Electronic Supplementary Information for

In proteins, the structural responses of a position to mutation rely on the Goldilocks principle: not too many links, not too few

Rodrigo Dorantes-Gilardi, Laëtitia Bourgeat, Lorenza Pacini, Laurent Vuillon, Claire Lesieur

Claire Lesieur
Email: claire.lesieur@ens-lyon.fr

ESI file includes:

Supplementary Methods
Figs. S1 to S4
Tables S1
References

Supplementary Methods

Box plot. A box plot divides data by fourth equal part. The first quartile Q1 is the values of the first 25 % of the data, the second quartile Q2 is the median, (50% of the data), and the third quartile Q3 is the values of 75 % of the data. Above Q3 are the values between Q3 and the maximum, and below Q1 are the value between Q1 value and the min value.
The degrees and weights probably overestimate the amino acid and the atomic packing of amino acids because the radius of van der Waals of atoms is ignored. Nevertheless, amino acids are composed of the same atoms, carbon, hydrogen (not included here), oxygen, nitrogen and sulfur (Met and Cys), and in the dataset same residue types have identical number of atoms, so the over estimation is likewise for every amino acid, making the approximation (ignoring van der walls volume) reasonable.

Torus. In order to compare the number of amino acids on the surface of a protein and the number of amino acids inside the protein (called buried amino acids), we made a theoretical model. As proteins in the dataset are oligomers their topology is a torus (a doughnut-shape object), they cannot be modelled by a sphere as are monomeric proteins. In order to define a torus, we need two quantities: the whole diameter 2R of the doughnut (from the two most opposite outside points) and the diameter 2r of the 'tube' of the doughnut (from an outside point to its closest opposite point inside point on the tube). The area (that is the contact surface of the doughnut) is calculated with the usual formula for a torus, namely $4\pi 2Rr$ x 0.9 where 0.9 is the density of spherical packing on the plane, because as a first approximation an amino acid is a sphere on the surface. The volume is computed with the usual volume of a torus namely $2\pi 2Rr2$ x 0.74 where 0.74 is the spherical packing in space. With this computation the ratio of the number of amino acids of the protein and the number of amino acids on the surface of the protein is between 0.2 and 2 when r varies from 3 to 8 nm. This means that the doughnut - shaped model gives a large possibility of ranges: from a number of amino acids twice as large as at the surface to a number of amino acids 5 times bigger on the inside of the protein (Fig. S1).

Accessibility Surface Area (ASA). ASA was calculated using the program available at http://cib.cf.ocha.ac.jp/bitool/ASA. This program is based on a method previously described in (1).

Link weight perturbation networks. The perturbation networks are built as follows:

1. The amino acids networks of the reference Gref = (Vref , Eref) and the mutant Gmut = (Vmut, Emut) are generated. G, V and E stand for Graph, Vertex (node) and Edge (link), respectively. The ref is CtxB5.

2. Initially, the perturbation network Gp = (Vp, Ep) contains all the nodes that appear in Gref and Gmut :

$$Vp = Vref \cup Vmut \qquad (1)$$

3. Ep contains all the links that have a weight difference between the two networks higher than 4. If a link is contained in only one network, it is considered as having null weight [w (u,v)=0]:

$$Ep = \{(i, j) \in Eref \cup Emut \text{ s.t. } |wmut (i, j)- wref (i, j)| > 4\} \qquad (2)$$

4. The link weights w (i, j) in the perturbation network are given by the absolute value of the difference in link weights between the two networks:

$$wp (i, j) = | \Delta w (i,j)| = |wmut (i, j)- wref (i, j)| \qquad (3)$$

5. A link color is assigned based on the sign of Δw (i, j):

$$\begin{aligned} color (i, j) = & \text{ red if } wmut (i, j) - wref (i, j) < 0 \\ & \text{ green if } wmut (i, j) - wref (i, j) > 0 \end{aligned} \qquad (4)$$

6. All nodes of degree zero are removed from the perturbation network (i.e, nodes for which there is no difference in link weights between the two networks).

Sphere of influence. The induced perturbation network from a source node v*, referred to as the sphere of influence of the position v*, is built as follows:

1. The perturbation tree is built by applying the Breadth-_rst search algorithm in the rooted cases to
the perturbation network, using v* as root (https://en.wikipedia.org/wiki/Breadth-first_search). The perturbation tree contains all the nodes that can be reached starting at the source v* following simple paths in the perturbation network. If v* is a mutated position, then the perturbation tree contains all the nodes that are affected by the mutation.

2. All links of the perturbation network whose end-points are in the perturbation tree are added to the
perturbation tree, if not present yet. In this way, the rescue mechanisms appear as cycles in the induced
perturbation network.
Perturbation networks and induced perturbation networks are classified as 1D, 2D, 3D, 4D, 3-4D etc. if they contain links representing 1D, 2D, 3D, 4D, 3 and 4D contacts, respectively.

a 1D relation means that the two nodes are first neighbors in the amino acids sequence;
a 2D relation means that the two nodes belong to the same secondary structure (the same α-helix, the same β-sheet or the same loop);
a 3D relation means that the two nodes do not belong to the same secondary structure but they belong to the same chain;
a 4D relation means that the two nodes belong to different chains.

Jaccard similarity measure. We made an algorithm to compare the environment (neighborhood) of every amino acid of the two toxins CtxB5 and hLTB5. We start with vectors of 20 counters associated with the 20 amino acid types, and we initialize each counter with a value equals to 0. Given an amino acid –i-, the vector gives the number each amino acid type in the environment of –i-, e.g. if Val is 3 times in the environment of –i-, then the entry corresponding to Val in the vector is 3.
To compare two environments, we calculate the Jaccard similarity measure on the pair of vectors. The Jaccard similarity is computed using the environment vectors as follows: the intersection of each entry of the vector, that is the number of amino acids in common in the two proteins for each amino acid type, e.g. if there are 5 Val in the environment of amino acid –i- in protein 1 and 3 in protein 2, then the intersection of the entry Val in the vectors is equal to the minimal value that is 3; the union of each entry of the vector, that is the maximal number of amino acids in the two proteins for each amino acid type, e.g. if there are 5 Val in the environment of amino acid –i- in protein 1 and 3 in protein 2, then the union of the entry Val in the vectors is equal to the maximal value that is 5. There is one intersection value per amino acid type and the sum of the twenty intersection values is noted inter(-i-). Likewise, we compute the union of each entry in the two vectors maximal value and the sum of the union is noted union(-i-). The Jaccard measure for amino acid –i- is the ratio inter(-i-) to union (-i-). Note that the Jaccard measure is a value in the interval [0, 1] because inter(-i-) is lower or equal to union (-i-). If Jaccard (-i-) equals to 0, this means that inter (-i-) equals to 0 and the environments of –i- of the two proteins are either composed of 0 or do not share an amino acid type in common. If on the other hand, Jaccard (-i-) equals to 1, then the two environments are identical.

References

1. Samanta U, Bahadur RP, Chakrabarti P (2002) Quantifying the accessible surface area of protein residues in their local environment. *Protein Eng* 15(8):659–667.

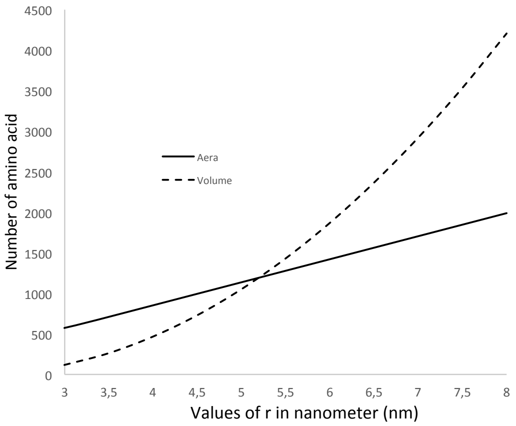

Fig. S1. Number of amino acids on the boundary (called area) and in global (called volume) of a torus shape molecule (doughnut-shape) with R=8 nm, the large radius of the torus and r, the small radius of the torus from 3 nm to 8 nm.

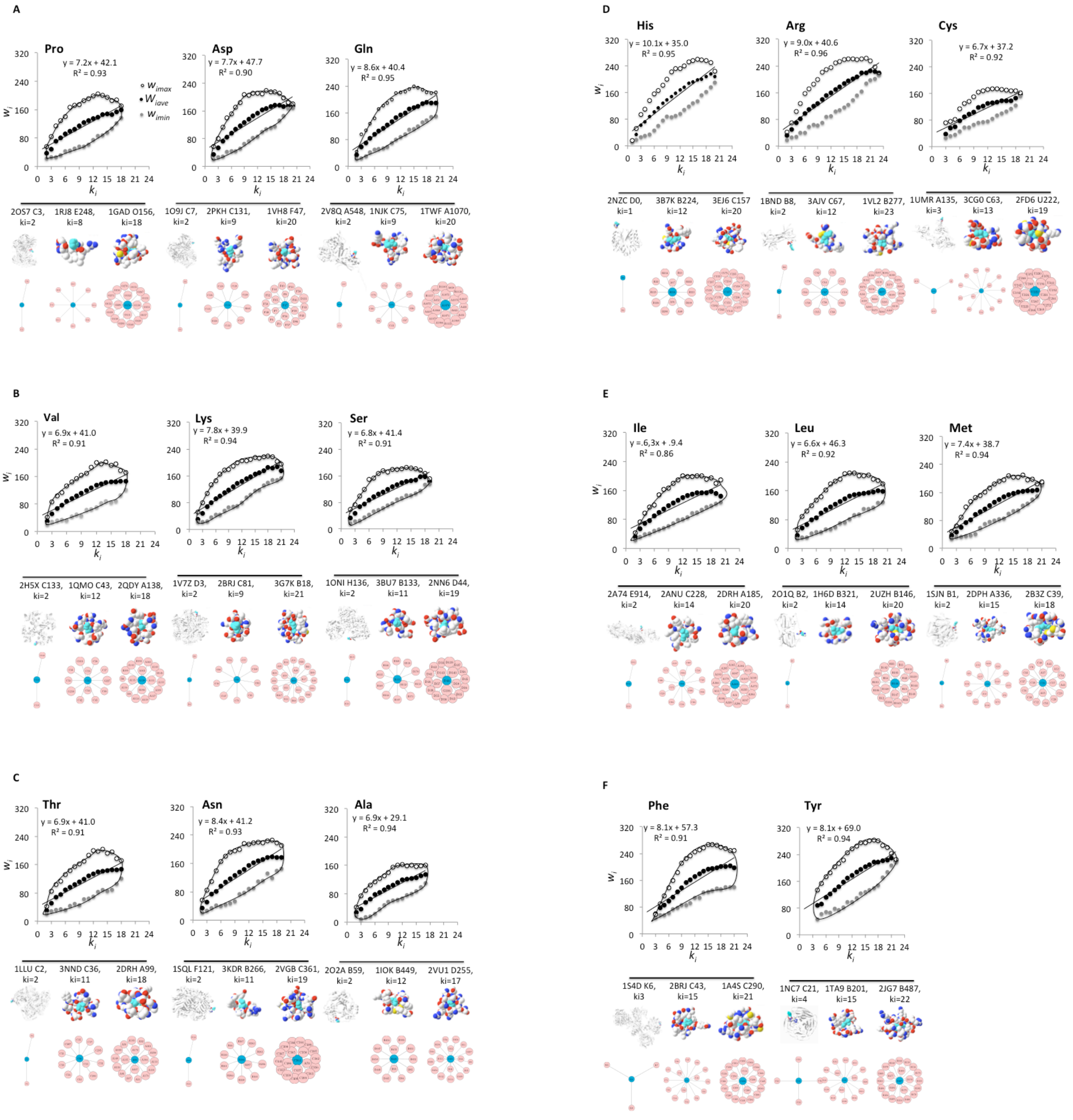

Figure S2A-F. Amino acid capacity of interactions. Upper panels: Weight versus degree of the amino acids. The continuous lines show the area (envelope) covered by the set of degrees and weights adopted by the amino acids. Middle panels. X-ray local structures of the amino acids (Atomic packing representation) for a min (left), a mode, i.e. most frequent (middle) and a max (right) degrees. The whole protein is shown for the min degrees but only the local structures are shown for the mode and max degrees. The residue –i- is indicated in cyan and the neighbors –jk- in CPK. The amino acids –i- and –jk- are shown in spacefill. The figure is generated with sPDB viewer. The PDB code, the chain, the position of the residue along the sequence and degree are indicated. Lower panels. Local Networks of the local structures (amino acid packing representation) as in middle panel. The residue –i- is indicated in cyan and the neighbors –jk- in pink. The nodes (circles) are the residues and the links between amino acid pairs (lines) are based on the two residues having at least one atom each within a 5Å distance.

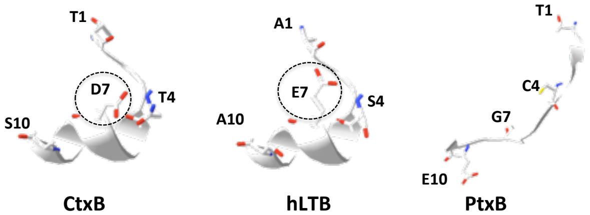

Fig. S3. X-ray structures of the first ten residues of the N-termini of CtxB5 (PDB 1EEI), hLTB5 (1LTR) and PtxB5 (2XSC). The N-termini are shown in ribbon with the mutated residues in sticks (sPDB viewer).

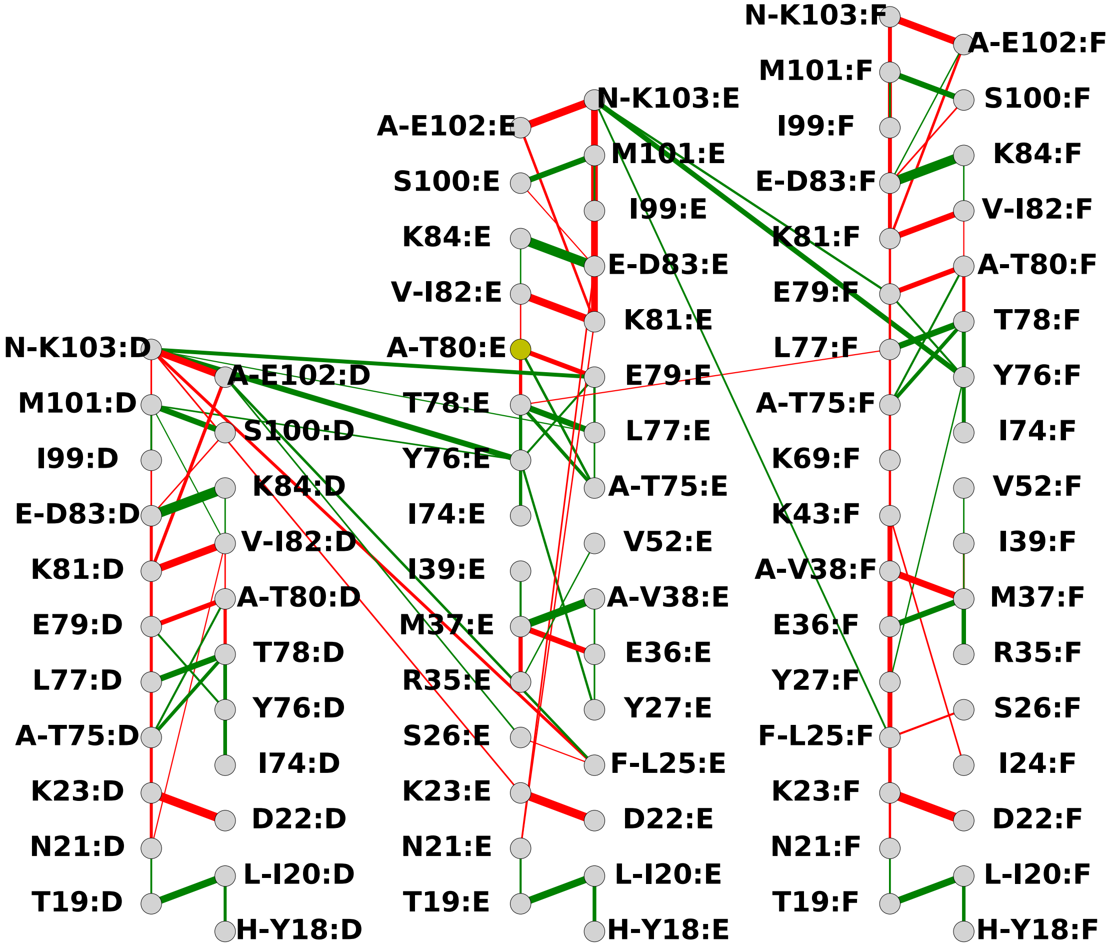

Fig. S4. Sphere of influence of the position 80. A. Sphere of influence of position 80. Nodes are amino acids, with K84:E for Lysine at position 84 in chain E. Mutated position are A-T80:E for A (Ala) in CtxB$_5$ and T (Thr) in LTB$_5$. Yellow node is the source of the perturbation. Green and red links are for lower and higher link weights in CtxB$_5$, respectively. Link thickness is proportional to Δw_{ij}.

Table S1. Local amino acid interaction measures of the amino acids of CtxB$_5$ (1EEI), hTLB$_5$ (1LTR) and PTX$_5$ (2XSC).

pi	1EEI	1LTR	ki1EEI	ki1LTR	Δki	wi1EEI	wi1LTR	Δwi	Nw1EEI	Nw1LTR	Nw2XSC
1*	T	A	7	7	0	81	68	13	12	10	12
2*	P	P	10	10	0	118	123	5	12	12	14
3*	Q	Q	9	9	0	108	95	13	12	11	12
4*	N	S	7	7	0	131	115	16	19	16	10
5	I	I	17	14	3	151	151	0	9	11	10
6*	T	T	8	8	0	120	116	4	15	15	11
7*	D	E	9	9	0	128	125	3	14	14	11
8	L	L	16	15	1	147	144	3	9	10	8
9*	C	C	12	12	0	135	141	6	11	12	9
10*	A	S	8	7	1	84	88	4	11	13	12
11*	E	E	8	8	0	141	126	15	18	16	
12	Y	Y	12	12	0	168	175	7	14	15	
13*	H	H	5	5	0	77	80	3	15	16	
14*	N	N	8	8	0	120	119	1	15	15	
15	T	T	11	12	1	153	151	2	14	13	
16	Q	Q	11	10	1	151	138	13	14	14	
17*	I	I	10	10	0	124	133	9	12	13	
18	H	Y	10	11	1	154	162	8	15	15	
19*	T	T	6	6	0	99	93	6	17	16	
20	L	I	12	10	2	145	152	7	12	15	
21*	N	N	7	7	0	123	102	21	18	15	
22	D	D	9	9	0	142	148	6	16	16	
23*	K	K	10	10	0	129	139	10	13	14	
24	I	I	14	15	1	140	139	1	10	9	
25	F	L	11	15	4	163	165	2	15	11	
26	S	S	10	10	0	145	140	5	15	14	
27	Y	Y	17	17	0	209	220	11	12	13	
28	T	T	11	12	1	157	161	4	14	13	
29	E	E	15	16	1	170	179	9	11	11	
30*	S	S	12	12	0	137	137	0	11	11	
31	L	M	15	16	1	144	145	1	10	9	
32*	A	A	11	11	0	134	129	5	12	12	
33*	G	G	8	9	1	90	90	0	11	10	
34*	K	K	8	8	0	91	92	1	11	12	
35	R	R	13	13	0	183	182	1	14	14	
36	E	E	17	17	0	186	191	5	11	11	
37	M	M	15	15	0	137	156	19	9	10	
38	A	V	11	13	2	110	144	34	10	11	
39	I	I	16	15	1	149	160	11	9	11	
40	I	I	14	16	2	155	154	1	11	10	
41	T	T	9	10	1	151	156	5	17	16	
42	F	F	14	14	0	212	213	1	15	15	
43*	K	K	6	9	3	83	99	16	14	11	
44*	N	S	5	4	1	95	79	16	19	20	
45*	G	G	5	5	0	64	64	0	13	13	
46*	A	A	8	7	1	112	106	6	14	15	
47*	T	T	10	10	0	127	129	2	13	13	
48	F	F	15	15	0	205	208	3	14	14	
49	Q	Q	16	16	0	200	204	4	13	13	
50*	V	V	12	14	2	123	128	5	10	9	
51*	E	E	12	12	0	129	134	5	11	11	
52*	V	V	11	10	1	113	123	10	10	12	
53*	P	P	11	9	2	89	82	7	8	9	
54*	G	G	5	5	0	90	58	32	18	12	
55*	S	S	4	5	1	60	56	4	15	11	

56*	Q	Q	8	8	0	127	101	26	16	13
57	H	H	10	11	1	185	174	11	19	16
58*	I	I	9	7	2	126	88	38	14	13
59*	D	D	6	6	0	81	81	0	14	14
60*	S	S	8	8	0	108	86	22	14	11
61	Q	Q	15	14	1	200	186	14	13	13
62*	K	K	9	8	1	108	104	4	12	13
63*	K	K	9	11	2	91	110	19	10	10
64*	A	A	12	12	0	114	126	12	10	11
65	I	I	14	13	1	174	175	1	12	13
66	E	E	12	11	1	172	161	11	14	15
67	R	R	17	18	1	214	224	10	13	12
68	M	M	17	17	0	182	177	5	11	10
69	K	K	16	15	1	190	186	4	12	12
70	D	D	10	11	1	160	163	3	16	15
71	T	T	14	14	0	154	167	13	11	12
72	L	L	15	18	3	145	151	6	10	8
73	R	R	15	15	0	187	199	12	13	13
74*	I	I	11	12	1	125	137	12	11	11
75	A	T	11	15	4	123	162	39	11	11
76	Y	Y	16	18	2	202	246	44	13	14
77*	L	L	10	13	3	122	132	10	12	10
78*	T	T	7	8	1	119	116	3	17	15
79*	E	E	8	10	2	107	138	31	13	14
80*	A	T	10	14	4	88	128	40	9	9
81	K	K	12	10	2	172	119	53	14	12
82	V	I	14	17	3	147	152	5	11	9
83	E	D	12	11	1	174	152	22	15	14
84	K	K	13	13	0	166	153	13	13	12
85	L	L	16	16	0	145	148	3	9	9
86	C	C	15	16	1	153	149	4	10	9
87	V	V	14	14	0	182	175	7	13	13
88	W	W	19	19	0	187	217	30	10	11
89	N	N	8	8	0	144	138	6	18	17
90*	N	N	5	6	1	104	109	5	21	18
91*	K	K	10	10	0	126	140	14	13	14
92*	T	T	7	7	0	93	94	1	13	13
93	P	P	11	11	0	165	170	5	15	15
94	H	N	14	13	1	169	168	0	12	13
95	A	S	10	11	1	126	155	29	13	14
96	I	I	16	16	0	138	139	1	9	9
97*	A	A	12	13	1	131	133	2	11	10
98*	A	A	13	13	0	129	130	1	10	10
99	I	I	16	17	1	142	143	1	9	8
100	S	S	12	11	1	149	140	9	12	13
101	M	M	17	16	1	140	161	21	8	10
102*	A	E	8	9	1	106	120	14	13	13
103	N	N	5	10	5	74	153	79	15	15

pi stands for position of amino acid –i- in the sequence. Red is for mutated positions, stars for amino acids whose degrees and weights are within the intersecting envelop commons to the twenty amino acids.

References

1. Samanta U, Bahadur RP, Chakrabarti P (2002) Quantifying the accessible surface area of protein residues in their local environment. *Protein Eng* 15(8):659–667.