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 Cα 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 Cα coordinates onto the best-fit line and **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 Cα positions. **VL** is required to be in the same direction as **VR**. Initially it is defined as $\overrightarrow{P_N P_C}$; if the angle between **VR** and **VL** is >90°, **VL** is redefined as $\overrightarrow{P_C P_N}$.

To do this, a C program, `pdbline` (Martin, A.C.R., Raghavan, A. and Ferdous, S. (2014) "pdbline V1.2", Software) was written which draws a best fit line through a specified set of Cα 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 Cα closest to the midpoint is identified and defined as C_i by calculating the Euclidean distance between the Cα of each amino acid and the mid point using Equation S2.

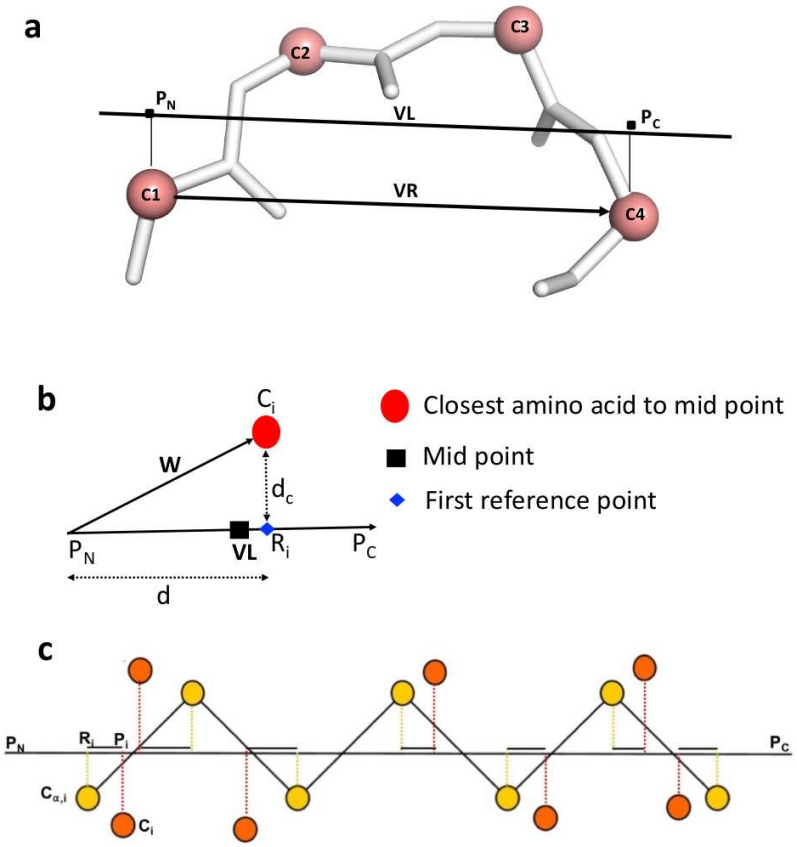

Figure S1: a) The vector **VL**, is defined by projections of the N-terminal and C-terminal Cα atoms (P_N and P_C) onto the best-fit line through the Cα positions (shown as spheres) — in this example through a set of four residues. The region vector $\mathbf{VR} = \overrightarrow{C_N C_C}$ is defined between the N-terminal and C-terminal Cα atoms. The direction of **VL** ($\overrightarrow{P_N P_C}$ or $\overrightarrow{P_C P_N}$) is required to be the same as the direction of **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 **VL**; C_i is the closest Cα to the mid point (M indicated by a square) and separated by the distance d_c. d is a line segment on **VL**. c) Yellow points (Cα_i) represent the Cα positions of an ideal extended peptide with their projections onto **VL** indicated by R_i with a spacing of 3.5 Å. Red points (C_i) represent the actual Cα positions with their projections onto **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 Cα 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 Cα (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 **VL**; C_i is the closest Cα to the mid point (M) separated from **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 **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 **VL** at a distance d from P_N, **VL** was normalised to **U** by using Equation S3:

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

Step 5: The ideal reference points are mapped onto VL. These reference points are the positions at which one would expect Cα atoms (Cα_i) to be projected for an ideal extended 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 **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 **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, **U**.

Step 6: The actual Cα points are mapped. These points (P_i) corresponding to the projection of each Cα (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 **W** corresponds to each Cα projection onto **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:

$$t_{r++} = \sum_{c=1}^{C} \sum_{p=1}^{P} o_{rcp} \tag{S7}$$

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

Similarly for columns and planes:

$$t_{+c+} = \sum_{r=1}^{R} \sum_{p=1}^{P} o_{rcp} \tag{S8}$$

$$t_{++p} = \sum_{r=1}^{R} \sum_{c=1}^{C} o_{rcp} \tag{S9}$$

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

$$e_{rcp} = \frac{t_{r++} \times t_{+c+} \times t_{++p}}{N^2} \tag{S10}$$

(where N is the total number of observations).

The χ^2 value is then calculated as normal:

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

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

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

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.

```
procedure
    Input
    PeptideResidues[]
    Initialization
    C_T ← 0
    C_L ← 0
    C_D ← 0
    len ← length of peptide
    Procedure
    if len <= 12 then
        T_C ← len/2
    else
        T_C ← 5
    end if
    for n ← 0, n < len do
        for i ← 3, i < len − 3 do
            for d ← 0, d < len do
                res1 ← PeptideResidues[n − d]
                res2 ← PeptideResidues[n + i + d]
                if res1 ≥ 0 & res2 < len then
                    CalculateAtomicDistance(res1, res2)
                    if distance ≤ 4.0 then
                        if i ≤ T_C then
                            C_L++
                        else
                            C_D++
                        end if
                    end if
                end if
            end for
        end for
    end for
    C_T ← C_L + C_D
    if C_L ≥ 3‖C_D ≥ 2‖C_T ≥ 3 then
        shape ← folded
    else
        shape ← curved
    end if
end procedure
```

Supplementary Figures

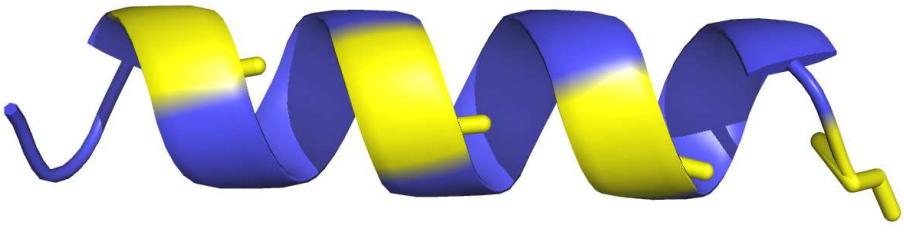

Figure S2: A gap of up to three residue in the epitope region. Amino acid residues which lie on the same face of an α-helix are shown in yellow with a spacing of 3 between them.

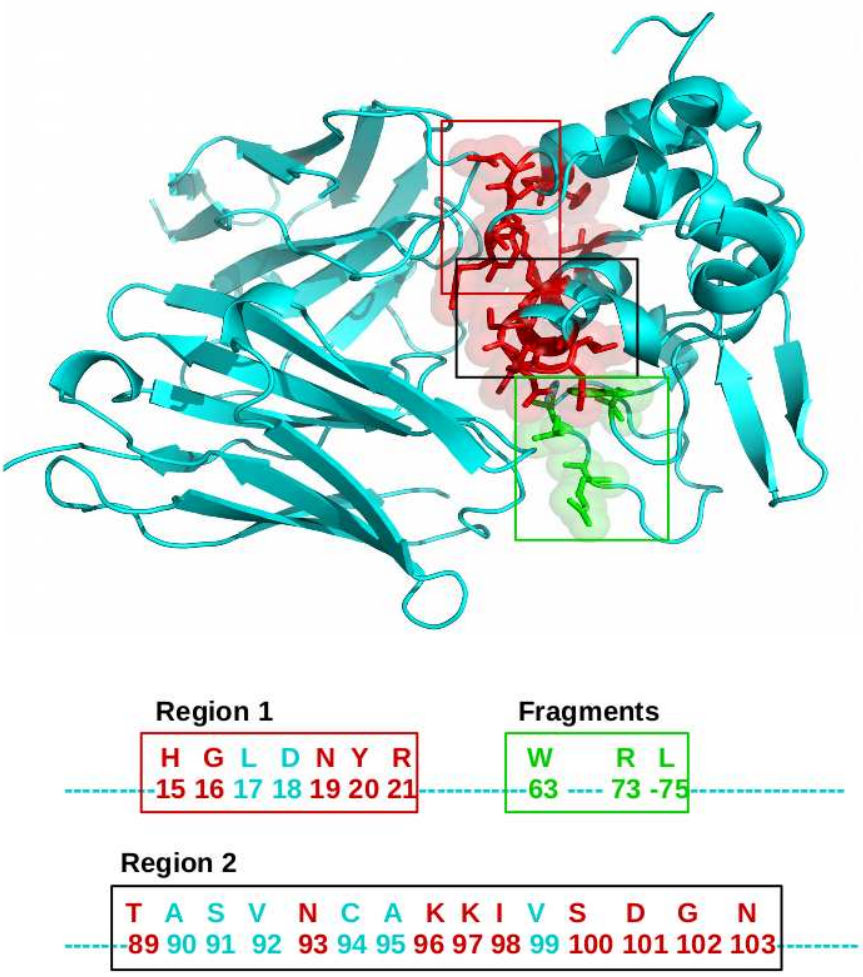

Figure S3: Regions and Fragments — This epitope is comprised of two regions and three fragments. Region 1 has a gap of two non-contacting residues (shown in blue) while Region 2 has three gaps of up to three non-contacting residues. The contacting residues in the epitope are shown in red.

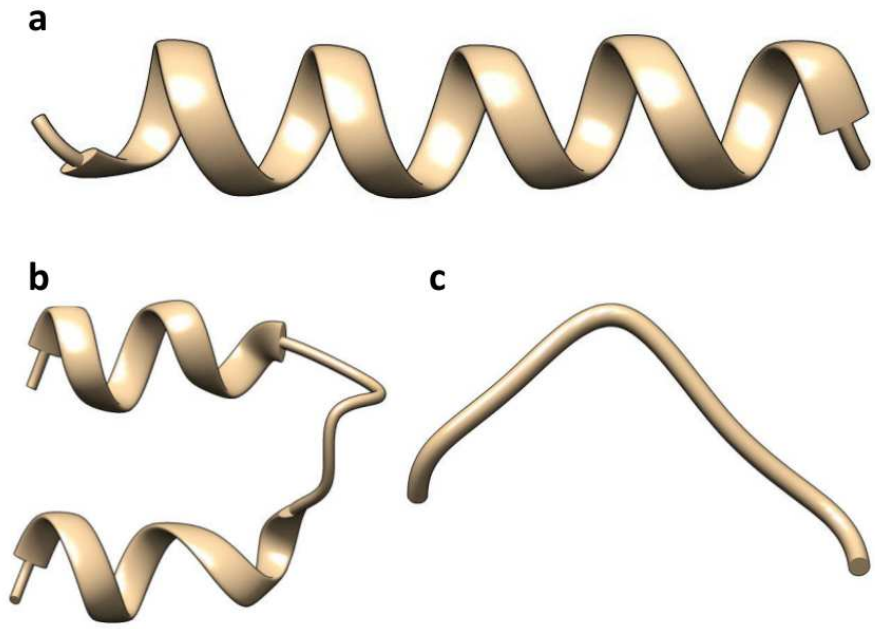

Figure S4: Region shapes: a) Extended, b) Folded, and c) Curved regions.

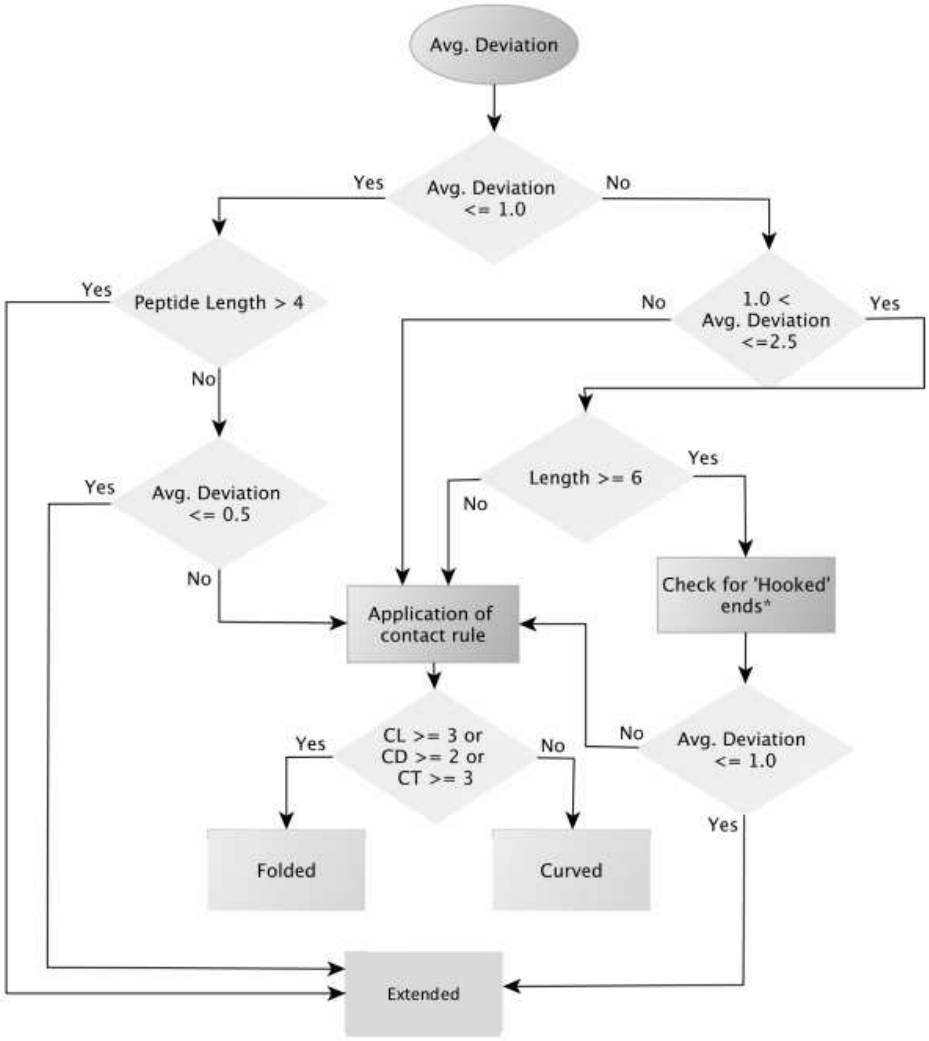

Figure S5: Flow chart of the peptide shape classification protocol: C_L, C_D and C_T refer to the number of local, distant and total contacts, respectively, between pairs of residues in the peptide. *Hooked peptides are described in 'Classification Protocol' in the Materials and Methods.

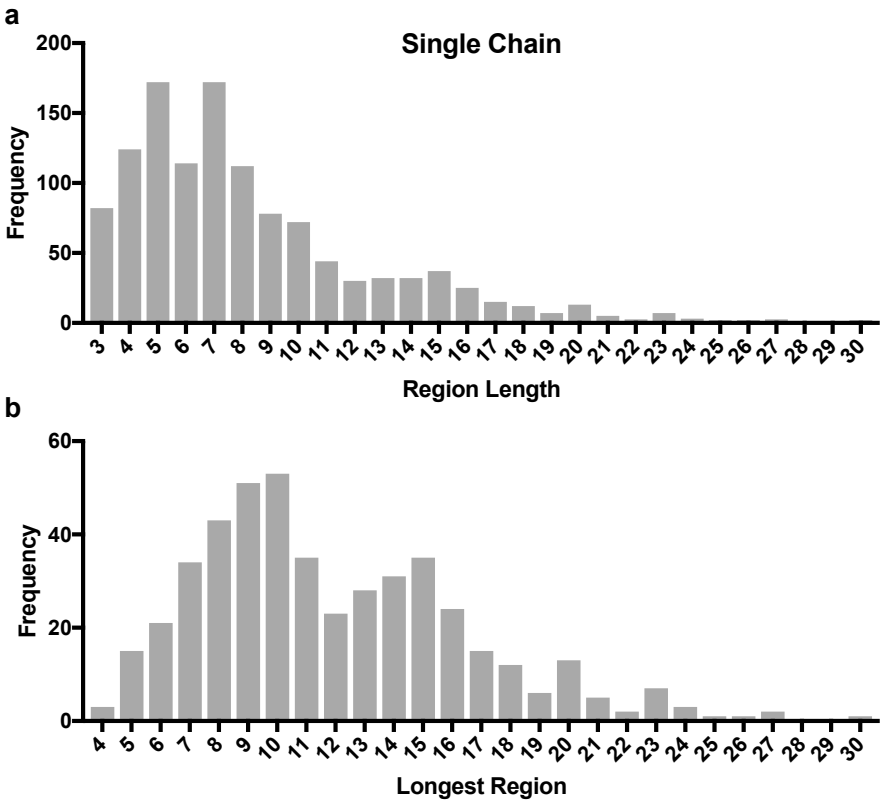

Figure S6: a) The distribution of the lengths of 1195 regions in 464 epitopes (single-chain antigen dataset). Regions range from 3–30 residues in length. b) The distribution of the length of the longest region in 464 epitopes which ranges from 4–30 residues in length.

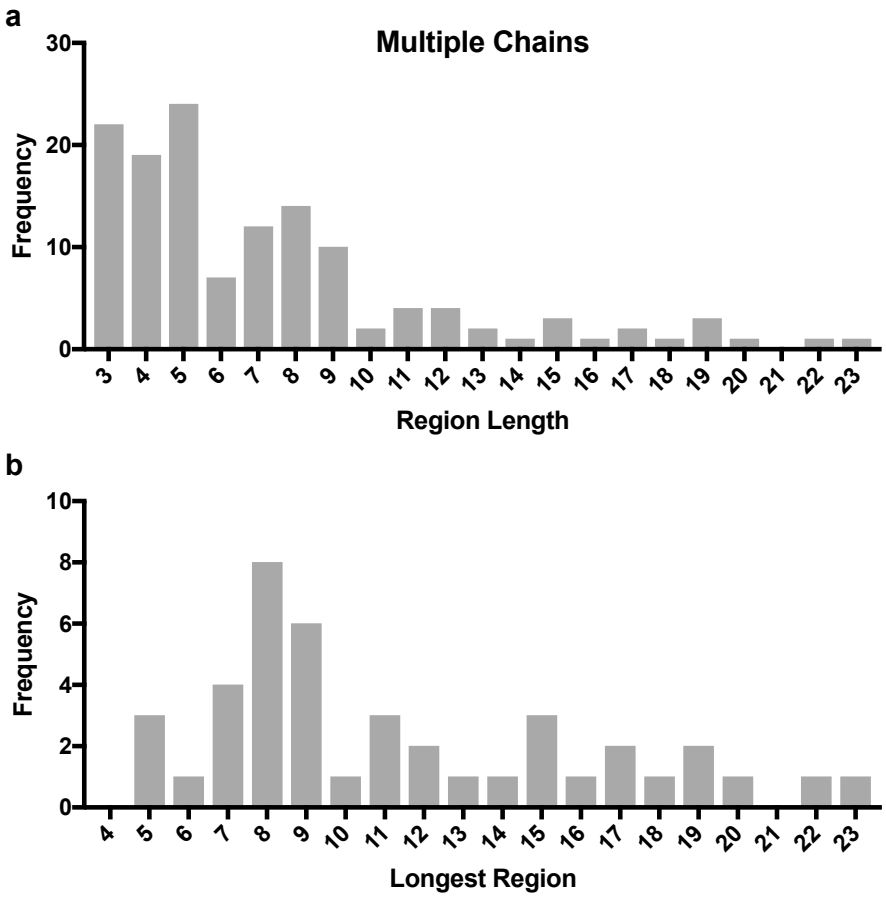

Figure S7: a) The distribution of the lengths of 134 regions in 42 epitopes (multiple-chain antigen dataset). Regions range from 3–23 residues in length. b) The distribution of the length of the longest region in 42 epitopes which ranges from 5–23 residues in length.

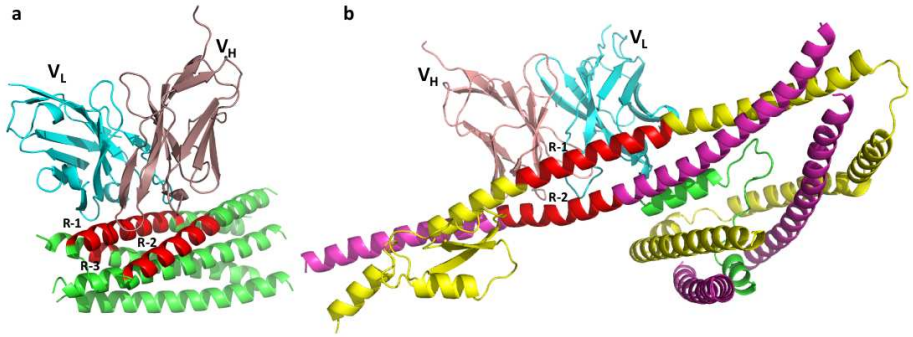

Figure S8: Unusual examples in which regions of 19 residues are accompanied by longer regions. a) An epitope with three regions; R-1 (19 residues), R-2 (20 residues) and R-3 (4 residues) from PDB file 3MA9 in the single-chain dataset. b) An epitope with two regions; R-1 (23 residues), R-2 (19 residues) from PDB file 5CWS in the multiple-chain dataset.

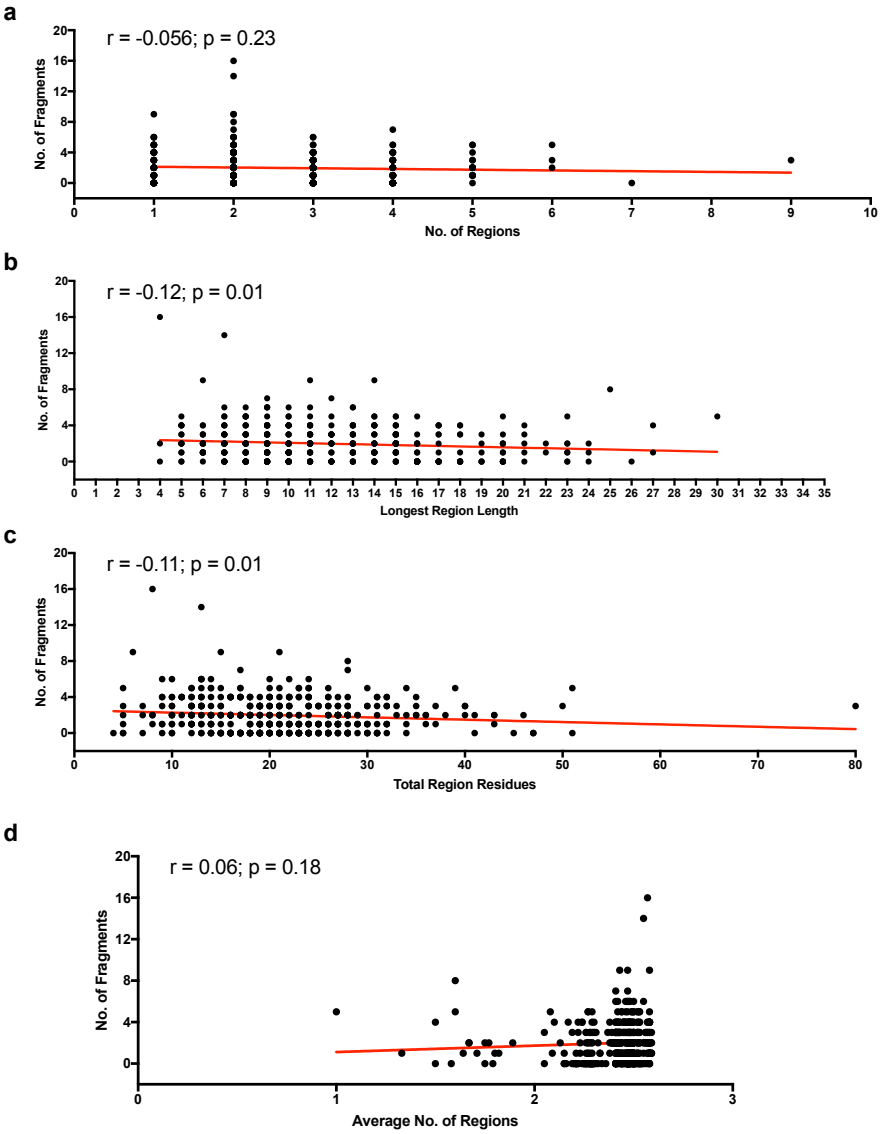

Figure S9: The correlation between the number of fragments and a) the number of regions, b) the longest region, c) the number of residues in regions and d) the average number of regions in an epitope in the single-chain dataset. No statistically significant correlations were observed.

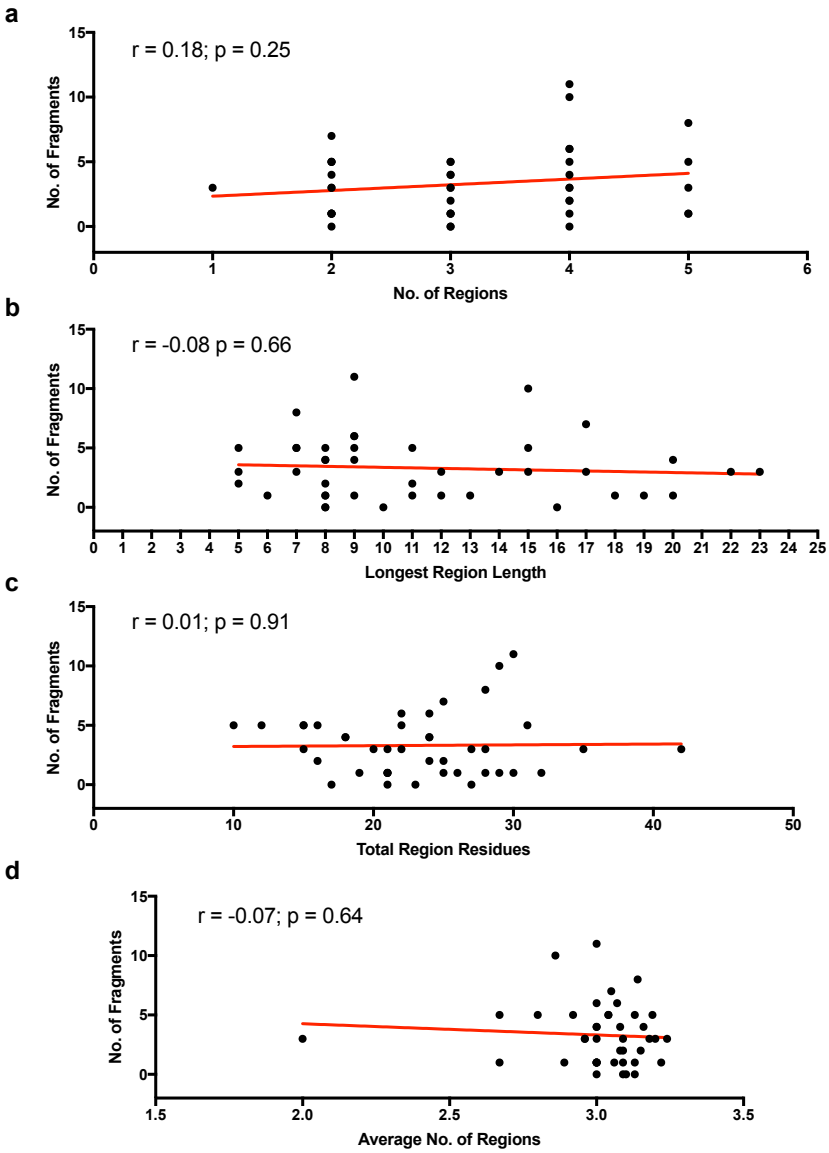

Figure S10: The correlation between the number of fragments and a) the number of regions, b) the longest region, c) the number of residues in regions and d) the average number of regions in an epitope in the multiple-chain dataset. No statistically significant correlations were observed for a) and d) while only very weak correlations were observed for b) and c).

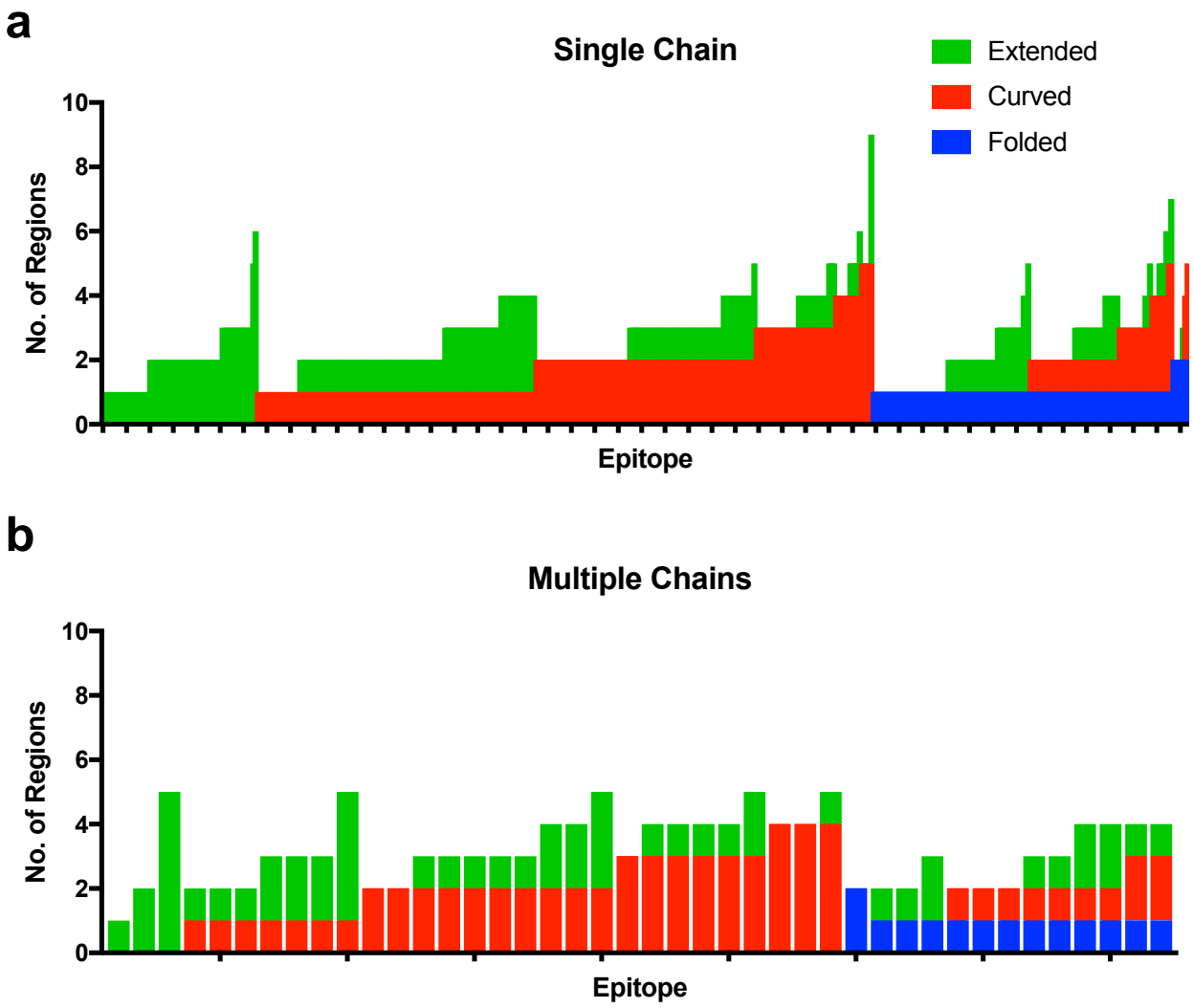

Figure S11: The distribution of extended, curved and folded regions in a) the single-chain (464 unique epitopes) and b) the multiple-chain (42 unique epitopes) epitope datasets.

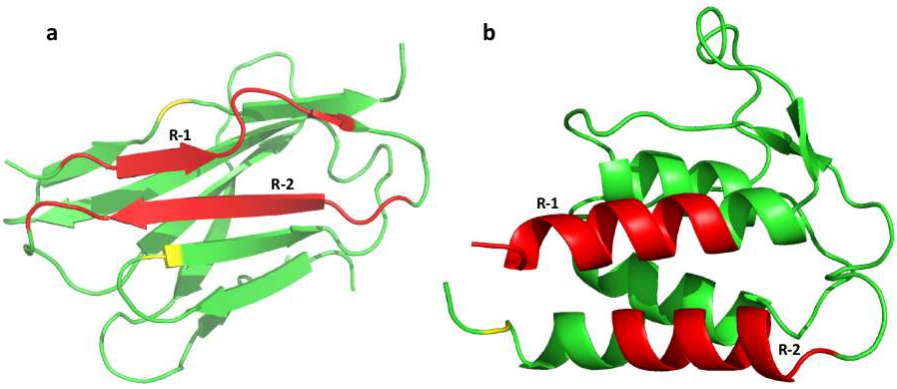

Figure S12: Example epitopes having two extended regions. a) An epitope from PDB code 2ARJ having two extended regions (red) and two fragments (yellow). b) An epitope from PDB code 4JLR having two extended regions (red) and one fragment (yellow).

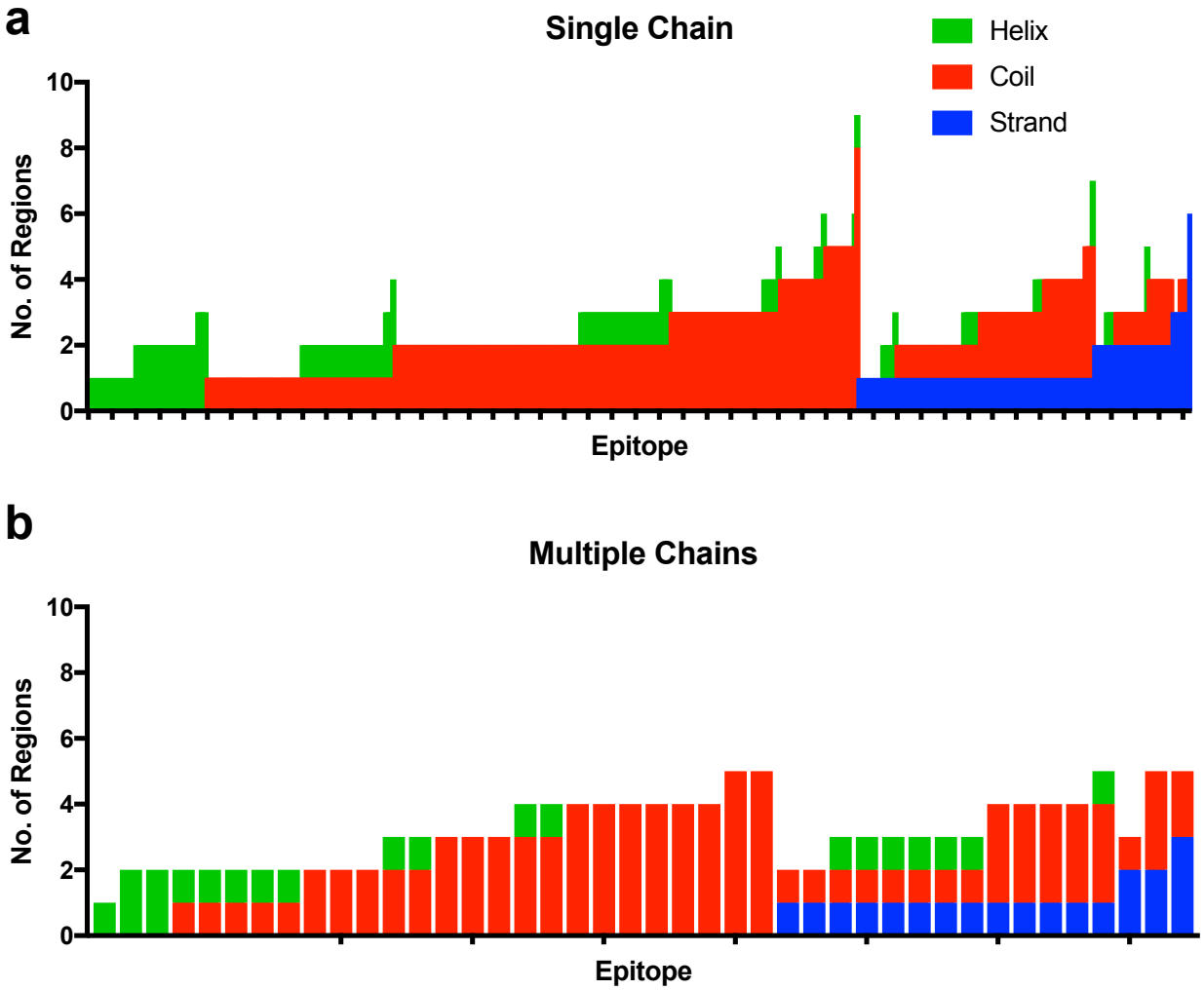

Figure S13: Analysis of secondary structure in epitope regions. The distribution of regions classified as predominantly helix, strand or coil in a) the single-chain dataset (464 unique epitopes) and b) the multiple-chain dataset (42 unique epitopes).

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	23	45	35	21	1	0	1	0	0	126
F1	13	34	24	19	7	0	0	0	0	97
F2	10	39	41	9	9	1	0	0	0	109
F3	7	25	22	10	3	1	0	0	1	69
F4	9	24	15	6	1	0	0	0	0	55
F5	5	11	4	3	3	1	0	0	0	27
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	71	189	143	73	25	3	1	0	1	506

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	23	44	33	20	1	0	1	0	0	122
F1	13	30	21	18	5	0	0	0	0	87
F2	10	39	40	7	9	1	0	0	0	106
F3	6	23	20	8	2	1	0	0	1	61
F4	9	23	13	5	1	0	0	0	0	51
F5	5	8	2	2	2	1	0	0	0	20
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	1	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	70	177	131	61	20	3	1	0	1	464

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	36	0
R2	F0-F1	74	5
R3-R9	F0-F1	99	9
R1	F2-F16	34	1
R2	F2-F16	103	7
R3-R9	F2-F16	118	20

Table S5: Data grouped from the complete dataset (containing both single- and multiple-chain epitopes, as shown in Supplementary Table S1) to satisfy the requirements of a χ^2 test to test whether the number of fragments is dependent on the number of regions. Each cell shows the observed and expected values.

	R0-R1	R2	R3-R9	Total
F0-F1	36/31.73	79/83.74	108/107.53	223
F2-F16	36/40.27	111/106.26	136/136.47	283
Total	72	190	244	506

Table S6: Data grouped from the single-chain dataset shown in Supplementary Table S2 to satisfy the requirements of a χ^2 test. Each cell shows the observed and expected values, the latter being calculated from the observed values in the combined dataset (Supplementary Table S1). The χ^2 test is designed to determine whether the single-chain dataset is representative of the combined set.

	R1	R2	R3-R9	R4-R9	Total
F0-F1	36/31.53	74/79.73	54/59.01	45/38.74	209
F2-F5	30/35.91	93/90.79	75/67.19	40/44.11	238
F6-F16	4/2.56	10/6.48	2/4.80	1/3.15	17
Total	70	177	133	86	464

Table S7: Data grouped from the multiple-chain dataset shown in Supplementary Table S3 to satisfy the requirements of a χ^2 test. Each cell shows the observed and expected values, the latter being calculated from the observed values in the combined dataset (Supplementary Table S1). The χ^2 test is designed to determine whether the multiple-chain dataset is representative of the combined set.

	R1	R2	R3-R9	Total
F0-F1	0/3.26	5/6.70	9/8.96	14
F2-F16	1/3.08	7/9.32	20/10.68	28
Total	1	12	29	42

Table S8: Frequency of combinations of folded (f), curved (c) and extended (e) regions in the single-chain dataset. f0–f2 represents epitopes containing 0–2 folded regions, c0–c5 represents 0–5 curved regions and e0–e6 represents 0–6 extended regions. This represents a three-dimensional contingency table containing $3 \times 6 \times 7$ cells.

Folded	Curved	Extended	Count	Folded	Curved	Extended	Count	Folded	Curved	Extended	Count	Folded	Curved	Extended	Count
f0	c0	e0	0	f0	c4	e4	0	f1	c3	e1	3	f2	c1	e5	0
f0	c0	e1	20	f0	c4	e5	0	f1	c3	e2	1	f2	c1	e6	0
f0	c0	e2	31	f0	c4	e6	0	f1	c3	e3	0	f2	c2	e0	1
f0	c0	e3	13	f0	c5	e0	4	f1	c3	e4	0	f2	c2	e1	0
f0	c0	e4	0	f0	c5	e1	0	f1	c3	e5	0	f2	c2	e2	0
f0	c0	e5	1	f0	c5	e2	0	f1	c3	e6	0	f2	c2	e3	0
f0	c0	e6	1	f0	c5	e3	0	f1	c4	e0	1	f2	c2	e4	0
f0	c1	e0	18	f0	c5	e4	1	f1	c4	e1	0	f2	c2	e5	0
f0	c1	e1	62	f0	c5	e5	0	f1	c4	e2	1	f2	c2	e6	0
f0	c1	e2	24	f0	c5	e6	0	f1	c4	e3	0	f2	c3	e0	1
f0	c1	e3	15	f1	c0	e0	32	f1	c4	e4	0	f2	c3	e1	0
f0	c1	e4	0	f1	c0	e1	21	f1	c4	e5	0	f2	c3	e2	0
f0	c1	e5	0	f1	c0	e2	11	f1	c4	e6	0	f2	c3	e3	0
f0	c1	e6	0	f1	c0	e3	2	f1	c5	e0	0	f2	c3	e4	0
f0	c2	e0	40	f1	c0	e4	1	f1	c5	e1	0	f2	c3	e5	0
f0	c2	e1	40	f1	c0	e5	0	f1	c5	e2	0	f2	c3	e6	0
f0	c2	e2	13	f1	c0	e6	0	f1	c5	e3	0	f2	c4	e0	0
f0	c2	e3	1	f1	c1	e0	19	f1	c5	e4	0	f2	c4	e1	0
f0	c2	e4	0	f1	c1	e1	13	f1	c5	e5	0	f2	c4	e2	0
f0	c2	e5	0	f1	c1	e2	6	f1	c5	e6	0	f2	c4	e3	0
f0	c2	e6	0	f1	c1	e3	0	f2	c0	e0	4	f2	c4	e4	0
f0	c3	e0	18	f1	c1	e4	0	f2	c0	e1	1	f2	c4	e5	0
f0	c3	e1	13	f1	c1	e5	0	f2	c0	e2	0	f2	c4	e6	0
f0	c3	e2	3	f1	c1	e6	0	f2	c0	e3	0	f2	c5	e0	0
f0	c3	e3	0	f1	c2	e0	11	f2	c0	e4	0	f2	c5	e1	0
f0	c3	e4	0	f1	c2	e1	2	f2	c0	e5	0	f2	c5	e2	0
f0	c3	e5	0	f1	c2	e2	1	f2	c0	e6	0	f2	c5	e3	0
f0	c3	e6	0	f1	c2	e3	0	f2	c1	e0	0	f2	c5	e4	0
f0	c4	e0	6	f1	c2	e4	0	f2	c1	e1	0	f2	c5	e5	0
f0	c4	e1	4	f1	c2	e5	0	f2	c1	e2	0	f2	c5	e6	0
f0	c4	e2	1	f1	c2	e6	0	f2	c1	e3	0				
f0	c4	e3	0	f1	c3	e0	3	f2	c1	e4	0				

Table S9: Grouped data from the three-dimensional contingency table of the distribution of region shapes (folded (f), curved (c) and extended (e)) in the single-chain dataset (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 extended was calculated using a 2x2x2 χ^2 test and a Bonferroni correction for multiple testing was applied to the resultant p-values.

Folded	Curved	Extended	Observed	Expected	Corrected p-value
f0	c0	e1	20	40.6	6.88×10^{-14}
f0	c0	e2	31	20.9	0
f0	c0	e3-e6	15	8	0
f0	c1	e0	18	40.8	2.49×10^{-10}
f0	c1	e1	62	46.2	5.95×10^{-4}
f0	c1	e2	24	23.8	0.3998
f0	c1	e3-e6	15	9.1	0.0092
f0	c2	e0	40	28.3	6.88×10^{-14}
f0	c2	e1	40	32.1	5.33×10^{-5}
f0	c2	e2	13	16.5	1.93×10^{-5}
f0	c2	e3-e6	1	6.2	1.44×10^{-8}
f0	c3-c5	e0	28	15.6	8.95×10^{-14}
f0	c3-c5	e1	17	17.7	0.0077
f0	c3-c5	e2	4	9.1	0.0017
f0	c3-c4	e3-e6	1	3.5	6.11×10^{-4}
f1-f2	c0	e0	36	14.8	0
f1-f2	c0	e1	22	16.7	6.88×10^{-14}
f1-f2	c0	e2	11	8.6	0
f1-f2	c0	e3-e6	3	3.2	0
f1-f2	c1	e0	19	16.8	2.49×10^{-10}
f1-f2	c1	e1	13	19	5.95×10^{-4}
f1-f2	c1	e2	6	9.8	0.3998
f1-f2	c1	e3-e6	0	3.7	0.0092
f1-f2	c2	e0	12	11.6	6.88×10^{-14}
f1-f2	c2	e1	2	13.1	5.33×10^{-5}
f1-f2	c2	e2	1	6.8	1.93×10^{-5}
f1-f2	c2	e3-e6	0	2.6	1.44×10^{-8}
f1-f2	c3-c5	e0	5	6.4	8.95×10^{-14}
f1-f2	c3-c5	e1	3	7.2	0.0077
f1-f2	c3-c5	e2	2	3.8	0.0017
f1-f2	c3-c5	e3-e6	0	1.4	6.11×10^{-4}

Table S10: Frequency of combinations of α-helix (H), β-strand (E) and coil (C) regions in the single-chain dataset. 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	Helix	Strand	Coil	Count
H0	E0	C0	0	H0	E3	C5	0	H1	E0	C1	35	H1	E3	C6	0
H0	E0	C1	40	H0	E3	C6	0	H1	E0	C2	34	H1	E3	C7	0
H0	E0	C2	78	H0	E3	C7	0	H1	E0	C3	6	H1	E3	C8	0
H0	E0	C3	39	H0	E3	C8	0	H1	E0	C4	3	H1	E4	C0	0
H0	E0	C4	15	H0	E4	C0	0	H1	E0	C5	1	H1	E4	C1	0
H0	E0	C5	12	H0	E4	C1	0	H1	E0	C6	0	H1	E4	C2	0
H0	E0	C6	0	H0	E4	C2	0	H1	E0	C7	0	H1	E4	C3	0
H0	E0	C7	0	H0	E4	C3	0	H1	E0	C8	1	H1	E4	C4	0
H0	E0	C8	0	H0	E4	C4	0	H1	E1	C0	5	H1	E4	C5	0
H0	E1	C0	10	H0	E4	C5	0	H1	E1	C1	7	H1	E4	C6	0
H0	E1	C1	28	H0	E4	C6	0	H1	E1	C2	4	H1	E4	C7	0
H0	E1	C2	23	H0	E4	C7	0	H1	E1	C3	0	H1	E4	C8	0
H0	E1	C3	17	H0	E4	C8	0	H1	E1	C4	0	H1	E5	C0	0
H0	E1	C4	3	H0	E5	C0	0	H1	E1	C5	0	H1	E5	C1	0
H0	E1	C5	0	H0	E5	C1	0	H1	E1	C6	0	H1	E5	C2	0
H0	E1	C6	0	H0	E5	C2	0	H1	E1	C7	0	H1	E5	C3	0
H0	E1	C7	0	H0	E5	C3	0	H1	E1	C8	0	H1	E5	C4	0
H0	E1	C8	0	H0	E5	C4	0	H1	E2	C0	4	H1	E5	C5	0
H0	E2	C0	5	H0	E5	C5	0	H1	E2	C1	0	H1	E5	C6	0
H0	E2	C1	13	H0	E5	C6	0	H1	E2	C2	0	H1	E5	C7	0
H0	E2	C2	10	H0	E5	C7	0	H1	E2	C3	0	H1	E5	C8	0
H0	E2	C3	0	H0	E5	C8	0	H1	E2	C4	0	H1	E6	C0	0
H0	E2	C4	0	H0	E6	C0	1	H1	E2	C5	0	H1	E6	C1	0
H0	E2	C5	0	H0	E6	C1	0	H1	E2	C6	0	H1	E6	C2	0
H0	E2	C6	0	H0	E6	C2	0	H1	E2	C7	0	H1	E6	C3	0
H0	E2	C7	0	H0	E6	C3	0	H1	E2	C8	0	H1	E6	C4	0
H0	E2	C8	0	H0	E6	C4	0	H1	E3	C0	0	H1	E6	C5	0
H0	E3	C0	3	H0	E6	C5	0	H1	E3	C1	0	H1	E6	C6	0
H0	E3	C1	4	H0	E6	C6	0	H1	E3	C2	0	H1	E6	C7	0
H0	E3	C2	0	H0	E6	C7	0	H1	E3	C3	0	H1	E6	C8	0
H0	E3	C3	0	H0	E6	C8	0	H1	E3	C4	0	H2	E0	C0	26
H0	E3	C4	0	H1	E0	C0	20	H1	E3	C5	0	H2	E0	C1	3
H2	E0	C2	4	H2	E3	C6	0	H3	E0	C1	1	H3	E3	C5	0
H2	E0	C3	1	H2	E3	C7	0	H3	E0	C2	0	H3	E3	C6	0
H2	E0	C4	1	H2	E3	C8	0	H3	E0	C3	0	H3	E3	C7	0

Continued on next page...

Table S10 – *Continued from previous page*

Helix	Strand	Coil	Count	Helix	Strand	Coil	Count	Helix	Strand	Coil	Count	Helix	Strand	Coil	Count
H2	E0	C5	0	H2	E4	C0	0	H3	E0	C4	0	H3	E3	C8	0
H2	E0	C6	0	H2	E4	C1	0	H3	E0	C5	0	H3	E4	C0	0
H2	E0	C7	0	H2	E4	C2	0	H3	E0	C6	0	H3	E4	C1	0
H2	E0	C8	0	H2	E4	C3	0	H3	E0	C7	0	H3	E4	C2	0
H2	E1	C0	1	H2	E4	C4	0	H3	E0	C8	0	H3	E4	C3	0
H2	E1	C1	0	H2	E4	C5	0	H3	E1	C0	0	H3	E4	C4	0
H2	E1	C2	0	H2	E4	C6	0	H3	E1	C1	0	H3	E4	C5	0
H2	E1	C3	0	H2	E4	C7	0	H3	E1	C2	0	H3	E4	C6	0
H2	E1	C4	1	H2	E4	C8	0	H3	E1	C3	0	H3	E4	C7	0
H2	E1	C5	0	H2	E5	C0	0	H3	E1	C4	0	H3	E4	C8	0
H2	E1	C6	0	H2	E5	C1	0	H3	E1	C5	0	H3	E5	C0	0
H2	E1	C7	0	H2	E5	C2	0	H3	E1	C6	0	H3	E5	C1	0
H2	E1	C8	0	H2	E5	C3	0	H3	E1	C7	0	H3	E5	C2	0
H2	E2	C0	0	H2	E5	C4	0	H3	E1	C8	0	H3	E5	C3	0
H2	E2	C1	1	H2	E5	C5	0	H3	E2	C0	0	H3	E5	C4	0
H2	E2	C2	0	H2	E5	C6	0	H3	E2	C1	0	H3	E5	C5	0
H2	E2	C3	0	H2	E5	C7	0	H3	E2	C2	0	H3	E5	C6	0
H2	E2	C4	0	H2	E5	C8	0	H3	E2	C3	0	H3	E5	C7	0
H2	E2	C5	0	H2	E6	C0	0	H3	E2	C4	0	H3	E5	C8	0
H2	E2	C6	0	H2	E6	C1	0	H3	E2	C5	0	H3	E6	C0	0
H2	E2	C7	0	H2	E6	C2	0	H3	E2	C6	0	H3	E6	C1	0
H2	E2	C8	0	H2	E6	C3	0	H3	E2	C7	0	H3	E6	C2	0
H2	E3	C0	0	H2	E6	C4	0	H3	E2	C8	0	H3	E6	C3	0
H2	E3	C1	0	H2	E6	C5	0	H3	E3	C0	0	H3	E6	C4	0
H2	E3	C2	0	H2	E6	C6	0	H3	E3	C1	0	H3	E6	C5	0
H2	E3	C3	0	H2	E6	C7	0	H3	E3	C2	0	H3	E6	C6	0
H2	E3	C4	0	H2	E6	C8	0	H3	E3	C3	0	H3	E6	C7	0
H2	E3	C5	0	H3	E0	C0	4	H3	E3	C4	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 single-chain dataset (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	40	64.8	2.10×10^{-9}
H0	E0	C2	78	75.1	1.53×10^{-9}
H0	E0	C3-C8	66	49.1	2.55×10^{-15}
H0	E1	C0	10	11.8	0
H0	E1	C1	28	19.8	3.09×10^{-4}
H0	E1	C2	23	23	1.14×10^{-5}
H0	E1	C3-C8	20	14.9	2.25×10^{-9}
H0	E2-E6	C0	9	4.9	0
H0	E2-E6	C1	17	8.2	1.88×10^{-4}
H0	E2-E6	C2	10	9.5	5.10×10^{-4}
H0	E2-E6	C3-C8	0	6.2	5.34×10^{-13}
H1-H3	E0	C0	50	21	0
H1-H3	E0	C1	39	35.1	2.10×10^{-9}
H1-H3	E0	C2	38	40.6	1.53×10^{-9}
H1-H3	E0	C3-C8	13	26.6	2.55×10^{-15}
H1-H3	E1	C0	6	6.4	0
H1-H3	E1	C1	7	10.7	3.09×10^{-4}
H1-H3	E1	C2	4	12.5	1.14×10^{-5}
H1-H3	E1	C3-C8	1	8.1	2.25×10^{-9}
H1-H3	E2-E6	C0	4	2.7	0
H1-H3	E2-E6	C1	1	4.4	1.88×10^{-4}
H1-H3	E2-E6	C2	0	5.1	5.10×10^{-4}
H1-H3	E2-E6	C3-C8	0	3.4	5.34×10^{-13}