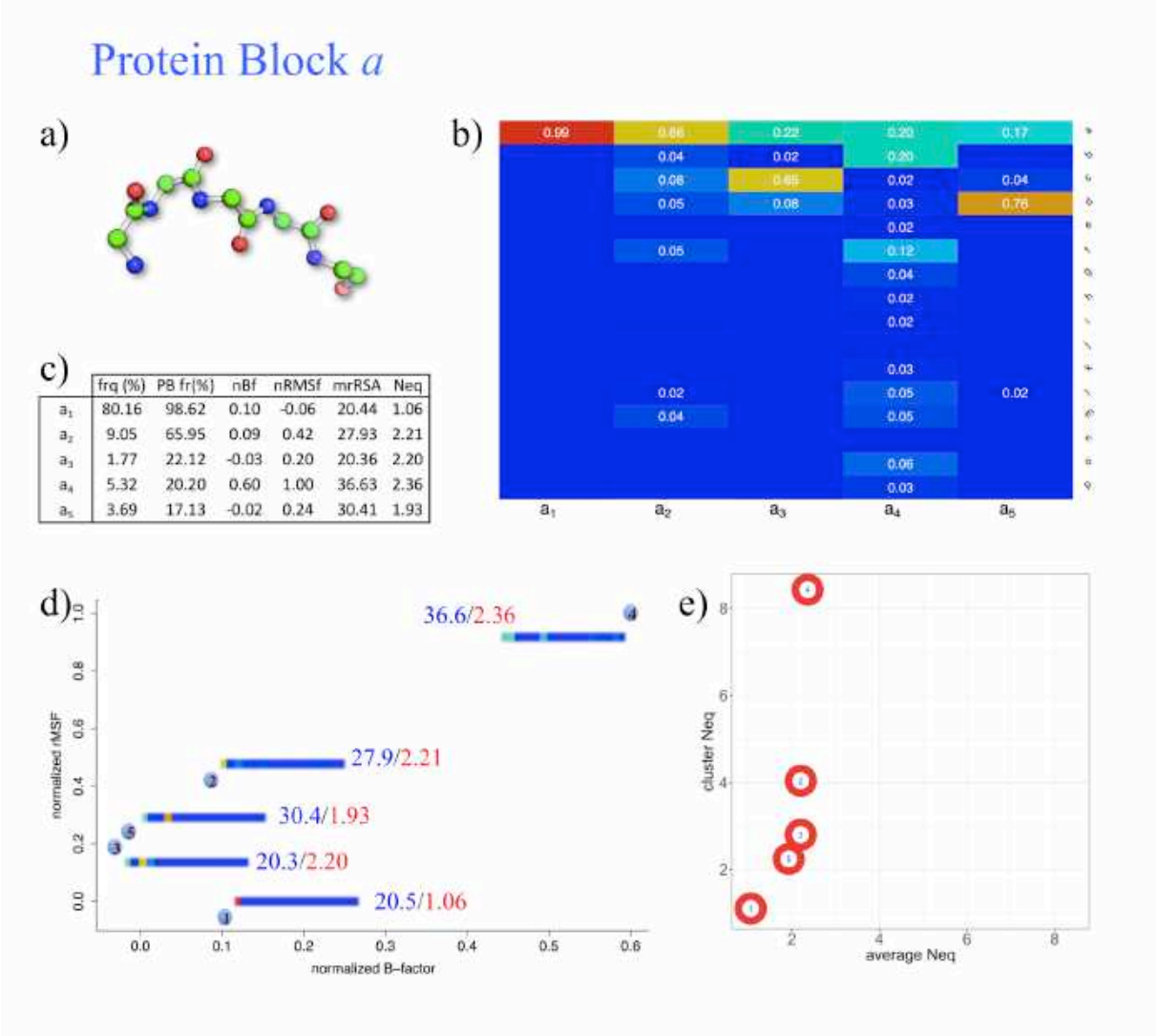
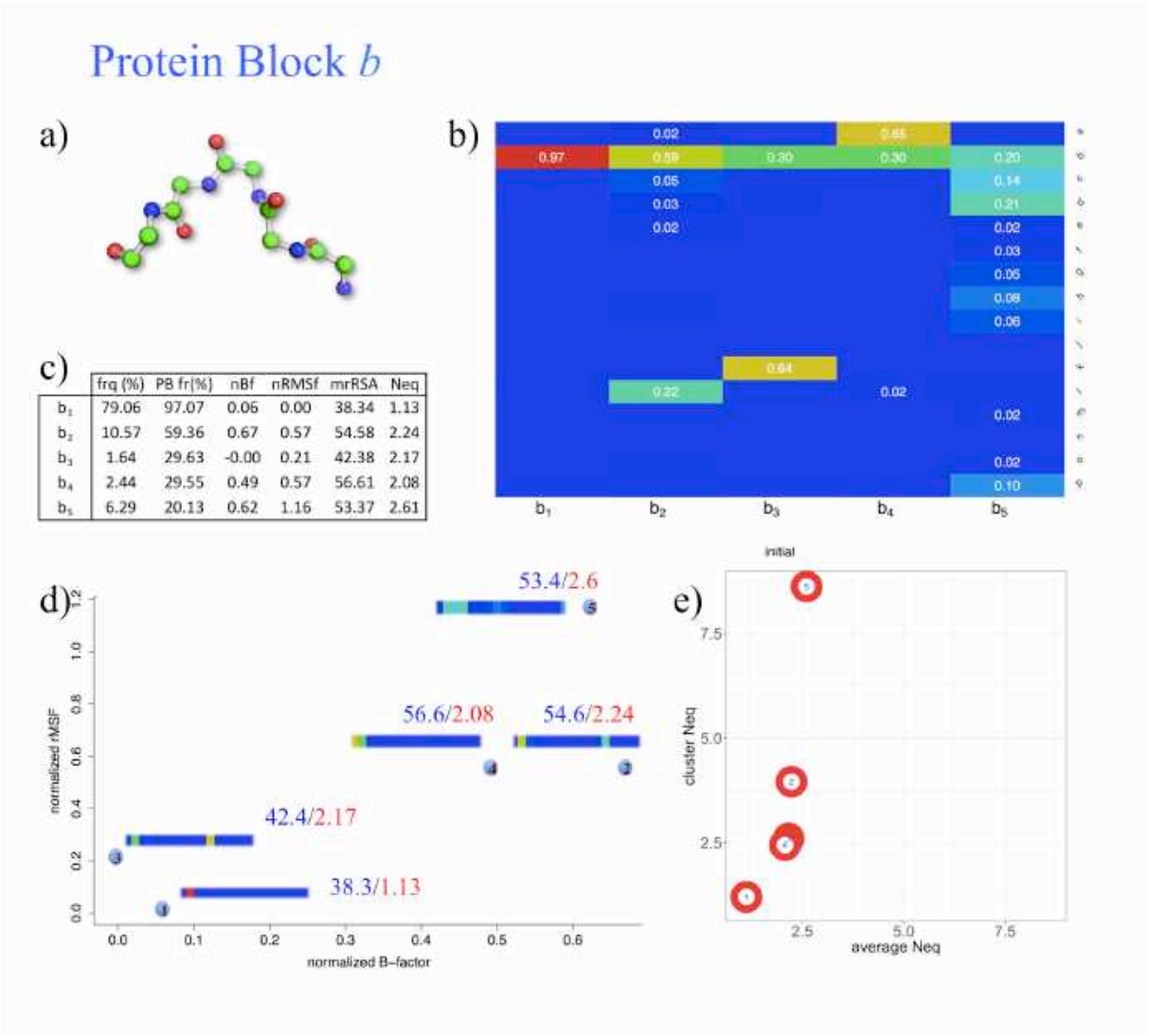
Sup data 8. *Clustering of each PB*. For each of the 16 Protein Block is given (a) a 3D representation of the Protein Blocks using PyMOL, (b) a visualization with heatmap.2 from R of the clustering obtained with *k-means*, (c) a Table summarizing the features of each cluster with associated occurrences (%), mean normalized B-factor, mean normalized RMSf, mean relative solvent accessibility and average N_{eq} (from PB contents of the simulations associated to the cluster, see Sup data 9 for more information). (d) A plot of mean normalized B-factor *vs.* mean normalized RMSf for the 5 clusters with a direct coloured visualisation of the PB cluster distribution and mean relative solvent accessibility (in blue) and mean N_{eq} (in red). (e) A plot of the average N_{eq} and cluster N_{eq} (corresponding to the N_{eq} computed from the PB content of the cluster, see Figure b).



	frq (%)	PB fr(%)	nBf	nRMSf	mrRSA	Neq
a_1	80.16	98.62	0.10	-0.06	20.44	1.06
a_2	9.05	65.95	0.09	0.42	27.93	2.21
a_3	1.77	22.12	-0.03	0.20	20.36	2.20
a_4	5.32	20.20	0.60	1.00	36.63	2.36
a_5	3.69	17.13	-0.02	0.24	30.41	1.93

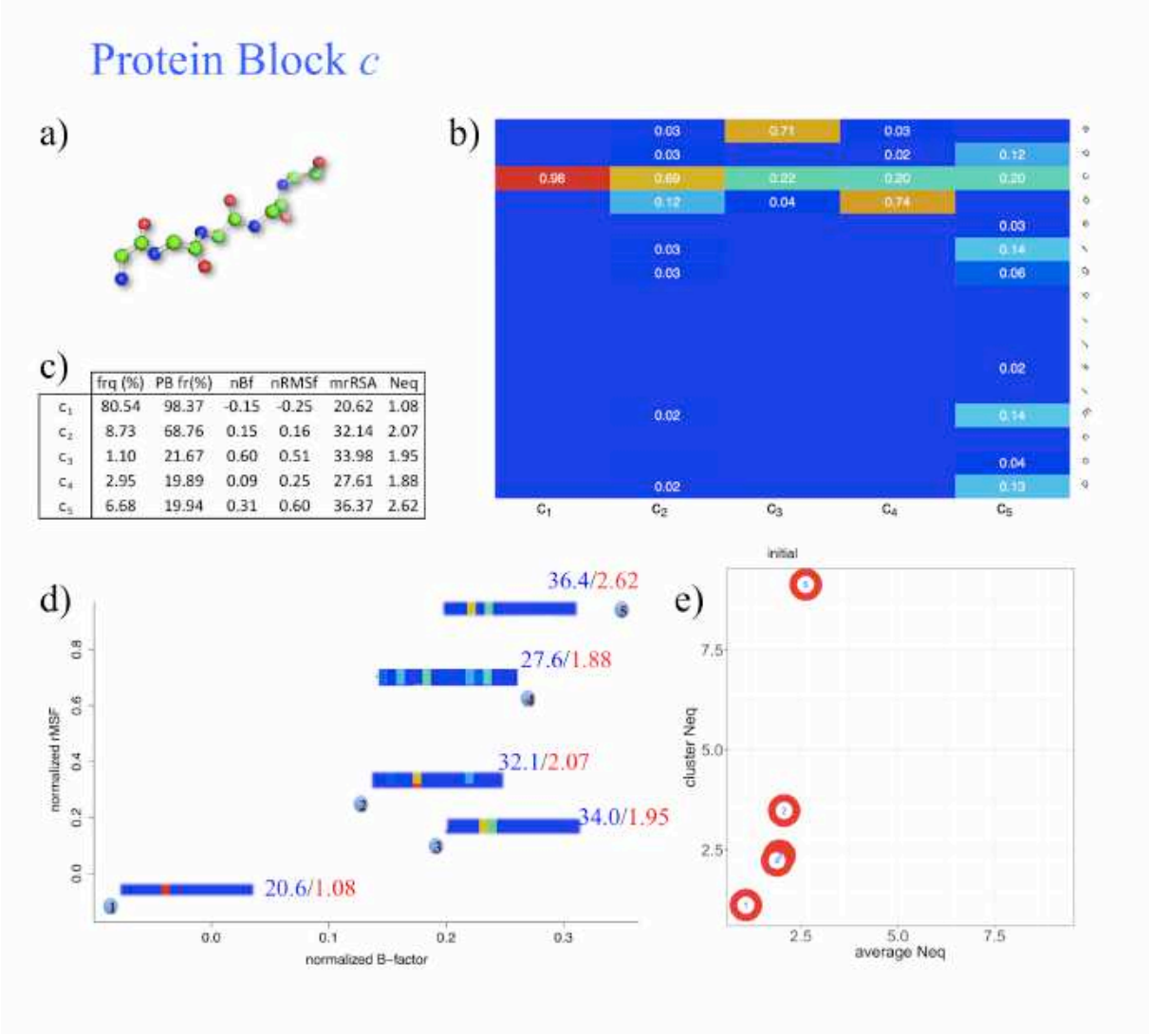
Protein Block *a* represents 3.9% of the protein structure dataset, it is 75.8% coil and 24.1 β-sheet (1). In a previous study, we have analysed the major geometrical transition to another PB, i.e. if this PB was not here which can be the second best one (2). The major transitions were so PB *c* (51%), PB *f* (17%) and PB *d* (9.4%)

The five clusters of PB *a* gives interesting results as cluster a_1 (>98% of PB *a*) that represents 4^{th}/5 of initial had in fact the 4^{th} highest nBf (0.10 vs. -0.03 for cluster a_3) but the lowest nRMSf and one of the lowest rSA. Clusters a_3 and a_5 followed this idea of geometrical resemblance presented before with high presence of PB *c* (65%) and PB *d* (76%). Cluster a_2 represents the variation around original PB *a* (still representing 67% of the occurrences) with 6 PBs at more than 2%. Cluster a_4 represented the 5.3% of original PB that must be the more deformable (average N_{eq} of 2.36, cluster N_{eq} of 8.43) associated to highest accessibility, highest nBf and highest nRMSf.



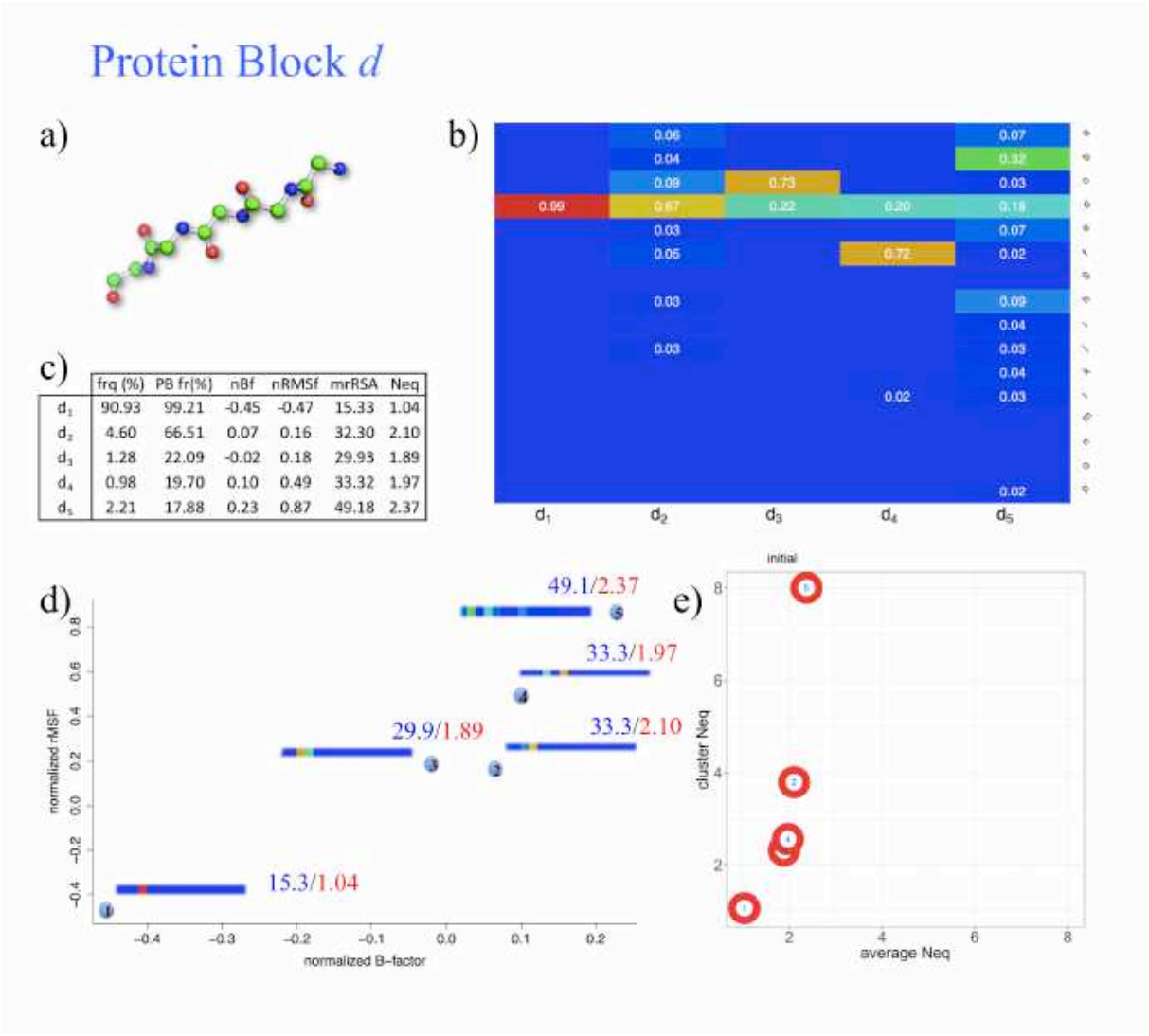
Protein Block *b* represents 4.4% of the protein structure dataset, it is 85.3% coil and 14.6 β-sheet (1). The major transitions as described in (2) were PB *d* (48%), PB *c* (16%) and PB *f* (13%).

Cluster b_1 (>97% of PB *b*) represents 79% of original PB *b* positions had lowest rSA and lowest nRMSf but the second lowest nBf. Interestingly not one the three following clusters have used the expected geometrical transitions (PBs *d*, *c* and *f*), but PB *l* for cluster b_2 (22%), PB *k* for cluster b_3 (64%) and PB *a* for cluster b_4 (65%). Only cluster b_3 can be considered as comparable with cluster b_1 in terms of nRMSf and nBf and the closest rSA. Cluster b_5 represented the 6.3% of original PB that must be the more deformable (average N_{eq} of 2.61, cluster N_{eq} of 8.62) associated to highest accessibility, high nBf and highest nRMSf.



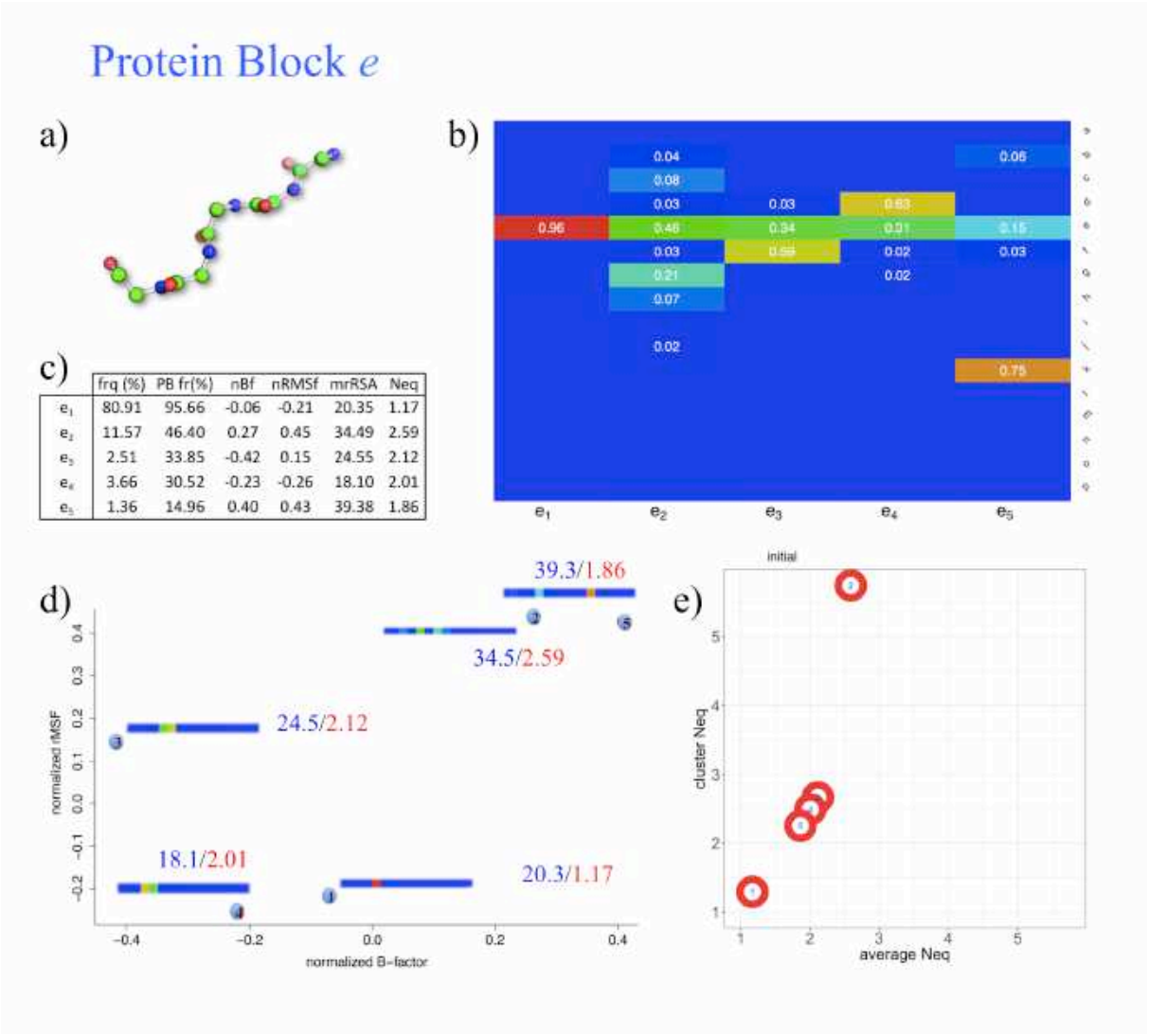
Protein Block *c* represents 8.1% of the protein structure dataset, it is 57.6% coil and 42.4 β-sheet (1). The major transitions as described in (2) were PB *d* (62%), PB *f* (23%) and PB *e* (6%).

Cluster c_1 (>98% of PB *c*) represents 81% of original PB *c* positions and is and is by far associated with the lowest values of nBf, nRMSf and rSA. The four others are characterized by slightly similar rSA and nBf, but evolved quite differently in terms of nRMSf. Hence, the most stable in regard to dynamics is cluster c_3 characterized by a high propensity of PB *a* (71%), a PB not considered as geometrically equivalent (2). Cluster c_2 is still highly populated with PB *c* and a little with PB *d* (12%) while Cluster c_4 is a real PB *d* cluster (74%), PB *d* is the best transition, these clusters were so expected. Cluster c_5 represented the 6.7% of original PB that must be the more deformable (average N_{eq} of 2.62, cluster N_{eq} of 9.14) associated to highest accessibility, highest nBf and highest nRMSf.



Protein Block *d* represents 19% of the protein structure dataset, it is 29% coil and 71% β-sheet (1). The major transitions as described in (2) were PB *f* (50%), PB *c* (26%) and PB *e* (20%).

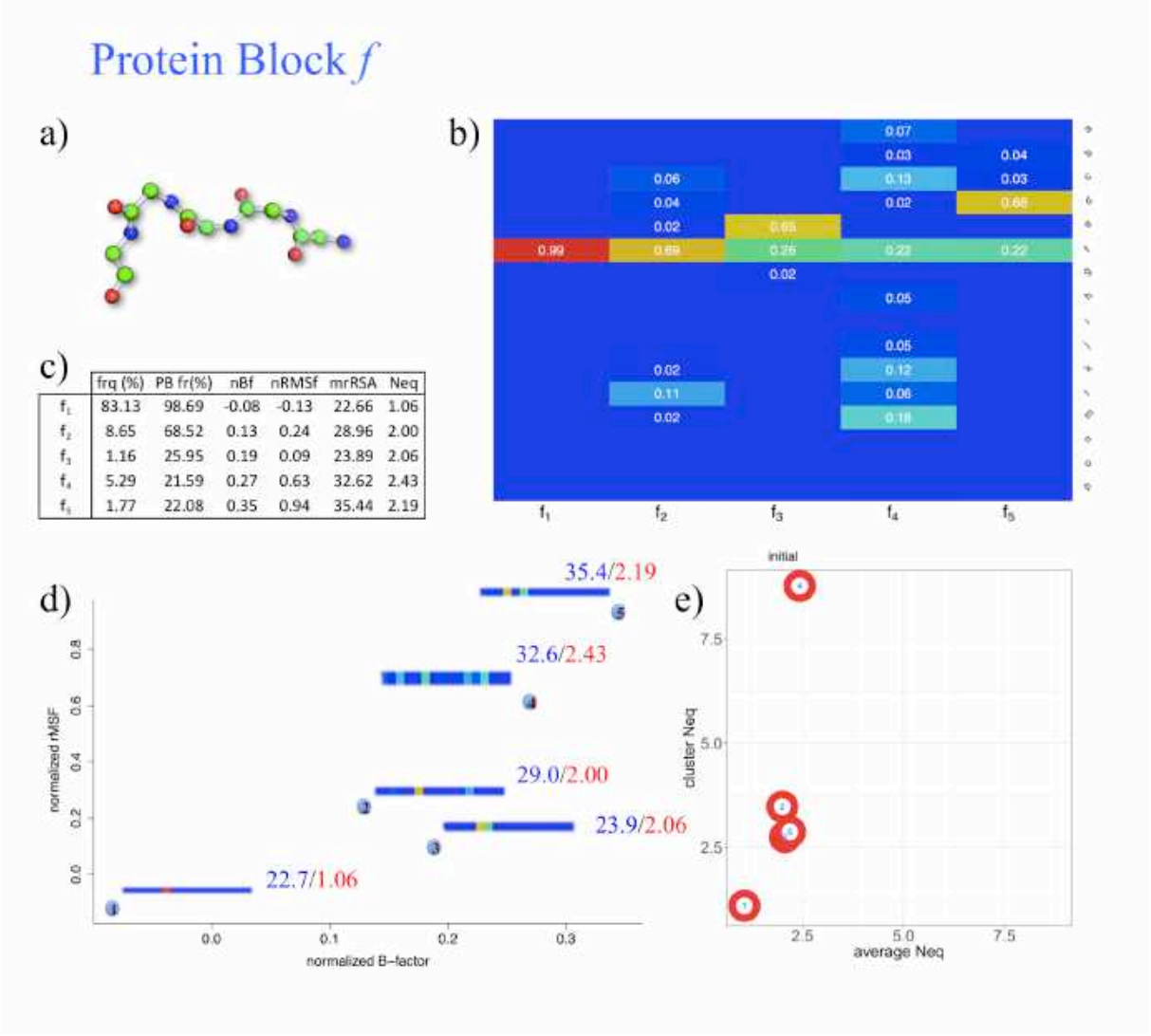
Cluster d_1 (>99% of PB *d*) represents 90% of original PB *d* positions and is and is by far associated with the lowest values of nBf, nRMSf and rSA. The three following clusters are characterized by slightly similar rSA and nBf, but evolved quite differently in terms of nRMSf. Clusters d_3 and d_4 are guided by PBs with high geometrical transition rates, namely PB *c* for cluster d_3 (73%) and PB *f* for cluster d_4 (72%). Cluster d_2 is still highly PB d (67%) associated to higher accessibility and flexibility than cluster d_1. Cluster d_5 represented the 2.2% of original PB that must be the more deformable (average N_{eq} of 2.37, cluster N_{eq} of 8.00) associated to highest accessibility, highest nBf and highest nRMSf.



	frq (%)	PB fr(%)	nBf	nRMSf	mrRSA	Neq
e_1	80.91	95.66	-0.06	-0.21	20.35	1.17
e_2	11.57	46.40	0.27	0.45	34.49	2.59
e_3	2.51	33.85	-0.42	0.15	24.55	2.12
e_4	3.66	30.52	-0.23	-0.26	18.10	2.01
e_5	1.36	14.96	0.40	0.43	39.38	1.86

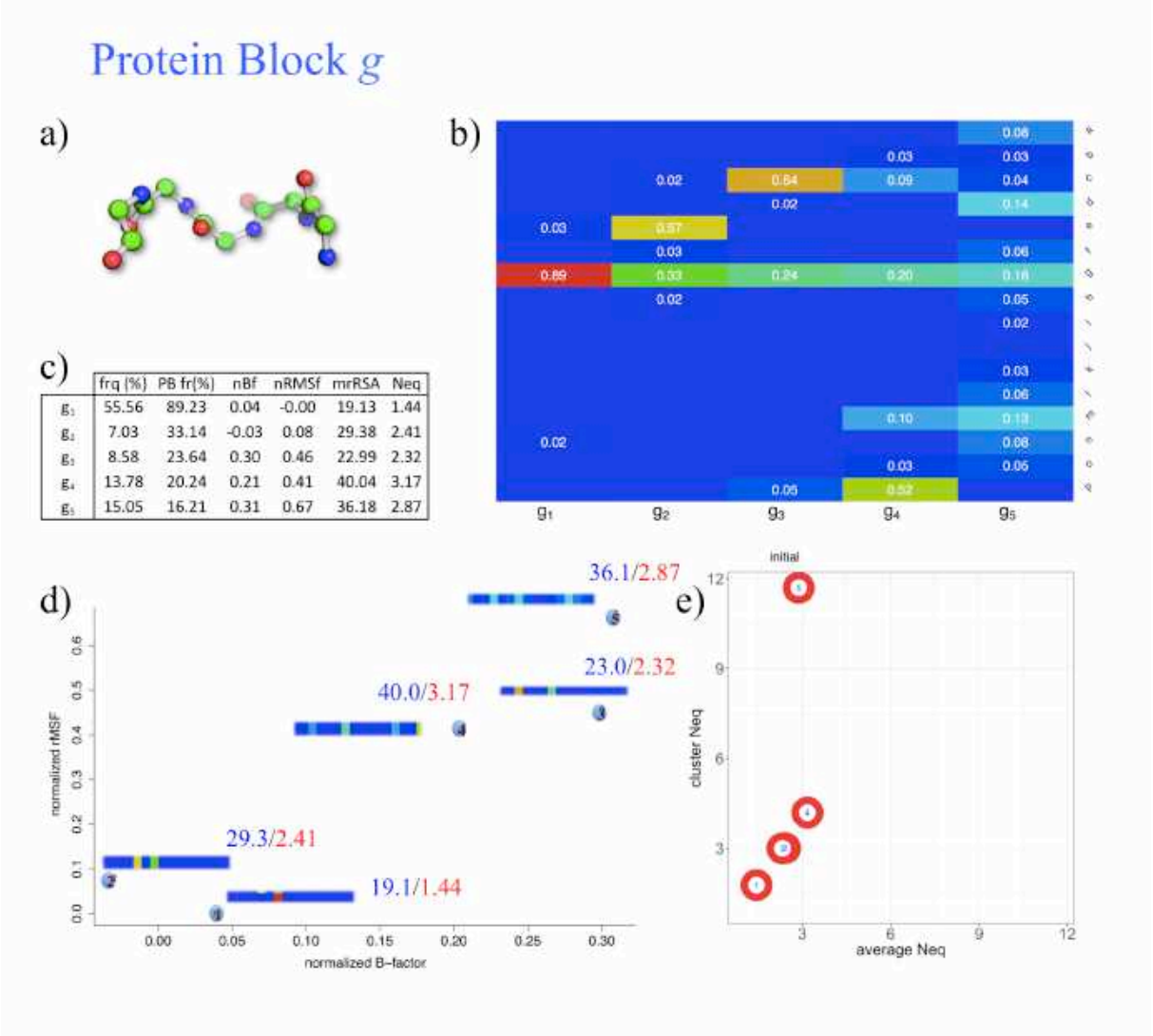
Protein Block *e* represents 2.45% of the protein structure dataset, it is 45.5% coil and 54.4 β-sheet (1). The major transitions as described in (2) were PB *h* (81%) and PB *d* (9%).

Interestingly cluster e_1 (>95% of PB *e*) represents 80% of original PB *e* positions is not associated with the lowest values of nBf, nRMSf and rSA, it is the cluster e_4. This last is mainly directed by PB *d* (63%) that is the second best geometrically compatible. But strangely, the best one PB *h* is found only with very low occurrence in cluster e_2 that is associated to very nBf, nRMSf, rSA and N_{eq}. Cluster e_2 is slightly comparable with only a lower cluster N_{eq}. It is mainly associated with unexpected PB *k* (75%). PB *f* also was unexpected but control 59% of cluster e_3 with low nBf and rSA, but a medium nRMSf.



Protein Block *f* represents 6.7% of the protein structure dataset, it is 73.3% coil and 26.7 β-sheet (1). The major transitions as described in (2) were PB *k* (61%) and PB *b* (35%).

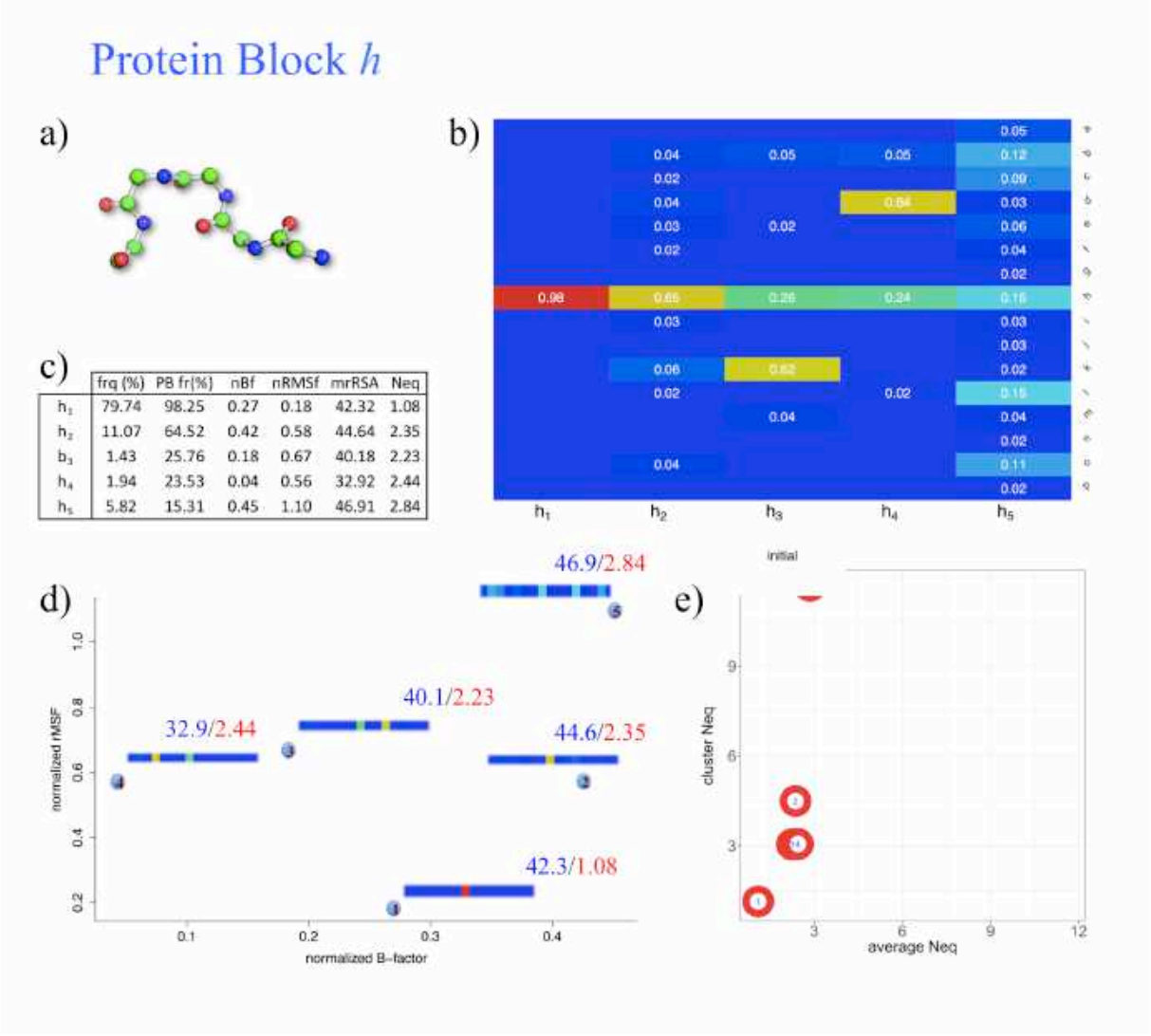
As often seen the majority cluster, namely cluster f_1 (>98% of PB *f*) represents 83% of original PB *b* positions had lowest nBf, nRMSf and rSA. The expected geometrical transitions (PBs *b* and *k*) have not been used during dynamics to substitute PB *f*. They have been replaced by PB *e* for cluster f_3 (65%) and PB *d* for cluster f_5 (66%). Interestingly, this last is associated to high nBf, high nRMSf and high rSA that is quite uncommon for PB *d*. Cluster f_2 is less stable than cluster b_1 with a lower PB content of PB *f* (69%). Cluster f_4 represented the 1.8% of original PB that must be the more deformable (average N_{eq} of 2.43, cluster N_{eq} of 8.78) associated to high accessibility, high nBf and high nRMSf (but not the highest).



	frq (%)	PB fr(%)	nBf	nRMSf	mrRSA	Neq
g_1	55.56	89.23	0.04	-0.00	19.13	1.44
g_2	7.03	33.14	-0.03	0.08	29.38	2.41
g_3	8.58	23.64	0.30	0.46	22.99	2.32
g_4	13.78	20.24	0.21	0.41	40.04	3.17
g_5	15.05	16.21	0.31	0.67	36.18	2.87

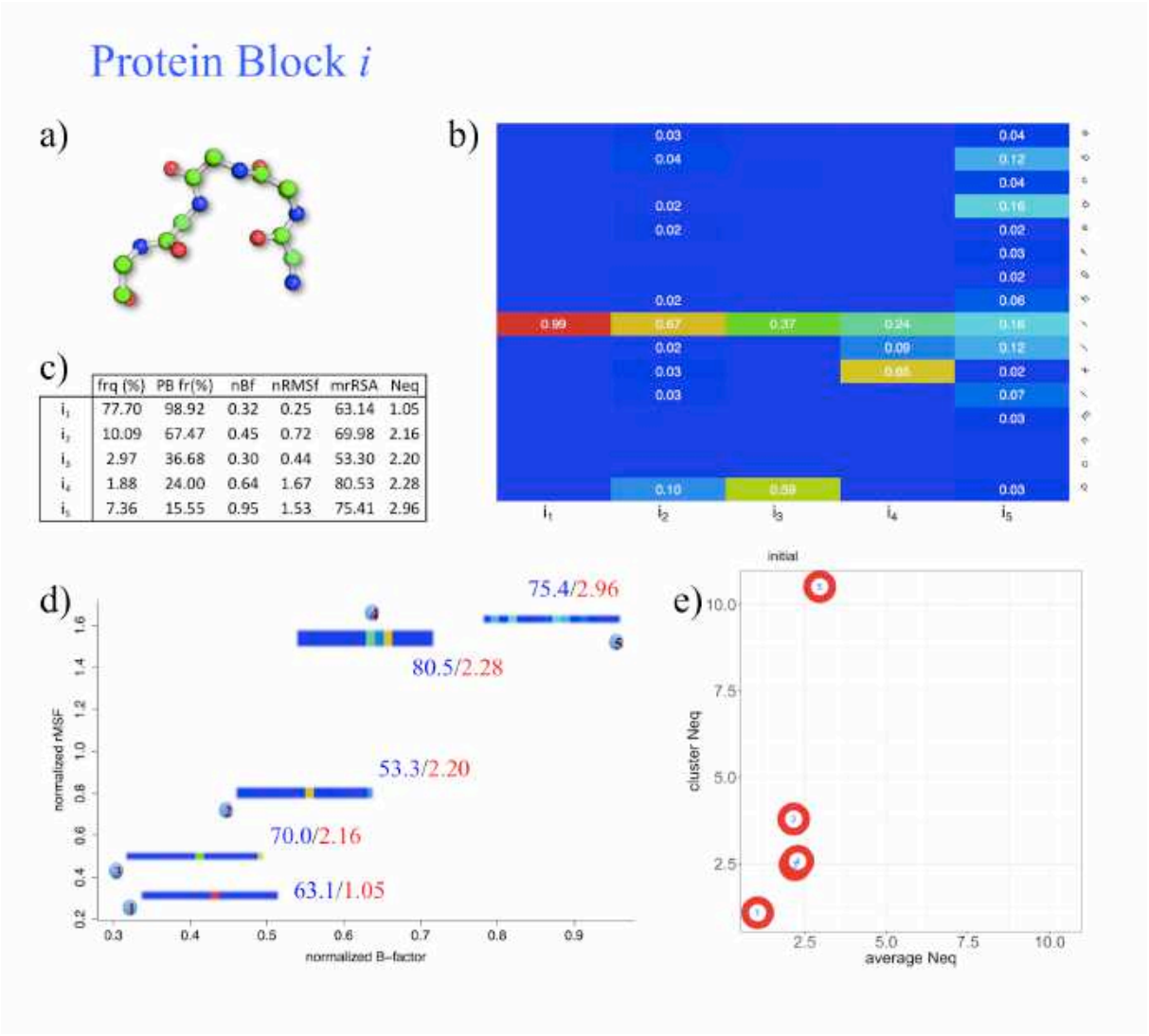
Protein Block *g* represents 1.15% of the protein structure dataset, it is 80.2% coil, 13.3% α-helix and 6.4 β-sheet (1). The major transitions as described in (2) were PB *h* (38%), PB *c* (30%) and PB *o* (16%).

Cluster g_1 (>89% of PB *g*) represents 55% of original PB *g* positions had lowest rSA and lowest nRMSf but the second lowest nBf. As seen in the first section on the Protein Blocks, PB g does not stay as PB *g* as often that other PBs. It is seen again here, the following clusters are only composed of 33%, 24%, 20% and 16% of PB *g*. The first surprise cluster is cluster g_2 directed by PB *e* (57%) that is quite comparable to cluster g_1 in terms of protein flexibility characteristics. Cluster g_3 was more expected as PB *c* is an expected geometrical transition; it represents 64% of the cluster. The second surprise cluster is so cluster g_4 controlled by PB *p* (52%). Cluster g_5 is the second most occurring cluster (15.1% of original PB), it encompass the more deformable regions with high average N_{eq} (2.61) and highest cluster N_{eq} (11.69) associated to high accessibility, highest nBf and highest nRMSf.



Protein Block *h* represents 2.4% of the protein structure dataset, it is 76.2% coil, 2.0% α-helix and 21.4 β-sheet (1). The major transitions as described in (2) were PB *i* (68%), PB *j* (14%) and PB *k* (9%).

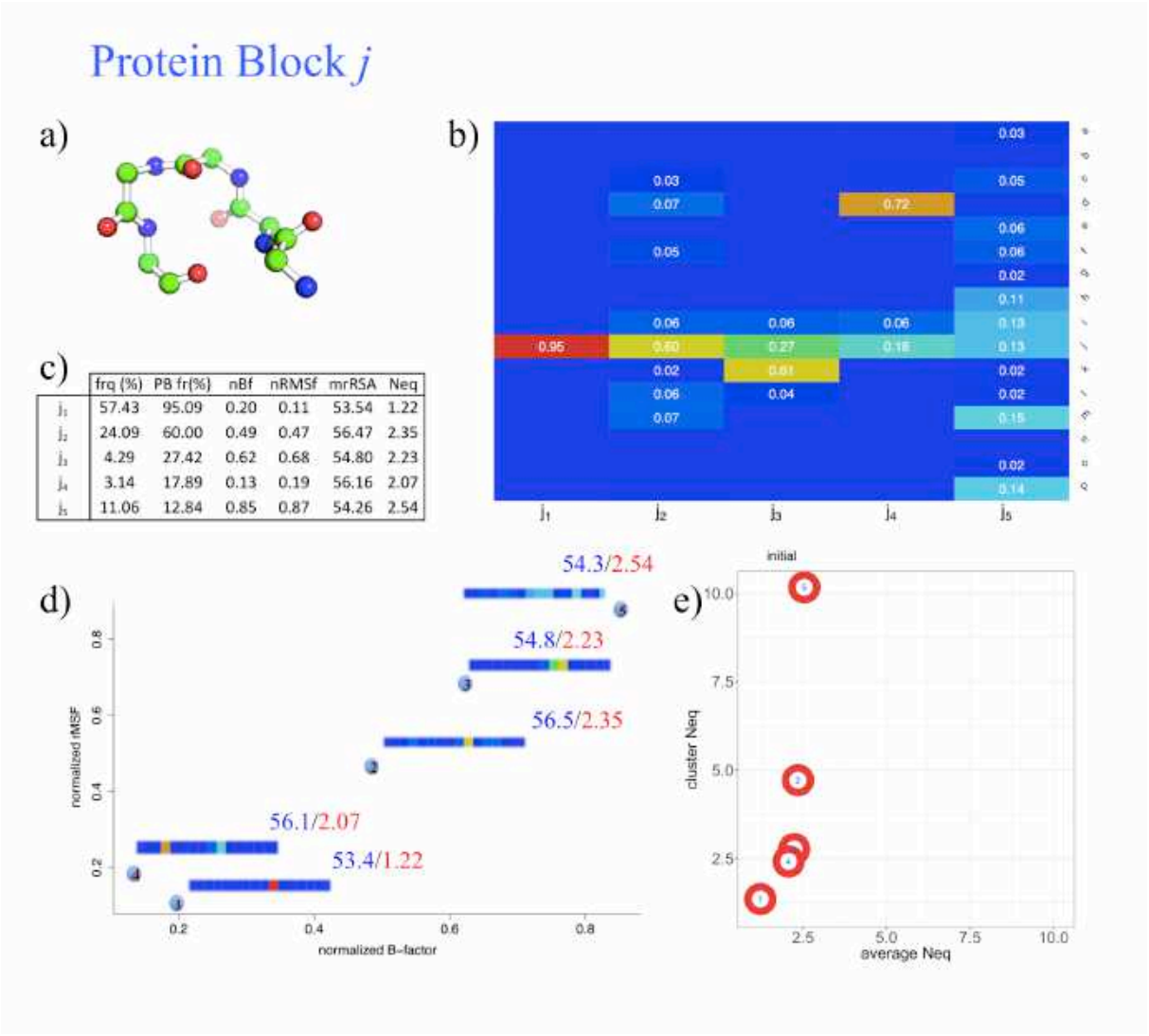
Cluster h_1 (>98% of PB *h*) represents 79% of original PB *g* positions had only the lowest nRMSf and the third lowest nBf and far beyond the first one (cluster h_3). It is the best nRMSf. Cluster g_2 corresponds to high PB *g* content (65%), but with a higher nRMSf. Cluster g_3 is controlled by compatible (geometrical transition) PB *k* content (65%), with a higher nRMSf nbut lower nBf. Cluster g_4 is controlled by unexpected PB *d* content (64%), associated to a higher nRMSf but a very low nBf. Cluster h_5 is the third most occurring cluster (5.8% of original PB), it encompass the more deformable regions with high average N_{eq} (2.84) and highest cluster N_{eq} (11.68) associated to high accessibility, highest nBf and highest nRMSf. It had high similarity with cluster g_5.



	frq (%)	PB fr(%)	nBf	nRMSf	mrRSA	Neq
i_1	77.70	98.92	0.32	0.25	63.14	1.05
i_2	10.09	67.47	0.45	0.72	69.98	2.16
i_3	2.97	36.68	0.30	0.44	53.30	2.20
i_4	1.88	24.00	0.64	1.67	80.53	2.28
i_5	7.36	15.55	0.95	1.53	75.41	2.96

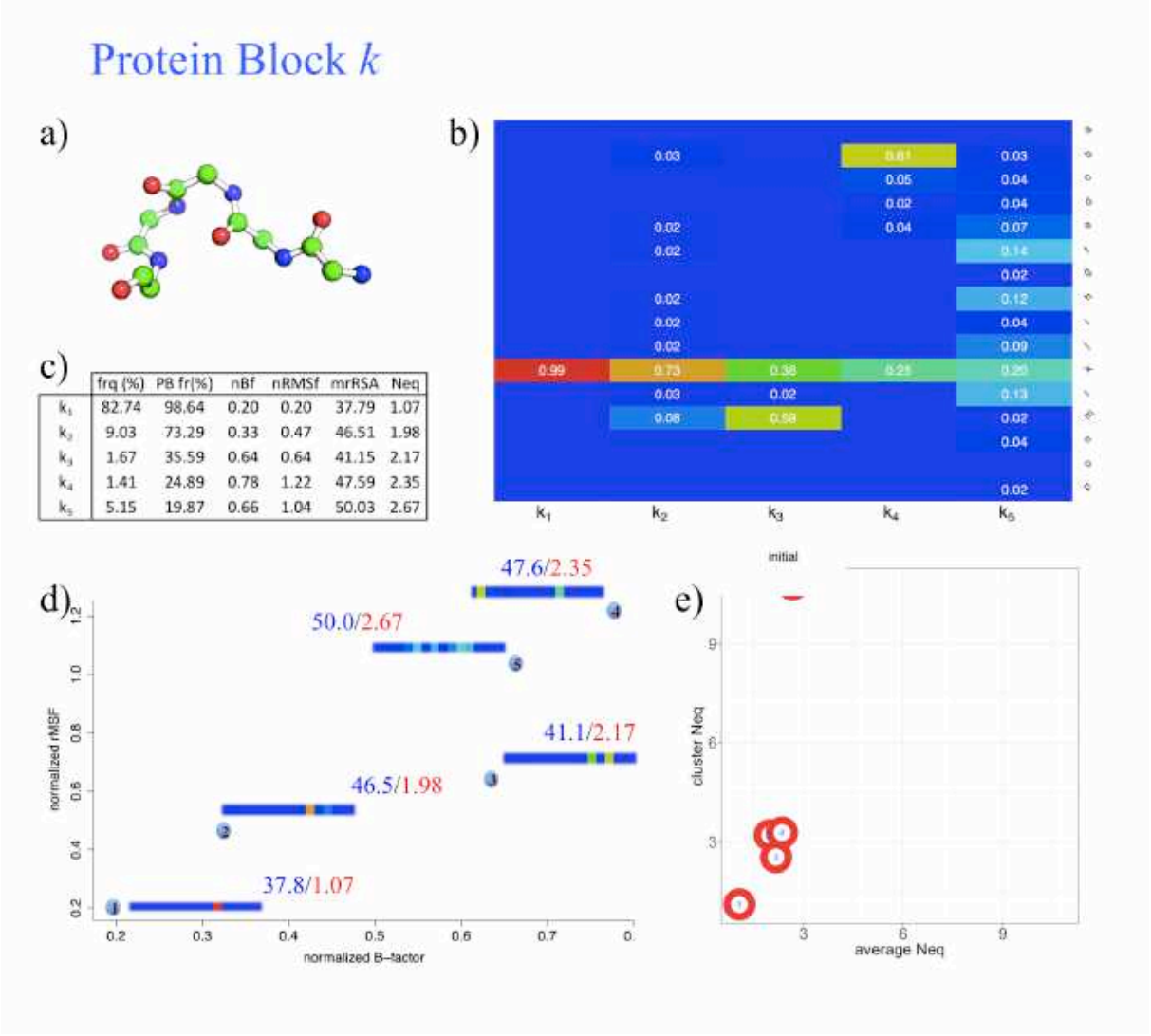
Protein Block *i* represents 1.9% of the protein structure dataset, it is 90.3% coil, 2.0% α-helix and 7.7% β-sheet (1). The major transitions as described in (2) were PB *a* (83%) and PB *l* (6%).

Cluster i_1 (>99% of PB *i*) represents 78% of original PB *i* positions had only the lowest nRMSf and nBf, but a higher rSA than cluster i_2. This last represents the second cluster with high PB frequency (PB *i* is 67%) here, but highly disperse after (9 PBs with frequency higher than 2%). It is so more deformable than cluster i_1 but with a lower accessibility. Surprisingly cluster i_3 is more accessible than the two first but (i) is quite comparable to cluster i_1 in terms of nBf and nRMSf, and (ii) is controlled by unexpected PB *p* (59%). Clusters i_4 and i_5 are very different but both with high nBf, high nRMSf and high rSA. While cluster i_4 is controlled by unexpected PB *k* (65%), cluster i_5 is associated to N_{eq} values close to those of clusters g_5 and h_5 (average N_{eq} of 2.96 and cluster N_{eq} of 10.51).



Protein Block *j* represents 0.83% of the protein structure dataset, it is 81.6% coil, 8.0% α-helix and 10.4% β-sheet (1). The major transitions as described in (2) were PB *b* (22%), PB *a* (15%) and PB *k* (15%). It is the fuzziest of all PBs are in fact only PBs *g* and *p* have no transition from it (2).

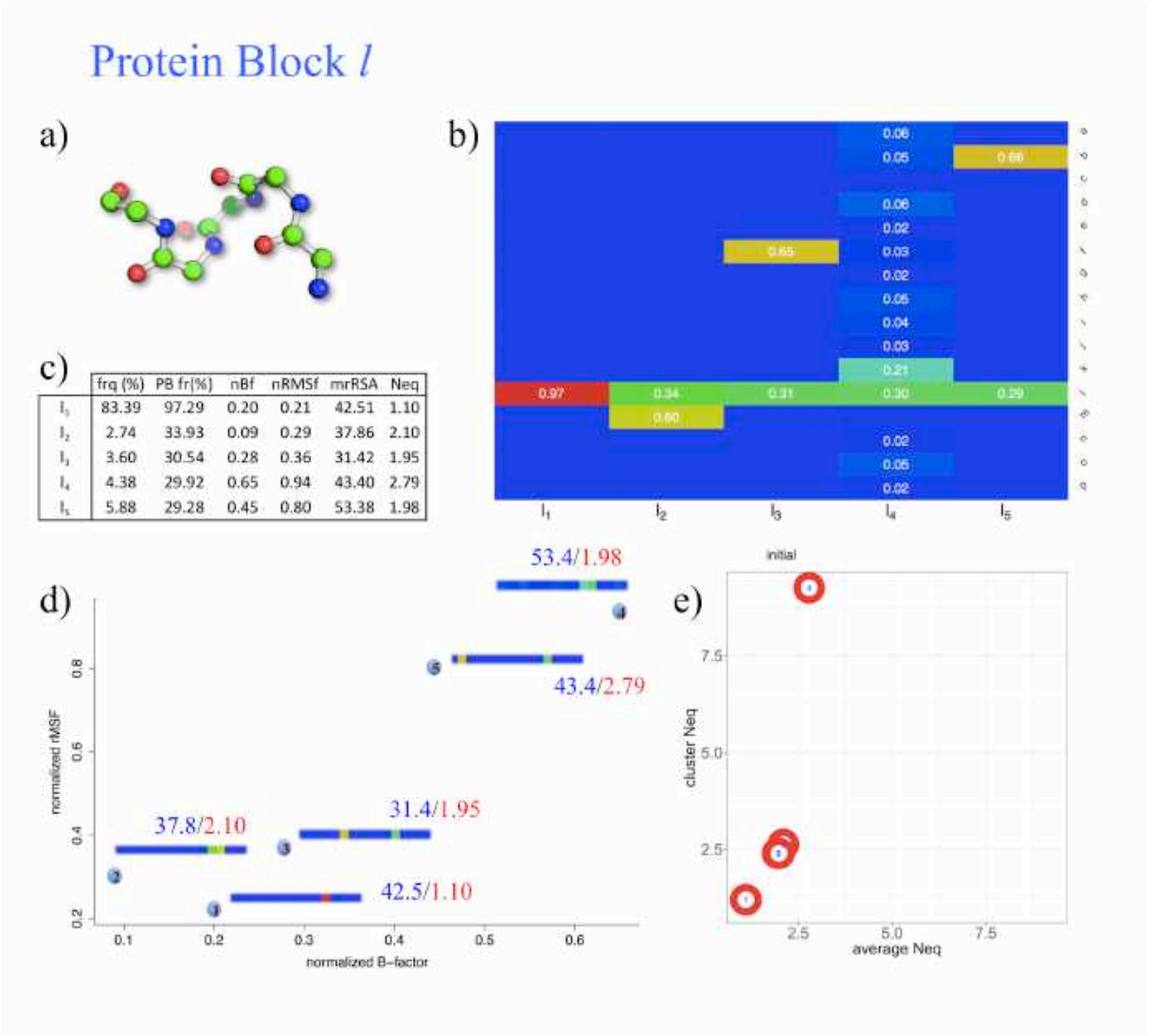
As seen in Supplementary Data 6, the relative solvent accessibility values of PB *j* are high. rSAs are similar in all 5 clusters. Cluster j_1 (>95% of PB *g*) represents 57% of original PB *j* positions had the lowest nRMSf and nBf as cluster j_4 that is directed by PB *d* (74%), something logical for this PB. Cluster j_2 is this typical type of second cluster with a high PB content of the original PB (here 60%) and large number of small frequencies for many PBs (7 PBs with a frequency higher than 2%). In a recurrent way as we have seen previous PBs, the most expected (geometrical transition) PBs, namely *a* and *b* are not found. Clusters j_3 and j_5 are very different but both with high nBf and high RMSf. While cluster j_3 is controlled by PB *k* (61%), cluster j_5 is associated to N_{eq} values close to the previous fifth (average N_{eq} of 2.54 and cluster N_{eq} of 10.17).



	frq (%)	PB fr(%)	nBf	nRMSf	mrRSA	Neq
k_1	82.74	98.64	0.20	0.20	37.79	1.07
k_2	9.03	73.29	0.33	0.47	46.51	1.98
k_3	1.67	35.59	0.64	0.64	41.15	2.17
k_4	1.41	24.89	0.78	1.22	47.59	2.35
k_5	5.15	19.87	0.66	1.04	50.03	2.67

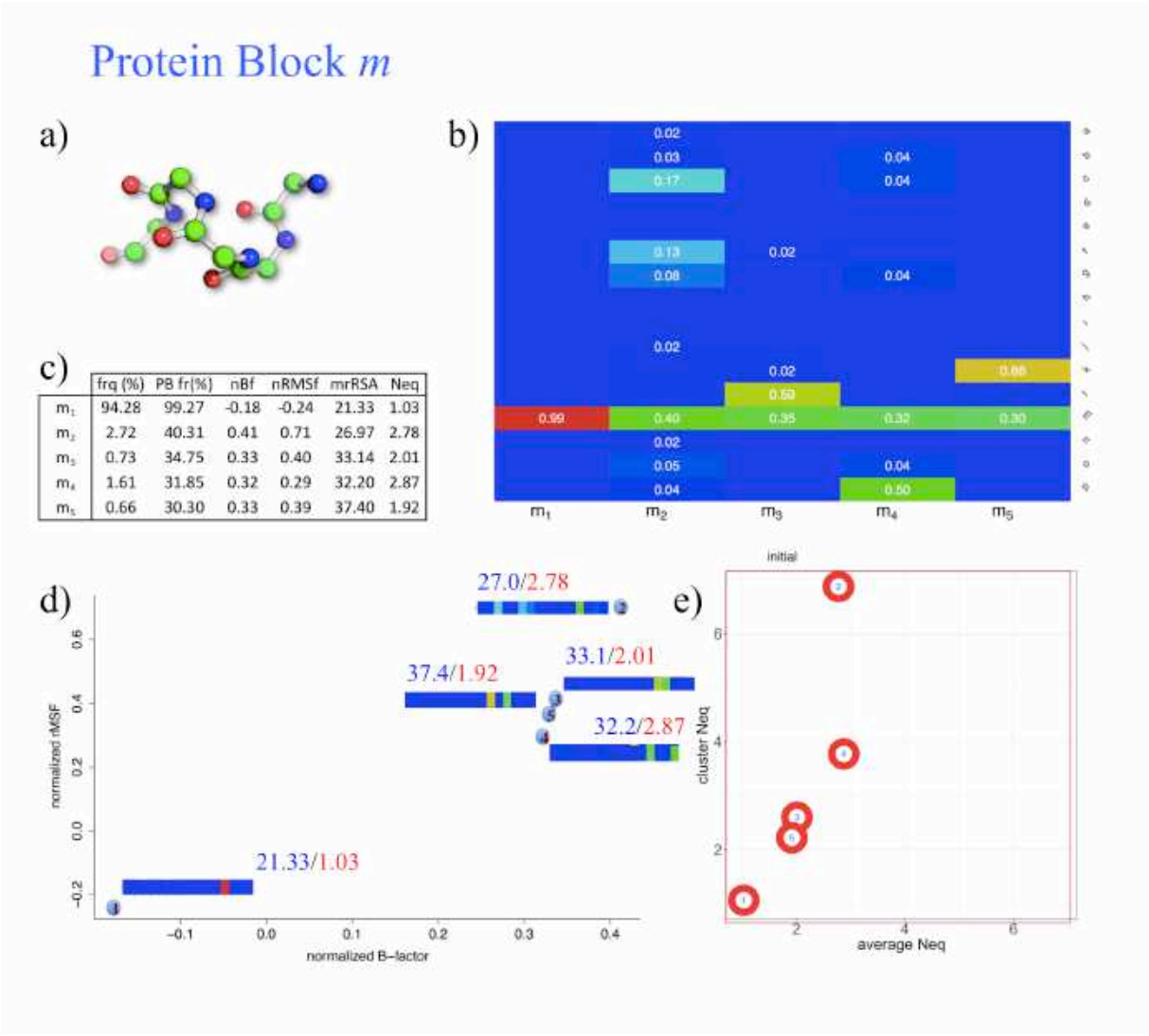
Protein Block *k* represents 5.45% of the protein structure dataset, it is 50.2% coil, 49.3% α-helix and 0.4% β-sheet (1). The major transitions as described in (2) were PB *l* (78%), PB *b* (11%) and PB *o* (6%).

Cluster k_1 (>98% of PB *k*) represents 83% of original PB *k* positions had the lowest nBf, the lowest nRMSf and the lowest rSA, while cluster k_2 is a degenerated version of it. Cluster k_2 (73% of PB *k*) has a large number of small frequencies for many PBs (8 PBs with a frequency higher than 2%) and higher nBf, lowest nRMSf and rSA that cluster k_1. Cluster k_3 is mainly directed by PB *m* (59%) but has a higher nBf and a medium nRMSf. Clusters k_4 and k_5 are very different but both with high nBf and high RMSf. While cluster k_4 is controlled by PB *b* (61%) that is geometrically compatible, cluster j_5 is associated to N_{eq} values close to the previous fifth (average N_{eq} of 2.67 and cluster N_{eq} of 10.80).



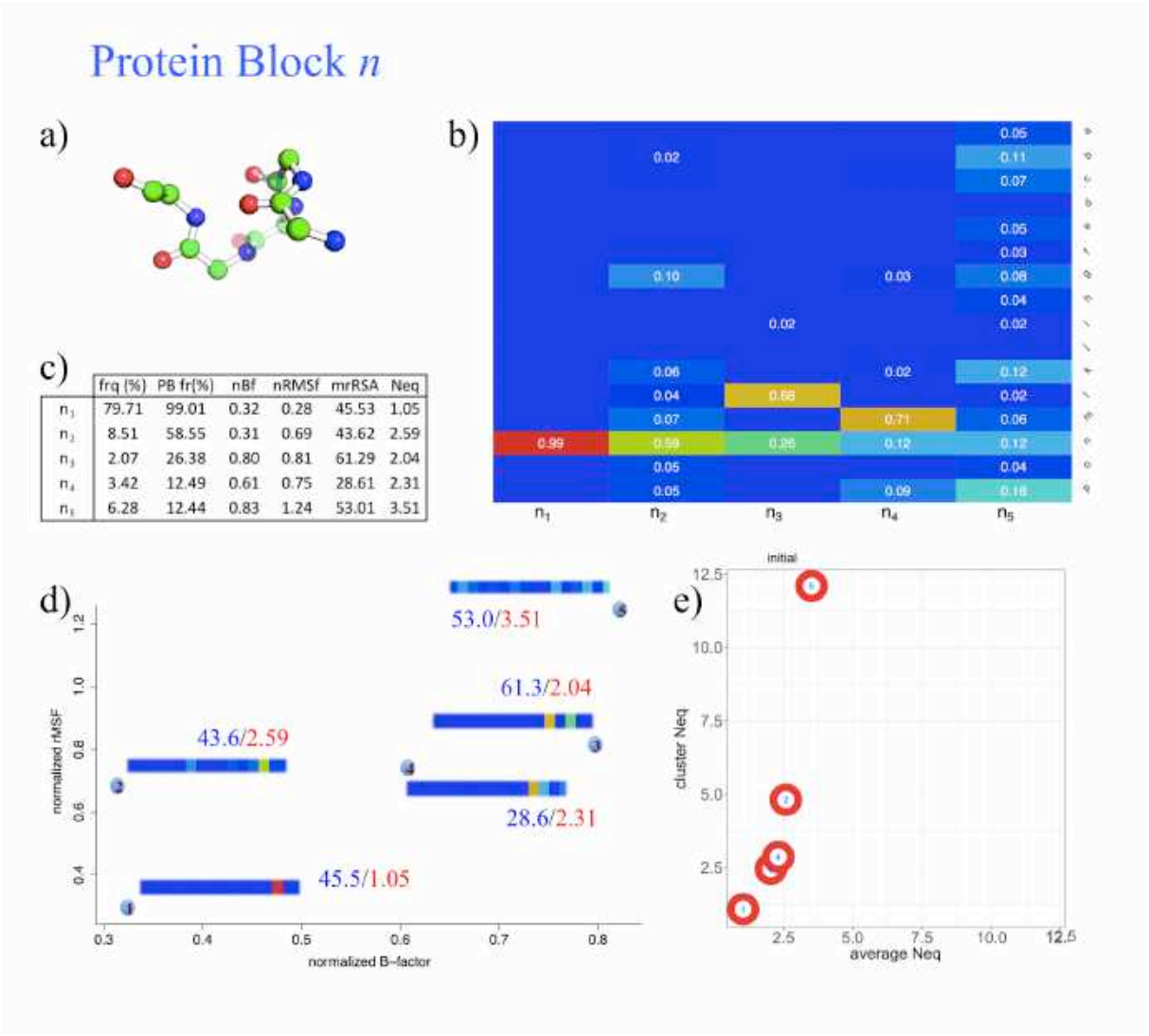
Protein Block *l* represents 5.46% of the protein structure dataset, it is 38.6% coil, 61.0% α-helix and 0.4% β-sheet (1). The major transitions as described in (2) were PB *m* (68%), PB *p* (9%) and PB *c* (7%).

Cluster l_1 (>97% of PB *l*) represents 83% of original PB *l* positions had the lowest nRMSf, the second lowest nBf and third lowest rSA. Three of the four remaining clusters are highly direct: (i) cluster l_2 (34% of PB *l*) is directed by expected PB *m* (60%), (ii) cluster l_3 (31% of PB *l*) is directed by un-expected PB *f* (65%), and (iii) cluster l_5 (29% of PB *l*) is directed by un-expected PB *b* (60%); this last is associated with highest nRMSf, highest nBf and highest rSA. Cluster k_4 is associated to N_{eq} values close to the previous fifth (average N_{eq} of 2.79 and cluster N_{eq} of 9.25).



Protein Block *m* represents 30% of the protein structure dataset, it is 7.6% coil and 92.3% α-helix (1). The major transitions as described in (2) were PB *n* (35%), PB *p* (16%) and PB *k* (11%).

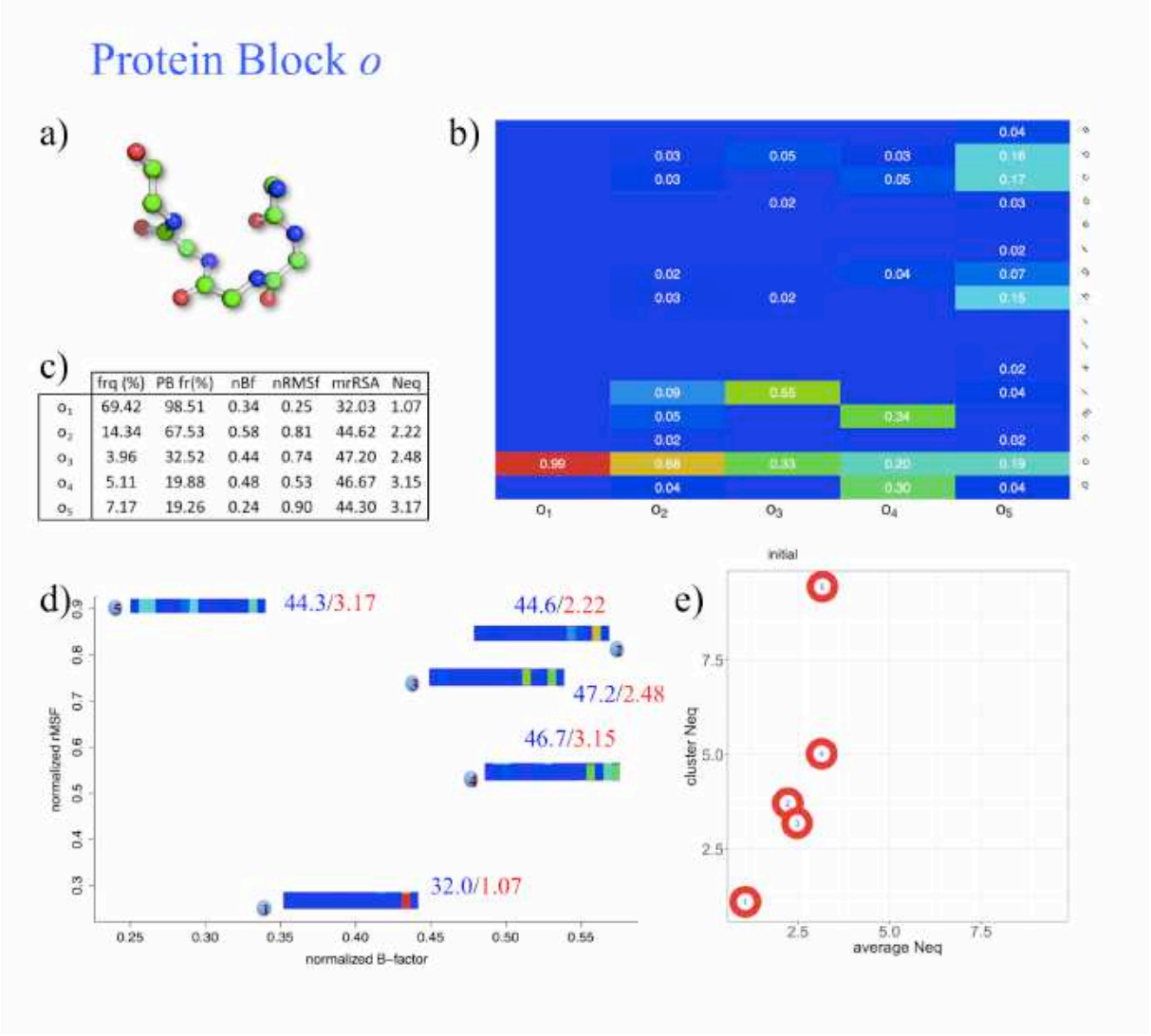
PB m is the most stable PB during dynamics (see Figure 5). So its main cluster, namely cluster m_1 (>99% of PB *m*) represents 94% of original PB *m* positions had the lowest nRMSf, the lowest nBf and the lowest rSA. The four others have clearly higher nRMSf, the higher nBf and higher rSA. Cluster m_2 is the one associated to high N_{eq} values (but quite lower than the previous fifth with average N_{eq} of 2.78 and cluster N_{eq} of 6.87). The three remaining clusters are highly direct: (i) cluster m_3 (35% of PB *m*) is directed by un-expected PB *l* (59%), (ii) cluster m_4 (32% of PB *m*) is directed by expected PB *p* (50%), and (iii) cluster m_5 (30% of PB *m*) is directed by expected PB *k* (66%); they are all associated to low N_{eq} values.



Protein Block *n* represents 2.0% of the protein structure dataset, it is 24.0% coil, 75.7% α-helix and 0.3% β-sheet (1). The major transition as described in (2) was PB *o* (92%).

PB *n* has some behaviour similar to PB m.

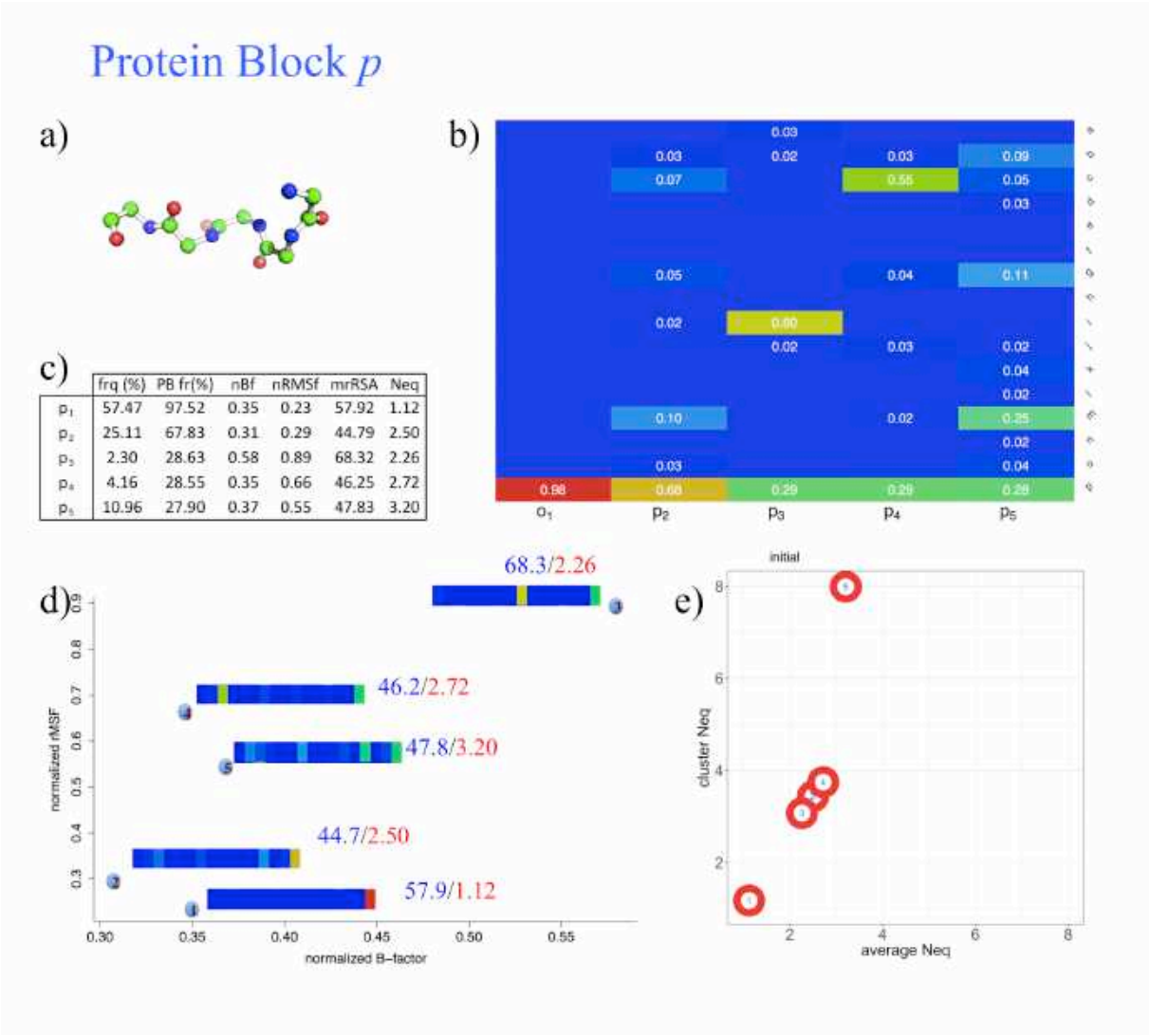
Cluster n_1 (>99% of PB *m*) represents 79% of original PB *n* positions had the lowest nRMSf, the lowest nBf and the lowest rSA. Cluster n_2 is a degenerated version of it with PB *n* frequency of 59% and Neq of 2.59. It was considered has highly rigid with nBf (such as cluster n_1), but its nRMSf is largely higher. The two following clusters, namely cluster n_3 and n_4, are controlled by PB l for the first one (68%) and PB m for the second (71%). Interestingly cluster n_4 has very lower rSA (highly comparable with previous PB m clusters). As seen with many PBs, the most expected (geometrical transition) PB *o* is not found. Cluster n_5 is associated to N_{eq} values close to the previous fifth (average N_{eq} of 3.51 and cluster N_{eq} of 12.09).



	frq (%)	PB fr(%)	nBf	nRMSf	mrRSA	Neq
o_1	69.42	98.51	0.34	0.25	32.03	1.07
o_2	14.34	67.53	0.58	0.81	44.62	2.22
o_3	3.96	32.52	0.44	0.74	47.20	2.48
o_4	5.11	19.88	0.48	0.53	46.67	3.15
o_5	7.17	19.26	0.24	0.90	44.30	3.17

Protein Block *g* represents 2.8% of the protein structure dataset, it is 49.0% coil, 50.8% α-helix and 0.2% β-sheet (1). The major transitions as described in (2) were PB *p* (78%), PB *m* (7%) and PB *i* (6%).

Cluster o_1 (>98% of PB *o*) represents 69% of original PB *o* positions had the lowest nRMSf, the second lowest nBf and the lowest rSA. Cluster o_2 is a degenerated version of it with PB *o* frequency of 68% and Neq of 2.22. Surprisingly, it is associated to very high nRMSf and nBf values and a high rSA. It is so clearly different. But a really more surprising cluster is cluster n_5 that represents 7.2% of the PB *o* initial observations and is highly fuzzy as its N_{eq} is of 3.17, but it is also associated to the lowest nBf of this PB. Cluster o_3 (PB *o* 33%) is mainly controlled by un-expected PB *l* (55%), and cluster o_4 (PB *o* 20%) by expected PB *m* (34%) and PB p (30%).



Protein Block *p* represents 3.5% of the protein structure dataset, it is 81.3% coil, 17.1% α-helix and 1.6 β-sheet (1). The major transitions as described in (2) were PB *a* (59%), PB *c* (24%) and PB *m* (8%).

The two clusters associated with high PB *p* contents (cluster p_1 and p_2, >97% and 68%, resp.) have low nBf and low nRMSf. The first is associated to a largely higher rSA (57.9 vs. 44.7). Surprisingly the most deformable cluster, namely cluster p_3, is directed by un-expected PB *i* (60%), it is associated to very high rSA values, classical for this last PB. Cluster p_4, is directed by expected PB *c* (55%) and is slightly more deformable than cluster p_5, is directed is associated to N_{eq} values close to the highest values (average N_{eq} of 3.20 and cluster N_{eq} of 7.99).

References

1. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins. 1995;23(4):566-79.

2. de Brevern AG. New assessment of a structural alphabet. In Silico Biol. 2005;5(3):283-9.