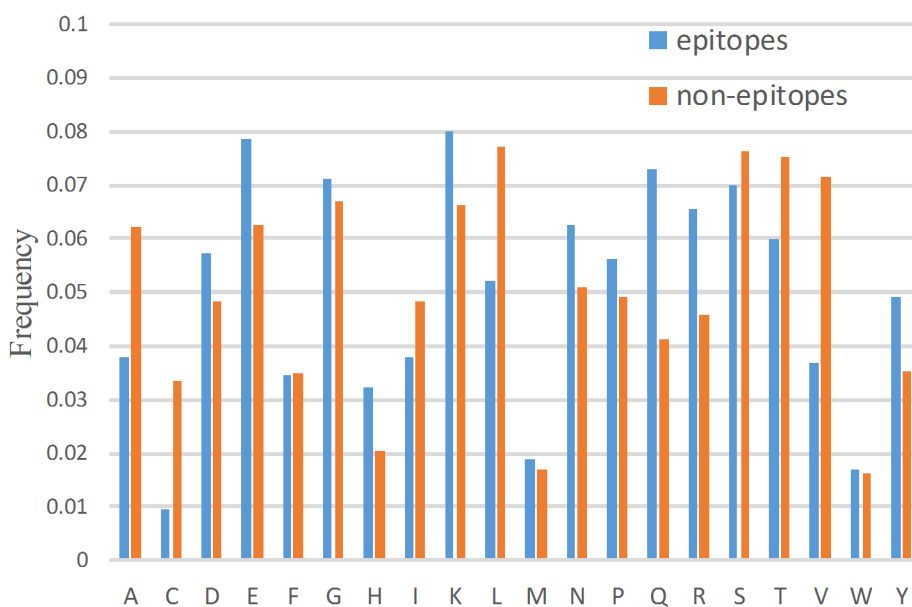
Fig 2. Amino acid composition of epitopes and non-epitopes.

the average amino acid composition for 57 antigens. It shows that conformational b-cell epitopes are rich in negatively charged residues Asp(D) and Glu(E); positively charged residues Lys(K), Arg(R), and His(H); non-polar, aliphatic residues Gly(G) and Pro(P); polar, non-charged residues Asn(N) and Gln(Q), and the aromatic residue Tyr(Y). Compared to the epitopes, non-epitopes are rich in non-polar, aliphatic residues Ala(A), Leu(L), and Val(V), and the polar, non-charged residue Ser(S). These results are consistent with previous findings that epitopes are rich in polar amino acids and aromatic amino acids but depleted in aliphatic amino acids [33, 50, 51].

Secondary Structure of Epitopes and Non-Epitopes

Secondary structure was computed by the DSSP program, and then eight types of secondary structure are combined into three types: (1) Helices, which groups α-helices, 3-helices and π-helices. (2) Strands, that is, isolated β–bridges and extended strands participate inβ–ladders. (3) Coils, consisting of hydrogen-bonded turns, bends and others. The secondary structures of epitopes and non-epitopes are shown in Fig 3. Conformational b-cell epitopes are rich in coils. In contrast, the non-epitopes are rich in strands and helices. Further analyzing the eight types of secondary structure from DSSP, we see that epitopes are rich in bends (S) and hydrogen-bonded turns (T). In contrast, non-epitopes are rich in extended strands which participate inβ–ladders (E), andα-helices (H). (See S1 Fig).

RSA Values of Epitopes and Non-Epitopes

Fig 4 shows the histogram of RSA values of both epitopes and non-epitopes. Fig 4A shows the histogram for epitopes alone. The average and standard deviation of the RSA values of epitopes are 50.0 and 24.8, respectively (see Table 1). Fig 4B shows the histogram for non-epitopes alone. The average and standard deviation of the RSA values are 35.4 and 27.4, respectively. The distributions for epitopes and non-epitopes do not follow the normal distribution, based on the Shapiro-Wilk normality test. The average RSA of epitopes is significantly greater than the average RSA of non-epitopes, based on the Wilcoxon rank-sum test (p-value<2.2e-16).

Correlations between RSA and depth functions are also considered. Table 1 shows average 57 antigen proteins Pearson correlation coefficient (PCC) between RSA and depth. Chakravarty depth, DPX, HSEAU and HSEBU obtain higher correlation with absolute PCC of (> 0.70). HSEAD, HSEBD and HSD obtain lower correlation with absolute PCC of (<0.55). Half-sphere exposure using up half-sphere (HSEAU, HSEBU) achieves higher correlation coefficient than the using down half-sphere (HSEAD, HSEBD).

Depth Functions for Epitopes and Non-Epitopes

Table 1 shows the average values of different depth functions for epitopes and non-epitopes. The average epitope depth values, i.e. the Chakravarty depth, DPX, Half Space Exposure (HSE) and Half Space Depth (HSD), are smaller than the average non-epitope depth values. Take the Chakravarty depth for example; the average depth for epitopes is 4.15, which is lower than average depth for non-epitopes, 5.05. This indicates that epitopes prefer the outer layer of the protein surface, making it easier for epitopes to interact with antibodies.

## Conservation, Depth Functions, and RSA for Epitopes and Non-Epitopes

The conservation of epitopes is widely used in the prediction of epitopes. The average conservation scores for epitopes are not significantly lower than for non-epitopes. We further analyze the relationship between RSA, Chakravarty depth and conservation (Fig 5). It shows that
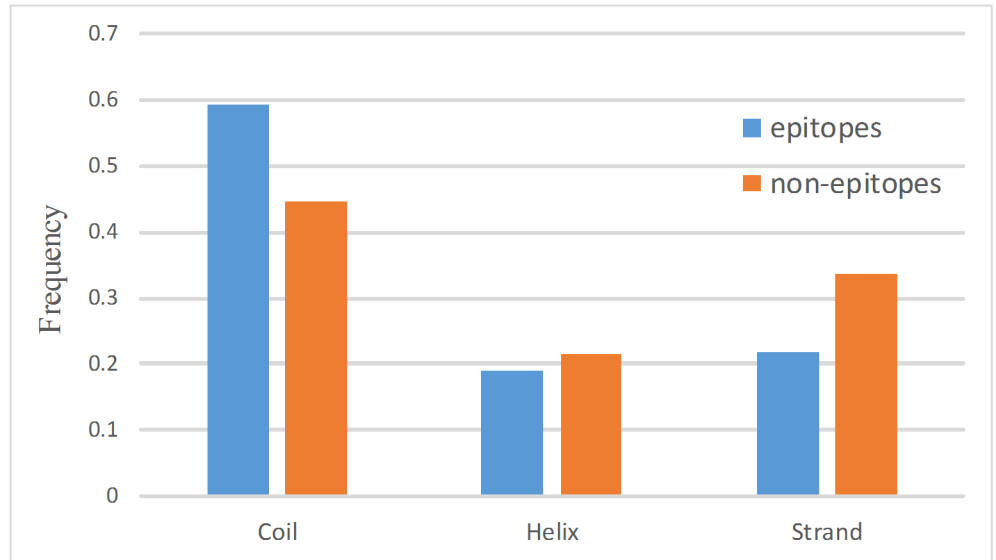
Fig 3. Secondary structures of epitopes and non-epitopes.

doi:10.1371/journal.pone.0134835.g003

Chakravarty depth for all epitopes is below 8Å. Further examination of non-epitopes with residue depth above 8Å shows that the median RSA and the conservation of these residues are 1.4 and 0.84, respectively, In contrast, the median of the RSA and the conservation for epitopes are 49.5 and 0.36. With information about the depth function, the epitopes are easily distinguished from non-epitopes.
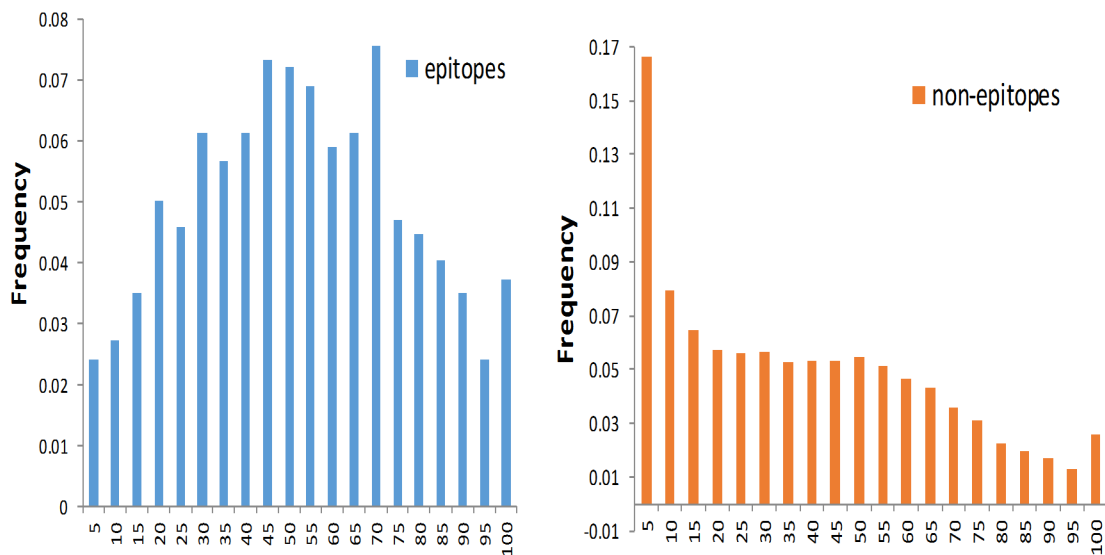
Fig 4. Histogram of RSA values for epitopes and non-epitopes. (A) RSA values of epitopes (red) (B) RSA values of non-epitopes (blue).

doi:10.1371/journal.pone.0134835.g004

Table 1. Average values and standard deviation (in bracket) of depth function for epitopes and non-epitopes.

Depth	Epitopes	Non-epitopes	PCC between depth and RSA	Significance difference between epitopes and non-epitopes (p-value<0.05)
Chakravarty depth[a]	4.15(0.69)	5.05(1.79)	-0.78(0.04)	-[b] (p-value< 2.2e-16)
DPX[c]	0.41(0.41)	0.67(0.62)	-0.70(0.04)	- (p-value< 2.2e-16)
HSEAU[d]	8.46(6.76)	13.1(9.1)	-0.78(0.08)	- (p-value< 2.2e-16)
HSEAD[d]	17.97 (6.55)	19.2(7.3)	-0.01(0.14)	- (p-value = 2.8e-06)
HSEBU[d]	9.64(6.83)	14.2(8.7)	-0.79(0.04)	- (p-value< 2.2e-16)
HSEBD[d]	17.11 (6.51)	18.4(7.3)	-0.07(0.14)	- (p-value = 1.0e-06)
HSD[e]	0.02(0.04)	0.05(0.07)	-0.55(0.13)	- (p-value< 2.2e-16)
RSA	50.0(24.8)	35.4(27.4)	1.00	+ (p-value< 2.2e-16)

[a]Chakravarty depth is introduced by [37] and computed following [46].

[b] means that the average of conservation scores is significantly smaller than the average of conservation scores of non-epitopes at p-value<0.05 using Wilcoxon rank sum test with one-side test.

[c] DPX is the average of constituent atom DPX values, DPX is calculated following [38].

[d]HSEAU/HSEAD: Half-Sphere Exposure using C_α position and selected Up/Down half-sphere, HSEBU/HSEBD: Half-Sphere Exposure using C_β position and selected Up/Down half sphere. Radius of sphere 13Å is used.

[e]HSD: half space depth which was firstly introduced by [41], and used in [47, 48].

doi:10.1371/journal.pone.0134835.t001

Analysis of Epitopes Using the Convex Hull (CH$_k$)

We calculate the level of the convex hull for each epitope. The smaller the level, the more external the layer in which the epitope is found. The average level of the convex hull for epitopes is 4.5, while the non-epitopes are at 8.0. The level of the convex hull is significantly smaller (p-value < 2.2e-16) than for non-epitopes, using the Wilcoxon rank-sum test. This indicates that
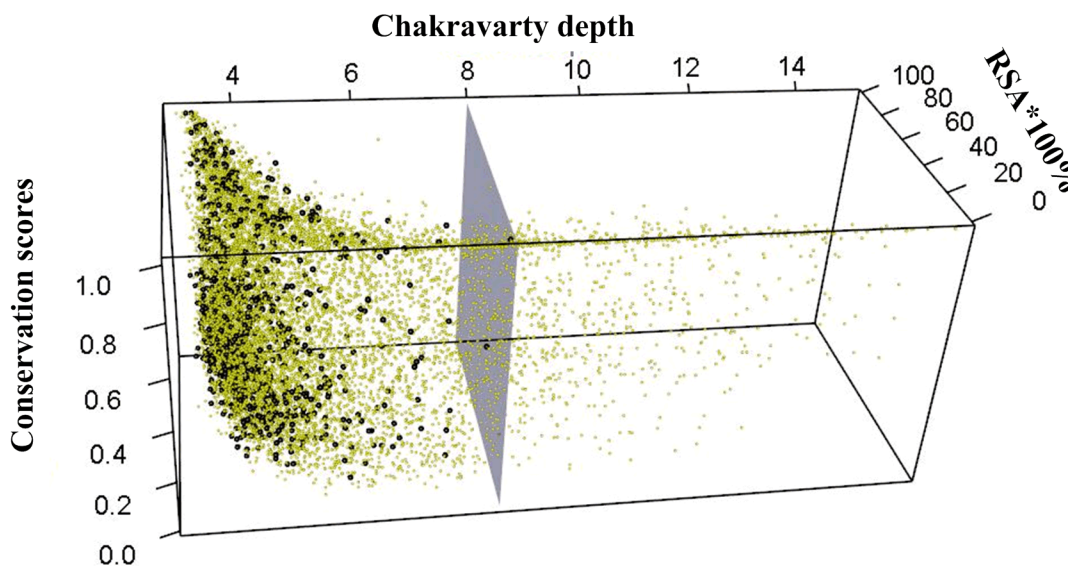
Fig 5. Chakravarty depth, RSA and conservation of epitopes (black points) and non-epitopes (yellow points). Depths of all epitopes are less than 8Å (gray plane).

doi:10.1371/journal.pone.0134835.g005

epitopes are closer to the convex hull. It is also consistent with the results in Rubinstein et al. [51] which found that the distance between epitope site and protein convex residue is less than the distance between non-epitope site and convex residue.

Epitopes prefer the outer layer of the protein surface, but not all epitopes are in the first convex hull ($CH_1(X)$) of the protein. Fig 6 shows the average $CREPI_k$, $CREXP_k$ and $PROP_k$ in the k-th surface convex hull, where k = 1,...15. There are 42.2% of the epitopes in 1st convex hull of the protein ($CREPI_1$ = 42.2%). These 42.2% of epitopes cover 26.0% of the surface area of the protein ($CREXP_1$ = 26.0%). There is approximately one epitope per six residues of CH_1 ($PROP_1$ = 17.9%, 1/0.179≈6). We also noted that $CREPI_k$, $CREXP_k$ and $PROP_k$ are decreased while k is increasing. This indicates that the percentage of epitopes would be decreased when the residue in the protein interior. Take the k = 2 for example; there are 12.6% of the epitopes in the secondary convex hull (CH_2). These residues in CH_2 cover 7.76% of the protein surface. Each epitope is around 4 non-epitopes.

We also analyzed the influence of the depth function according to the k-th convex hull. Fig 7 shows these results. For DPX (Fig 7A), the values for epitopes are smaller than the values for non-epitopes, except on CH_2. For Chakravarty depth (Fig 7B), all values for epitopes are smaller than the values for non-epitopes. There are two types of HSE, i.e. HSEA, HSEB. There is no significant rule for the HSEA down sphere (HSEAD) (See S2 Fig), but all values of HSEAU for epitopes are smaller than the values of HSEAU for non-epitopes, except CH_4. On the other hand, there is no significant rule for the HSEB down sphere (HSEBD), all values of HSEBU for epitopes are smaller than the values of HSEBU for non-epitopes (S3 Fig). This indicates that the DPX, Chakravarty depth and HSEAU, HSEBU may be useful to classify epitopes and non-epitopes on the surface.

Minimal Level of the Convex Hull for Antigen Proteins

K_0 is the minimal level of the convex hull such that all epitopes are on the cumulate k-th convex hull (CH^k). We analyzed the cumulate k-th convex hull of different antigen protein chains. Fig 8 shows the results. For 24.6% ((1+1+1+3+8)/57, $K_0 \leq 5$) of the antigen chains, all epitopes are covered in the top five layers of the convex hull of the antigen. There are a total of 86.0% ((3+8 +7+6+5+7+4+2+3+4)/57 = 86.0%) proteins for which all epitopes are located in top 4~13 layers of the convex hull. This also indicated that there is only one protein for which all epitopes are located in the first layer of the convex hull (CH^1). From these results, we can see that the convex hull functions can further describe the distribution of epitopes.

Choose K_0 for the Cumulate K-Th Surface Convex Hull

For a given protein, we do not know which K_0 of the cumulate k-th convex hull can contain all the epitope sites. If the K_0 value is too small, many epitopes will not be considered. On the other hand, if the K_0 value is too big, too many non-epitopes will be considered. To estimate a proper K_0 value, 7 chains (protein IDs: 1AFV:A, 1FSK:A, 1IAI:M, 1KB5:A, 1NFD:D, 1OTS:A, 1QFU:A) were randomly selected as a test set, and the remaining 50 chains were used as training set. We calculated $CREPI^k$, $CREXP^k$ and $PROP^k$ of CH^k (k = 1, 2,..., 20) for each protein in the training set. Then, averages for three kinds of ratios in CH^k were computed. Fig 9 shows the results.

As the k value is increased, CH^k will contain many more residues, including both epitopes and non-epitopes. The more epitopes included in CH^k, the larger $CREPI^k$ will be. The more non-epitope sites included in CH^k, the larger $CREXP^k$ will be, and the smaller $PROP^k$ will be (see Fig 9). If we want to obtain a proper k value, we must make sure $CREPI^k$ and $PROP^k$ are as large as possible, and $CREXP^k$ is as small as possible at the same time. For this training dataset,
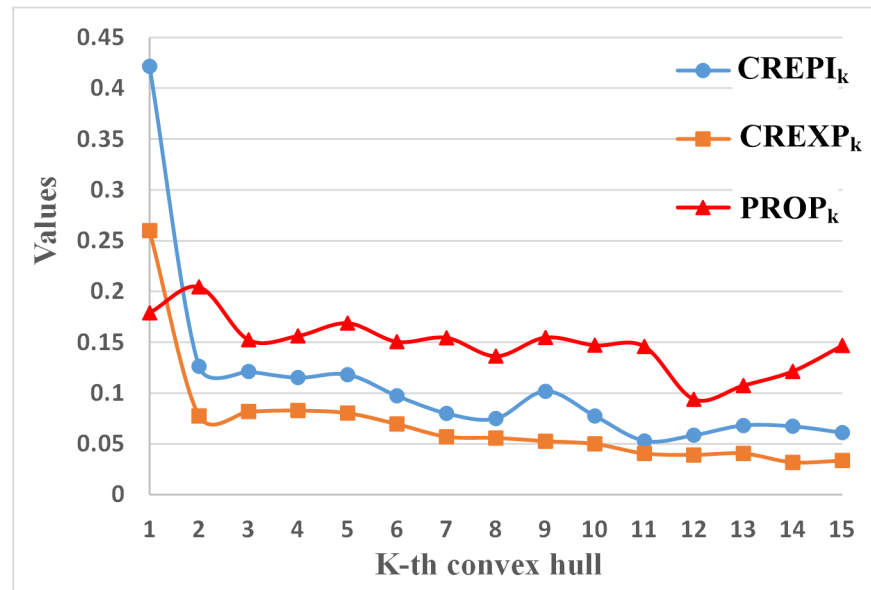
Fig 6. Average CREPI$_k$, CREXP$_k$ and PROP$_k$ in different k-th surface convex hull layers CH$_k$. CREPI$_k$ is the number of epitopes in the k-th convex hull divided by the total number of *epitopes* in this protein. CREXP$_k$ is the number of *exposed residues* in the k-th convex hull divided by the number of total *exposed residues* in the protein. PROP$_k$ is defined as the number of *epitopes* in the k-th convex hull divided by the number of *all residues* in the k-th convex hull.

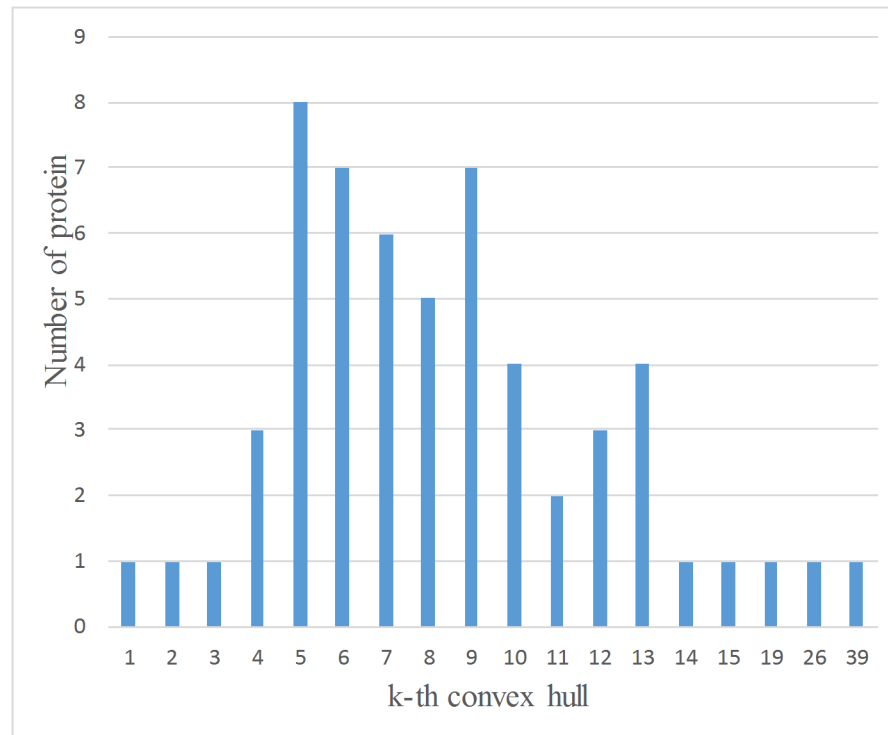doi:10.1371/journal.pone.0134835.g006

$K_0 = 10$ is selected. Generally, CH^{10} covers 84.6% of the residues on the protein surface area, and nearly 95% of the epitope sites. In CH^{10} of a protein, 13.1% of the residues are epitopes.

We test our results on the test dataset. $CREPI^{10}$, $CREXP^{10}$ and $PROP^{10}$ of CH^{10} for each protein are calculated. The average of $CREPI^{10}$s is 96.7%. The $CREPI^{10}$ values for five proteins are above 95%, except for 1AFV:A (92.8%) and 1QFU:A(84.2%). The average of the $CREXP^{10}$s is 77.3%. This indicates that just 77.3% of the exposed residues are covered in CH^{10}. The
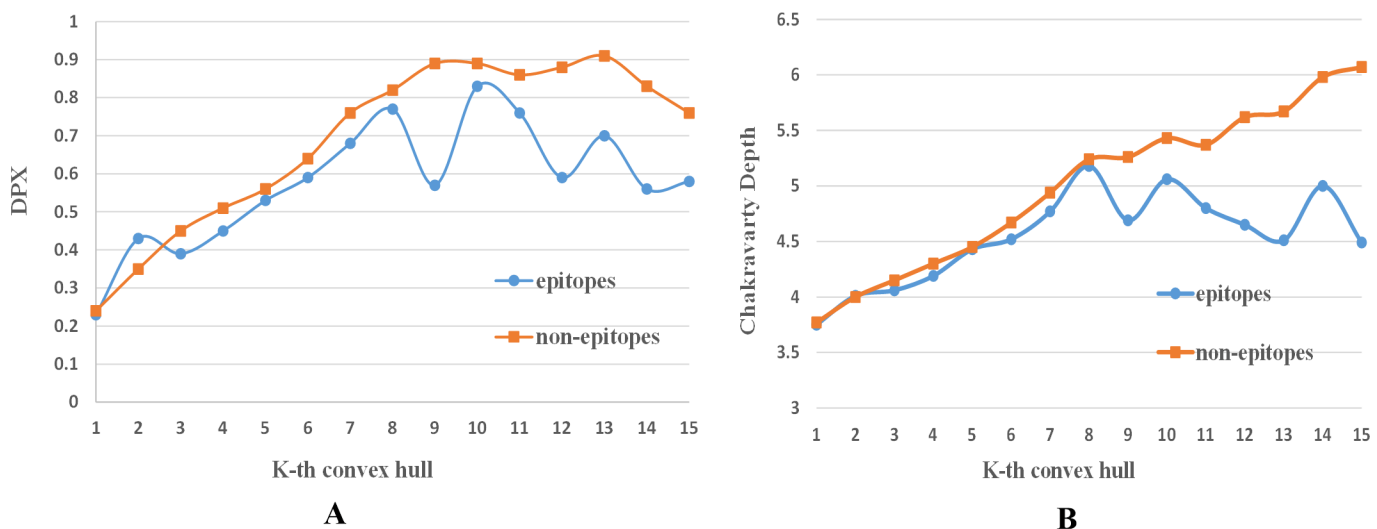
Fig 7. Different depth functions according to the k-th convex hull layers CH$_k$ (k = 1, 2,. . ., 15). (A)DPX (B) Chakravarty Depth.

doi:10.1371/journal.pone.0134835.g007

Fig 8. Minimal level of the convex hull for antigen proteins. Take $K_0 = 7$ for example, there are six proteins for which all epitopes are located in the cumulate 7-th convex hull (CH^7).

average value of $PROP^{10}$ is 8.2%. We further analyze CH^6 for the test dataset, and CH^6 for 1FSKA, 1IAIM, 1KB5A, 1OTSA covers 100% of the epitopes. The remaining 7th, 8th, 9th, 10th layers of the convex hull contain zero epitopes. This indicates that the cutoff of $K_0 = 10$ is probably robust for the test data.

Conclusions

Relative Solvent Accessibility or solvent surface is widely used in the analysis of proteins and protein functions. The accessible surfaces are always the residues for which RSA >5%. If the cutoff 5% is used, about 2.4% of the epitopes are buried, in the dataset used in this paper. We use the k-th convex hull and the cumulate k-th convex hull to categorize the protein residues, and analyze the location of epitopes on different layers of the proteins. On each layer of the protein, we compute the four different types of depth functions to analyze the location of epitopes on different layers of the proteins.

Based on RSA of the epitopes and non-epitopes on the protein surface, the average RSA for epitopes is significantly greater than the average RSA for non-epitopes. Nevertheless, there is no significantly difference between RSA values of epitopes and non-epitopes which are in top eight convex hull layers (see S4 Fig). It may be the reason why the b-cell prediction performance is moderate using monotonous RSA-based features. For Chakravarty depth, DPX, Half Sphere Exposure, and HSD, the average values for epitopes are significantly lower than the average values for non-epitopes on the surface. Take Chakravarty depth for example; all epitopes have depth below 8Å. This indicates that epitopes may be distinguished from non-epitopes on the basis of the depth function.
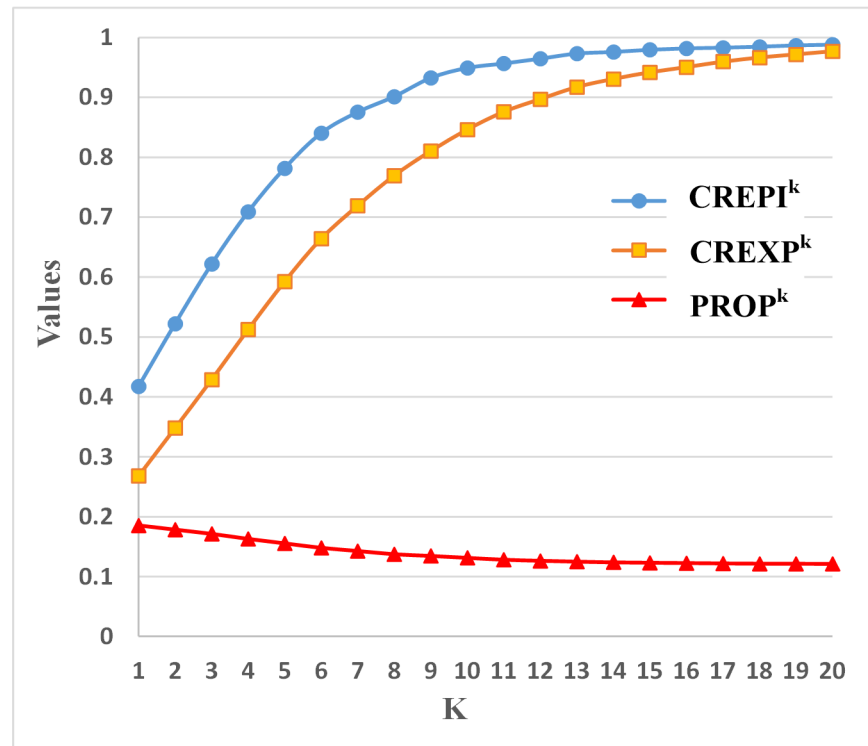
Fig 9. CREPI^k, CREXP^k and PROP^k curve of CH^k with different k values. CREPI^k is the number of epitopes on the cumulate k-th convex hull divided by the total number of *epitopes* in the protein. CREXP^k is the number of *exposed residues* on the *cumulate* k-th convex hull divided by the number of total *exposed residues* in the protein. PROP^k is defined as the number of *epitopes* in the cumulate k-th surface convex hull divided by the number of *all residues* in the cumulate k-th convex hull.

doi:10.1371/journal.pone.0134835.g009

Correlation between RSA and different depth functions are also analyzed. Chakravarty depth, DPX, Half-sphere exposure using up half-sphere(HSEAU, HSEBU) achieve higher absolute Person correlation coefficients. HSD, half-sphere exposure using down half-sphere (HSEAD, HSEBD) and half space depth(HSD) achieve lower absolute Person correlation coefficients. Depth functions provide more detailed description of the b-cell epitopes It may provide useful clues for b-cell epitope prediction.

The conservation for epitopes is not significantly lower than that for non-epitopes. This is due to the fact that some non-epitopes may play important biological functions, such as glycosylation sites and pockets, giving higher conservation. For example, for the residue LEU259 of hiv-1 JR-RF gp120 core protein (2B4C:G), its Chakravarty depth is 13Å, RSA is just 0.2, while its conservation is 0.98; this residue is a glycosylation site.

Epitopes prefer to be located in the outer layer of the protein surface, but not all epitopes are in the convex hull of the protein. For Chakravarty depth, HSEAU, and HSEBU, the average depth function values for epitopes are smaller than the average values for non-epitopes on the surface. On the benchmark dataset, CH^{10} just covers 84.6% of the residues on protein surface area, but nearly 95% of the epitope sites.

Our software for calculating the Convex Hull of Protein Surface (CHOPS) can be downloaded from <www.sourceforge.net/projects/chops>. As demonstrated in a series of recent publications [52–63] in developing new prediction or analysis methods, user-friendly and publicly accessible web-servers will significantly enhance their impacts [64]. We shall make efforts in our future work to provide a web-server for the method presented in this paper.

Supporting Information

S1 Fig. Secondary structure of epitopes and non-epitopes. B: residue in isolated β-bridge; E: extended strand, participates in β-ladder; G: 3-helix; H: α-helix; S: bend; T: hydrogen bonded turn; I: 5-helix.
(DOC)

S2 Fig. HSEA depth functions according to the k-th convex hull layers CH_k (k = 1, 2,..., 15). HSEA:Half-Sphere Exposure using information only about the C_α position.
(DOC)

S3 Fig. HSEB depth functions according to the k-th convex hull layers CH_k (k = 1, 2,..., 15). HSEB: Half-Sphere Exposure using information about both the C_α and C_β positions.
(DOC)

S4 Fig. RSA according to k-th convex hull layers CH_k (k = 1, 2,..., 15). There is no significantly difference between epitopes and non-epitopes in the top 8 convex hull layers (p-values are 0.82, 0.08, 0.60, 0.37, 0.53, 0.42, 0.24 and 0.64 using Wilcoxon rank sum test with two sides).
(DOC)

S1 Table. RSA and different depth values for epitopes and exposed non-epitopes.
(XLSX)

Acknowledgments

Author Contributions

Conceived and designed the experiments: JG. Performed the experiments: WZ. Analyzed the data: JG WZ JR GH KW MH. Contributed reagents/materials/analysis tools: JG WZ JR GH KW. Wrote the paper: JG WZ JR GH KW MH. Designed the software used in analysis: ZW JG.

References

1. Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput Biol. 2012; 8(12):e1002829. doi: 10.1371/journal.pcbi.1002829 PMID: 23300419; PubMed Central PMCID: PMC3531324.

2. Pellequer JL, Westhof E, Van Regenmortel MH. Predicting location of continuous epitopes in proteins from their primary structures. Methods Enzymol. 1991; 203:176–201. PMID: 1722270.

3. Reineke U, Schutkowski M. Epitope mapping protocols. Preface. Methods in molecular biology. 2009; 524:v–vi. PMID: 19514158.

4. El-Manzalawy Y, Honavar V. Recent advances in B-cell epitope prediction methods. Immunome Res. 2010; 6 Suppl 2:S2. doi: 10.1186/1745-7580-6-S2-S2 PMID: 21067544; PubMed Central PMCID: PMC2981878.

5. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proceedings of the National Academy of Sciences of the United States of America. 1981; 78(6):3824–8. Epub 1981/06/01. PMID: 6167991; PubMed Central PMCID: PMCPmc319665.

6. Welling GW, Weijer WJ, van der Zee R, Welling-Wester S. Prediction of sequential antigenic regions in proteins. FEBS Lett. 1985; 188(2):215–8. PMID: 2411595.

7. Karplus P, Schulz G. Prediction of chain flexibility in proteins. Naturwissenschaften. 1985; 72(4):212–3.

8. Parker JM, Guo D, 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):5425–32. PMID: 2430611.

9. Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett. 1990; 276(1–2):172–4. PMID: 1702393.

10. Pellequer J-L, Westhof E, Van Regenmortel MH. Correlation between the location of antigenic sites and the prediction of turns in proteins. Immunology letters. 1993; 36(1):83–99. PMID: 7688347

11. Alix AJ. Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine. 1999; 18(3–4):311–4. PMID: 10506656.

12. Odorico M, Pellequer JL. BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. J Mol Recognit. 2003; 16(1):20–2. doi: 10.1002/jmr.602 PMID: 12557235.

13. Saha S, Raghava G. BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. Artificial Immune Systems: Springer; 2004. p. 197–204.

14. Chang HT, Liu CH, Pai TW. Estimation and extraction of B-cell linear epitopes predicted by mathematical morphology approaches. J Mol Recognit. 2008; 21(6):431–41. doi: 10.1002/jmr.910 PMID: 18680207.

15. Saha S, Raghava GP. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins. 2006; 65(1):40–8. doi: 10.1002/prot.21078 PMID: 16894596.

16. Larsen JE, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res. 2006; 2:2. doi: 10.1186/1745-7580-2-2 PMID: 16635264; PubMed Central PMCID: PMC1479323.

17. Chen J, Liu H, Yang J, Chou KC. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids. 2007; 33(3):423–8. doi: 10.1007/s00726-006-0485-9 PMID: 17252308.

18. El-Manzalawy Y, Dobbs D, Honavar V. Predicting linear B-cell epitopes using string kernels. J Mol Recognit. 2008; 21(4):243–55. doi: 10.1002/jmr.893 PMID: 18496882; PubMed Central PMCID: PMC2683948.

19. El-Manzalawy Y, Dobbs D, Honavar V. Predicting flexible length linear B-cell epitopes. Comput Syst Bioinformatics Conf. 2008; 7:121–32. PMID: 19642274; PubMed Central PMCID: PMC3400678.

20. Sweredoski MJ, Baldi P. COBEpro: a novel system for predicting continuous B-cell epitopes. Protein Eng Des Sel. 2009; 22(3):113–20. doi: 10.1093/protein/gzn075 PMID: 19074155; PubMed Central PMCID: PMC2644406.

21. Wee LJ, Simarmata D, Kam YW, Ng LF, Tong JC. SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction. BMC Genomics. 2010; 11 Suppl 4:S21. doi: 10.1186/1471-2164-11-S4-S21 PMID: 21143805; PubMed Central PMCID: PMC3005920.

22. Wang HW, Lin YC, Pai TW, Chang HT. Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification. J Biomed Biotechnol. 2011; 2011:432830. doi: 10.1155/2011/432830 PMID: 21876642; PubMed Central PMCID: PMC3163029.

23. Wang Y, Wu W, Negre NN, White KP, Li C, Shah PK. Determinants of antigenicity and specificity in immune response for protein sequences. BMC Bioinformatics. 2011; 12:251. doi: 10.1186/1471-2105-12-251 PMID: 21693021; PubMed Central PMCID: PMC3133554.

24. Yao B, Zhang L, Liang S, Zhang C. SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. PloS one. 2012; 7(9):e45152. doi: 10.1371/journal.pone.0045152 PMID: 22984622; PubMed Central PMCID: PMC3440317.

25. Singh H, Ansari HR, Raghava GP. Improved method for linear B-cell epitope prediction using antigen's primary sequence. PloS one. 2013; 8(5):e62216. doi: 10.1371/journal.pone.0062216 PMID: 23667458; PubMed Central PMCID: PMC3646881.

26. Kulkarni-Kale U, Bhosle S, Kolaskar AS. CEP: a conformational epitope prediction server. Nucleic Acids Res. 2005; 33(Web Server issue):W168–71. doi: 10.1093/nar/gki460 PMID: 15980448; PubMed Central PMCID: PMC1160221.

27. Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. 2006; 15(11):2558–67. doi: 10.1110/ps.062405906 PMID: 17001032; PubMed Central PMCID: PMC2242418.

28. Sun J, Wu D, Xu T, Wang X, Xu X, Tao L, et al. SEPPA: a computational server for spatial epitope prediction of protein antigens. Nucleic Acids Res. 2009; 37(Web Server issue):W612–6. doi: 10.1093/nar/gkp417 PMID: 19465377; PubMed Central PMCID: PMC2703964.

29. Sweredoski MJ, Baldi P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics. 2008; 24(12):1459–60. doi: 10.1093/bioinformatics/btn199 PMID: 18443018.

30. Liang S, Zheng D, Standley DM, Yao B, Zacharias M, Zhang C. EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. BMC bioinformatics. 2010; 11(1):381.

31. Zhang W, Xiong Y, Zhao M, Zou H, Ye X, Liu J. Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. BMC Bioinformatics. 2011; 12:341. doi: 10.1186/1471-2105-12-341 PMID: 21846404; PubMed Central PMCID: PMC3228550.

32. Liu R, Hu J. Prediction of discontinuous B-cell epitopes using logistic regression and structural information. J Proteomics Bioinform. 2011; 4:010–5.

33. Zhao L, Li J. Mining for the antibody-antigen interacting associations that predict the B cell epitopes. BMC structural biology. 2010; 10(Suppl 1):S6. doi: 10.1186/1472-6807-10-S1-S6 PMID: 20487513

34. Rubinstein ND, Mayrose I, Martz E, Pupko T. Epitopia: a web-server for predicting B-cell epitopes. BMC Bioinformatics. 2009; 10:287. doi: 10.1186/1471-2105-10-287 PMID: 19751513; PubMed Central PMCID: PMC2751785.

35. Rubinstein ND, Mayrose I, Pupko T. A machine-learning approach for predicting B-cell epitopes. Mol Immunol. 2009; 46(5):840–7. doi: 10.1016/j.molimm.2008.09.009 PMID: 18947876.

36. Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins: Structure, Function, and Bioinformatics. 2005; 59(1):38–48.

37. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure. 1999; 7(7):723–32. PMID: 10425675.

38. Pintar A, Carugo O, Pongor S. DPX: for the analysis of the protein core. Bioinformatics. 2003; 19 (2):313–4. PMID: 12538266.

39. Pintar A, Carugo O, Pongor S. Atom depth as a descriptor of the protein interior. Biophysical journal. 2003; 84(4):2553–61. PMID: 12668463

40. Pintar A, Carugo O, Pongor S. Atom depth in protein structure and function. Trends Biochem Sci. 2003; 28(11):593–7. doi: 10.1016/j.tibs.2003.09.004 PMID: 14607089.

41. Tukey JW, editor Mathematics and the picturing of data. Proceedings of the international congress of mathematicians; 1975.

42. Ansari HR, Raghava GP. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. Immunome Res. 2010; 6:6. doi: 10.1186/1745-7580-6-6 PMID: 20961417; PubMed Central PMCID: PMC2974664.

43. Hubbard SJ, Thornton JM. Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London. 1993; 2(1).

44. Valdar WS. Scoring residue conservation. Proteins. 2002; 48(2):227–41. doi: 10.1002/prot.10146 PMID: 12112692.

45. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–402. PMID: 9254694; PubMed Central PMCID: PMC146917.

46. Tan KP, Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS. Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Res. 2013; 41(Web Server issue):W314–21. doi: 10.1093/nar/gkt503 PMID: 23766289; PubMed Central PMCID: PMC3692129.

47. Shen S, Hu G, Tuszynski JA. Analysis of protein three-dimension structure using amino acids depths. The protein journal. 2007; 26(3):183–92. PMID: 17557208

48. Wang K, Gao J, Shen S, Tuszynski JA, Ruan J, Hu G. An accurate method for prediction of protein-ligand binding site on protein surface using SVM and statistical depth function. Biomed Res Int. 2013; 2013:409658. doi: 10.1155/2013/409658 PMID: 24195070; PubMed Central PMCID: PMC3806129.

49. de Berg M, van Kreveld M, Overmars M, Schwarzkopf O. Computational Geometry. Computational Geometry: Springer Berlin Heidelberg; 2000. p. 1–17.

50. Kringelum JV, Nielsen M, Padkjaer SB, Lund O. Structural analysis of B-cell epitopes in antibody:protein complexes. Mol Immunol. 2013; 53(1–2):24–34. doi: 10.1016/j.molimm.2012.06.001 PMID: 22784991; PubMed Central PMCID: PMC3461403.

51. Rubinstein ND, Mayrose I, Halperin D, Yekutieli D, Gershoni JM, Pupko T. Computational characterization of B-cell epitopes. Mol Immunol. 2008; 45(12):3477–89. doi: 10.1016/j.molimm.2007.10.016 PMID: 18023478.

52. Feng P-M, Chen W, Lin H, Chou K-C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Analytical Biochemistry. 2013; 442(1):118–25. http://dx.doi.org/10.1016/j.ab.2013.05.024. doi: 10.1016/j.ab.2013.05.024 PMID: 23756733

53. Chen W, Lin H, Feng P, Ding C, Zuo Y, Chou K. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PloS one. 2012; 7(10):e47843. doi: 10.1371/journal.pone.0047843 PMID: 23144709

54.   Lin H, Deng E-Z, Ding H, Chen W, Chou K-C. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Research. 2014; 42:12961–72. doi: 10.1093/nar/gku1019 PMID: 25361964

55.   Chen W, Feng P-M, Deng E-Z, Lin H, Chou K-C. iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Analytical Biochemistry. 2014; 462(0):76–83. http://dx.doi.org/10.1016/j.ab.2014.06.022.

56.   Ding H, Deng E-Z, Yuan L-F, Liu L, Lin H, Chen W, et al. iCTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels. BioMed Research International. 2014; 2014:10. doi: 10.1155/2014/286419

57.   Guo S-H, Deng E-Z, Xu L-Q, Ding H, Lin H, Chen W, et al. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics. 2014; 30(11):1522–9. doi: 10.1093/bioinformatics/btu083 PMID: 24504871

58.   Liu Z, Xiao X, Qiu W-R, Chou K-C. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Analytical Biochemistry. 2015; 474(0):69–77. http://dx.doi.org/10.1016/j.ab.2014.12.009.

59.   Jia J, Liu Z, Xiao X, Liu B, Chou K-C. iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. Journal of Theoretical Biology. 2015; 377(0):47–56. http://dx.doi.org/10.1016/j.jtbi.2015.04.011.

60.   Liu B, Liu F, Wang X, Chen J, Fang L, Chou K-C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Research. 2015. doi: 10.1093/nar/gkv458

61.   Gao J, Faraggi E, Zhou Y, Ruan J, Kurgan L. BEST: improved prediction of B-cell epitopes from antigen sequences. PloS one. 2012; 7(6):e40104. doi: 10.1371/journal.pone.0040104 PMID: 22761950; PubMed Central PMCID: PMC3384636.

62.   Gao J, Kurgan L. Computational prediction of B cell epitopes from antigen sequences. Methods in molecular biology. 2014; 1184:197–215. doi: 10.1007/978-1-4939-1115-8_11 PMID: 25048126.

63.   Zheng W, Zhang C, Hanlon M, Ruan J, Gao J. An ensemble method for prediction of conformational B-cell epitopes from antigen sequences. Computational biology and chemistry. 2014; 49:51–8. doi: 10.1016/j.compbiolchem.2014.02.002 PMID: 24607818.

64.   Chou K-C. Impacts of bioinformatics to medicinal chemistry. Medicinal chemistry (Shariqah (United Arab Emirates)). 2015; 11(3):218–34. PMID: 25548930.