An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier

Jiaqi Xia, Zhenling Peng, Dawei Qi, Hongbo Mu and Jianyi Yang

Supplementary Materials

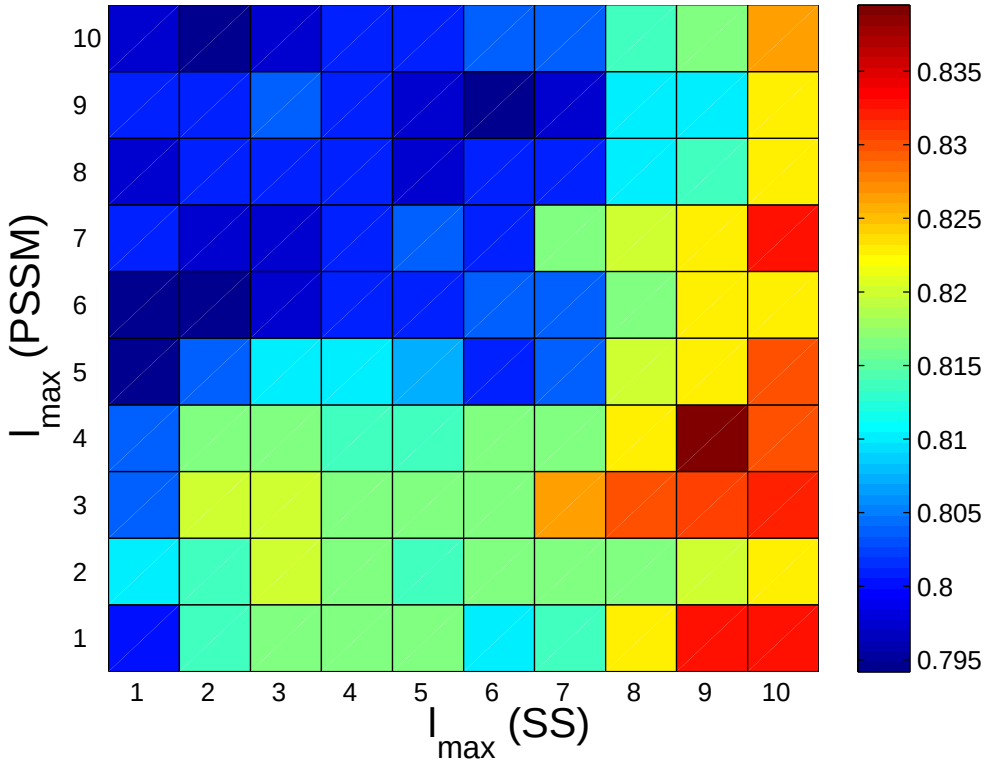

Figure S1: The overall accuracy of SVM-fold on the RDD training set with different values of l_{max}.

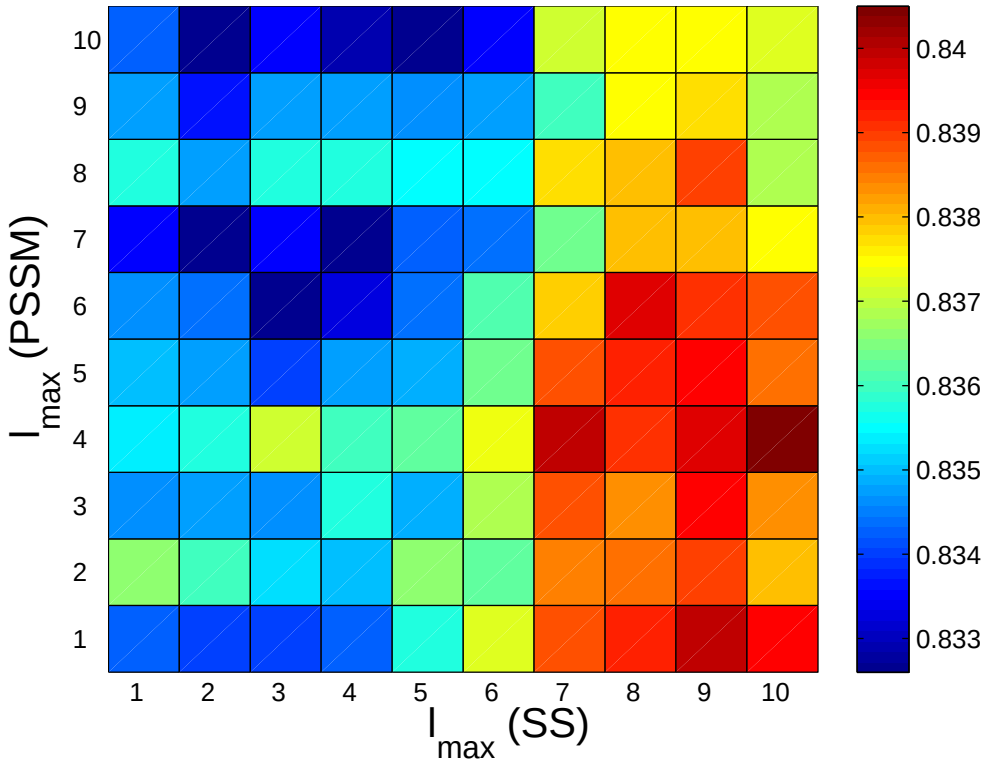

Figure S2: The overall accuracy of SVM-fold on the F184 dataset with different values of l_{max}.

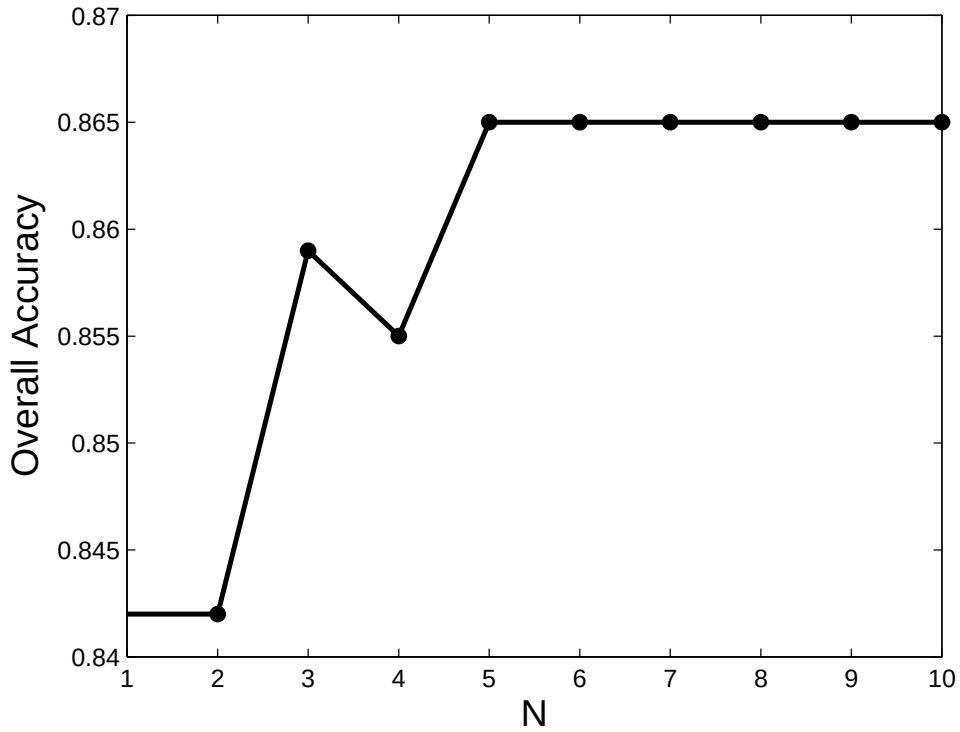

Figure S3: The overall accuracy of HH-fold on the DD training set with different numbers of top-ranked templates.

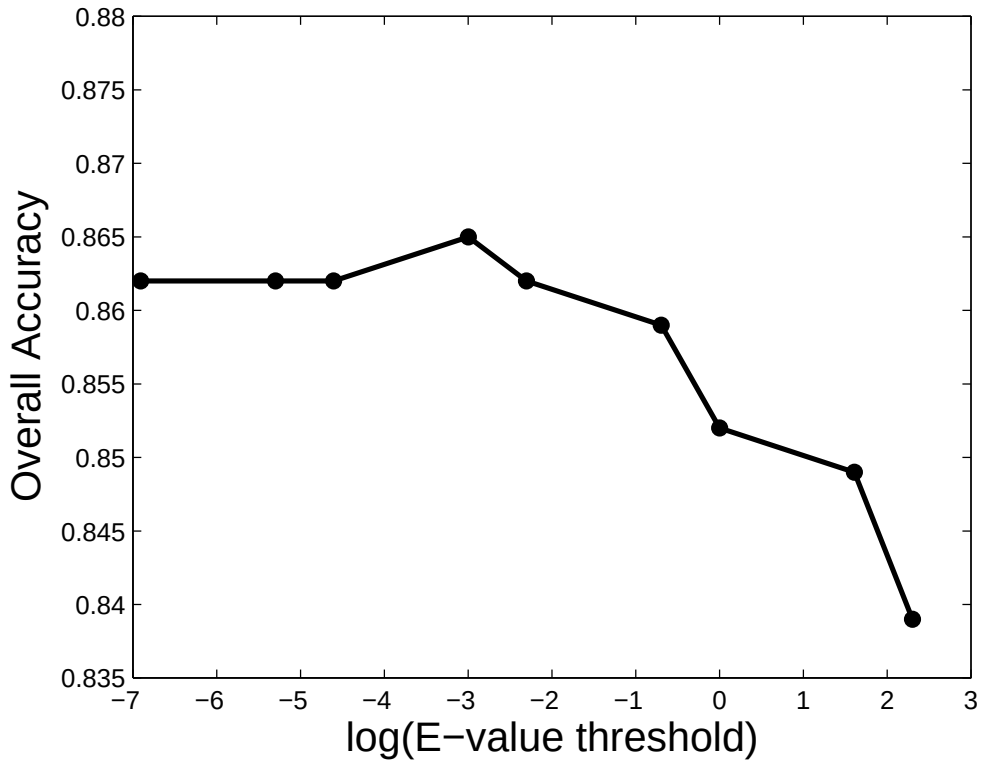

Figure S4: The overall accuracy of TA-fold on the RDD training set with different E-value cutoffs.

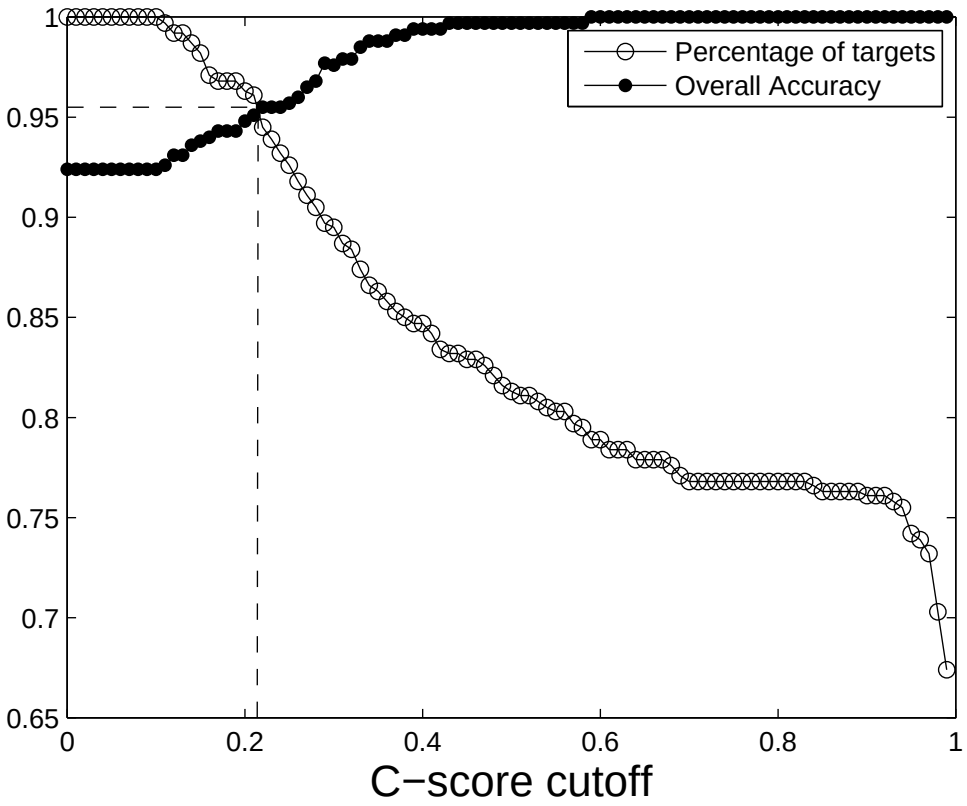

Figure S5: The overall accuracy of TA-fold predictions and the percentage of targets in the RDD test set predicted at different C-score cutoffs.

Table S1. The overall accuracy of SVM-fold by adding the AC features of the HHblits profile.

Feature	Dataset				
	DD	RDD	EDD	TG	F184
574	**0.773**	**0.9**	**0.945**	**0.865**	**0.84**
574+AC features on HHblits profile	0.76	0.892	0.945	0.862	0.838

Table S2. The overall accuracy of SVM-fold with different kernel functions in LIBSVM. For all kernel functions, only the two parameters *C* and *γ* were optimized. For other parameters, the default values in the LIBSVM package were used.

Kernel function	Dataset				
	DD	RDD	EDD	TG	F184
linear	0.766	0.876	0.941	0.855	0.775
polynomial	0.745	0.861	0.935	0.851	0.77
sigmoid	0.763	0.876	0.908	0.826	0.811
RBF	**0.773**	**0.9**	**0.945**	**0.865**	**0.84**

Table S3. The fold-specific accuracy of the predictions on the DD dataset. The optimized values for *C* and γ are 16 and 0.00390625, respectively.

No.	Fold	#Samples	HH-fold	SVM-fold	TA-fold
1	Globin-like	6	1.000	1.000	1.000
2	Cytochrome C	9	1.000	1.000	1.000
3	DNA-binding 3-helical bundle	20	0.750	0.750	0.800
4	4-Helical up-and-down bundle	8	0.875	1.000	1.000
5	4-Helical cytokines	9	1.000	1.000	1.000
6	α EF-hand	9	0.889	0.889	0.889
7	Immunoglobulin-like β-sandwich	44	0.773	0.864	0.864
8	Cupredoxins	12	0.917	0.833	0.917
9	Viral coat and capsid proteins	13	0.769	0.769	0.846
10	ConA-like lectins/glucanases	6	0.833	1.000	1.000
11	SH3-like serine proteases	8	0.500	0.750	0.625
12	OB-fold	19	0.684	0.737	0.789
13	Trefoil	4	1.000	1.000	1.000
14	Trypsin-like serine	4	1.000	1.000	1.000
15	Lipocalins	7	1.000	1.000	1.000
16	(TIM）-barrel	48	0.750	0.729	0.750
17	FAD (also NAD)-binding motif	12	0.917	0.917	0.917
18	Flavodoxin-like	13	0.615	0.538	0.615
19	NAD (P)-bindingRossmann-fold	27	0.815	0.704	0.815
20	P-loop containing nucleotide	12	0.583	0.500	0.583
21	Thioredoxin-like	8	0.500	0.375	0.500
22	Ribonuclease H-like motif	12	0.667	0.667	0.833
23	Hydrolases	7	0.857	0.857	0.857
24	Periplasmic binding protein-like	4	1.000	0.750	1.000
25	β-Grasp	8	0.500	0.375	0.375

26	Ferredoxin-like	27	0.593	0.593	0.556
27	Small inhibitors,toxins,lectins	28	0.893	0.929	0.929
Overall		384	0.763	0.773	0.799

Table S4. The fold-specific accuracy of the predictions on the RDD dataset. The optimized values for *C* and γ are 16 and 0.0078125, respectively.

No.	Fold	#Samples	HH-fold	SVM-fold	TA-fold
1	Globin-like	6	1.000	1.000	1.000
2	Cytochrome C	9	1.000	1.000	1.000
3	DNA-binding 3-helical bundle	20	0.950	0.950	1.000
4	4-Helical up-and-down bundle	7	0.857	1.000	1.000
5	4-Helical cytokines	9	1.000	1.000	1.000
6	α EF-hand	9	0.889	0.778	0.889
7	Immunoglobulin-like β-sandwich	44	0.909	0.977	0.955
8	Cupredoxins	12	1.000	1.000	1.000
9	Viral coat and capsid proteins	13	0.846	0.846	0.923
10	ConA-like lectins/glucanases	6	1.000	1.000	1.000
11	SH3-like serine proteases	8	0.375	0.500	0.500
12	OB-fold	19	0.737	0.895	0.947
13	Trefoil	4	1.000	1.000	1.000
14	Trypsin-like serine	4	1.000	1.000	1.000
15	Lipocalins	7	1.000	1.000	1.000
16	(TIM）-barrel	48	1.000	0.958	0.979
17	FAD (also NAD)-binding motif	12	1.000	0.917	1.000
18	Flavodoxin-like	13	1.000	0.769	0.846
19	NAD (P)-bindingRossmann-fold	27	0.963	0.889	0.963
20	P-loop containing nucleotide	12	0.917	0.833	0.917
21	Thioredoxin-like	7	0.857	0.857	0.857
22	Ribonuclease H-like motif	12	0.667	0.833	0.833
23	Hydrolases	7	1.000	1.000	1.000
24	Periplasmic binding protein-like	4	1.000	0.750	1.000
25	β-Grasp	8	0.500	0.375	0.500

26	Ferredoxin-like	27	0.704	0.778	0.778
27	Small inhibitors,toxins,lectins	26	0.923	1.000	1.000
Overall		380	0.887	0.900	0.932

Table S5. The fold-specific accuracy of the predictions on the EDD dataset. The optimized values for *C* and γ are 64 and 0.0078125, respectively.

No.	Fold	#Samples	HH-fold	SVM-fold	TA-fold
1	Globin-like	41	0.927	0.976	1.000
2	Cytochrome C	35	1.000	0.971	1.000
3	DNA-binding 3-helical bundle	321	0.988	0.991	0.994
4	4-Helical up-and-down bundle	69	0.855	0.899	0.913
5	4-Helical cytokines	30	0.933	0.967	1.000
6	α EF-hand	59	0.949	0.966	0.966
7	Immunoglobulin-like β-sandwich	391	0.959	0.964	0.967
8	Cupredoxins	47	0.979	0.957	0.957
9	Viral coat and capsid proteins	60	0.783	0.833	0.833
10	ConA-like lectins/glucanases	57	0.895	0.930	0.930
11	SH3-like serine proteases	129	0.953	0.907	0.946
12	OB-fold	156	0.942	0.821	0.936
13	Trefoil	45	1.000	0.978	1.000
14	Trypsin-like serine	45	1.000	0.933	0.978
15	Lipocalins	37	0.892	0.946	0.946
16	(TIM)-barrel	336	0.997	0.973	0.991
17	FAD (also NAD)-binding motif	73	1.000	0.959	0.986
18	Flavodoxin-like	130	0.915	0.946	0.931
19	NAD (P)-bindingRossmann-fold	194	1.000	0.979	0.995
20	P-loop containing nucleotide	239	0.992	0.975	0.987
21	Thioredoxin-like	111	0.991	0.892	0.973
22	Ribonuclease H-like motif	128	0.961	0.922	0.953
23	Hydrolases	83	1.000	0.988	1.000
24	Periplasmic binding protein-like	16	1.000	1.000	1.000
25	β-Grasp	121	0.959	0.926	0.959

26	Ferredoxin-like	339	0.941	0.929	0.956
27	Small inhibitors,toxins,lectins	105	0.990	1.000	1.000
Overall		3397	0.966	0.945	0.971

Table S6. The fold-specific accuracy of the predictions on the TG dataset. The optimized values for C and γ are 16 and 0.015625, respectively.

No.	Fold	#Samples	HH-fold	SVM-fold	TA-fold
1	Cytochrome C	25	1.000	0.920	1.000
2	DNA/RNA binding 3-helical bundle	103	0.971	0.913	0.961
3	Four Helical up-and-down bundle	26	0.692	0.808	0.846
4	EF hand-like fold	25	0.880	0.880	0.960
5	SAM domain-like	26	0.885	0.731	0.846
6	α-α super helix	47	0.766	0.851	0.894
7	Immunoglobulin-like β-sandwich	173	0.931	0.919	0.931
8	Common fold of diphtheria toxin/transcription/factors/cytochrome	28	0.464	0.536	0.571
9	Cupredoxins-like	30	1.000	0.967	1.000
10	Galactose-binding domain-like	25	0.840	0.800	0.840
11	Concanavalin A-like lectins/glucanases	26	0.923	0.885	0.885
12	SH3-like barrel	42	0.976	0.690	0.857
13	OB-fold	78	0.872	0.731	0.885
14	Double-stranded α-helix	34	0.912	0.853	0.912
15	Nucleoplasmin-like	42	0.667	0.738	0.786
16	TIM α/β-barrel	145	0.993	0.966	0.986
17	NAD(P)-binding Rossmann-fold domains	77	0.987	0.935	0.987
18	FAD/NAD(P)-binding domain	31	1.000	0.903	1.000
19	Flavodoxin-like	55	0.818	0.818	0.836
20	Adenine nucleotide a hydrolase-like	34	0.941	0.912	0.941
21	P-loop containing nucleoside triphosphate hydrolases	95	0.979	0.958	0.989
22	Thioredoxin-fold	32	0.969	0.813	0.938
23	Ribonuclease H-like motif	49	0.898	0.612	0.878
24	S-adenosyl-L-menthionine-dependent methyltransferases	34	1.000	1.000	1.000
25	α/β-Hydrolases	37	1.000	1.000	1.000

26	β-Grasp,ubiquitin-like	42	0.857	0.714	0.857
27	Cystatin-like	25	0.880	0.800	0.840
28	Ferredoxin-like	118	0.873	0.831	0.881
29	Knottins	80	0.988	1.000	1.000
30	Rubredoxin-like	28	0.964	0.750	0.964
Overall		1612	0.915	0.865	0.927

Table S7. The fold-specific accuracy of the predictions on the F184 dataset. The optimized values for *C* and γ are 128 and 0.00390625, respectively.

No.	Fold	#Samples	HH-fold	SVM-fold	TA-fold
1	Globin-like	25	0.96	0.96	0.96
2	Cytochrome c	26	1	0.923	1
3	DNA/RNA-binding 3-helical bundle	265	0.962	0.94	0.97
4	Four-helical up-and-down bundle	58	0.69	0.638	0.741
5	4-helical cytokines	30	0.8	0.933	0.933
6	EF Hand-like	43	0.953	0.814	0.93
7	Immunoglobulin-like beta-sandwich	292	0.908	0.914	0.932
8	Cupredoxin-like	34	0.941	0.853	0.941
9	Nucleoplasmin-like/VP (viral coat and capsid proteins)	49	0.592	0.714	0.755
10	Concanavalin A-like lectins/glucanases	50	0.82	0.86	0.9
11	SH3-like barrel	104	0.894	0.827	0.933
12	OB-fold	132	0.871	0.773	0.886
13	beta-Trefoil	35	1	0.914	1
14	Trypsin-like serine proteases	19	0.947	0.842	0.947
15	Lipocalins	24	0.708	0.833	0.833
16	TIM beta/alpha-barrel	247	0.98	0.96	0.984
17	FAD/NAD(P)-binding domain	47	1	0.936	1
18	Flavodoxin-like	107	0.907	0.897	0.907
19	NAD(P)-binding Rossmann-fold domains	122	1	0.943	0.992
20	P-loop containing nucleoside triphosphate hydrolases	177	0.977	0.938	0.983
21	Thioredoxin fold	98	0.99	0.908	0.99
22	Ribonuclease H-like motif	121	0.959	0.843	0.95
23	alpha/beta-Hydrolases	68	1	0.985	1
24	Periplasmic binding protein-like I	40	1	1	1

25	beta-Grasp (ubiquitin-like)	92	0.924	0.848	0.902
26	Ferredoxin-like	271	0.882	0.838	0.893
27	Knottins (small inhibitors, toxins, lectins)	121	0.95	0.975	1
28	Long alpha-hairpin	36	0.667	0.667	0.806
29	RuvA C-terminal domain-like	37	0.784	0.676	0.811
30	Putative DNA-binding domain	14	0.857	0.5	0.857
31	Spectrin repeat-like	34	0.735	0.618	0.706
32	immunoglobulin/albumin-binding domain-like	21	0.571	0.381	0.476
33	Histone-fold	18	1	0.667	0.944
34	Ferritin-like	46	0.978	0.87	0.978
35	Acyl carrier protein-like	11	0.818	0.818	0.818
36	Bromodomain-like	33	0.727	0.636	0.727
37	lambda repressor-like DNA-binding domains	27	1	0.889	1
38	CH domain-like	13	0.923	0.769	0.923
39	Ribbon-helix-helix	14	0.857	0.571	0.714
40	GST C-terminal domain-like	24	1	1	1
41	STAT-like	11	0.364	0.364	0.455
42	SAM domain-like	58	0.897	0.724	0.914
43	Cyclin-like	18	0.944	0.944	0.944
44	DEATH domain	17	1	1	1
45	post-AAA+ oligomerization domain-like	11	0.818	0.727	0.818
46	6-phosphogluconate dehydrogenase C-terminal domain-like	22	0.818	0.864	0.955
47	alpha/alpha toroid	31	0.968	0.903	0.935
48	Cytochrome P450	15	1	1	1
49	alpha-alpha superhelix	116	0.828	0.862	0.888
50	Tetracyclin repressor-like, C-terminal domain	34	1	0.882	1
51	Nuclear receptor ligand-binding domain	10	1	1	1
52	Heme oxygenase-like	12	1	1	1

53	Non-globular all-alpha subunits of globular proteins	15	0.133	0.133	0.267
54	Multiheme cytochromes	14	1	0.857	0.929
55	LEM/SAP HeH motif	14	0.929	0.714	0.786
56	HD-domain/PDEase-like	12	1	0.667	0.833
57	Common fold of diphtheria toxin/transcription factors/cytochrome f	42	0.667	0.643	0.762
58	Prealbumin-like	20	0.9	0.75	0.9
59	C2 domain-like	19	0.842	0.737	0.842
60	gamma-Crystallin-like	10	0.7	0.7	0.7
61	Galactose-binding domain-like	46	0.891	0.783	0.891
62	SMAD/FHA domain	12	0.917	0.667	0.833
63	Supersandwich	22	0.773	0.864	0.909
64	ISP domain	10	1	1	1
65	GroES-like	12	0.917	0.75	0.917
66	PDZ domain-like	30	1	1	1
67	Sm-like fold	17	0.765	0.529	0.765
68	Reductase/isomerase/elongation factor common domain	34	0.912	0.647	0.912
69	Split barrel-like	29	0.897	0.759	0.897
70	Domain of alpha and beta subunits of F1 ATP synthase-like	12	0.75	0.75	0.75
71	Acid proteases	10	1	0.9	1
72	Double psi beta-barrel	17	1	0.706	1
73	PH domain-like barrel	61	0.984	0.885	0.967
74	Streptavidin-like	13	0.538	0.308	0.308
75	6-bladed beta-propeller	22	0.682	0.773	0.682
76	7-bladed beta-propeller	36	0.861	0.833	0.917
77	Glycosyl hydrolase domain	36	0.75	0.722	0.806
78	Single-stranded right-handed beta-helix	26	0.654	0.808	0.808
79	Single-stranded left-handed beta-helix	16	0.75	0.625	0.75
80	Double-stranded beta-helix	86	0.965	0.907	0.942

81	Barrel-sandwich hybrid	18	0.944	0.5	0.889
82	beta-clip	22	0.909	0.773	0.864
83	Composite domain of metallo-dependent hydrolases	17	0.941	0.647	0.941
84	Phage tail proteins	11	0.818	0.545	0.818
85	PUA domain-like	26	0.885	0.654	0.885
86	7-stranded beta/alpha barrel	14	1	0.643	1
87	The "swivelling" beta/beta/alpha domain	19	0.737	0.684	0.789
88	Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)	20	0.9	0.9	0.9
89	ClpP/crotonase	24	1	0.875	1
90	Adenine nucleotide alpha hydrolase-like	46	1	0.913	1
91	PreATP-grasp domain	15	0.8	0.467	0.667
92	DHS-like NAD/FAD-binding domain	14	0.929	0.714	0.929
93	Thiamin diphosphate-binding fold (THDP-binding)	18	1	0.944	1
94	CoA-dependent acyltransferases	14	0.929	0.857	0.857
95	Phosphotyrosine protein phosphatases I-like	11	1	0.818	1
96	(Phosphotyrosine protein) phosphatases II	15	1	0.733	1
97	Rhodanese/Cell cycle control phosphatase	14	1	1	1
98	Anticodon-binding domain-like	19	0.895	0.474	0.789
99	Restriction endonuclease-like	43	0.674	0.581	0.674
100	Phosphorylase/hydrolase-like	41	0.927	0.854	0.927
101	Aminoacid dehydrogenase-like, N-terminal domain	10	0.9	0.7	0.9
102	PRTase-like	20	1	1	1
103	vWA-like	13	1	1	1
104	S-adenosyl-L-methionine-dependent methyltransferases	87	0.989	0.92	0.989
105	PLP-dependent transferase-like	47	0.957	0.957	0.957
106	Nucleotide-diphospho-sugar transferases	29	0.931	0.897	0.931
107	Ribokinase-like	24	0.958	0.875	0.958
108	ATC-like	11	1	0.818	1

109	SIS domain	11	1	1	1
110	UDP-Glycosyltransferase/glycogen phosphorylase	17	0.882	0.706	0.824
111	Chelatase-like	18	1	0.889	0.944
112	Periplasmic binding protein-like II	65	1	0.938	0.985
113	Thiolase-like	19	1	0.842	1
114	Cytidine deaminase-like	20	0.9	0.8	0.9
115	HAD-like	51	1	0.98	1
116	alpha/beta knot	10	1	0.4	0.9
117	NagB/RpiA/CoA transferase-like	22	0.955	0.818	0.909
118	Lysozyme-like	15	0.8	0.667	0.867
119	Cysteine proteinases	31	0.839	0.613	0.774
120	Ribosomal protein S5 domain 2-like	33	0.788	0.697	0.697
121	FAD-linked reductases, C-terminal domain	21	0.81	0.667	0.762
122	Cystatin-like	78	0.91	0.833	0.91
123	MHC antigen-recognition domain	10	1	1	1
124	UBC-like	10	1	0.9	1
125	FKBP-like	17	0.824	0.765	0.824
126	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	28	1	0.964	0.964
127	CBS-domain pair	19	1	1	1
128	Thioesterase/thiol ester dehydrase-isomerase	44	1	0.977	1
129	alpha/beta-Hammerhead	12	0.917	0.75	0.917
130	dsRBD-like	17	0.588	0.529	0.529
131	Eukaryotic type KH-domain (KH-domain type I)	12	1	0.833	1
132	Alpha-lytic protease prodomain-like	21	0.667	0.476	0.714
133	Enolase N-terminal domain-like	13	1	0.923	1
134	IF3-like	16	0.813	0.563	0.813
135	DCoH-like	12	0.917	0.583	0.75
136	Bacillus chorismate mutase-like	31	0.839	0.677	0.806

137	Tautomerase/MIF	12	1	0.917	1
138	FwdE/GAPDH domain-like	26	0.731	0.615	0.615
139	CO dehydrogenase flavoprotein C-domain-like	12	0.833	0.75	0.833
140	FMN-dependent nitroreductase-like	11	1	1	1
141	Zincin-like	30	0.967	0.7	0.867
142	SH2-like	10	1	0.9	1
143	Homing endonuclease-like	12	0.917	0.833	0.917
144	T-fold	14	0.929	0.786	0.929
145	Class II aaRS and biotin synthetases	22	1	0.955	1
146	Acyl-CoA N-acyltransferases (Nat)	65	0.969	0.862	0.969
147	Gelsolin-like	13	0.923	0.769	0.923
148	Profilin-like	41	0.927	0.829	0.878
149	Nudix	26	0.962	0.885	0.962
150	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase	11	1	1	1
151	TBP-like	47	0.894	0.723	0.872
152	DNA clamp	19	0.789	0.789	0.895
153	Phospholipase D/nuclease	11	1	0.909	1
154	ATP-grasp	22	0.909	0.864	0.909
155	Protein kinase-like (PK-like)	27	0.963	0.852	0.926
156	FAD-binding/transporter-associated domain-like	14	1	1	1
157	Ntn hydrolase-like	25	0.96	0.8	0.92
158	Metallo-hydrolase/oxidoreductase	20	1	0.95	1
159	Metallo-dependent phosphatases	18	1	1	1
160	LDH C-terminal domain-like	10	1	1	1
161	ADP-ribosylation	16	0.75	0.563	0.75
162	C-type lectin-like	22	0.909	0.955	0.955
163	LuxS/MPP-like metallohydrolase	15	0.933	0.867	0.933
164	Chorismate lyase-like	12	1	1	1

165	Secretion chaperone-like	18	0.889	0.556	0.722
166	Nucleotidyltransferase	19	0.947	0.842	0.895
167	beta-lactamase/transpeptidase-like	17	1	0.941	1
168	DNA/RNA polymerases	19	0.947	0.789	0.947
169	Toxins' membrane translocation domains	15	0.8	0.667	0.867
170	Transmembrane beta-barrels	23	0.913	0.913	0.913
171	Transmembrane helix hairpin	11	0.455	0.636	0.727
172	Heme-binding four-helical bundle	11	0.909	1	1
173	Single transmembrane helix	53	0.585	0.868	0.887
174	Snake toxin-like	11	1	1	1
175	Defensin-like	14	1	0.643	0.857
176	Cystine-knot cytokines	10	1	0.7	0.9
177	Complement control module/SCR domain	14	0.929	0.929	0.929
178	TNF receptor-like	15	1	0.867	1
179	beta-beta-alpha zinc fingers	60	0.95	0.9	0.95
180	Glucocorticoid receptor-like (DNA-binding domain)	53	0.811	0.774	0.83
181	Rubredoxin-like	48	0.854	0.75	0.813
182	RING/U-box	24	1	1	1
183	Metallothionein	10	0.8	0.8	0.8
184	FYVE/PHD zinc finger	15	1	0.867	1
Overall		6451	0.906	0.840	0.913