

B-cell Epitopes: Discontinuity and Conformational Analysis

Supplementary Materials

Saba Ferdous, Sebastian Kelm, Terry S. Baker,
Jiye Shi and Andrew C.R. Martin

Supplementary Sections

S1 Evaluation of Linearity

Step 1: A best-fit straight line is calculated through all $C\alpha$ atoms of a region. $P_N = (x_N, y_N, z_N)$ and $P_C = (x_C, y_C, z_C)$ are defined as the projections of the N-terminal and C-terminal $C\alpha$ coordinates onto the best-fit line and \mathbf{VL} is defined as a vector between P_N and P_C as shown in Supplementary Figure S1a. A ‘region vector’ ($\mathbf{VR} = \overrightarrow{C_N C_C}$) is also calculated using the N-terminal and C-terminal $C\alpha$ positions. \mathbf{VL} is required to be in the same direction as \mathbf{VR} . Initially it is defined as $\overrightarrow{P_N P_C}$; if the angle between \mathbf{VR} and \mathbf{VL} is $>90^\circ$, \mathbf{VL} is redefined as $\overrightarrow{P_C P_N}$.

To do this, a C program, `pdpline` (Martin, A.C.R., Raghavan, A. and Ferdous, S. (2014) “pdpline V1.2”, Software) was written which draws a best fit line through a specified set of $C\alpha$ atoms. This program works as follows:

1. Extract the structure (coordinates) of the zone of interest from the PDB file.
2. The centroid of the given set of coordinates is calculated.
3. The covariance matrix and Eigen vectors are computed.
4. Finally, the Eigen vector components of the regression line are used to find other points on the line that pass through the centroid. The Eigen vector with the largest value represents the best fit line passing through the centroid.

Step 2: The mid-point of the best fit line is calculated by taking the average of points P_N and P_C using Equation S1:

$$M = (x_M, y_M, z_M) = \frac{P_N + P_C}{2} = \left(\frac{x_N + x_C}{2}, \frac{y_N + y_C}{2}, \frac{z_N + z_C}{2} \right) \quad (\text{S1})$$

Step 3: The $C\alpha$ closest to the midpoint is identified and defined as C_i by calculating the Euclidean distance between the $C\alpha$ of each amino acid and the mid point using Equation S2.

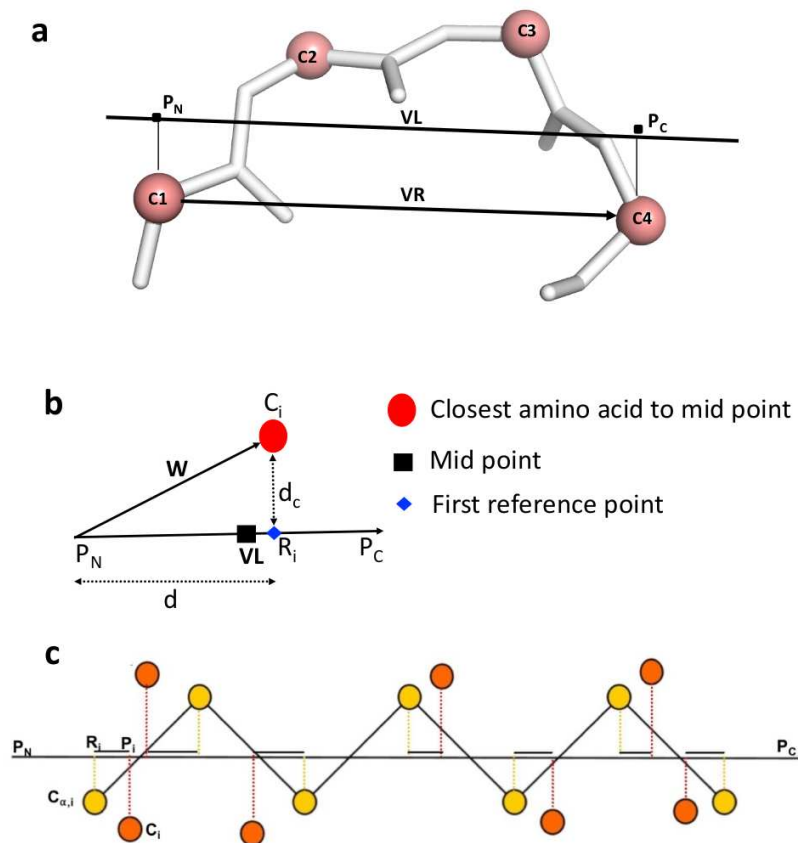


Figure S1: a) The vector \mathbf{VL} , is defined by projections of the N-terminal and C-terminal $C\alpha$ atoms (P_N and P_C) onto the best-fit line through the $C\alpha$ positions (shown as spheres) — in this example through a set of four residues. The region vector $\mathbf{VR} = \overrightarrow{C_N C_C}$ is defined between the N-terminal and C-terminal $C\alpha$ atoms. The direction of \mathbf{VL} ($\overrightarrow{P_N P_C}$ or $\overrightarrow{P_C P_N}$) is required to be the same as the direction of \mathbf{VR} (see text). b) Computation of the first reference point on the best fit line: P_N and P_C are the start and end points of \mathbf{VL} ; C_i is the closest $C\alpha$ to the mid point (M indicated by a square) and separated by the distance d_c . d is a line segment on \mathbf{VL} . c) Yellow points ($C_{\alpha,i}$) represent the $C\alpha$ positions of an ideal extended/straight peptide with their projections onto \mathbf{VL} indicated by R_i with a spacing of 3.5 Å. Red points (C_i) represent the actual $C\alpha$ positions with their projections onto \mathbf{VL} indicated by P_i . The distance between R_i P_i is calculated and used to find the average deviation of actual and ideal $C\alpha$ positions as a measure of linearity.

$$d_{min} = \min_{i=1}^n \left(\sqrt{(x_i - x_M)^2 + (y_i - y_M)^2 + (z_i - z_M)^2} \right) \quad (\text{S2})$$

Step 4: The first reference point (R_i) is determined. The closest $C\alpha$ (C_i) to the mid point identified in Step 3 is used to map the first reference point (R_i) onto the best-fit line. As shown in Supplementary Figure S1b, P_N and P_C are the start and end points of \mathbf{VL} ; C_i is the closest $C\alpha$ to the mid point (M) separated from \mathbf{VL} by distance d_c . The vector \mathbf{W} ($= \overrightarrow{P_N C_i}$) is then calculated.

The line segment distance d on the best fit line is computed using Pythagoras where the length of \mathbf{W} and of vector $\overrightarrow{C_i R_i}$ (length d_c) are taken as two sides of a right angled triangle. In order to find the point R_i along \mathbf{VL} at a distance d from P_N , \mathbf{VL} was normalised to \mathbf{U} by using Equation S3:

$$\mathbf{U} = \frac{\mathbf{VL}}{\|\mathbf{VL}\|} \quad (\text{S3})$$

Step 5: The ideal reference points are mapped onto \mathbf{VL} . These reference points are the positions at which one would expect $C\alpha$ atoms ($C\alpha_i$) to be projected for an ideal extended/straight β -strand (at a spacing of 3.5 Å) or ideal α -helix (spacing of 1.5 Å). In total, n reference points are mapped onto the best fit line where n is the number of residues in the peptide.

Given the first reference point (R_i) mapped in Step 4 for residue i , $i - 1$ reference points, R_{i-j} (where $1 \leq j \leq i - 1$), are mapped onto \mathbf{VL} in the direction of P_N from R_i as:

$$R_{i-j} = P_i - j \times \delta \mathbf{U} \quad (\text{S4})$$

and $n - i$ reference points, R_{i+j} (where $1 \leq j \leq n - i$), are mapped onto \mathbf{VL} in the direction of P_C from R_i as:

$$R_{i+j} = P_i + j \times \delta \mathbf{U} \quad (\text{S5})$$

δ , which is represented by d in Supplementary Figure S1b, is the ideal spacing between the projections of the atoms expressed as a fraction of the unit vector, \mathbf{U} .

Step 6: The actual $C\alpha$ points are mapped. These points (P_i) corresponding to the projection of each $C\alpha$ (C_i) in the peptide onto the best fit line. The procedure used above to calculate the first reference point is used where point C_i in vector \mathbf{W} corresponds to each $C\alpha$ projection onto \mathbf{VL} (Supplementary Figure S1c).

Step 7: The average deviation (D) is calculated between the reference projections (R_i) and the actual projections (P_i). This is used as a measure of the linearity as shown in Equation S6 and Supplementary Figure S1c:

$$D = \frac{\sum_{i=1}^n |R_i - P_i|}{n} \quad (\text{S6})$$

S2 Calculation of 3D χ^2

First, for the null hypothesis, complete independence was assumed between the three variables.

If rows, columns and planes are referred as r, c, p , (with dimensions R, C, P) with each cell containing the observed value o_{rcp} then a total, t , can be defined for a particular row, r , as:

$$t_{r++} = \sum_{c=1}^C \sum_{p=1}^P o_{rcp} \quad (\text{S7})$$

(where a subscript of + indicates summation over the appropriate index)

Similarly for columns and planes:

$$t_{+c+} = \sum_{r=1}^R \sum_{p=1}^P o_{rcp} \quad (\text{S8})$$

$$t_{++p} = \sum_{r=1}^R \sum_{c=1}^C o_{rcp} \quad (\text{S9})$$

The expected value for a given cell, e_{rcp} is then:

$$e_{rcp} = \frac{t_{r++} \times t_{+c+} \times t_{++p}}{N^2} \quad (\text{S10})$$

(where N is the total number of observations).

The χ^2 value is then calculated as normal:

$$\chi^2 = \sum_{r=1}^R \sum_{c=1}^C \sum_{p=1}^P \frac{(o_{rcp} - e_{rcp})^2}{e_{rcp}} \quad (\text{S11})$$

The number of degrees of freedom, D , is simply:

$$D = (R - 1)(C - 1)(P - 1) \quad (\text{S12})$$

The calculation of the expected values is based on information from Lienert, G.A. and Wolfrum, C. (1980) ‘‘Simplified formulas for three-way chi-square testing’’, *Biometrical Journal* **22**:159–167, Lin, C.F.J. (2006) ‘‘Analysis of three-way contingency table’’ (available online at <http://web.ntpu.edu.tw/~cflin/Teach/Cate/06CateUEN05ThreeWayPPT.pdf>, accessed 12th February 2018) and Li, Q. (2012) ‘‘STAT 504 — Analysis of Discrete Data. Penn State Science’’ (available online at <http://onlinecourses.science.psu.edu/stat504/book/export/html/102>, accessed 12th February 2018).

Algorithm S1 Calculation of contacts for folded and curved shape classification.

```
1: procedure
   Input
2:   PeptideResidues[]
   Initialization
3:    $C_T \leftarrow 0$ 
4:    $C_L \leftarrow 0$ 
5:    $C_D \leftarrow 0$ 
6:    $len \leftarrow$  length of peptide
   Procedure
7:   if  $len \leq 12$  then
8:      $T_C \leftarrow len/2$ 
9:   else
10:     $T_C \leftarrow 5$ 
11:   end if
12:   for  $n \leftarrow 0, n < len$  do
13:     for  $i \leftarrow 3, i < len - 3$  do
14:       for  $d \leftarrow 0, d < len$  do
15:          $res1 \leftarrow$  PeptideResidues[ $n - d$ ]
16:          $res2 \leftarrow$  PeptideResidues[ $n + i + d$ ]
17:         if  $res1 \geq 0$  &  $res2 < len$  then
18:           CalculateAtomicDistance( $res1, res2$ )
19:           if  $distance \leq 4.0$  then
20:             if  $i \leq T_C$  then
21:                $C_L++$ 
22:             else
23:                $C_D++$ 
24:             end if
25:           end if
26:         end if
27:       end for
28:     end for
29:   end for
30:    $C_T \leftarrow C_L + C_D$ 
31:   if  $C_L \geq 3 \parallel C_D \geq 2 \parallel C_T \geq 3$  then
32:      $shape \leftarrow$  folded
33:   else
34:      $shape \leftarrow$  curved
35:   end if
36: end procedure
```

Supplementary Figures

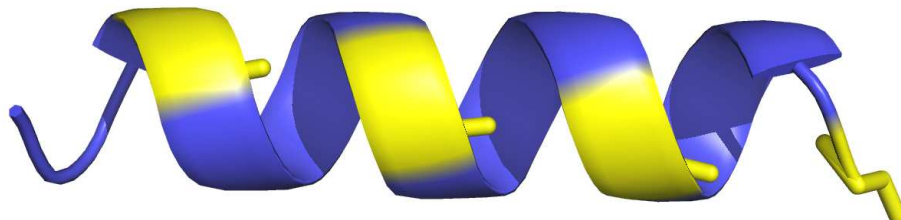


Figure S2: A gap of up to three residue in the epitope region (straight conformation). Amino acid residues which lie on the same face of an α -helix are shown in yellow with a spacing of 3 between them.

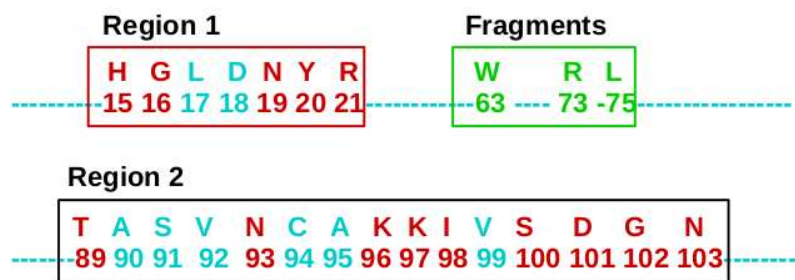
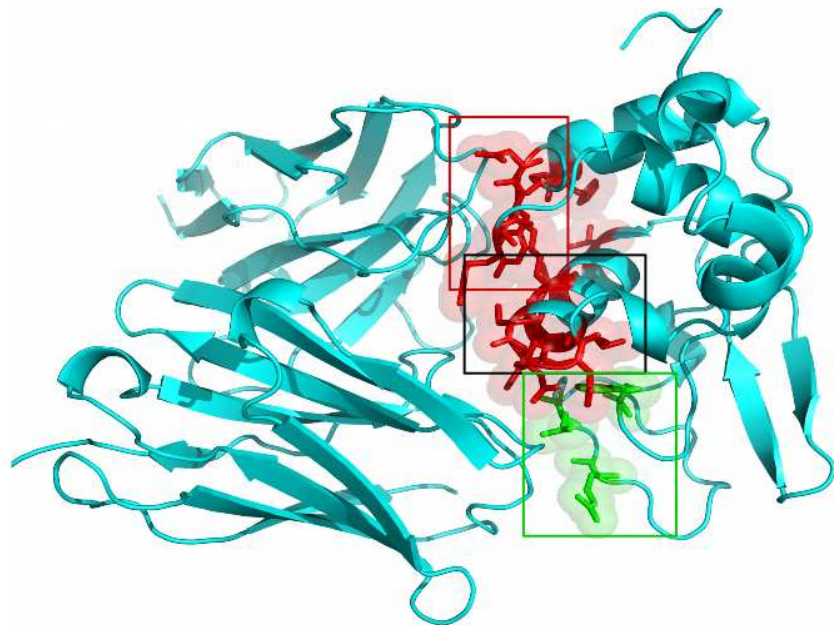


Figure S3: Regions and Fragments — This epitope is comprised of two regions and three fragments. Region 1 has a gap of two non-contacting residues (shown in blue) while Region 2 has three gaps of up to three non-contacting residues. The contacting residues in the epitope are shown in red.

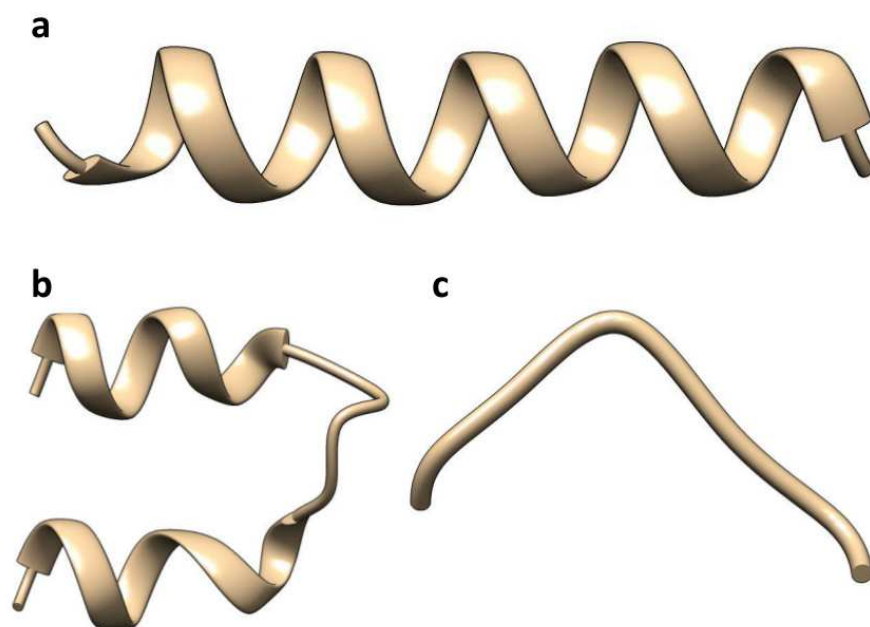


Figure S4: Region shapes: a) Straight, b) Folded, and c) Curved regions.

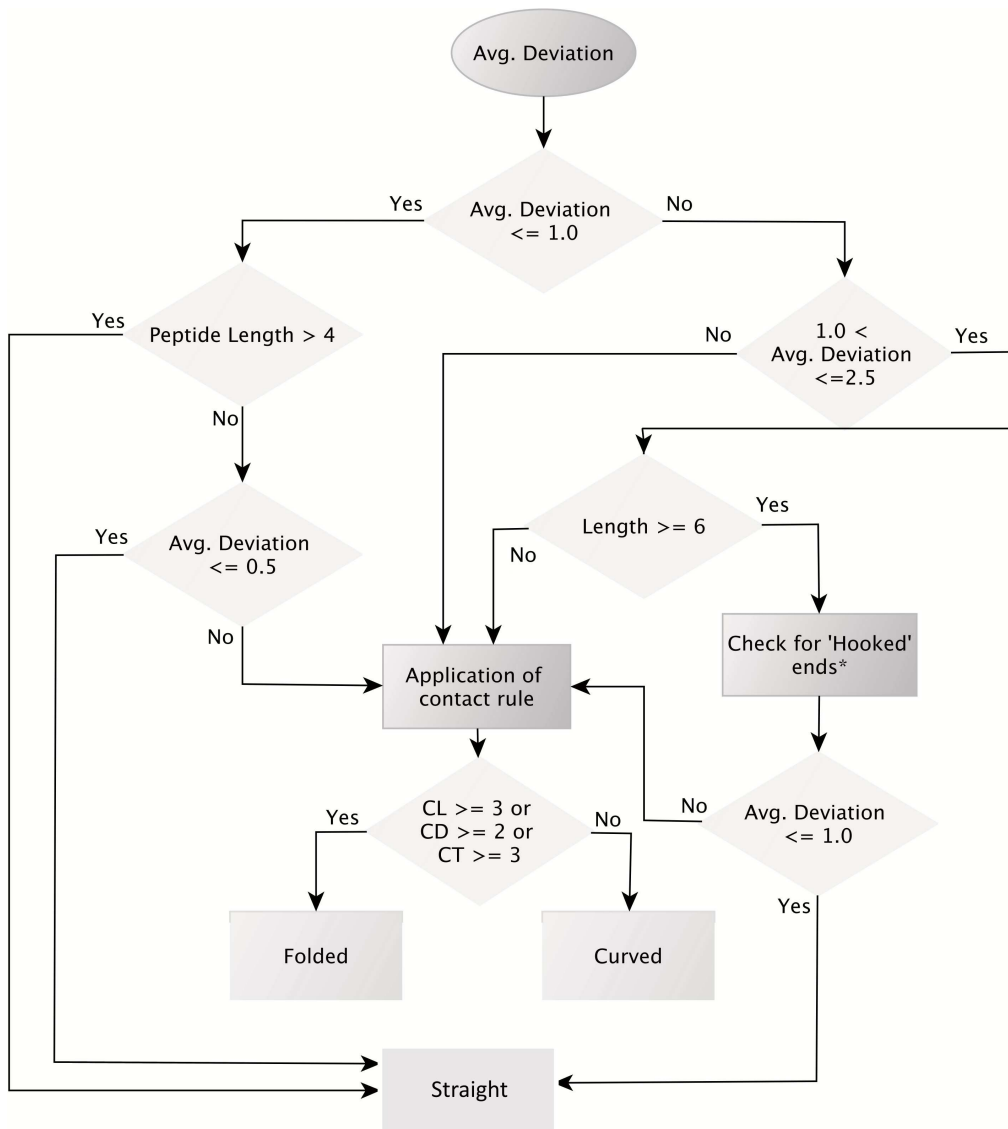


Figure S5: Flow chart of the peptide shape classification protocol: C_L , C_D and C_T refer to the number of local, distant and total contacts, respectively, between pairs of residues in the peptide. *Hooked peptides are described in ‘Classification Protocol’ in the Materials and Methods.

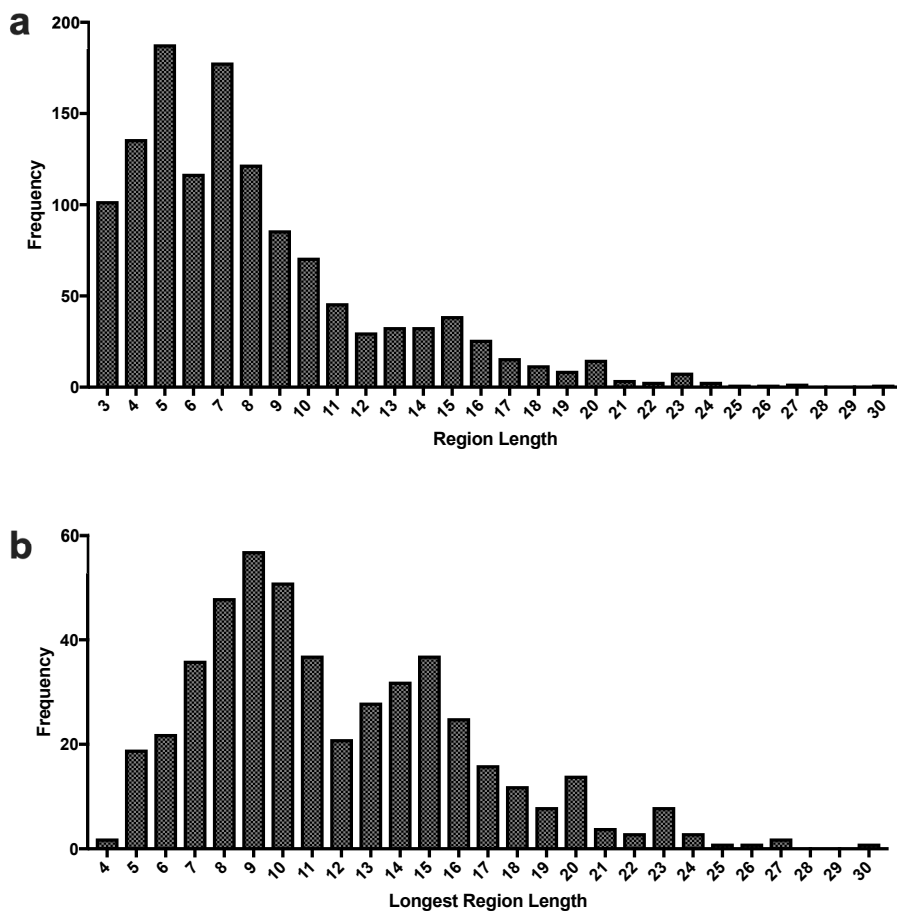


Figure S6: a) The distribution of the lengths of 1282 regions in 488 epitopes. Regions range from 3–30 residues in length. b) The distribution of the length of the longest region in 488 epitopes which ranges from 4–30 residues in length.

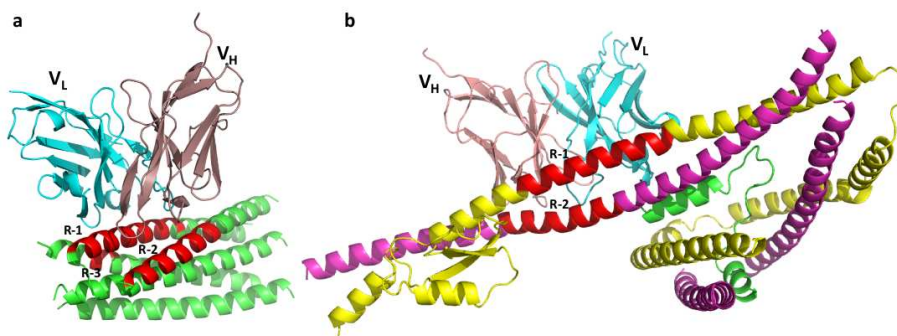


Figure S7: Unusual examples in which regions of 19 residues are accompanied by longer regions. a) An epitope with three regions; R-1 (19 residues), R-2 (20 residues) and R-3 (4 residues) from PDB file 3MA9. b) An epitope with two regions; R-1 (23 residues), R-2 (19 residues) from PDB file 5CWS.

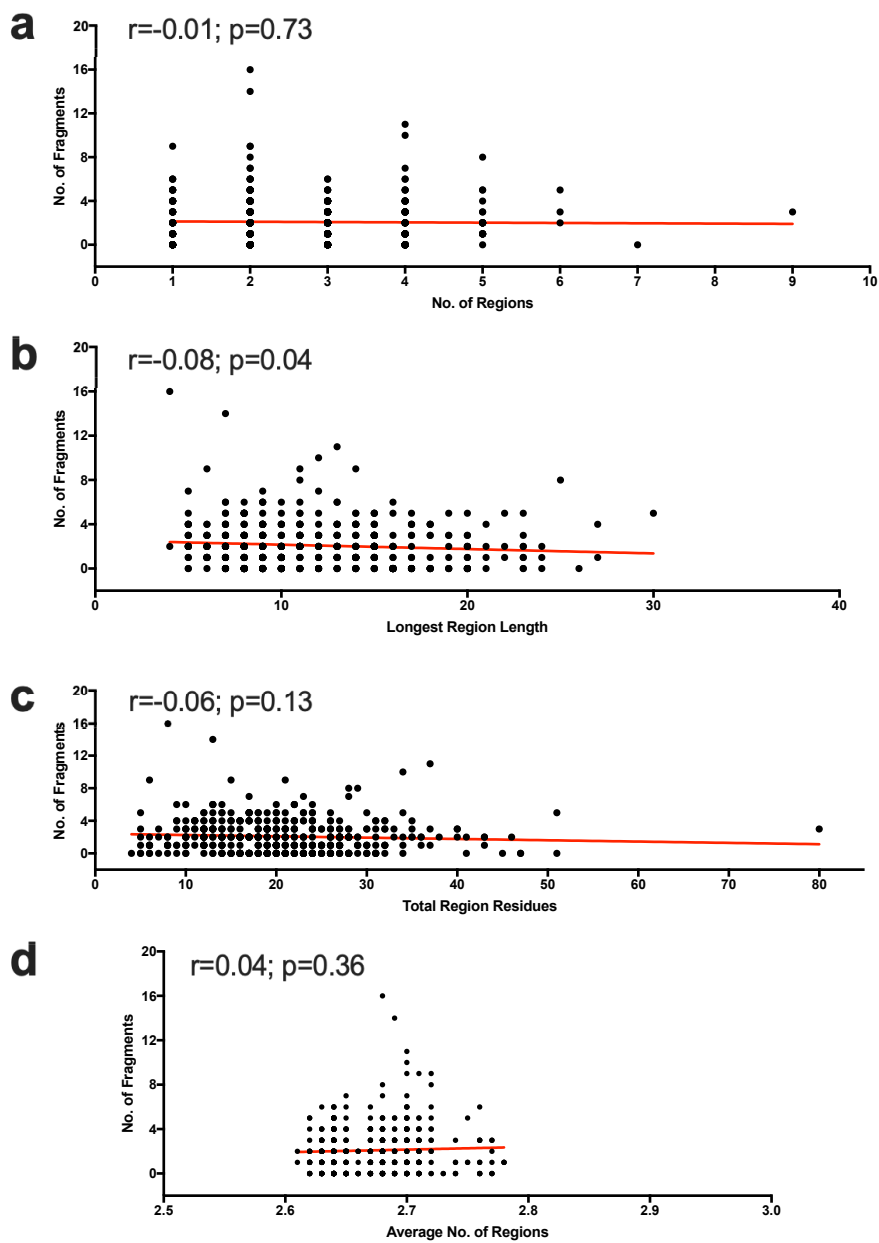
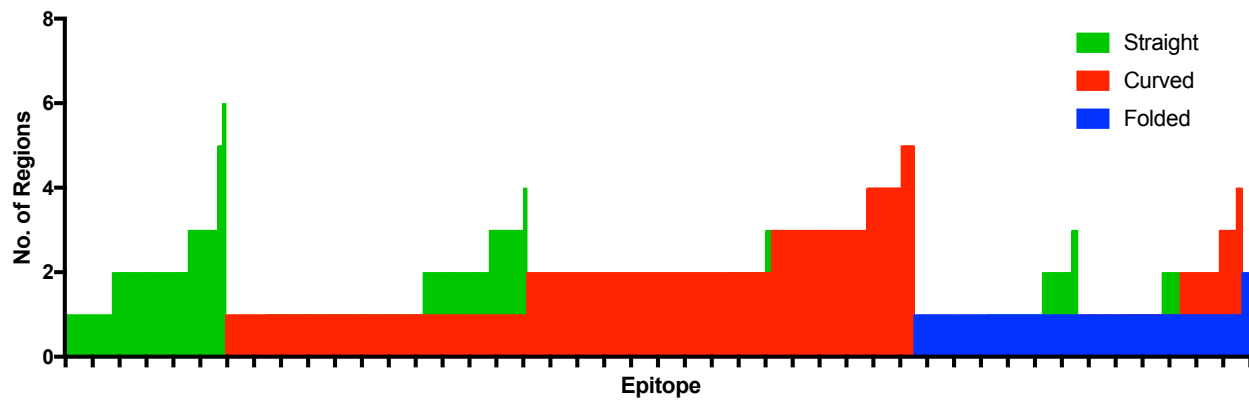


Figure S8: The correlation between the number of fragments and a) the number of regions, b) the longest region, c) the number of residues in regions and d) the average number of regions in an epitope in the single-chain dataset. No statistically significant correlations were observed.



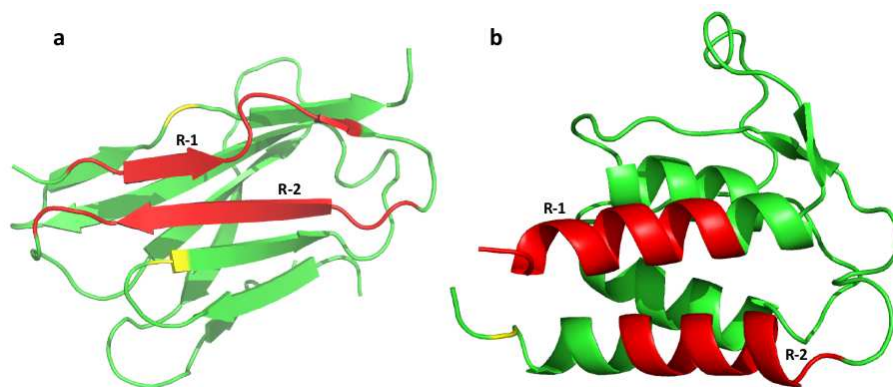
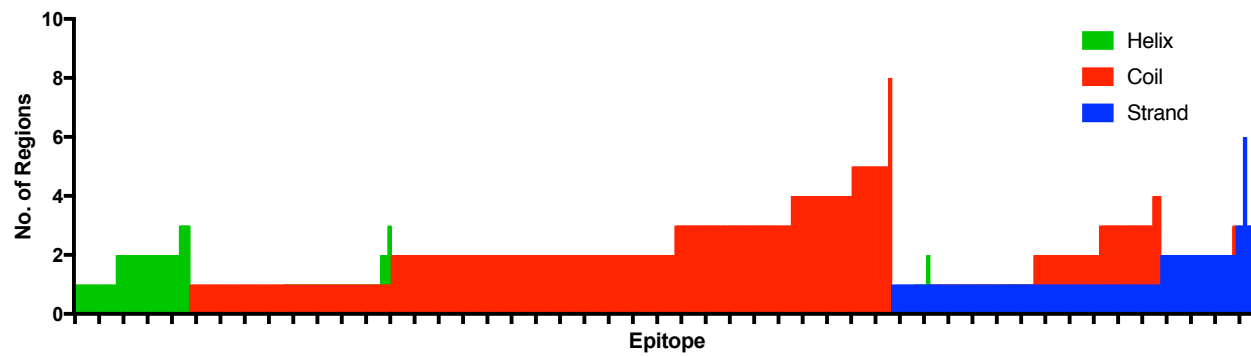


Figure S10: Example epitopes having two straight regions. a) An epitope from PDB code 2ARJ having two straight regions (red) and two fragments (yellow). b) An epitope from PDB code 4JLR having two straight regions (red) and one fragment (yellow).



Supplementary Tables

Table S1: Distribution of regions (R) and fragments (F) in the complete dataset containing both single- and multiple-chain epitopes.

| | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | Total |
|--------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|--------------|
| F0 | 20 | 45 | 34 | 21 | 1 | 0 | 1 | 0 | 0 | 122 |
| F1 | 13 | 33 | 24 | 17 | 7 | 0 | 0 | 0 | 0 | 94 |
| F2 | 10 | 37 | 40 | 9 | 9 | 1 | 0 | 0 | 0 | 106 |
| F3 | 7 | 25 | 21 | 10 | 2 | 1 | 0 | 0 | 1 | 67 |
| F4 | 8 | 22 | 13 | 6 | 1 | 0 | 0 | 0 | 0 | 50 |
| F5 | 5 | 11 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 26 |
| F6 | 3 | 4 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 11 |
| F7 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 |
| F8 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| F9 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| F10 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| F11 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| F12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F14 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| F15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F16 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Total | 67 | 184 | 138 | 71 | 23 | 3 | 1 | 0 | 1 | 488 |

Table S2: Distribution of regions (R) and fragments (F) in the single-chain dataset.

| | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | Total |
|--------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|--------------|
| F0 | 20 | 44 | 32 | 20 | 1 | 0 | 1 | 0 | 0 | 118 |
| F1 | 13 | 29 | 21 | 16 | 5 | 0 | 0 | 0 | 0 | 84 |
| F2 | 10 | 37 | 39 | 7 | 9 | 1 | 0 | 0 | 0 | 103 |
| F3 | 6 | 23 | 19 | 8 | 1 | 1 | 0 | 0 | 1 | 59 |
| F4 | 8 | 21 | 11 | 5 | 1 | 0 | 0 | 0 | 0 | 46 |
| F5 | 5 | 8 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 19 |
| F6 | 3 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| F7 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |
| F8 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| F9 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| F10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F14 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| F15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F16 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Total | 66 | 172 | 126 | 59 | 18 | 3 | 1 | 0 | 1 | 446 |

Table S3: Distribution of regions (R) and fragments (F) in the multiple-chain dataset.

| | R1 | R2 | R3 | R4 | R5 | Total |
|--------------|-----------|-----------|-----------|-----------|-----------|--------------|
| F0 | 0 | 1 | 2 | 1 | 0 | 4 |
| F1 | 0 | 4 | 3 | 1 | 2 | 10 |
| F2 | 0 | 0 | 1 | 2 | 0 | 3 |
| F3 | 1 | 2 | 2 | 2 | 1 | 8 |
| F4 | 0 | 1 | 2 | 1 | 0 | 4 |
| F5 | 0 | 3 | 2 | 1 | 1 | 7 |
| F6 | 0 | 0 | 0 | 2 | 0 | 2 |
| F7 | 0 | 1 | 0 | 0 | 0 | 1 |
| F8 | 0 | 0 | 0 | 0 | 1 | 1 |
| F9 | 0 | 0 | 0 | 0 | 0 | 0 |
| F10 | 0 | 0 | 0 | 1 | 0 | 1 |
| F11 | 0 | 0 | 0 | 1 | 0 | 1 |
| Total | 1 | 12 | 12 | 12 | 5 | 42 |

Table S4: Data grouped from the single- and multiple-chain datasets (as shown in Supplementary Tables S2 and S3) to satisfy the requirements of a χ^2 test to compare the two datasets.

| | | Single | Multiple |
|-------|--------|--------|----------|
| R1 | F0-F1 | 33 | 0 |
| R2 | F0-F1 | 73 | 5 |
| R3-R9 | F0-F1 | 96 | 9 |
| R1 | F2-F16 | 33 | 1 |
| R2 | F2-F16 | 99 | 7 |
| R3-R9 | F2-F16 | 112 | 20 |

Table S5: Distribution of regions (R) and fragments (F) in the single-chain dataset when R1 is removed

| | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | Total |
|--------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|--------------|
| F0 | 44 | 32 | 20 | 1 | 0 | 1 | 0 | 0 | 98 |
| F1 | 29 | 21 | 16 | 5 | 0 | 0 | 0 | 0 | 71 |
| F2 | 37 | 39 | 7 | 9 | 1 | 0 | 0 | 0 | 93 |
| F3 | 23 | 19 | 8 | 1 | 1 | 0 | 0 | 1 | 53 |
| F4 | 21 | 11 | 5 | 1 | 0 | 0 | 0 | 0 | 38 |
| F5 | 8 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 14 |
| F6 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| F7 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |
| F8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| F9 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| F10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F14 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| F15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F16 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Total | 172 | 126 | 59 | 18 | 3 | 1 | 0 | 1 | 380 |

Table S6: Distribution of regions (R) and fragments (F) in the multi-chain dataset when R1 is removed

| | R2 | R3 | R4 | R5 | Total |
|--------------|-----------|-----------|-----------|-----------|--------------|
| F0 | 1 | 2 | 1 | 0 | 4 |
| F1 | 4 | 3 | 1 | 2 | 10 |
| F2 | 0 | 1 | 2 | 0 | 3 |
| F3 | 2 | 2 | 2 | 1 | 7 |
| F4 | 1 | 2 | 1 | 0 | 4 |
| F5 | 3 | 2 | 1 | 1 | 7 |
| F6 | 0 | 0 | 2 | 0 | 2 |
| F7 | 1 | 0 | 0 | 0 | 1 |
| F8 | 0 | 0 | 0 | 1 | 1 |
| F9 | 0 | 0 | 0 | 0 | 0 |
| F10 | 0 | 0 | 1 | 0 | 1 |
| F11 | 0 | 0 | 1 | 0 | 1 |
| Total | 12 | 12 | 12 | 5 | 41 |

Table S7: Data grouped from the single- and multiple-chain datasets (as shown in Supplementary Tables S5 and S6) to satisfy the requirements of a χ^2 test to compare the two datasets.

| | Single | Multiple |
|--------------------|---------------|-----------------|
| R2_F0F1 | 73 | 5 |
| R3-R9_F0F1 | 96 | 9 |
| R2_F2F16 | 99 | 7 |
| R3-R9_F2F16 | 112 | 20 |

Table S8: Frequency of combinations of folded (f), curved (c) and straight (s) regions in the dataset of 488 epitopes. f0–f2 represents epitopes containing 0–2 folded regions, c0–c5 represents 0–5 curved regions and s0–s6 represents 0–6 straight regions. This represents a three-dimensional contingency table containing $3 \times 6 \times 7$ cells.

| Folded | Curved | Straight | Count | Folded | Curved | Straight | Count | Folded | Curved | Straight | Count | Folded | Curved | Straight | Count |
|--------|--------|----------|-------|--------|--------|----------|-------|--------|--------|----------|-------|--------|--------|----------|-------|
| f0 | c0 | s0 | 0 | f0 | c4 | s4 | 0 | f1 | c3 | s1 | 3 | f2 | c1 | s5 | 0 |
| f0 | c0 | s1 | 20 | f0 | c4 | s5 | 0 | f1 | c3 | s2 | 1 | f2 | c1 | s6 | 0 |
| f0 | c0 | s2 | 31 | f0 | c4 | s6 | 0 | f1 | c3 | s3 | 0 | f2 | c2 | s0 | 1 |
| f0 | c0 | s3 | 12 | f0 | c5 | s0 | 4 | f1 | c3 | s4 | s4 | f2 | c2 | s1 | 0 |
| f0 | c0 | s4 | 0 | f0 | c5 | s1 | 0 | f1 | c3 | s5 | 0 | f2 | c2 | s2 | 0 |
| f0 | c0 | s5 | 2 | f0 | c5 | s2 | 0 | f1 | c3 | s6 | 0 | f2 | c2 | s3 | 0 |
| f0 | c0 | s6 | 1 | f0 | c5 | s3 | 0 | f1 | c4 | s0 | 1 | f2 | c2 | s4 | 0 |
| f0 | c1 | s0 | 17 | f0 | c5 | s4 | 1 | f1 | c4 | s1 | 0 | f2 | c2 | s5 | 0 |
| f0 | c1 | s1 | 64 | f0 | c5 | s5 | 0 | f1 | c4 | s2 | 1 | f2 | c2 | s6 | 0 |
| f0 | c1 | s2 | 27 | f0 | c5 | s6 | 0 | f1 | c4 | s3 | 0 | f2 | c3 | s0 | 0 |
| f0 | c1 | s3 | 14 | f1 | c0 | s0 | 30 | f1 | c4 | s4 | 0 | f2 | c3 | s1 | 0 |
| f0 | c1 | s4 | 1 | f1 | c0 | s1 | 23 | f1 | c4 | s5 | 0 | f2 | c3 | s2 | 0 |
| f0 | c1 | s5 | 0 | f1 | c0 | s2 | 12 | f1 | c4 | s6 | 0 | f2 | c3 | s3 | 0 |
| f0 | c1 | s6 | 0 | f1 | c0 | s3 | 2 | f1 | c5 | s0 | 0 | f2 | c3 | s4 | 0 |
| f0 | c2 | s0 | 40 | f1 | c0 | s4 | 0 | f1 | c5 | s1 | 0 | f2 | c3 | s5 | 0 |
| f0 | c2 | s1 | 43 | f1 | c0 | s5 | 0 | f1 | c5 | s2 | 0 | f2 | c3 | s6 | 0 |
| f0 | c2 | s2 | 15 | f1 | c0 | s6 | 0 | f1 | c5 | s3 | 0 | f2 | c4 | s0 | 0 |
| f0 | c2 | s3 | 2 | f1 | c1 | s0 | 21 | f1 | c5 | s4 | 0 | f2 | c4 | s1 | 0 |
| f0 | c2 | s4 | 0 | f1 | c1 | s1 | 14 | f1 | c5 | s5 | 0 | f2 | c4 | s2 | 0 |
| f0 | c2 | s5 | 0 | f1 | c1 | s2 | 7 | f1 | c5 | s6 | 0 | f2 | c4 | s3 | 0 |
| f0 | c2 | s6 | 0 | f1 | c1 | s3 | 0 | f2 | c0 | s0 | 5 | f2 | c4 | s4 | 0 |
| f0 | c3 | s0 | 18 | f1 | c1 | s4 | 0 | f2 | c0 | s1 | 1 | f2 | c4 | s5 | 0 |
| f0 | c3 | s1 | 17 | f1 | c1 | s5 | 0 | f2 | c0 | s2 | 0 | f2 | c4 | s6 | 0 |
| f0 | c3 | s2 | 4 | f1 | c1 | s6 | 0 | f2 | c0 | s3 | 0 | f2 | c5 | s0 | 0 |
| f0 | c3 | s3 | 0 | f1 | c2 | s0 | 11 | f2 | c0 | s4 | 0 | f2 | c5 | s1 | 0 |
| f0 | c3 | s4 | 0 | f1 | c2 | s1 | 4 | f2 | c0 | s5 | 0 | f2 | c5 | s2 | 0 |
| f0 | c3 | s5 | 0 | f1 | c2 | s2 | 1 | f2 | c0 | s6 | 0 | f2 | c5 | s3 | 0 |
| f0 | c3 | s6 | 0 | f1 | c2 | s3 | 0 | f2 | c1 | s0 | 0 | f2 | c5 | s4 | 0 |
| f0 | c4 | s0 | 8 | f1 | c2 | s4 | 0 | f2 | c1 | s1 | 0 | f2 | c5 | s5 | 0 |
| f0 | c4 | s1 | 5 | f1 | c2 | s5 | 0 | f2 | c1 | s2 | 0 | f2 | c5 | s6 | 0 |
| f0 | c4 | s2 | 1 | f1 | c2 | s6 | 0 | f2 | c1 | s3 | 0 | | | | |
| f0 | c4 | s3 | 0 | f1 | c3 | s0 | 3 | f2 | c1 | s4 | 0 | | | | |

Table S9: Grouped data from the three-dimensional contingency table of the distribution of region shapes (folded (f), curved (c) and straight (s)) in the dataset of 488 epitopes (Supplementary Table S8). Grouping was performed to satisfy the requirements of a χ^2 test. The significance of individual under- or over-represented combinations of folded, curved and straight was calculated using a 2x2x2 χ^2 test and a Bonferroni correction for multiple testing was applied to the resultant p -values.

| Folded | Curved | Straight | Observed | Expected | Corrected p -value |
|--------|--------|----------|----------|----------|-------------------------|
| - | | | | | |
| f0 | c0 | s1 | 20 | 39.9 | 2.07×10^{-14} |
| f0 | c0 | s2 | 31 | 19.6 | 0 |
| f0 | c0 | s3-s6 | 15 | 7.1 | 0 |
| f0 | c1 | s0 | 17 | 37.5 | 1.85×10^{-09} |
| f0 | c1 | s1 | 64 | 46.5 | 4.77×10^{-03} |
| f0 | c1 | s2 | 24 | 22.9 | 1.71 |
| f0 | c1 | s3-s6 | 15 | 8.3 | 0.00693 |
| f0 | c2 | s0 | 40 | 26.9 | 3.06×10^{-13} |
| f0 | c2 | s1 | 42 | 33.3 | 1.47×10^{-04} |
| f0 | c2 | s2 | 15 | 16.4 | 1.26×10^{-04} |
| f0 | c2 | s3-s6 | 2 | 5.9 | 3.71×10^{-08} |
| f0 | c3-c5 | s0 | 30 | 16.4 | 6.88×10^{-14} |
| f0 | c3-c5 | s1 | 26 | 20.4 | 0.00124 |
| f0 | c3-c5 | s2 | 5 | 10 | 4.71×10^{-05} |
| f0 | c3-c5 | s3-s6 | 1 | 3.6 | 3.40×10^{-06} |
| f1-f2 | c0 | s0 | 35 | 13.1 | 0 |
| f1-f2 | c0 | s1 | 24 | 16.2 | 2.07×10^{-14} |
| f1-f2 | c0 | s2 | 12 | 8 | 0 |
| f1-f2 | c0 | s3-s6 | 2 | 2.9 | 0 |
| f1-f2 | c1 | s0 | 21 | 15.3 | 1.85×10^{-09} |
| f1-f2 | c1 | s1 | 14 | 18.9 | 0.00477 |
| f1-f2 | c1 | s2 | 7 | 9.3 | 1.71 |
| f1-f2 | c1 | s3-s6 | 0 | 3.4 | 6.93×10^{-03} |
| f1-f2 | c2 | s0 | 12 | 10.9 | 3.06×10^{-13} |
| f1-f2 | c2 | s1 | 4 | 13.5 | 1.47×10^{-04} |
| f1-f2 | c2 | s2 | 1 | 6.7 | 1.26×10^{-04} |
| f1-f2 | c2 | s3-s6 | 0 | 2.4 | 3.71×10^{-08} |
| f1-f2 | c3-c5 | s0 | 4 | 6.7 | 6.88×10^{-14} |
| f1-f2 | c3-c5 | s1 | 3 | 8.3 | 0.00124 |
| f1-f2 | c3-c5 | s2 | 2 | 4.1 | 4.71×10^{-05} |
| f1-f2 | c3-c5 | s3-s6 | 0 | 1.5 | 3.40×10^{-06} |

Table S10: Frequency of combinations of α -helix (H), β -strand (E) and coil (C) regions in the dataset of 488 epitopes. H0–H3 represents epitopes containing 0–3 α -helical regions, E0–E6 represents 0–6 β -strand regions and C0–C8 represents 0–8 coil regions. This represents a three-dimensional contingency table containing $4 \times 7 \times 9$ cells.

| Helix | Strand | Coil | Count | Helix | Strand | Coil | Count | Helix | Strand | Coil | Count |
|-------|--------|------|-------|-------|--------|------|-------|-------|--------|------|-------|
| H0 | E0 | C0 | 0 | H0 | E4 | C6 | 0 | H1 | E2 | C3 | 0 |
| H0 | E0 | C1 | 39 | H0 | E4 | C7 | 0 | H1 | E2 | C4 | 0 |
| H0 | E0 | C2 | 78 | H0 | E4 | C8 | 0 | H1 | E2 | C5 | 0 |
| H0 | E0 | C3 | 39 | H0 | E5 | C0 | 0 | H1 | E2 | C6 | 0 |
| H0 | E0 | C4 | 21 | H0 | E5 | C1 | 0 | H1 | E2 | C7 | 0 |
| H0 | E0 | C5 | 14 | H0 | E5 | C2 | 0 | H1 | E2 | C8 | 0 |
| H0 | E0 | C6 | 0 | H0 | E5 | C3 | 0 | H1 | E3 | C0 | 0 |
| H0 | E0 | C7 | 0 | H0 | E5 | C4 | 0 | H1 | E3 | C1 | 0 |
| H0 | E0 | C8 | 0 | H0 | E5 | C5 | 0 | H1 | E3 | C2 | 0 |
| H0 | E1 | C0 | 10 | H0 | E5 | C6 | 0 | H1 | E3 | C3 | 0 |
| H0 | E1 | C1 | 30 | H0 | E5 | C7 | 0 | H1 | E3 | C4 | 0 |
| H0 | E1 | C2 | 23 | H0 | E5 | C8 | 0 | H1 | E3 | C5 | 0 |
| H0 | E1 | C3 | 21 | H0 | E6 | C0 | 0 | H1 | E3 | C6 | 0 |
| H0 | E1 | C4 | 2 | H0 | E6 | C1 | 0 | H1 | E3 | C7 | 0 |
| H0 | E1 | C5 | 0 | H0 | E6 | C2 | 0 | H1 | E3 | C8 | 0 |
| H0 | E1 | C6 | 0 | H0 | E6 | C3 | 0 | H1 | E4 | C0 | 0 |
| H0 | E1 | C7 | 0 | H0 | E6 | C4 | 0 | H1 | E4 | C1 | 0 |
| H0 | E1 | C8 | 0 | H0 | E6 | C5 | 0 | H1 | E4 | C2 | 0 |
| H0 | E2 | C0 | 5 | H0 | E6 | C6 | 0 | H1 | E4 | C3 | 0 |
| H0 | E2 | C1 | 14 | H0 | E6 | C7 | 0 | H1 | E4 | C4 | 0 |
| H0 | E2 | C2 | 8 | H0 | E6 | C8 | 0 | H1 | E4 | C5 | 0 |
| H0 | E2 | C3 | 1 | H1 | E0 | C0 | 19 | H1 | E4 | C6 | 0 |
| H0 | E2 | C4 | 0 | H1 | E0 | C1 | 40 | H1 | E4 | C7 | 0 |
| H0 | E2 | C5 | 0 | H1 | E0 | C2 | 35 | H1 | E4 | C8 | 0 |
| H0 | E2 | C6 | 0 | H1 | E0 | C3 | 8 | H1 | E5 | C0 | 0 |
| H0 | E2 | C7 | 0 | H1 | E0 | C4 | 3 | H1 | E5 | C1 | 0 |
| H0 | E2 | C8 | 0 | H1 | E0 | C5 | 1 | H1 | E5 | C2 | 0 |
| H0 | E3 | C0 | 4 | H1 | E0 | C6 | 0 | H1 | E5 | C3 | 0 |
| H0 | E3 | C1 | 4 | H1 | E0 | C7 | 0 | H1 | E5 | C4 | 0 |
| H0 | E3 | C2 | 1 | H1 | E0 | C8 | 1 | H1 | E5 | C5 | 0 |
| H0 | E3 | C3 | 0 | H1 | E1 | C0 | 5 | H1 | E5 | C6 | 0 |
| H0 | E3 | C4 | 0 | H1 | E1 | C1 | 13 | H1 | E5 | C7 | 0 |
| H0 | E3 | C5 | 0 | H1 | E1 | C2 | 4 | H1 | E5 | C8 | 0 |
| H0 | E3 | C6 | 0 | H1 | E1 | C3 | 1 | H1 | E6 | C0 | 0 |
| H0 | E3 | C7 | 0 | H1 | E1 | C4 | 0 | H1 | E6 | C1 | 0 |

Continued on next page...

Table S10 – *Continued from previous page*

| Helix | Strand | Coil | Count | Helix | Strand | Coil | Count | Helix | Strand | Coil | Count |
|-------|--------|------|-------|-------|--------|------|-------|-------|--------|------|-------|
| H0 | E3 | C8 | 0 | H1 | E1 | C5 | 0 | H1 | E6 | C2 | 0 |
| H0 | E4 | C0 | 0 | H1 | E1 | C6 | 0 | H1 | E6 | C3 | 0 |
| H0 | E4 | C1 | 0 | H1 | E1 | C7 | 0 | H1 | E6 | C4 | 0 |
| H0 | E4 | C2 | 0 | H1 | E1 | C8 | 0 | H1 | E6 | C5 | 0 |
| H0 | E4 | C3 | 0 | H1 | E2 | C0 | 3 | H1 | E6 | C6 | 0 |
| H0 | E4 | C4 | 0 | H1 | E2 | C1 | 0 | H1 | E6 | C7 | 0 |
| H0 | E4 | C5 | 0 | H1 | E2 | C2 | 0 | H1 | E6 | C8 | 0 |
| H2 | E0 | C0 | 26 | H2 | E4 | C6 | 0 | H3 | E2 | C3 | 0 |
| H2 | E0 | C1 | 3 | H2 | E4 | C7 | 0 | H3 | E2 | C4 | 0 |
| H2 | E0 | C2 | 4 | H2 | E4 | C8 | 0 | H3 | E2 | C5 | 0 |
| H2 | E0 | C3 | 1 | H2 | E5 | C0 | 0 | H3 | E2 | C6 | 0 |
| H2 | E0 | C4 | 1 | H2 | E5 | C1 | 0 | H3 | E2 | C7 | 0 |
| H2 | E0 | C5 | 0 | H2 | E5 | C2 | 0 | H3 | E2 | C8 | 0 |
| H2 | E0 | C6 | 0 | H2 | E5 | C3 | 0 | H3 | E3 | C0 | 0 |
| H2 | E0 | C7 | 0 | H2 | E5 | C4 | 0 | H3 | E3 | C1 | 0 |
| H2 | E0 | C8 | 0 | H2 | E5 | C5 | 0 | H3 | E3 | C2 | 0 |
| H2 | E1 | C0 | 1 | H2 | E5 | C6 | 0 | H3 | E3 | C3 | 0 |
| H2 | E1 | C1 | 0 | H2 | E5 | C7 | 0 | H3 | E3 | C4 | 0 |
| H2 | E1 | C2 | 0 | H2 | E5 | C8 | 0 | H3 | E3 | C5 | 0 |
| H2 | E1 | C3 | 0 | H2 | E6 | C0 | 0 | H3 | E3 | C6 | 0 |
| H2 | E1 | C4 | 1 | H2 | E6 | C1 | 0 | H3 | E3 | C7 | 0 |
| H2 | E1 | C5 | 0 | H2 | E6 | C2 | 0 | H3 | E3 | C8 | 0 |
| H2 | E1 | C6 | 0 | H2 | E6 | C3 | 0 | H3 | E4 | C0 | 0 |
| H2 | E1 | C7 | 0 | H2 | E6 | C4 | 0 | H3 | E4 | C1 | 0 |
| H2 | E1 | C8 | 0 | H2 | E6 | C5 | 0 | H3 | E4 | C2 | 0 |
| H2 | E2 | C0 | 0 | H2 | E6 | C6 | 0 | H3 | E4 | C3 | 0 |
| H2 | E2 | C1 | 0 | H2 | E6 | C7 | 0 | H3 | E4 | C4 | 0 |
| H2 | E2 | C2 | 0 | H2 | E6 | C8 | 0 | H3 | E4 | C5 | 0 |
| H2 | E2 | C3 | 0 | H3 | E0 | C0 | 4 | H3 | E4 | C6 | 0 |
| H2 | E2 | C4 | 0 | H3 | E0 | C1 | 1 | H3 | E4 | C7 | 0 |
| H2 | E2 | C5 | 0 | H3 | E0 | C2 | 0 | H3 | E4 | C8 | 0 |
| H2 | E2 | C6 | 0 | H3 | E0 | C3 | 0 | H3 | E5 | C0 | 0 |
| H2 | E2 | C7 | 0 | H3 | E0 | C4 | 0 | H3 | E5 | C1 | 0 |
| H2 | E2 | C8 | 0 | H3 | E0 | C5 | 0 | H3 | E5 | C2 | 0 |
| H2 | E3 | C0 | 0 | H3 | E0 | C6 | 0 | H3 | E5 | C3 | 0 |
| H2 | E3 | C1 | 0 | H3 | E0 | C7 | 0 | H3 | E5 | C4 | 0 |
| H2 | E3 | C2 | 0 | H3 | E0 | C8 | 0 | H3 | E5 | C5 | 0 |
| H2 | E3 | C3 | 0 | H3 | E1 | C0 | 0 | H3 | E5 | C6 | 0 |
| H2 | E3 | C4 | 0 | H3 | E1 | C1 | 0 | H3 | E5 | C7 | 0 |
| H2 | E3 | C5 | 0 | H3 | E1 | C2 | 0 | H3 | E5 | C8 | 0 |

Continued on next page...

Table S10 – *Continued from previous page*

| Helix | Strand | Coil | Count | Helix | Strand | Coil | Count | Helix | Strand | Coil | Count |
|-------|--------|------|-------|-------|--------|------|-------|-------|--------|------|-------|
| H2 | E3 | C6 | 0 | H3 | E1 | C3 | 0 | H3 | E6 | C0 | 0 |
| H2 | E3 | C7 | 0 | H3 | E1 | C4 | 0 | H3 | E6 | C1 | 0 |
| H2 | E3 | C8 | 0 | H3 | E1 | C5 | 0 | H3 | E6 | C2 | 0 |
| H2 | E4 | C0 | 0 | H3 | E1 | C6 | 0 | H3 | E6 | C3 | 0 |
| H2 | E4 | C1 | 0 | H3 | E1 | C7 | 0 | H3 | E6 | C4 | 0 |
| H2 | E4 | C2 | 0 | H3 | E1 | C8 | 0 | H3 | E6 | C5 | 0 |
| H2 | E4 | C3 | 0 | H3 | E2 | C0 | 0 | H3 | E6 | C6 | 0 |
| H2 | E4 | C4 | 0 | H3 | E2 | C1 | 0 | H3 | E6 | C7 | 0 |
| H2 | E4 | C5 | 0 | H3 | E2 | C2 | 0 | H3 | E6 | C8 | 0 |

Table S11: Grouped data from the three-dimensional contingency table of the distribution of region secondary structure (α -helix (H), β -strand (E) and coil (C)) in the dataset of 488 epitopes (Supplementary Table S10). Grouping was performed to satisfy the requirements of a χ^2 test. The significance of individual under- or over-represented combinations of helix, strand and coil was calculated using a 2x2x2 χ^2 test and a Bonferroni correction for multiple testing was applied to the resultant p -values.

| Helix | Strand | Coil | Observed | Expected | Corrected p -value |
|-------|--------|-------|----------|----------|-------------------------|
| - | | | | | |
| H0 | E0 | C1 | 39 | 63.5 | 1.47×10^{-09} |
| H0 | E0 | C2 | 78 | 68 | 4.18×10^{-09} |
| H0 | E0 | C3-C8 | 74 | 51.1 | 0 |
| H0 | E1 | C0 | 10 | 11.1 | 0 |
| H0 | E1 | C1 | 29 | 20.7 | 6.10×10^{-04} |
| H0 | E1 | C2 | 23 | 22.1 | 1.70×10^{-04} |
| H0 | E1 | C3-C8 | 23 | 16.6 | 3.89×10^{-09} |
| H0 | E2-E6 | C0 | 9 | 4 | 0 |
| H0 | E2-E6 | C1 | 18 | 7.5 | 1.48×10^{-06} |
| H0 | E2-E6 | C2 | 9 | 8 | 2.75×10^{-05} |
| H0 | E2-E6 | C3-C8 | 1 | 6 | 5.11×10^{-15} |
| H1-H3 | E0 | C0 | 49 | 19.1 | 0 |
| H1-H3 | E0 | C1 | 44 | 35.5 | 1.47×10^{-09} |
| H1-H3 | E0 | C2 | 39 | 38 | 4.18×10^{-09} |
| H1-H3 | E0 | C3-C8 | 15 | 28.6 | 0 |
| H1-H3 | E1 | C0 | 6 | 6.2 | 0 |
| H1-H3 | E1 | C1 | 13 | 11.6 | 6.10×10^{-04} |
| H1-H3 | E1 | C2 | 4 | 12.4 | 1.70×10^{-04} |
| H1-H3 | E1 | C3-C8 | 2 | 9.3 | 3.89×10^{-09} |
| H1-H3 | E2-E6 | C0 | 3 | 2.3 | 0 |
| H1-H3 | E2-E6 | C1 | 0 | 4.2 | 1.48×10^{-06} |
| H1-H3 | E2-E6 | C2 | 0 | 4.5 | 2.75×10^{-05} |
| H1-H3 | E2-E6 | C3-C8 | 0 | 3.4 | 5.11×10^{-15} |