B-cell Epitopes: Discontinuity and Conformational Analysis

Saba Ferdous[a,b], Sebastian Kelm[c], Terry S. Baker[c], Jiye Shi[c,d], and Andrew C.R. Martin[a,*]

[a]Institute of Structural and Molecular Biology, Division of Biosciences, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK
[b]Present address: Cancer Research UK Manchester Institute, The University of Manchester, Alderley Park, SK10 4TG, UK
[c]UCB Pharma, 216 Bath Road, Slough SL1 4EN, UK;
[d]Shanghai Institute of Applied Physics, Chinese Academy of Sciences, 201800 Shanghai, China

[*]To whom correspondence should be addressed (andrew@bioinf.org.uk –or– andrew.martin@ucl.ac.uk)

Abstract

Peptide vaccines have many potential advantages over conventional ones including low cost, lack of need for cold-chain storage, safety and specificity. However, it is well known that approximately 90% of B-cell Epitopes (BCEs) are discontinuous in nature making it difficult to mimic them for creating vaccines. In this study, the degree of discontinuity in B-cell epitopes and their conformational nature is examined. The discontinuity of B-cell epitopes is analyzed by defining 'regions' (consisting of at least three antibody-contacting residues each separated by ≤ 3 residues) and small fragments (antibody-contacting residues that do not satisfy the requirements for a region). Secondly, an algorithm has been developed that classifies each region's shape as straight, curved or folded on the basis that straight and folded regions are more likely to retain their native conformation as isolated peptides. We have investigated the structures of 488 B-cell epitopes from which 1282 regions and 1018 fragments have been identified. 90% of epitopes have five or fewer regions and five or fewer fragments with 14% containing only one region and 4% being truly linear (i.e. having one region and no fragments). Of the 1282 regions, 508 are straight in shape, 626 are curved and 148 are folded.

Highlights

- A comprehensive analysis of epitope discontinuity from 488 B-cell epitopes.

- Regions defined as ≥3 antibody-contacting residues with gaps of ≤3 residues.
- Fragments defined as antibody-contacting residues that are not in regions.
- 14% of epitopes have only one region; only 4% are truly linear, having one region and no fragments.
- 39.6% of regions are straight, 48.8% are curved and 11.5% are folded.

Keywords: antigen; protein conformation; protein structure; antibody-antigen interactions; discontinuous epitopes; epitope conformation

1 Introduction

Despite the successes of vaccination against major infectious diseases and the resulting global eradication of diseases such as small pox and poliomyelitis (Henderson, 1998; Hovi, 2001; John, 2000; Breman and Arita, 1980), vaccines for many infectious diseases remain elusive. Traditionally, vaccine development has involved the delivery of live attenuated or inactivated viruses or bacteria by injection. However, vaccines that include the whole organisms may cause a detrimental immune response owing to unnecessary proteins present in the vaccine formulation (Thompson and Staats, 2011) which may result in unwanted host responses such as allergenic and/or reactogenic immune responses (Petrovsky and Aguilar, 2004). This has led to a focus on using a single protein (or a few proteins) to induce the protective immune response (Thompson and Staats, 2011; Petrovsky and Aguilar, 2004). However, even a single protein contains many epitopes, some of which may lead to an undesired immune response. 'Peptide vaccines' are capable of inducing more specific immune responses that cross-react with intact protein, avoiding allergenic and/or reactogenic responses (Nemchinov *et al.*, 2000; Arthur *et al.*, 1987; Sun *et al.*, 1991; Purcell *et al.*, 2007).

Unlike T-cell epitopes that are linear continuous stretches of residues, B-cell epitopes are generally conformational (discontinuous) being comprised of multiple sequential segments that are in close spatial proximity in the 3D fold of an antigen. This discontinuous nature of B-cell epitopes has made identification and prediction from sequence challenging (Haste Andersen *et al.*, 2006; Kulkarni-Kale *et al.*, 2005; Lo *et al.*, 2013; Moreau *et al.*, 2008; Zhao *et al.*, 2012; Ponomarenko *et al.*, 2008; Sun *et al.*, 2009; Sweredoski and Baldi, 2008). The increase in structural data available for antibody/antigen complexes has provided new opportunities for conformational analysis and characterization of epitopes to understand their properties in detail. Thus far, structural characterization of epitopes has been performed on the basis of solvent accessibility (Novotnỳ *et al.*, 1986; Lollier *et al.*, 2011), amino acid composition, size (Haste Andersen *et al.*, 2006; Ofran *et al.*, 2008; Rubinstein *et al.*, 2008; Zhao and Li, 2010; Sun *et al.*, 2011), secondary structure (Ofran *et al.*, 2008; Rubinstein *et al.*, 2008; Liang *et al.*, 2010), location on the antigen (Haste Andersen *et al.*, 2006; Rubinstein *et al.*, 2008; Thornton *et al.*, 1986) and geometry (Rubinstein *et al.*, 2008). More recently, Sivalingam and Shepherd (2012) investigated discontinuous epitopes defining regions with no gaps, gaps of three and gaps of five non-contacting

residues. Their findings suggest that with the gap of three or five residues, 85–88% epitopes are comprised of multiple regions.

A recent study by Kringelum *et al.* (2013) was performed on a relatively large dataset (107 unique antibody-antigen complex structures) compared with previous studies which used smaller datasets (up to 53 unique antibody-antigen complex structures (Sivalingam and Shepherd, 2012; Rubinstein *et al.*, 2008)). They present a detailed analysis of antigen-antibody interaction surfaces and described the epitope in terms of its size, shape, segmentation, secondary structure, location, orientation relative to the antibody, amino acid composition, amino acid 'co-operativeness' (particular amino acid pairs mediating cooperative antibody-antigen binding) and spatial amino acid composition. In terms of shape, Kringelum *et al.* described B-cell epitopes as flat, oblong or oval based on an analysis of epitope and paratope residues.

Several methods (Kulkarni-Kale *et al.*, 2005; Haste Andersen *et al.*, 2006; Sweredoski and Baldi, 2008; Ponomarenko *et al.*, 2008; Sun *et al.*, 2009; Schneidman-Duhovny *et al.*, 2005; Comeau *et al.*, 2004; Rubinstein *et al.*, 2009a,b; Liang *et al.*, 2009, for example) have been developed for the prediction of conformational B-cell epitopes. These methods used 3D structural information of an epitope along with several other features that include amino acid properties, spatial information, surface accessibility and residue clustering. Unfortunately, none of these methods is able to provide good predictions of conformational B-cell epitopes, but an understanding of the 3D structural shape of epitopes may aid in their prediction.

In this paper, we take a different approach to analyzing the structures of epitopes to look at the level of discontinuity and the conformational nature of continuous stretches. Unlike previous work, we also consider antigens formed from multiple protein chains.

2 Materials and Methods

All code was implemented in Perl and C. Data and code are available at `github.com/ACRMGroup/epitopes`. The code also makes use of programs from BiopTools (Porter and Martin, 2015) available at `www.bioinf.org.uk/software/bioptools` and `github.com/ACRMGroup/bioptools`.

2.1 Dataset Preparation

A dataset of 673 unique antibody-antigen structures was obtained from AbDb (`www.bioinf.org.uk/abs/abdb`) (Ferdous and Martin, 2018) in December 2016. The dataset was filtered to remove peptide antigens of <30 amino acids reducing the dataset to 520 unique antibody-protein antigen complexes. A further 11 antibody-antigen complexes solved using electron diffraction were excluded because of their low resolution and three complexes were removed owing to incorrect pairing resulting from single-chain Fabs in AbDb, missing structural information and incorrect symmetry. Of the remaining 506 complexes, after performing redundancy test using cdhit (Li and Godzik, 2006) (100% cut-off), the dataset was reduced to 488 complexes. Among these 488 complexes, 446 had a single antigen chain associated with the antibodies while 42 had multiple chains bound.

2.2 Defining Epitopes

Epitope residues were defined as the set of antigen residues having any atoms in contact with the CDR region of an antibody where a contact was defined as a centre-to-centre distance less than 4 Å (Ponomarenko and Bourne, 2007; Haste Andersen *et al.*, 2006; Sun *et al.*, 2011) and implemented in the program `chaincontacts` from BiopTools (Porter and Martin, 2015).

2.3 Epitope Structural Discontinuity Determination

Epitopes were separated into two types of structural elements: regions (R) and fragments (F). Regions were defined as continuous stretches of antigen sequence having at least three residues in contact with antibody. As with several previous studies (Haste Andersen *et al.*, 2006; Sun *et al.*, 2011; Kringelum *et al.*, 2013; Sivalingam and Shepherd, 2012), gaps between contacting residues were allowed and a gap size of up to three non-contacting residues was chosen on the basis of the structure of α-helices allowing the inclusion of amino acids which lie on the same face of an α-helix (Supplementary Figure S2).

Note that terminology varies between different studies. We use the term 'region' for extended sets of contacting residues (with gaps) where other studies have used terms such as 'segment' or 'fragment' (Rubinstein *et al.*, 2008; Sivalingam and Shepherd, 2012; Kringelum *et al.*, 2013); we reserve the term 'fragment' (F) for single amino acids that make contact with antibody, but which do not form part of a region (Supplementary Figure S3).

2.4 Conformational Analysis of Epitope Regions

A method was developed to classify regions into straight, curved and folded conformations (Supplementary Figure S4).

Each peptide region was first classified as predominantly alpha, beta or coil based on secondary structure assignments performed using the program `pdbsecstr` from BiopTools (Porter and Martin, 2015), an implementation of the Kabsch and Sander method (Kabsch and Sander, 1983) as modified by Smith and Thornton (1989). A threshold of >60% occurrence of a given secondary structure was used to classify a region as helix, strand or coil.

The shape classification algorithm uses a measure of linearity by comparing a given region with an ideal β-strand or α-helix. A best-fit line is calculated through the Cα positions of the peptide using `pdbline` from BiopTools and the Cα closest to the midpoint of this line is identified. The position of that Cα is then projected onto the line and reference positions for the projections of the other Cα atoms are calculated based on a spacing of 3.5 Å for an ideal β-strand and 1.5 Å for an ideal α-helix. Regions classified as coil also use the β-strand spacing. The remaining actual Cα atoms are also projected onto the line and the mean absolute deviation in their positions from the ideal reference positions is calculated as a descriptor of linearity. The method is described in detail in Supplementary Section 'Evaluation of Linearity'.

2.5 Classification Protocol

Classification cut-offs for the linearity descriptor were explored using a visual analysis. The following cut-offs were selected to distinguish straight and non-

straight (i.e. curved or folded) peptides:

$$\text{Class} = \begin{cases} \text{Straight}, & \text{if } (L > 4) \text{ and } (D \leq 1.0 \text{ Å}), \\ \text{Straight}, & \text{if } (L \leq 4) \text{ and } (D \leq 0.5 \text{ Å}), \\ \text{Non-straight}, & \text{otherwise.} \end{cases} \quad (1)$$

where L is the peptide length and D is the linearity descriptor (the mean absolute deviation in projected Cα positions from ideal positions) as defined in Equation S6 of Supplementary Section 'Evaluation of Linearity'.

Peptides classified as non-straight, but with $L \geq 6$ and (1.0 Å $\leq D \leq$ 2.5 Å) are sometimes essentially straight, but with a 'hooked' end. To check for the presence of a 'hook', the N-terminal residue's deviation from its ideal position was compared with the C-terminal residue's deviation. If the deviation of the N-terminal residue was more than the C-terminal residue then there is potentially a hook at the N-terminus of the peptide; otherwise there is a potential hook at the C-terminus. The average deviation of the whole peptide is then recalculated excluding the possible hooked terminal residue. If the average deviation is still more than 1.0 Å, then the process is repeated excluding up to three terminal residues. If, during this process, the average deviation for the peptide falls to $\leq$1.0 Å, the peptide is defined as having a hook and is reclassified as straight.

Non-straight peptides are further classified into curved and folded on the basis of the number of contacts among the residues along the peptide using a 'contact rule' as described below.

A flow chart of the classification protocol is shown in Supplementary Figure S5.

The contact rule

Non-straight peptides, identified as described above, are classified into curved or folded classes using a 'contact rule' which counts the number of contacts (defined as a distance of $\leq$ 4 Å between any pairs of atom centres) among residues as we walk along the peptide. The number of contacts is calculated between pairs of residues defined as:

$$n - d \Leftrightarrow n + i + d \quad (2)$$

where $\Leftrightarrow$ represents a contact being made, n is the current reference position in the peptide, i is the minimum separation between sets of residues making contact ($i \geq 3$) and d is a step along the residues of the peptide ($d \geq 0$). Thus the rule does not just identify local and distant contacts, but identifies sets of contacts made by residues that are paired with one another as one walks along the sequence. This equation is iterated over n, d and i as shown in Figure 1.

Local contacts are defined as those where the minimum separation is less than (or equal to) a 'contact threshold', T_C (i.e. $i \leq T_C$), while distant contacts have $i > T_C$. T_C is defined as:

$$T_C = \begin{cases} N/2 & \text{if } (N \leq 12), \\ 5 & \text{otherwise.} \end{cases} \quad (3)$$

where N is the length of the peptide.

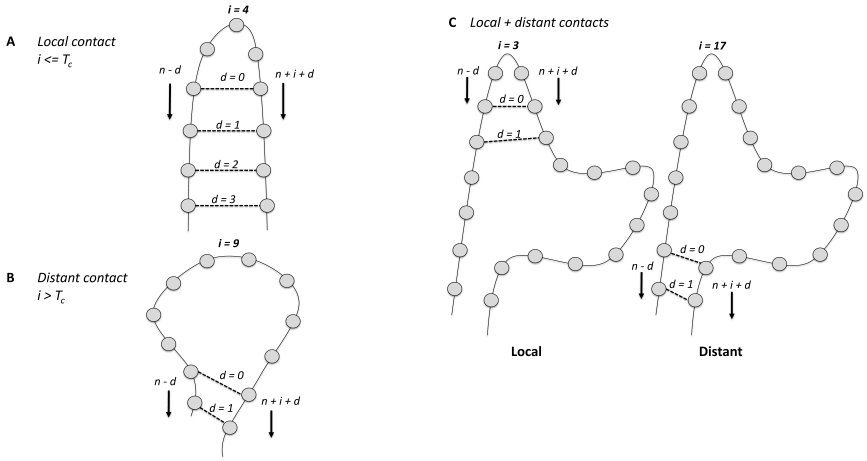

Figure 1. Contacts in folded peptides are defined between positions $n - d$ and $n + i + d$ where n is a reference position in the peptide to start identifying contacting pairs of amino acids, i is the minimum separation between residues making contact ($i \geq 3$) and d is a step along the residues of the peptide ($d \geq 0$). Iterations are performed over n, d and i. As defined in Equation 3, when $i \leq T_C$, contacts are defined as local while $i > T_C$ defines distant contacts. a) An example of a peptide with 11 residues which has local contacts; at this stage of the algorithm, $i = 4$ with $d = 1 \ldots 4$. b) An example of a 12 residue peptide which has distant contacts; at this stage of the algorithm, $i = 9$ with $d = 0 \ldots 2$ c) An example of a 20 residue peptide which has both local and distant contacts; at the stage of the algorithm in the left-hand panel, $i = 3$ and $d = 0 \ldots 1$ identifies two local contacts while in the right-hand panel, $i = 17$ and $d = 0 \ldots 1$ identifies two distant contacts.

Table 1. The number of regions and fragments in the dataset of 488 epitopes

	Epitopes	Regions	Fragments
Single-chain	446	1148	879
Multiple-chain	42	134	139
Combined	488	1282	1018

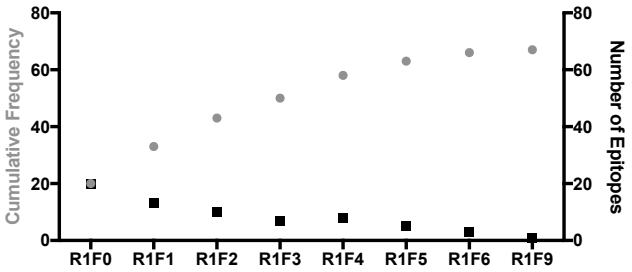

Figure 2. Number (black squares) and cumulative frequency (grey circles) of epitopes having a single region and different number of fragments.

Non-straight peptides are classified as folded or curved based on the number of local contacts (C_L, where $i \leq T_C$), distant contacts (C_D, where $i > T_C$) and total contacts ($C_T = C_L + C_D$) as shown in Equation 4.

$$\text{Class} = \begin{cases} \text{Folded}, & \text{if } C_L \geq 3 \\ \text{Folded}, & \text{if } C_D \geq 2 \\ \text{Folded}, & \text{if } C_T \geq 3 \\ \text{Curved}, & \text{otherwise.} \end{cases} \tag{4}$$

See also Algorithm 1 in Supplementary Material.

3 Results and Discussion

Epitopes were analyzed in terms of the number of regions, number of fragments, region length, longest region length, probability of having other regions given a region of a certain length, the relationship between region length and either the number of regions or the number of fragments, the epitope size, the shape and the secondary structure composition.

In the dataset of 488 epitopes, a total of 1282 regions and 1018 fragments were observed as shown in Table 1. In describing epitopes, we adopt a nomenclature of RxFy where x is the number of regions (three or more contacting residues with gaps of up to three residues between contacting residues) and y is the number of fragments (individual contacting residues that are not part of a region).

3.1 Distribution of Regions and Fragments

Among the 488 distinct B-cell epitopes most (∼91%) were composed of a single antigen chain while the remaining ∼9% were composed of multiple chains. Epitopes were composed of one to nine regions and zero to sixteen fragments with the most frequent compositions being R2F0 > R3F2 > R2F2 > R3F0≈R2F1 (See Supplementary Table S1).

Only 20 epitopes out of 488 (4%) were truly linear (R1F0) agreeing with several studies that report over 90% of B-cell epitopes are conformational (Van Regenmortel, 2001; Haste Andersen *et al.*, 2006; Theisen *et al.*, 2000). However, approximately 14% are comprised of a single region and up to 9 fragments (R1F0–R1F9) as shown in Figure 2. This is in agreement with the work of Sivalingam and Shepherd (Sivalingam and Shepherd, 2012) who found that 12–15% of epitopes are contain a single region.

The full dataset (Supplementary Table S1) was divided into epitopes containing only single chains (Supplementary Table S2) and those containing multiple chains (Supplementary Table S3). A χ^2 test shows that the region/fragment distribution in these datasets is significantly different ($p = 0.026$). See Supplementary Table S4 for the grouping performed to satisfy the requirements of a χ^2 test (no expected <1 and <20% less than 5).

The statistical analysis from Supplementary Table S4 shows that the statistical difference between the two datsets stems from the absence of R1 in the multiple-chain dataset. When R1 is removed from both the single (Supplementary Table S5) and multiple chain (Supplementary Table S6) dataset, the χ^2 test shows that the data are no longer significantly different ($p = 0.08$) and we have therefore combined the datasets for all the further analysis. See Supplementary Table S7 for the grouping of data.

3.2 Lengths of Regions

In the epitope dataset, the region length ranges from 3–30 residues (mean=8.15, σ=4.44) with 94% $\leq$ 16 residues (Supplementary Figure S6a). A similar trend was observed by Kringelum *et al.* (2013), where regions of up to 15 residues were seen, but in the present study, about 7.8% have a length of more than 15 residues. This is likely to be explained by the larger dataset and the fact that a gap of up to 3 non-epitope residues is allowed in regions in the present study compared with only one amino acid in the work of Kringelum *et al.*

The distribution of the longest region was calculated for the epitope dataset and was found to range from 4–30 residues (Supplementary Figure S6b).

Probability of a Region Being the Longest

Given the scenario that a region of a certain length is being analysed as a candidate immunogen, one needs to know whether there are likely to be other longer regions within the same epitope. In other words, for a given region length, what is the chance of that being the longest region and therefore the major structural component of the epitope? This would allow us to extrapolate the results to epitopes where the antigen structure and only the rough epitope is known (perhaps by alanine scanning mutagenesis). The fraction of epitopes having region length X and also having regions longer than X was calculated

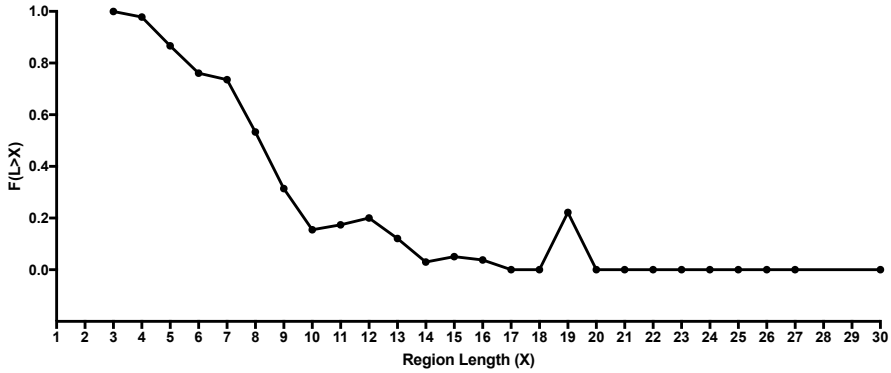

Figure 3. Given a region length X, the probability that there are also regions of length $> X$ in the epitope is plotted against X.

as follows and plotted for each possible length of a region in the observed data (Figure 3).

$$F_{(L>X)} = \frac{N_{R_{l=X}} \cap N_{R_{l>X}}}{N_{R_{l=X}}} \tag{5}$$

where $F_{(L>X)}$ is the probability of having a region of length $> X$ given a region of length X, $N_{R_{l=X}}$ is the number of epitopes having regions of length X, $N_{R_{l>X}}$ is the number of epitopes having regions of length $> X$, and the intersection represents the number of epitopes having a region of length X also having regions of length $> X$.

The data show that epitope regions of length 3 or 4 will almost always be accompanied by longer regions. This falls off gradually as region length increases and it becomes statistically unlikely to see longer regions accompanying regions of 14 amino acids or more ($F_{(L>X)}$ falls below 0.05). However there is an unexpected peak at length 19 showing that epitopes of this length do tend to be accompanied by a longer region. Looking at these examples, it was found that the dataset contains 2 such examples (Supplementary Figure S7). In general, however, it can be concluded that if an epitope has a region of length 14 residues or more, it is most likely that this is the longest region and it is likely to be a linear epitopes.

The Relationship Between the Length of Regions and the Number of Regions and Fragments

It was hypothesized that the length of a region would be inversely correlated with the number of regions. In other words, given the limited dimensions of an epitope, if it includes a long region, it is less likely that there would be other regions present. Thus, again, with the aim of identifying regions that are likely to be dominant within epitopes, the correlation between region length and either the number of regions or of fragments was investigated. For $R_{X_{\max}}$, a given maximum region length X, the fraction of epitopes also having N_R regions ($F_{(R_{X_{\max}},N_R)}$) and the fraction of epitopes also having N_F fragments ($F_{(R_{X_{\max}},N_F)}$) was calculated as:

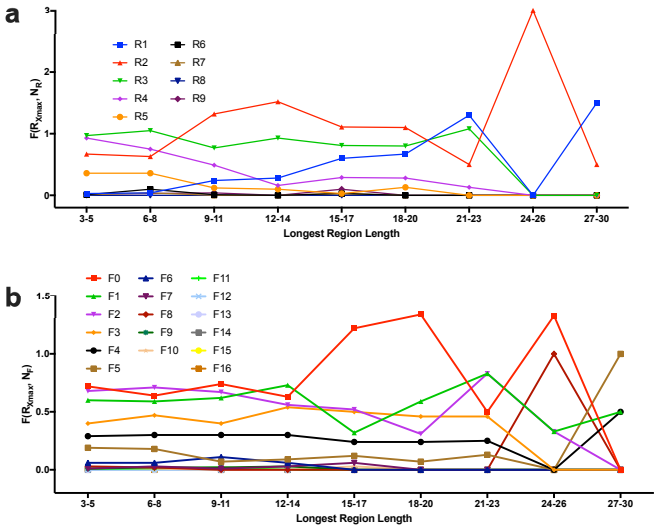

Figure 4. The fraction of epitopes having a) N_R regions ($F_{(R_{X_{\max}},N_R)}$) and b) N_F fragments ($F_{(R_{X_{\max}},N_F)}$), for a given length maximum region length ($R_{X_{\max}}$) in the epitope dataset. R1–R9 represent the number of regions (N_R) while F1–F16 represent the number of fragments (N_F). The peaks for region length $\geq$24 in the dataset are artefacts of the very small number of epitopes having such long regions.

$$F_{(R_{X_{\max}},N_R)} = \frac{N_{R_{X_{\max}}} \cap N_R}{N_{R_{X_{\max}}}} \tag{6}$$

and

$$F_{(R_{X_{\max}},N_F)} = \frac{N_{R_{X_{\max}}} \cap N_F}{N_{R_{X_{\max}}}} \tag{7}$$

where $N_{R_{X_{\max}}}$ is the number of epitopes having maximum region length $X_{\max}$ while N_R and N_F are the number of regions and fragments respectively ($1 \leq N_R \leq 9$ and $0 \leq N_F \leq 16$). The intersection is thus the number of epitopes with maximum region length $X_{\max}$ also having N_R regions or N_F fragments respectively.

In the epitope dataset, epitopes having smaller maximum region lengths ($N_{R_{X_{\max}}}$) tend to have more regions, while epitopes having longer maximum region lengths (12–23 residues) generally have one to three region (Figure 4a). Epitopes having a maximum region length of 15–20 generally have no fragments (Figure 4b). In other words, truly linear epitopes (R1F0) mostly have a region of length 15–20 amino acids while those with a single region and potentially a small number of fragments have a length of 12–23 amino acids. Thus epitopes identified by methods such as alanine scanning mutagenesis that appear to contain a region of these lengths are likely candidates for linear, or near-linear, epitopes.

3.3 The Relationship Between the Number of Regions and the Number of Fragments

It was hypothesized that an epitope with fewer regions may be expected to have more fragments and *vice versa*. Similarly, the length of regions and the number of residues comprising an epitope might have a relationship with the number of fragments in an epitope. Pearson correlation coefficients were calculated for the dataset of 488 epitopes (see Supplementary Figure S8), but no evidence for a correlation was observed.

3.4 Epitope Size

The size of an epitope was defined as the total number of residues that constitute regions and fragments. In the combined dataset, the mean size was 23.36 residues ($\sigma = 8.52$) with a minimum observed size of 5 and maximum of 83 (Figure 5).

These results are similar to a study conducted by Rubinstein *et al.* (2009a) on a dataset of 53 epitopes which concluded that 75% of epitopes are 15–25 residues. This compares with 80% of epitopes containing 15–35 residues in the current study. Another analysis of 107 epitopes (Kringelum *et al.*, 2013) calculated the average size of an epitope to be 15 residues. Presumably the differences result in part from our 5-times larger dataset, but mostly from our different definition of epitope residues which include the non-contacting residues contained within our regions.

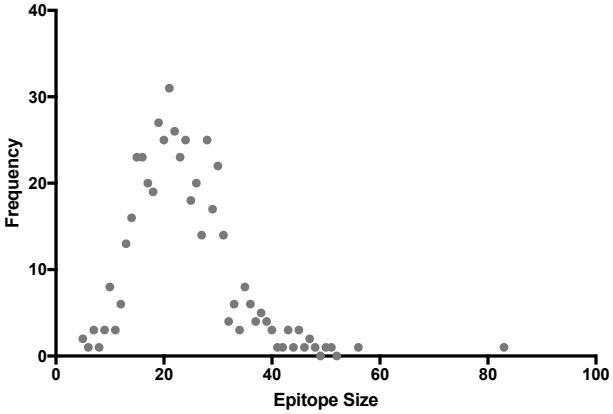

Figure 5. Distribution of epitope size in the dataset of 488 epitopes.

Table 2. Classification of regions into the three shape classes and sub-classification on the basis of secondary structure. The analysis is based on 1282 regions from 488 epitopes.

Shape	Structure	Number of Regions
Straight (s)	Total	508
	Helix	152
	Strand	127
	Coil	229
Curved (c)	Total	626
	Helix	40
	Strand	32
	Coil	554
Folded (f)	Total	148
	Helix	23
	Strand	33
	Coil	92
Total		1282

3.5 Shapes of Regions

Our dataset of 1282 regions was investigated on the basis of region shapes and secondary structures. The shape of each region was classified as either straight (s), curved (c) or folded (f). Each of the shapes was then further classified by secondary structure content as shown in Table 2.

Lengths of Each Region Shape

Most straight and curved regions are 4–9 residues long, whereas folded regions are comprised of 11–17 residues (Figure 6a). The data shows that the folded regions tend to be longer than the other shapes of regions. The difference in lengths of the folded with straight and curved was found to significant ($p < 0.0001$). However the difference between straight and curved was found insignificant ($p < 0.933$)

Correlation of Multiple Region Shapes

The distribution of each of the shapes in the region dataset is shown in Supplementary Figure S9. 66 epitopes ($\approx 14\%$) had only straight regions (up to 6); 87 epitopes ($\approx 18\%$) had only curved regions (up to 5); 35 epitopes ($\approx 7\%$) had only folded regions (up to 2).

In order to investigate all combinations of region shapes, a 3D contingency table was generated. The significance of particular combinations of straight, curved and folded regions was then calculated using a 3-way (2x2x2) χ^2 test as described in Supplementary Section 'Calculation of 3D χ^2'. A Bonferroni correction was applied to all p-values shown in this section (i.e. the p-values were multiplied by the number of tests rather than dividing the threshold for significance ($\alpha = 0.05$) by the number of tests).

A total of 126 ($7 \times 6 \times 3$) combinations were formed resulting from epitopes having 0–6 straight, 0–5 curved and 0–2 folded regions (Supplementary Table S8). The null hypothesis for this 3 way test is that there is no correlation between any of the shapes. For a χ^2 test to be valid, there should be no more than 20% of the expected values below five and no expected values below one (Dytham, 2010). Since the contingency table had many very low expected values, data were grouped as shown in Supplementary Table S9. A 3-way χ^2 test on the complete grouped table showed a p-value ≈ 0 indicating a strong correlation among the shapes. Below, the notation sx is used to indicate x straight regions, fx to indicate x folded regions and cx to indicate x curved regions.

The observed and expected values of different combinations showed clear trends. For example, the chance of having a single straight region ($s1$) when there are no curved or folded regions is much less likely than expected ($p = 2.07\times10^{-14}$). Straight regions are more frequently α-helix or β-strand secondary-structure elements (Table 2) and these are rarely observed alone. When there is a single straight region it is accompanied by one or two curved regions much more frequently than expected by chance ($f0/c1/s1$, $p = 4.77\times10^{-3}$; $f0/c2/s1$, $p = 1.47\times10^{-3}$).

On the other hand, when there are two straight regions, these occur in the absence of any curved or folded regions much more frequently than expected ($p \approx 0$). In other words, when two straight regions occur in an epitope they tend to be present without the contribution of regions of any other shape. As

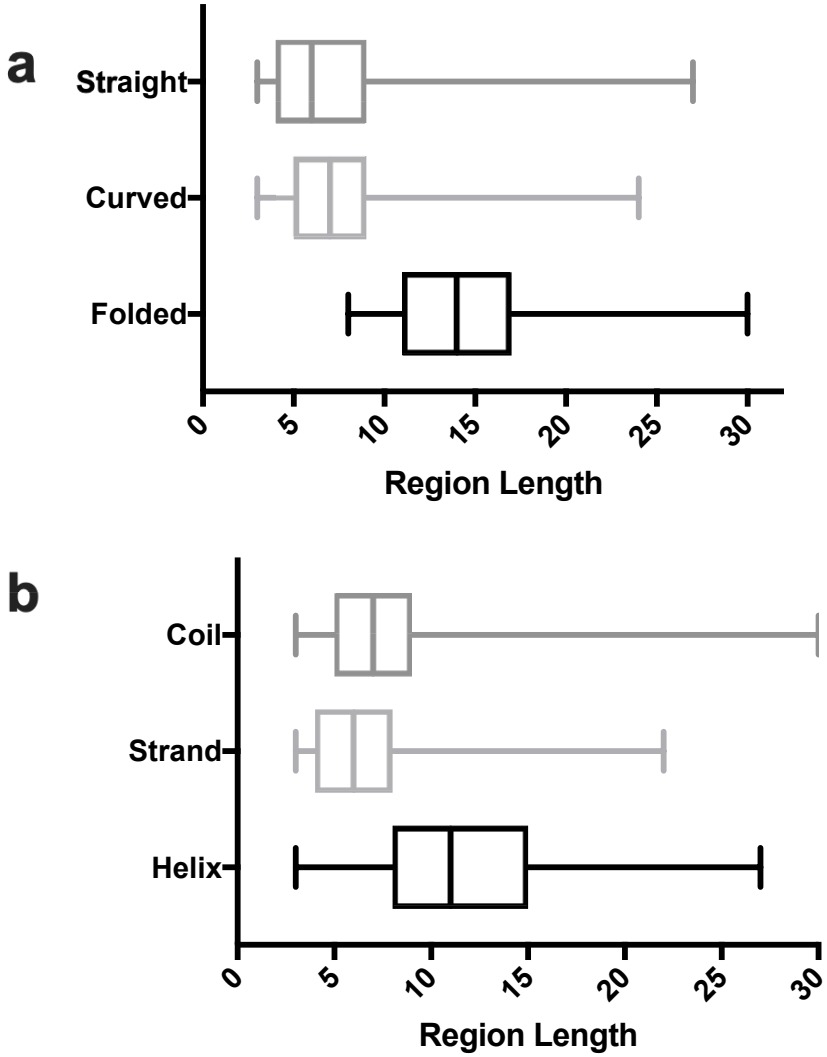

Figure 6. a) The distribution of region lengths in straight, curved and folded regions in the dataset of 488 epitopes. b) The distribution of region lengths in regions that are predominantly helix, strand or coil in the datasets of 488 epitopes.

noted above, straight regions are more frequently α-helix or β-strand secondary-structure elements and these are often observed as two parallel or anti-parallel strands or helices. Two examples are shown in Supplementary Figure S10. Similarly epitopes having more than three straight regions, have zero or one curved regions much more frequently than expected (*f0/c0/s3–s6*, $p \approx 0$; *f0/c1/s3–s6*, $p = 0.0069$). In the context of creating peptide vaccines, such straight regions could be coupled by peptide stapling (Fairlie and Dantas de Araujo, 2016) or using a suitable presentation scaffold (Gururaja *et al.*, 2000; Tiede *et al.*, 2014).

Supplementary Table S9 also shows that epitopes containing single curved regions (*c1*), in the absence of straight or folded regions, occur much less frequently than expected by chance ($p = 1.85\times10^{-9}$). Unlike linear regions (that are more frequently α-helices or β-strands, often occurring in pairs in epitopes) and folded regions, both of which are internally stabilized, curved regions need other elements to stabilize their conformation. However a single curved region accompanied by a single straight region (*c1* with *s1*) occurs much more than expected by chance ($p = 4.77\times10^{-3}$). Similarly, when there are two curved regions, these are most likely to be present together or in the presence of a single straight region. Again, it is likely that two curved regions stabilize each other or with the help of a single straight region (*f0/c2/s1*, $p = 1.47\times10^{-4}$; *f0/c2/s2*, $p = 1.26\times10^{-4}$; *f0/c2/s3–s6*, $p = 3.71\times10^{-8}$). When there are three or more curved regions, it is unusual to see any straight or folded regions ($p = 6.88\times10^{-14}$).

When there are either one or two folded regions, it is rare to see any curved or straight regions ($p = 0$), implying that one or two folded regions are enough to form an epitope without the contribution of other region shapes. The significance of any two shapes occurring together in one epitope was further evaluated using a 2x2 χ^2 test (with Yates correction) and shows that one or two folded regions (*f1* or *f2*) tend not to occur with straight regions ($p = 1.87\times10^{-10}$) or with curved regions ($p = 6.88\times10^{-14}$). Such folded regions tend to be self-stabilizing and are normally longer than the other two region shapes (Figure 6a) explaining why they are generally present in the absence of other region shapes.

14 of the epitopes in the dataset contain all three shapes (i.e. *s1/c1/f1*), but this is much less than expected by chance ($p = 4.7\times10^{-3}$). As noted above, folded regions tend not to be accompanied by any curved or straight regions.

3.6 Secondary Structures of Epitopes

Epitope regions were classified into helix, strand and coil. Helical regions tend to be longer than strand regions ($p < 0.0001$) and are also longer than coil regions ($p < 0.0001$). This may be a result of the gaps of 3 amino acids allowed between contacting residues (Figure 6b). Strand regions were found to be marginally shorter than coil regions ($p < 0.041$).

Table 2 shows that, in general, regions are predominantly formed from coil (68% of regions). This agrees with previous studies where it was reported that epitopes are enriched in loops and depleted of helices and strands (Rubinstein *et al.*, 2008; Ofran *et al.*, 2008). Supplementary Figure S11 shows the distribution of secondary structures across the epitopes (i.e. the number of epitope regions having predominantly a given secondary structure type). In the whole dataset, nearly 39% of epitopes (191 of 488) had only coil (from 1–8 regions) while nearly 10% had only helical regions (49 of 488 epitopes, 1–3 regions) and only 3.8% had

only β-strand regions (19 of 488 epitopes, 1–6 regions). The remaining ∼47% of epitopes contain a mixture of regions with different secondary structure classes.

A three-way χ^2 test confirmed that the presence of one type of secondary structure element influences the presence of the others. A three-way contingency table (4x7x9) was calculated to include each combination of helix, strand and coil (Supplementary Table S10) grouped as shown in Supplementary Table S11. The notation Hx is used to indicate x α-helical regions, Ex to indicate x β-strand regions and Cx to indicate x coil regions. A Bonferroni correction was applied to all p-values shown in this section (i.e. the p-values were multiplied by the number of tests).

The data showed that the chances of having 3–8 coiled regions in the absence of any helical or strand regions (i.e. *C3–C8/H0/E0*) is much more likely than expected by chance. (2x2x2 χ^2 test, $p \approx 0$). In other words, epitopes consisting only of coil regions occur much more frequently than expected by chance.

Similarly, helical regions tend to occur without the presence of regions having other secondary structure classes. The probability of having 1–3 helical regions in the absence of any coiled and strand region is much more likely than expected by chance ($p \approx 0$). In summary, most epitopes tend to have regions with the same type of secondary structure element. While combinations of regions with different secondary structure classes are seen, this occurs less frequently than expected by chance.

4 Conclusions

Peptide vaccines have numerous potential advantages over conventional vaccines produced from killed or attenuated pathogens or pathogen proteins. However, the majority of natural B-cell epitopes (sites where antibodies bind) are discontinuous making it more difficult to create peptide vaccines. We define 'regions' (R) as stretches of at least three amino acids making contact with antibody with gaps of up to three residues between contacting residues and 'fragments' (F) as other contacting residues. In our analysis, only 4% of epitopes are truly linear and consist of just a single region with no other 'fragments' (R1F0 in our nomenclature). However, ∼14% have only one region with up to 9 individual residue fragments (R1F1–R1F9) and these may also make suitable immunogens. This analysis ('Distribution of Regions and Fragments' and Supplementary Table S1) broadly agrees with previous analyses of smaller datasets (Rubinstein *et al.*, 2008; Sivalingam and Shepherd, 2012).

In addition ∼38% of epitopes have 2 regions with up to 16 fragments (R2F0–R2F16) and it is possible that these regions could be artificially coupled to produce more complex peptide immunogens. For example, peptides can be stapled (Fairlie and Dantas de Araujo, 2016) or presented on a suitable scaffold (Gururaja *et al.*, 2000; Tiede *et al.*, 2014). Thus, development of peptide-based vaccines may be possible for ∼51% of antigens.

MacRaild *et al.* (2016) have shown that linear epitopes are enriched in disordered and flexible antigens and since these proteins cannot be crystallized, the protein databank has an inherent bias against these linear epitopes and thus our analysis will under-estimate the number of linear epitopes. However this is also a problem with all previous structural analysis of epitopes. It could also be argued that excluding peptides of <30 amino acids in curating the dataset

may exclude proteins with disordered regions which have been truncated in order to aid in crystallization. Only 10 of the 153 antigen structures (∼6.5%) rejected for being <30 amino acids in length have complete sequences longer than 30 amino acids (34–108 residues). While these might represent truncation of intrinsically disordered regions, we do not believe that this will further bias the dataset because, if such regions represented epitopes, these would stabilized by their interaction with an antibody and thus would not be flexible in the complex.

If one does not have a structure of a protein of interest, computational prediction of immuno-dominant B-cell epitopes would be valuable. However, this is a very hard problem and prediction of such regions is difficult even when one does have a structure. Two factors contribute to this difficulty: (i) the discontinuity described above and (ii) the fact that immuno-dominant B-cell epitopes are generally protein surfaces that are not normally involved in protein-protein interactions and are therefore difficult to distinguish from the rest of a protein surface. Increasing our understanding of the structure of these regions may help in improving B-cell epitope prediction software.

From the perspective of vaccine development, where the structure of an antigen of interest is not known, techniques such as alanine scanning mutagenesis can be used to define a functional epitope given an antibody that binds to the protein. If one identifies an epitope that appears to be largely continuous, it would then be useful to know whether this is likely to be the only region. Our analysis of region lengths showed that, in general, if an epitope has a region of at least 14 residues, there are unlikely to be longer regions and consequently such regions are likely be the dominant component of the epitope.

Another problem is that a peptide epitope, taken out of the context of the whole protein, will not necessarily adopt the same conformation as it does in the whole protein. Consequently it may fail to induce an immune response which generates antibodies that cross-react with the native protein. If a peptide adopts a conformation more similar to the conformation that it has in the native protein, it is more likely to activate a B-cell response that generates specific antibodies that will bind to whole antigen. We classified the shape of the 1282 epitope regions and found that nearly 40% are straight in shape and 11% are folded. While regions mostly adopt coil conformations rather than ordered secondary structure, nearly 30% of straight regions are α-helical and 25% are β-strand (Table 2). Helical straight regions and folded regions are more likely to adopt the same conformation as an isolated peptide as they are more often internally stabilized. 49% of regions are curved in shape and would be unlikely to adopt the same conformation when isolated. Again presentation scaffolds may help with this problem.

94% of regions were seen to be up to 16 residues long, but ranged from 3 to 30 residues with an average size (i.e. total number of residues in regions and fragments) of 23. There was no correlation between the number of fragments and the number of regions, the longest region, the total region residues or the average number of regions. Nonetheless, regions of 13 residues or fewer tend to be accompanied by additional regions while regions of $\geq$14 residues are generally predominantly-continuous epitopes having only one region and, perhaps, some additional fragments.

In summary, we have provided a comprehensive structural analysis of 488 B-cell epitopes encompassing epitopes formed both by single chains and across

multiple-chains. We expect this analysis to be helpful in the design of peptide-based vaccines and in improving the prediction of B-cell epitopes.

5 Acknowledgements

SF thanks the UCL Overseas Research Scholarship, The Schlumberger Foundation and UCB Pharma for funding. Tom Northey is thanked for useful discussions.

6 Competing Interests

Declaration of Interests: None

References

Arthur, L. O., Pyle, S. W., Nara, P. L., Bess, J. W., Gonda, M. A., Kelliher, J. C., Gilden, R. V., Robey, W. G., Bolognesi, D. P. and Gallo, R. C. (1987) Serological responses in chimpanzees inoculated with human immunodeficiency virus glycoprotein (gp120) subunit vaccine, *Proceedings of the National Academy of Sciences, USA*, **84**, 8583–8587.

Breman, J. G. and Arita, I. (1980) The confirmation and maintenance of smallpox eradication, *The New England Journal of Medicine*, **303**, 1263–1273.

Comeau, S. R., Gatchell, D. W., Vajda, S. and Camacho, C. J. (2004) ClusPro: an automated docking and discrimination method for the prediction of protein complexes, *Bioinformatics*, **20**, 45–50.

Dytham, C., (2010). *Choosing and using statistics: a biologist's guide*. Wiley-Blackwell, New Jersey, USA, 3rd edition.

Fairlie, D. P. and Dantas de Araujo, A. (2016) Stapling peptides using cysteine crosslinking, *Peptide Science*, **106**, 843–852.

Ferdous, S. and Martin, A. C. R. (2018) AbDb: Antibody structure database — a database of PDB-derived antibody structures, *Database*, **2018**, bay040.

Gururaja, T. L., Narasimhamurthy, S., Payan, D. G. and Anderson, D. (2000) A novel artificial loop scaffold for the noncovalent constraint of peptides, *Cell Chem. Biol.*, **7**, 515–527.

Haste Andersen, P., Nielsen, M. and Lund, O. (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures, *Protein Science*, **15**, 2558–2567.

Henderson, D. A. (1998) Eradication: lessons from the past, *Bulletin of the World Health Organization*, **76**, 17.

Hovi, T. (2001) Inactivated poliovirus vaccine and the final stages of poliovirus eradication, *Vaccine*, **19**, 2268–2272.

John, T. J. (2000) The final stages of the global eradication of polio, *The New England Journal of Medicine*, **343**, 806.

Kabsch, W. and Sander, C. (1983) How good are predictions of protein secondary structure?, *FEBS letters*, **155**, 179–182.

Kringelum, J. V., Nielsen, M., Padkjær, S. B. and Lund, O. (2013) Structural analysis of B-cell epitopes in antibody: protein complexes, *Molecular Immunology*, **53**, 24–34.

Kulkarni-Kale, U., Bhosle, S. and Kolaskar, A. S. (2005) CEP: a conformational epitope prediction server, *Nucleic Acids Research*, **33**, W168–W171.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658–1659.

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

Liang, S., Zheng, D., Zhang, C. and Zacharias, M. (2009) Prediction of antigenic epitopes on protein surfaces by consensus scoring, *BMC Bioinformatics*, **10**, 302.

Lo, Y.-T., Pai, T.-W., Wu, W.-K. and Chang, H.-T. (2013) Prediction of conformational epitopes with the use of a knowledge-based energy function and geometrically related neighboring residue characteristics, *BMC Bioinformatics*, **14**, S3.

Lollier, V., Denery-Papini, S., Larré, C. and Tessier, D. (2011) A generic approach to evaluate how B-cell epitopes are surface-exposed on protein structures, *Molecular immunology*, **48**, 577–585.

MacRaild, C. A., Richards, J. S., F., A. R. and Norton, R. S. (2016) Antibody recognition of disordered antigens, *Structure*, **24**, 148–157.

Moreau, V., Fleury, C., Piquer, D., Nguyen, C., Novali, N., Villard, S., Laune, D., Granier, C. and Molina, F. (2008) PEPOP: computational design of immunogenic peptides, *BMC Bioinformatics*, **9**, 71.

Nemchinov, L., Liang, T., Rifaat, M., Mazyad, H., Hadidi, A. and Keith, J. (2000) Development of a plant-derived subunit vaccine candidate against hepatitis C virus, *Archives of Virology*, **145**, 2557–2573.

Novotnỳ, J., Handschumacher, M., Haber, E., Bruccoleri, R. E., Carlson, W. B., Fanning, D. W., Smith, J. A. and Rose, G. D. (1986) Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains), *Proceedings of the National Academy of Sciences, USA*, **83**, 226–230.

Ofran, Y., Schlessinger, A. and Rost, B. (2008) Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B-cell epitopes, *The Journal of Immunology*, **181**, 6230–6235.

Petrovsky, N. and Aguilar, J. C. (2004) Vaccine adjuvants: current state and future trends, *Immunology and Cell Biology*, **82**, 488–496.

Ponomarenko, J., Bui, H.-H., Li, W., Fusseder, N., Bourne, P. E., Sette, A. and Peters, B. (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes, *BMC Bioinformatics*, **9**, 514.

Ponomarenko, J. V. and Bourne, P. E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation, *BMC Structural Biology*, **7**, 64.

Porter, C. T. and Martin, A. C. R. (2015) BiopLib and BiopTools — a C programming library and toolset for manipulating protein structure, *Bioinformatics*, **31**, 4017–4019.

Purcell, A. W., McCluskey, J. and Rossjohn, J. (2007) More than one reason to rethink the use of peptides in vaccine design, *Nature Reviews Drug Discovery*, **6**, 404–414.

Rubinstein, N. D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J. M. and Pupko, T. (2008) Computational characterization of B-cell epitopes, *Molecular Immunology*, **45**, 3477–3489.

Rubinstein, N. D., Mayrose, I., Martz, E. and Pupko, T. (2009a) Epitopia: a web-server for predicting B-cell epitopes, *BMC Bioinformatics*, **10**, 287.

Rubinstein, N. D., Mayrose, I. and Pupko, T. (2009b) A machine-learning approach for predicting B-cell epitopes, *Molecular Immunology*, **46**, 840–847.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. and Wolfson, H. J. (2005) PatchDock and SymmDock: servers for rigid and symmetric docking, *Nucleic Acids Research*, **33**, W363–W367.

Sivalingam, G. N. and Shepherd, A. J. (2012) An analysis of B-cell epitope discontinuity, *Molecular Immunology*, **51**, 304–309.

Smith, D. K. and Thornton, J. M., (1989). SSTRUC: Computer program. Department of Biochemistry and Molecular Biology, University College London.

Sun, D., Seyer, J., Kovari, I., Sumrada, R. and Taylor, R. (1991) Localization of protective epitopes within the pilin subunit of the *Vibrio cholerae* toxin-coregulated pilus, *Infection and Immunity*, **59**, 114–118.

Sun, J., Wu, D., Xu, T., Wang, X., Xu, X., Tao, L., Li, Y. and Cao, Z.-W. (2009) SEPPA: a computational server for spatial epitope prediction of protein antigens, *Nucleic Acids Research*, **37**, W612–W616.

Sun, J., Xu, T., Wang, S., Li, G., Wu, D. and Cao, Z. (2011) Does difference exist between epitope and non-epitope residues?, *Immunome Research*, **201**, 1–11.

Sweredoski, M. J. and Baldi, P. (2008) PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure, *Bioinformatics*, **24**, 1459–1460.

Theisen, D., Bouche, F., El Kasmi, K., von der Ahe, I., Ammerlaan, W., Demotz, S. and Muller, C. (2000) Differential antigenicity of recombinant polyepitope-antigens based on loop-and helix-forming B and T cell epitopes, *Journal of Immunological Methods*, **242**, 145–157.

Thompson, A. L. and Staats, H. F. (2011) Cytokines: the future of intranasal vaccine adjuvants, *Journal of Immunology Research*, **2011**.

Thornton, J., Edwards, M., Taylor, W. and Barlow, D. (1986) Location of 'continuous' antigenic determinants in the protruding regions of proteins., *The EMBO Journal*, **5**, 409.

Tiede, C., Tang, A. A. S., Deacon, S. E., Mandal, U., Nettleship, J. E., Owen, R. L., George, S. E., Harrison, D. J., Owens, R. J., Tomlinson, D. C. and McPherson, M. J. (2014) Adhiron: a stable and versatile peptide display scaffold for molecular recognition applications, *Prot. Eng. Des. Sel.*, **27**, 145–155.

Van Regenmortel, M. (2001) Antigenicity and immunogenicity of synthetic peptides, *Biologicals*, **29**, 209–213.

Zhao, L. and Li, J. (2010) Mining for the antibody-antigen interacting associations that predict the B-cell epitopes, *BMC Structural Biology*, **10**, S6.

Zhao, L., Wong, L., Lu, L., Hoi, S. C. and Li, J. (2012) B-cell epitope prediction through a graph model, *BMC Bioinformatics*, **13**, S20.