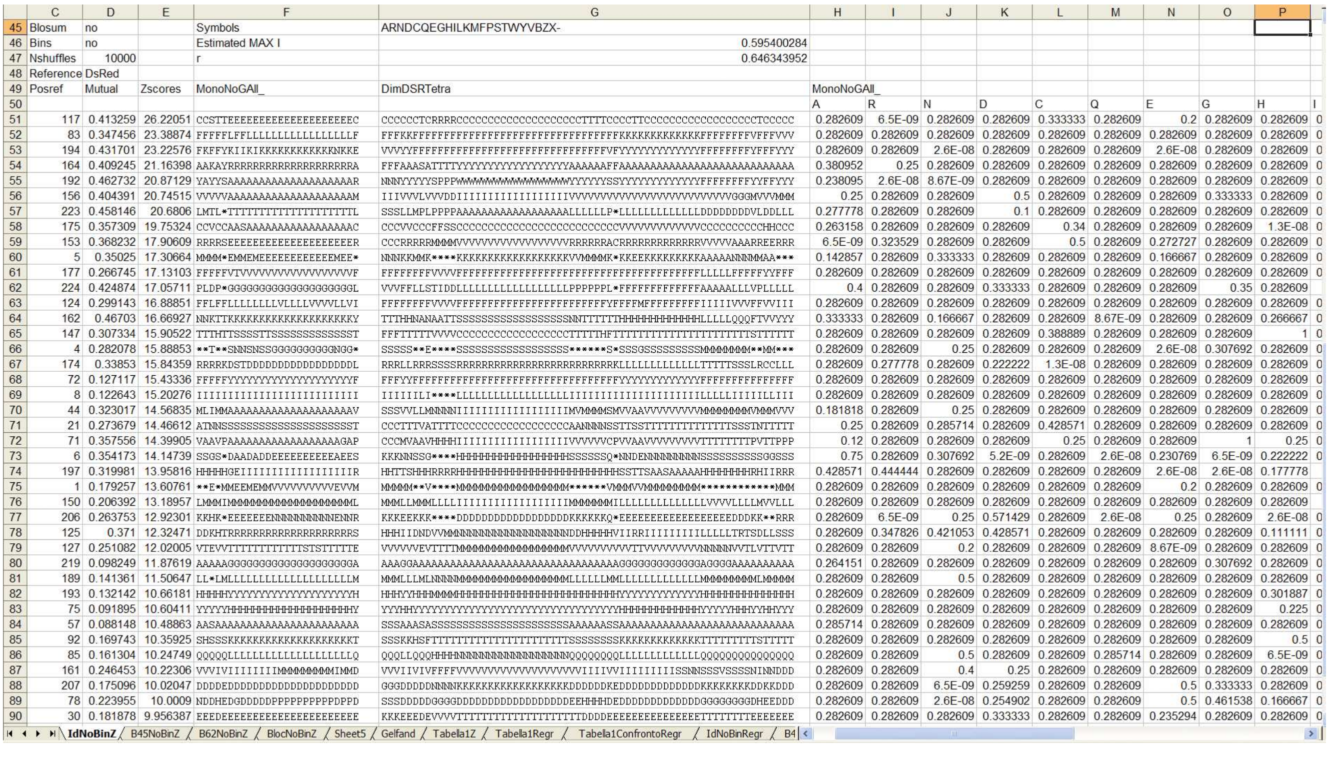
Supplementary Information

SDPhound, a Mutual Information-Based Method to Investigate Specificity-Determining Positions

Sara Bonella [1,†,‡], Walter Rocchia [1,2,†,⋆], Pietro Amat [1], Riccardo Nifosí [1] and Valentina Tozzini [1]

1. XLS and HTML reports

Figure 1. Snapshot of the Excel worksheet created by SDPhound in the case of a single position run.

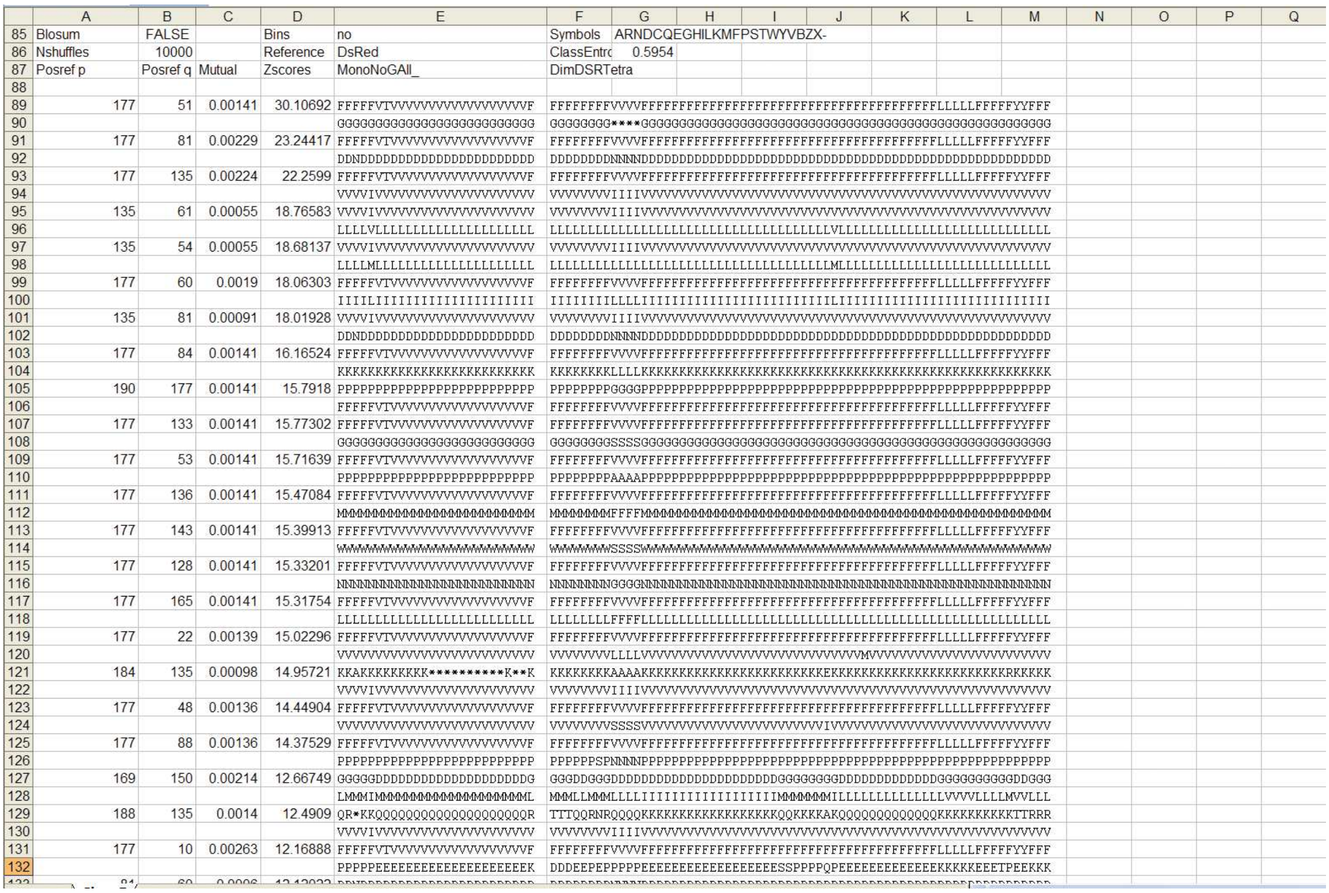


	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
45	Blosum	no		Symbols	ARNDCQEGHILKMFPSTWYVBZX-										
46	Bins	no		Estimated MAX I	0.595400284										
47	Nshuffles	10000		r	0.646343952										
48	Reference	DsRed													
49	Posref	Mutual	Zscores	MonoNoGAll_	DimDSRTetra	MonoNoGAll_									
50						A	R	N	D	C	Q	E	G	H	I
51	117	0.413259	26.22051	CCSTTEEEEEEEEEEEEEEEEEEEC	CCCCCCTCRRRRCCCCCCCCCCCCCCCCCCCCTTTTCCCCTTCCCCCCCCCCCCCCCCCCCCTCCCCC	0.282609	6.5E-09	0.282609	0.282609	0.333333	0.282609	0.2	0.282609	0.282609	0
52	83	0.347456	23.38874	FFFFFLFFLLLLLLLLLLLLLLLLF	FFFKKFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFKKKKKKKKKKKKKFFFFFFFFFVFFFVVV	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
53	194	0.431701	23.22576	FKFFYKIIKIKKKKKKKKKKKNKKE	VVVYYFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFVFYYYYYYYYYYYYYFFFFFFFFFYFFFYYY	0.282609	0.282609	2.6E-08	0.282609	0.282609	0.282609	2.6E-08	0.282609	0.282609	0
54	164	0.409245	21.16398	AAKAYRRRRRRRRRRRRRRRRRRRA	FFFAAASATTTTYYYYYYYYYYYYYYYYYYYAAAAAAFFAAAAAAAAAAAAAAAAAAAAAAAAAAAA	0.380952	0.25	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
55	192	0.462732	20.87129	YAYYSAAAAAAAAAAAAAAAAAAAR	NNNYYYYYSPPPWWWWWWWWWWWWWWWWWWWWYYYYYYSSYYYYYYYYYYYYYYFFFFFFFFYYFFYYY	0.238095	2.6E-08	8.67E-09	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
56	156	0.404391	20.74515	VVVVVAAAAAAAAAAAAAAAAAAAM	IIIVVVLVVVDDIIIIIIIIIIIIIIIIIIVVVVVVVVVVVVVVVVVVVVVVVVVVVVGGGMVVVMMM	0.25	0.282609	0.282609	0.5	0.282609	0.282609	0.282609	0.333333	0.282609	0
57	223	0.458146	20.6806	LMTL*TTTTTTTTTTTTTTTTTTTL	SSSLLMPLPPPPAAAAAAAAAAAAAAAAAAALLLLLLP*LLLLLLLLLLLLDDDDDDDDVLDDLLL	0.277778	0.282609	0.282609	0.1	0.282609	0.282609	0.282609	0.282609	0.282609	0
58	175	0.357309	19.75324	CCVCCAASAAAAAAAAAAAAAAAAC	CCCVVCCCFFSSCCCCCCCCCCCCCCCCCCCCCCCCCCVVVVVVVVVVVVCCCCCCCCCCHHCCC	0.263158	0.282609	0.282609	0.282609	0.34	0.282609	0.282609	0.282609	1.3E-08	0
59	153	0.368232	17.90609	RRRRSEEEEEEEEEEEEEEEEEEER	CCCRRRRRMMMMVVVVVVVVVVVVVVVVVVVVRRRRRRACRRRRRRRRRRRRRVVVVVAAARREERRR	6.5E-09	0.323529	0.282609	0.282609	0.5	0.282609	0.272727	0.282609	0.282609	0
60	5	0.35025	17.30664	MMMM*EMMEMEEEEEEEEEEEMEE*	NNNKKMMK****KKKKKKKKKKKKKKKKKKVVMMMMK*KKEEKKKKKKKKKKAAAAANNNMMAA***	0.142857	0.282609	0.333333	0.282609	0.282609	0.282609	0.166667	0.282609	0.282609	0
61	177	0.266745	17.13103	FFFFFVTVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
62	224	0.424874	17.05711	PLDP*GGGGGGGGGGGGGGGGGGGL	VVVFFLLSTIDDLLLLLLLLLLLLLLLLLLLPPPPPPL*FFFFFFFFFFFFFAAAAALLLVPLLLLL	0.4	0.282609	0.282609	0.333333	0.282609	0.282609	0.282609	0.35	0.282609	
63	124	0.299143	16.88851	FFLFFLLLLLLLVLLLLVVVVLLVI	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFYFFFFMFFFFFFFFFIIIIIVVVFFVVIII	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
64	162	0.46703	16.66927	NNKTTKKKKKKKKKKKKKKKKKKKY	TTTHHNANAATTSSSSSSSSSSSSSSSSSSSNNTTTTTTHHHHHHHHHHHHHHLLLLLQQQFTVVYYY	0.333333	0.282609	0.166667	0.282609	0.282609	8.67E-09	0.282609	0.282609	0.266667	0
65	147	0.307334	15.90522	TTTHTTSSSSTTSSSSSSSSSSSST	FFFTTTTTVVVVCCCCCCCCCCCCCCCCCCCCTTTTTHFTTTTTTTTTTTTTTTTTTTTTTTTSTTTTTT	0.282609	0.282609	0.282609	0.282609	0.388889	0.282609	0.282609	0.282609	1	0
66	4	0.282078	15.88853	**T**SNNSNSSGGGGGGGGGGNGG*	SSSSS**E****SSSSSSSSSSSSSSSSSSS******S*SSSGSSSSSSSSSSMMMMMMMM**MM***	0.282609	0.282609	0.25	0.282609	0.282609	0.282609	2.6E-08	0.307692	0.282609	0
67	174	0.33853	15.84359	RRRRKDSTDDDDDDDDDDDDDDDDL	RRRLLRRRSSSSRRRRRRRRRRRRRRRRRRRRRRRRRRRKLLLLLLLLLLLLLTTTTTSSSLRCCLLL	0.282609	0.277778	0.282609	0.222222	1.3E-08	0.282609	0.282609	0.282609	0.282609	0
68	72	0.127117	15.43336	FFFFFYYYYYYYYYYYYYYYYYYYF	FFFYYFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFYYYYYYYYYYYYYYFFFFFFFFFFFFFF	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
69	8	0.122643	15.20276	IIIIIIIIIIIIIIIIIIIIIIIII	IIIIIILI****LLLLLLLLLLLLLLLLLLLIIIIIIIIIIIIIIIIIIIIIIIIIILLLLLIIIIIILLIII	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
70	44	0.323017	14.56835	MLIMMAAAAAAAAAAAAAAAAAAAV	SSSVVLLMNNNNIIIIIIIIIIIIIIIIIIMVMMMMSMVVAAVVVVVVVVVVMMMMMMMMVMMMVVV	0.181818	0.282609	0.25	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
71	21	0.273679	14.46612	ATNNSSSSSSSSSSSSSSSSSSSST	CCCTTTVATTTTCCCCCCCCCCCCCCCCCCCAANNNNSSTTSSTTTTTTTTTTTTTTTTSSSTNTTTTT	0.25	0.282609	0.285714	0.282609	0.428571	0.282609	0.282609	0.282609	0.282609	0
72	71	0.357556	14.39905	VAAVPAAAAAAAAAAAAAAAAAGAP	CCCMVAAVHHHHIIIIIIIIIIIIIIIIIIVVVVVVCPVVAAVVVVVVVVVVTTTTTTTTTPVTTPPP	0.12	0.282609	0.282609	0.282609	0.25	0.282609	0.282609	1	0.25	0
73	6	0.354173	14.14739	SSGS*DAADADDEEEEEEEEEEAEES	KKKNNSSG****HHHHHHHHHHHHHHHHHHHSSSSSSQ*NNDENNNNNNNNNNSSSSSSSSSSGGSSS	0.75	0.282609	0.307692	5.2E-09	0.282609	2.6E-08	0.230769	6.5E-09	0.222222	0
74	197	0.319981	13.95816	HHHHHGEIIIIIIIIIIIIIIIIIR	HHTTSHHHRRRRHHHHHHHHHHHHHHHHHHHHHHHHHHSSTTSAASAAAAAHHHHHHHHRHIIRRR	0.428571	0.444444	0.282609	0.282609	0.282609	0.282609	2.6E-08	2.6E-08	0.177778	
75	1	0.179257	13.60761	**E*MMEEMEMMVVVVVVVVVVEVVM	MMMMM**V****MMMMMMMMMMMMMMMMMM******VMMMVVMMMMMMMMMM***********MMM	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.2	0.282609	0.282609	0
76	150	0.206392	13.18957	LMMMIMMMMMMMMMMMMMMMMMMML	MMMLLMMMLLLLIIIIIIIIIIIIIIIIIIIMMMMMMILLLLLLLLLLLLLLVVVVLLLLMVVLLL	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	
77	206	0.263753	12.92301	KKHK*EEEEEEENNNNNNNNNENNR	KKKEEKKK****DDDDDDDDDDDDDDDDDDDKKKKKKQ*EEEEEEEEEEEEEEEEEEEEDDDKK**RRR	0.282609	6.5E-09	0.25	0.571429	0.282609	2.6E-08	0.25	0.282609	2.6E-08	0
78	125	0.371	12.32471	DDKHTRRRRRRRRRRRRRRRRRRRS	HHHIIDNDVVMMNNNNNNNNNNNNNNNNNNNNNDDHHHHHVIIRRIIIIIIIIIIILLLLLTRTSDLLSSS	0.282609	0.347826	0.421053	0.428571	0.282609	0.282609	0.282609	0.282609	0.111111	0
79	127	0.251082	12.02005	VTEVVTTTTTTTTTTTSTSTTTTTE	VVVVVVEVTTTTMMMMMMMMMMMMMMMMMMVVVVVVVVVVVTTVVVVVVVVVVNNNNNVVTLVTTVTT	0.282609	0.282609	0.2	0.282609	0.282609	0.282609	8.67E-09	0.282609	0.282609	0
80	219	0.098249	11.87619	AAAAAGGGGGGGGGGGGGGGGGGGA	AAAGGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGGGGGGGGGGGGAGGGGAAAAAAAAAAA	0.264151	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.307692	0.282609	0
81	189	0.141361	11.50647	LL*LMLLLLLLLLLLLLLLLLLLLM	MMMLLLMLNNNNMMMMMMMMMMMMMMMMMMLLLLLLMMLLLLLLLLLLLLLMMMMMMMMMLMMMMM	0.282609	0.282609	0.5	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
82	193	0.132142	10.66181	HHHHHYYYYYYYYYYYYYYYYYYYH	HHHYYHHHMMMMHHHHHHHHHHHHHHHHHHHHHHHHHHHYYYYYYYYYYYYYYHHHHHHHHHHHHHH	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.301887	0
83	75	0.091895	10.60411	YYYYYHHHHHHHHHHHHHHHHHHHY	YYYHHYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYHHHHHHHHHHHHHYYYYYHHHYYHHYYY	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.225	0
84	57	0.088148	10.48863	AASAAAAAAAAAAAAAAAAAAAAAA	SSSAAASASSSSSSSSSSSSSSSSSSSSSSSSAAAAAASSAAAAAAAAAAAAAAAAAAAAAAAAAAAA	0.285714	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0
85	92	0.169743	10.35925	SHSSSKKKKKKKKKKKKKKKKKKKT	SSSKKHSFTTTTTTTTTTTTTTTTTTTTTTTTSSSSSSSSKKKKKKKKKKKKKTTTTTTTTTTSTTTTT	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.282609	0.5	0
86	85	0.161304	10.24749	QQQQQLLLLLLLLLLLLLLLLLLLQ	QQQLLQQQHHHHNNNNNNNNNNNNNNNNNNNQQQQQQQQLLLLLLLLLLLLLQQQQQQQQQQQQQQQ	0.282609	0.282609	0.5	0.282609	0.282609	0.285714	0.282609	0.282609	6.5E-09	0
87	161	0.246453	10.22306	VVVIVIIIIIIIMMMMMMMMIMMD	VVVIIVIVFFFFVVVVVVVVVVVVVVVVVVIIIIVVIIIIIIIIIIISSNNSSSVSSSSNINNDDD	0.282609	0.282609	0.4	0.25	0.282609	0.282609	0.282609	0.282609	0.282609	0
88	207	0.175096	10.02047	DDDDEDDDDDDDDDDDDDDDDDDDD	GGGDDDDDNNNNKKKKKKKKKKKKKKKKKKKDDDDDDKEDDDDDDDDDDDDDDKKKKKKKKDDKKDDD	0.282609	0.282609	6.5E-09	0.259259	0.282609	0.282609	0.5	0.333333	0.282609	0
89	78	0.223955	10.0009	NDDHEDGDDDDDPPPPPPPPPPDPPD	SSSDDDDDGGGGDDDDDDDDDDDDDDDDDDDDEEHHHDEDDDDDDDDDDDDDDGGGGGGGGDHEEDDD	0.282609	0.282609	2.6E-08	0.254902	0.282609	0.282609	0.5	0.461538	0.166667	0
90	30	0.181878	9.956387	EEEDEEEEEEEEEEEEEEEEEEEEE	KKKEEEDEVVVVTTTTTTTTTTTTTTTTTTTTDDDDEEEEEEEEEEEEEEEETTTTTTTTTEEEEEEE	0.282609	0.282609	0.282609	0.333333	0.282609	0.282609	0.235294	0.282609	0.282609	0

IdNoBinZ / B45NoBinZ / B62NoBinZ / BlocNoBinZ / Sheet5 / Gelfand / Tabella1Z / Tabella1Regr / Tabella1ConfrontoRegr / IdNoBinRegr / B4

Results of the application of the algorithm are reported in various forms, one of them is an Excel Worksheet that contains all the relevant information related to the run. Best ranking positions and run parameters are shown as well as the estimation of the conditional probability of belonging to a specificity class given that any specific symbol, amino acid or "pigeonhole", is found at the current position. Conditional probability is estimated from the alignment itself in the frequentist approximation. In case where the identity substitution matrix is used, a "-1.0" can appear in the conditional probability cells, indicating that the specific amino acid is not present in the alignment and therefore no information can be directly inferred for that mutation. Two typical examples are shown in Figure 1 and 2.

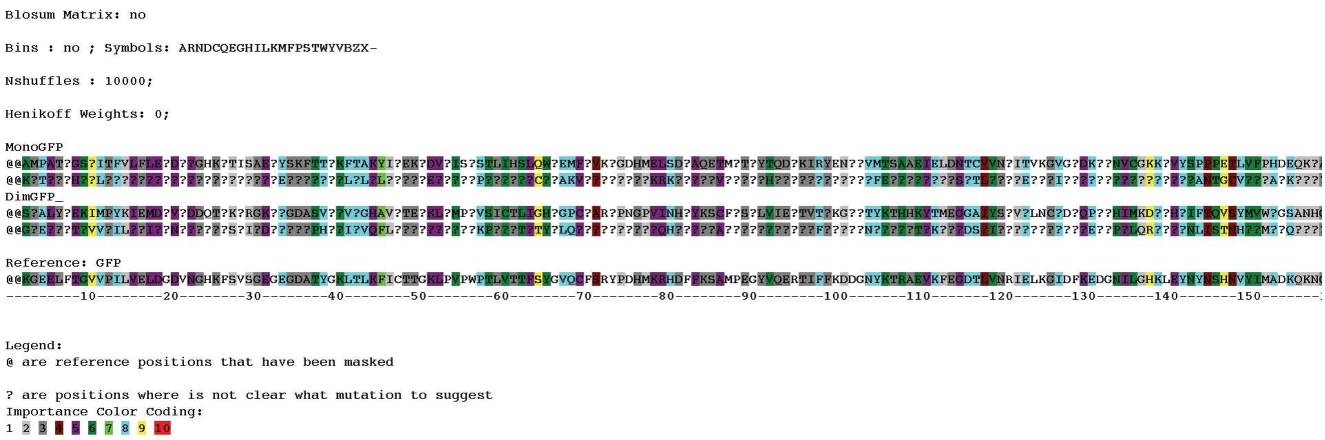
Figure 2. Snapshot of the Excel worksheet created by SDPhound in the case of a pairwise position correlation run.



	A	B	C	D	E	F	G
85	Blosum	FALSE		Bins	no	Symbols	ARNDCQEGHILKMFPSTWYVBZX-
86	Nshuffles	10000		Reference	DsRed	ClassEntrc	0.5954
87	Posref p	Posref q	Mutual	Zscores	MonoNoGAll_	DimDSRTetra	
88							
89	177	51	0.00141	30.10692	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
90					GGGGGGGGGGGGGGGGGGGGGGGGGG	GGGGGGGG****GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	
91	177	81	0.00229	23.24417	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
92					DDNDDDDDDDDDDDDDDDDDDDDDDD	DDDDDDDDNNNNDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD	
93	177	135	0.00224	22.2599	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
94					VVVVIVVVVVVVVVVVVVVVVVVVVV	VVVVVVVVIIIIVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV	
95	135	61	0.00055	18.76583	VVVVIVVVVVVVVVVVVVVVVVVVVV	VVVVVVVVIIIIVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV	
96					LLLLVLLLLLLLLLLLLLLLLLLLLL	LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLVLLLLLLLLLLLLLLLLLLLLLLLLLLLL	
97	135	54	0.00055	18.68137	VVVVIVVVVVVVVVVVVVVVVVVVVV	VVVVVVVVIIIIVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV	
98					LLLLMLLLLLLLLLLLLLLLLLLLLL	LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLMLLLLLLLLLLLLLLLLLLLLLLLLLLLL	
99	177	60	0.0019	18.06303	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
100					IIIILIIIIIIIIIIIIIIIIIIIII	IIIIIIIILLLLIIIIIIIIIIIIIIIIIIIIIIIILIIIIIIIIIIIIIIIIIIIIIIIIIIIII	
101	135	81	0.00091	18.01928	VVVVIVVVVVVVVVVVVVVVVVVVVV	VVVVVVVVIIIIVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV	
102					DDNDDDDDDDDDDDDDDDDDDDDDDD	DDDDDDDDNNNNDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD	
103	177	84	0.00141	16.16524	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
104					KKKKKKKKKKKKKKKKKKKKKKKKKK	KKKKKKKKLLLLKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKK	
105	190	177	0.00141	15.7918	PPPPPPPPPPPPPPPPPPPPPPPPPP	PPPPPPPPGGGGPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP	
106					FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
107	177	133	0.00141	15.77302	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
108					GGGGGGGGGGGGGGGGGGGGGGGGGG	GGGGGGGGSSSSGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	
109	177	53	0.00141	15.71639	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
110					PPPPPPPPPPPPPPPPPPPPPPPPPP	PPPPPPPPAAAAPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP	
111	177	136	0.00141	15.47084	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
112					MMMMMMMMMMMMMMMMMMMMMMMMMM	MMMMMMMMFFFFMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM	
113	177	143	0.00141	15.39913	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
114					WWWWWWWWWWWWWWWWWWWWWWWWWW	WWWWWWWWSSSSWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW	
115	177	128	0.00141	15.33201	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
116					NNNNNNNNNNNNNNNNNNNNNNNNNN	NNNNNNNNGGGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	
117	177	165	0.00141	15.31754	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
118					LLLLLLLLLLLLLLLLLLLLLLLLLL	LLLLLLLLFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL	
119	177	22	0.00139	15.02296	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
120					VVVVVVVVVVVVVVVVVVVVVVVVVV	VVVVVVVVLLLLVVVVVVVVVVVVVVVVVVVVVVVVVVVVVMVVVVVVVVVVVVVVVVVVVVVV	
121	184	135	0.00098	14.95721	KKAKKKKKKKKK**********K**K	KKKKKKKKAAAAKKKKKKKKKKKKKKKKKKKKEKKKKKKKKKKKKKKKKKKKKKKKKKKKKRKKKK	
122					VVVVIVVVVVVVVVVVVVVVVVVVVV	VVVVVVVVIIIIVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV	
123	177	48	0.00136	14.44904	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
124					VVVVVVVVVVVVVVVVVVVVVVVVVV	VVVVVVVVSSSSVVVVVVVVVVVVVVVVVVVVVVVVIVVVVVVVVVVVVVVVVVVVVVVVVVVVV	
125	177	88	0.00136	14.37529	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
126					PPPPPPPPPPPPPPPPPPPPPPPPPP	PPPPPPSPNNNNPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP	
127	169	150	0.00214	12.66749	GGGGGDDDDDDDDDDDDDDDDDDDDG	GGGDDGGGDDDDDDDDDDDDDDDDDDDDDDGGGGGGGDDDDDDDDDDDDDDGGGGGGGGGGDDGGG	
128					LMMMIMMMMMMMMMMMMMMMMMMMML	MMMLLMMMLLLLIIIIIIIIIIIIIIIIIIMMMMMMMILLLLLLLLLLLLLLVVVVLLLLMVVLLL	
129	188	135	0.0014	12.4909	QR*KKQQQQQQQQQQQQQQQQQQQQR	TTTQQRNRQQQQKKKKKKKKKKKKKKKKKKQQKKKAKQQQQQQQQQQQQQKKKKKKKKKKTTRRR	
130					VVVVIVVVVVVVVVVVVVVVVVVVVV	VVVVVVVVIIIIVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV	
131	177	10	0.00263	12.16888	FFFFFVTVVVVVVVVVVVVVVVVVVF	FFFFFFFFVVVVFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLFFFFFYYFFF	
132					PPPPPEEEEEEEEEEEEEEEEEEEEK	DDDEEPEPPPPPEEEEEEEEEEEEEEEEEESSPPPPQPEEEEEEEEEEEEEKKKKKEEETPEEKKK	

The same estimated probability is used to generate an html file, showing the reference sequence whose positions are color coded according to their statistical relevance as well as, for each specificity class, the most "promising" substitutions, as shown in Figure 3. Moreover, for each subclass the most promising substitutions are shown, when available, as suggested by conditional probability estimation done in the calculation.

Figure 3. Example of html reporting, GFP sequence along with most likely occurrences in Monomeric and Dimeric subclasses, positions marked with “@” are reference positions that have been masked, due for example to the fact that they have more than 30% of gaps. Positions marked with “?” are positions where the estimated conditional probability doesn’t give a clear indication concerning possible mutations.



```
Blosum Matrix: no

Bins : no ; Symbols: ARNDCQEGHILKMFPSTWYVBZX-

Nshuffles : 10000;

Henikoff Weights: 0;

MonoGFP
@@AMPAT?GS?ITFVLFLE?D??GHK?TISAE?YSKFTT?KFTAKYI?EK?DV?IS?STLIHSLQW?EMF?VK?GDHMELSD?AQETM?T?YTQD?KIRYEN??VMTSAAEIELDNTCVVN?ITVKGVG?DK??NVCGKK?VYSPPPETLVFPHDEQK?
@@K?T???H??L?????????????????????E???????L?L?L?????E?????P??????C??AKV????????KRK?????V?????H????????????FE????????S?TL????E???I????????????????ANTGCV???A?K???
DimGFP_
@@S?ALY?EKIMPYKIEMD?V?DDQT?K?RGK??GDASV??V?GHAV?TE?KL?MP?VSICTLIGH?GPC?AR?PNGPVINH?YKSCF?S?LVIE?TVT?KG??TYKTHHKYTMEGGAIYS?V?LNC?D?QP??HIMKD??H?IFTQVSYMVW?GSANH
@@G?E???T?VV?IL??I??N??????S?I?D?????PH??I?VQFL??????????KP???T?TY?LQ??????????QH?????A???????????F?????N?????T?K???DS?I???????????E??P?LQR????NLISTNH??M??Q???

Reference: GFP
@@KGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKN
---------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130-------140-------150-------


Legend:
@ are reference positions that have been masked

? are positions where is not clear what mutation to suggest
Importance Color Coding:
1 2 3 4 5 6 7 8 9 10
```

2. IFP Alignment

The complete alignment of the FPs is reported in the supplementary file A_FP.txt.

3. Physically based pigeonholes

The assignment of the residues to specific pigeonholes was performed according to the following groupings, derived from classical Taylor's Venn diagram [1]:

Hydrophobicity:

Pigeonhole W, hydrophilic aminoacids: R N D Q E G K P S T ;
Pigeonhole N, neutral aminoacids: A H;
Pigeonhole H, hydrophobic aminoacids: C I L M F W Y V.

Size:

Pigeonhole 1, extra extra small aminoacids: G;
Pigeonhole 2, extra small aminoacids: A S;
Pigeonhole 3, small aminoacids: C D P N T;
Pigeonhole 4, medium size aminoacids: Q E V;
Pigeonhole 5, large aminoacids: H M L I K R;
Pigeonhole 6, extra large aminoacids: F Y;
Pigeonhole 7, extra extra large aminoacids: W;

Charge:

Pigeonhole N, negative aminoacids: D E;
Pigeonhole P, positive aminoacids: R H K;
Pigeonhole L, polar aminoacids: N Q S T W Y;
Pigeonhole A, apolar aminoacids: P A G C I L M F V.

4. Extended IFP results

In the main text, we showed for space reasons only a subset of the results obtained in the runs to discriminate monomer vs multimer forms of IFPs. The tables here contain an extension to that set. In particular, Table 1 reports best ranking positions from 21 to 40, obtained in the different runs described in section 5.2 to show how many of the 33 positions experimentally identified were recognized by our approach. In Table 2 we show the full, i.e. 40 first positions, results of the same type of runs, but with inclusion of background removal correction.

Comparison of the two tables shows that, at least in this case, this procedure does not improve the performance of the method.

Table 1. SDPs inferred from 92 mono- and multimeric GFP homologs 40 best ranking SDPs derived with BLOSUM45 (B45), BLOSUM62 (B62), Identity (Id) and the local BLOSUM (Bloc) similarity matrices. No background correlation removal has been performed in these runs.

	B45	S	B62	S	Id	S	Bloc	S
21	78	4	83	2	21	1	223	1
22	85	3	85	3	71	2	147	*
23	6	1	78	4	6	1	125	1
24	83	2	118	4	197	2	1	4
25	118	4	5	1	1	4	92	3
26	193	4	193	4	150	2	79	3
27	30	3	6	1	206	3	127	1
28	216	3	30	3	125	1	21	1
29	5	1	216	3	127	1	85	3
30	203	4	98	3	219	4	6	1
31	11	3	21	1	189	4	218	4
32	207	4	207	4	193	4	118	4
33	218	4	195	2	75	4	11	3
34	184	4	71	2	57	4	30	3
35	195	2	11	3	92	3	189	4
36	98	3	203	4	85	3	207	4
37	71	2	184	4	161	4	203	4
38	49	3	218	4	207	4	219	4
39	21	1	147	*	78	4	161	4
40	36	4	206	3	30	3	193	4

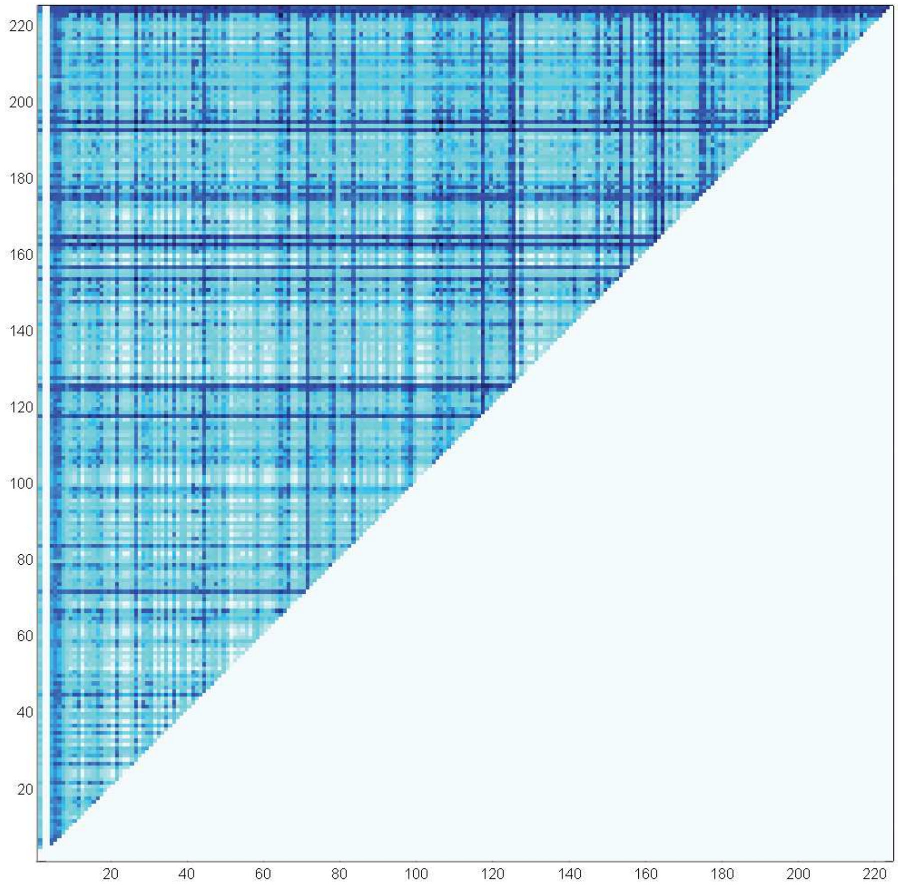
Table 2. SDPs inferred from 92 mono- and multimeric GFP homologs. Results reported are the same as Table 1 but with application of the background correlation removal and including the SDPpred results. Scores are defined as in the main text. Numbering refers to the "DsRed" sequence (1GGX PDB)

B45	S	B62	S	Id	S	Bloc	S	SDPpred	S
117	1	117	1	117	1	117	1	117	1
194	1	194	1	83	2	194	1	83	2
177	2	177	2	194	1	177	2	79	3
164	1	164	1	164	1	175	2	194	1
224	1	224	1	156	1	224	1	184	4
156	1	156	1	192	1	124	2	164	1
44	2	72	4	223	1	174	1	185	4
72	4	124	2	175	2	164	1	192	1
197	2	150	2	72	4	72	4	156	1
150	2	192	1	8	3	192	1	175	2
124	2	197	2	5	1	153	1	177	2
192	1	44	2	153	1	83	2	72	4
4	3	174	1	177	2	150	2	153	1
175	2	4	3	124	2	4	3	124	2
174	1	175	2	147	*	75	4	8	3
118	4	8	3	1	4	179	2	147	*
8	3	184	4	174	1	44	2	150	2
193	4	118	4	150	2	197	2	174	1
125	1	83	2	4	3	57	4	44	2
184	4	153	1	219	4	162	1	21	1
127	1	57	4	21	1	219	4	219	4
92	3	193	4	224	1	8	3	162	1
57	4	219	4	44	2	184	4	75	4
179	2	179	2	162	1	79	3	57	4
219	4	127	1	189	4	156	1	7	3
83	2	125	1	75	4	21	1	197	2
162	1	75	4	57	4	118	4	71	2
75	4	92	3	71	2	85	3	193	4
85	3	85	3	193	4	1	4	208	3
79	3	30	3	197	2	92	3	107	4
30	3	49	3	6	1	147	3	85	3
153	1	21	1	107	4	189	4	127	1
49	3	207	4	206	3	71	2	38	4
218	4	201	4	85	3	193	4	179	2
207	4	218	4	38	4	30	3	30	3
21	1	162	1	127	1	218	4	97	4
78	4	78	4	179	2	207	4	207	4
201	4	98	3	97	4	78	4	118	4
11	3	18	4	30	3	49	3	92	3
18	4	195	2	49	3	127	1	49	3

5. Pair correlation

Figure 4 reports a pictorial representation of the symmetric correlation matrix.

Figure 4. Pair correlation among positions is shown in a color coding ranging from white, indicating poor correlation, to intense blue. Rows and columns corresponding to positions 1,2 and 51 are white since they had a number of gaps larger than 30%.



6. MMPBSA details

The MMPBSA approach makes it possible to obtain solvation free energies of proteins through the combination of all-atoms molecular dynamics simulations of the solvated molecule and estimation of the terms accounting for solvent polarization due to solvent-solute interactions, changes in the conformational freedom of solvent upon solvation, and entropy of the solute.

In this context, the absolute free energy of the solute can be expressed as follows:

$$G = E_{MM} + G_{PB,polar} + G_{SA,nonpolar} - TS^{tr,rot,conf}_{solute},$$

where all quantities are averaged over a molecular mechanics trajectory, E_{MM} is the molecular mechanics energy, $G_{PB,polar}$ is the polarization free energy of the implicit solvent, which can be obtained through the solution of the Poisson-Boltzmann (PB) Equation, $G_{SA,nonpolar}$ is the nonpolar free energy estimated by scaling the solvent accessible surface area (SA) by an appropriate surface tension, and $S_{solute}^{tr,rot,conf}$ is the solute translational, rotational and conformational entropy, and T is the absolute temperature.

We applied this scheme to the interesting case of IFPs in order to compute tetramerization free energies of tetrameric Wild-Type DsRed and some tetrameric, dimeric or monomeric mutants as free energy differences of the tetramer with respect to the two dimers:

$$\Delta G_{i,tetramerization} = G_{i,tetramer} - 2 \cdot G_{i,dimer}.$$

$G_{i,tetramer}$ and $G_{i,dimer}$ are, respectively, the tetramer and dimer free energies of the i^{th} mutant.

In the present case, these quantities are computed over the same molecular dynamics trajectory of the tetramer for both the tetramer and the dimers; this is acceptable since, presumably, the conformations of these proteins in their dimeric and tetrameric state do not differ significantly.

Reproducing the trend of the relative stabilities of different mutants with respect to wild-type protein (as in the "virtual screening method" described in [2]) is of particular interest to select mutations that induce the stabilization of the dimers in an otherwise tetrameric protein. This can be accomplished by calculating

$$\Delta\Delta G_{i,WT} = \Delta G_{i,tetramerization} - \Delta G_{Wild-Type,tetramerization}.$$

The entropy contribution of the solute was not taken into account in the present study. Although the conformational $T\Delta S_i^{conf}$ is an important term in driving oligomer association/dissociation, we expect that this term is scarcely affected by mutations. Indeed, previous studies have shown that the conformational $T\Delta S^{conf}$ penalty upon side-chain burial is similar among different residues, with differences generally around 0.5 kcal/mol and always smaller than 2 kcal/mol at room temperature [3], and hence negligible with respect to the $\Delta\Delta G_{i,WT}$ obtained in the present study. Moreover, translational and rotational components of the entropy for the structurally homologous IFPs lead to $T\Delta\Delta S_{i,WT}^{tr,rot} \simeq 0$.

The starting structure for the MM dynamics was obtained by adding hydrogen atoms to the X-ray crystal structure of a DsRed tetramer (PDB code: 1GGX), which was solvated in a 85 Å-box of water molecules. All amino acid mutations were produced with Insight II©(Accelrys Inc.), starting from the 1GGX PDB structure for consistency. After equilibration of the system with restraints on the motion of non-hydrogen atoms (10 ps at 50 K, 20 ps at 150 K, 240 ps at 300 K), free molecular dynamics was performed for 400 ps at 300 K. Molecular Dynamics simulations and surface areas were computed with programs of the Amber 7 package, while Poisson-Boltzmann equation was solved with Delphi 4 [4].

E_{MM}, $G_{PB,polar}$ and $G_{SA,nonpolar}$ were calculated for 40 snapshots sampled from the trajectory at 10 ps intervals. The variance reported in Figure 3 of the main article is calculated on the ΔG taken at each snapshot, as usually carried out in MMPBSA calculations [2]. Each 670 ps Molecular Dynamics run (270 ps of equilibration and 400 ps of production), providing the 40 snapshots used to calculate the free energy, required 48 hours on four Intel® Xeon™ CPU 2.40GHz processors; the subsequent MMPBSA calculation, conversely, required 20 hours on a single processor.

References and Notes

1. Taylor, W. The classification of amino acid conservation. *J. Theor. Biol.* **1986**, *119*, 205-218.
2. Wang, W.; Kollman, P.A. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. *J. Mol. Biol.* **2000**, *303*, 567-582.

3. Doig, A.J.; Sternberg, M.J.E. Side-chain conformational entropy in protein folding. *Protein Sci.* **1995**, *4*, 2247-2251.
4. Rocchia, W.; Alexov, E.; Honig, B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem. B* **2001**, *105*, 6507-6514.