The Details of the four datasets used in this study

Table 1. Summary of the 27 fold classes in the DD set and the EDD_{new} set.

Index	Fold identifier	Fold name	No. of sequences			
			DD			EDD_{new}
			S_{Train}	S_{Test}	Total	
1	a.1	Globin-like	13	6	19	51
2	a.3	Cytochrome c	7	9	16	38
3	a.4	DNA/RNA-binding 3-helical bundle	12	30	32	335
4	a.24	4-helical up-and-down bundle	7	8	15	73
5	a.26	4-helical cytokines	9	9	18	30
6	a.39	EF hand-like	6	9	15	67
7	b.1	Immunoglobulin-like beta-sandwich	30	44	74	401
8	b.6	Cupredoxin-like	9	12	21	49
9	b.121	Nucleoplasmin-like/VP	16	13	29	62
10	b.29	ConA-like lectins/glucanases	7	6	13	63
11	b.34	SH3-like barrel	8	8	16	137
12	b.40	OB-fold	13	19	32	158
13	b.42	Beta-Trefoil	8	4	12	49
14	b.47	Trypsin-like serine proteases	9	4	13	50
15	b.60	Lipocalins	9	7	16	41
16	c.1	TIM beta/alpha-barrel	29	48	77	373
17	c.2	FAD/NAD(P)-binding domain	11	12	23	220
18	c.3	Flavodoxin-like	11	13	24	73
19	c.23	NAD(P)-binding Rossmann	13	27	40	150
20	c.37	P-loop containing NTH	10	12	22	252
21	c.47	Thioredoxin-fold	9	8	17	135
22	c.55	Ribonuclease H-like motif	10	12	22	130
23	c.69	alpha/beta-Hydrolases	11	7	18	91
24	c.93	Periplasmic binding protein-like	11	4	15	19
25	d.15	beta-grasp (ubiquitin-like)	7	8	15	124
26	d.58	Ferredoxin-like	13	27	40	347
27	g.3	Knottins (small inhibitors, toxins, lectins)	13	27	40	107
Total			**311**	**383**	**694**	**3,625**

Note that S_{Train} denotes the training sequences of the DD set, and S_{Train} denotes the testing sequences of the DD set.

Table 2. Summary of the 95 fold classes in the $F95_{new}$ set.

Index	Fold identifier	Fold name	No. of sequences
1	a.1	Globin-like	51
2	a.2	Long alpha-hairpin	41
3	a.3	Cytochrome c	38
4	a.4	DNA/RNA-binding 3-helical bundle	335
5	a.5	RuvA C-terminal domain-like	45
6	a.7	Spectrin repeat-like	46
7	a.24	Four-helical up-and-down bundle	73
8	a.25	Ferritin-like	56
9	a.26	4-helical cytokines	30
10	a.29	Bromodomain-like	39
11	a.35	lambda repressor-like DNA-binding domains	31
12	a.39	EF Hand-like	67
13	a.45	GST C-terminal domain-like	29
14	a.60	SAM domain-like	77
15	a.100	6-phosphogluconate dehydrogenase C-terminal domain-like	25
16	a.102	alpha/alpha toroid	45
17	a.118	alpha-alpha superhelix	110
18	a.121	Tetracyclin repressor-like, C-terminal domain	34
19	b.1	Immunoglobulin-like beta-sandwich	401
20	b.2	Common fold of diphtheria toxin/transcription factors/cytochrome f	49
21	b.6	Cupredoxin-like	49
22	b.7	C2 domain-like	30
23	b.121	Nucleoplasmin-like/VP (viral coat and capsid proteins)	62
24	b.18	Galactose-binding domain-like	61
25	b.29	Concanavalin A-like lectins/glucanases	63
26	b.30	Supersandwich	29
27	b.34	SH3-like barrel	137
28	b.36	PDZ domain-like	66
29	b.40	OB-fold	158
30	b.42	beta-Trefoil	49
31	b.43	Reductase/isomerase/elongation factor common domain	43
32	b.45	Split barrel-like	38
33	b.47	Trypsin-like serine proteases	50
34	b.55	PH domain-like barrel	78
35	b.60	Lipocalins	41

36	b.68	6-bladed beta-propeller	29
37	b.69	7-bladed beta-propeller	38
38	b.71	Glycosyl hydrolase domain	38
39	b.80	Single-stranded right-handed beta-helix	30
40	b.82	Double-stranded beta-helix	98
41	b.92	Composite domain of metallo-dependent hydrolases	28
42	b.122	PUA domain-like	32
43	c.1	TIM beta/alpha-barrel	373
44	c.2	NAD(P)-binding Rossmann-fold domains	220
45	c.3	FAD/NAD(P)-binding domain	73
46	c.10	Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)	29
47	c.14	ClpP/crotonase	34
48	c.23	Flavodoxin-like	150
49	c.26	Adenine nucleotide alpha hydrolase-like	73
50	c.108	HAD-like	75
51	c.124	NagB/RpiA/CoA transferase-like	30
52	c.36	Thiamin diphosphate-binding fold (THDP-binding)	30
53	c.37	P-loop containing nucleoside triphosphate hydrolases	252
54	c.47	Thioredoxin fold	135
55	c.51	Anticodon-binding domain-like	26
56	c.52	Restriction endonuclease-like	46
57	c.55	Ribonuclease H-like motif	130
58	c.56	Phosphorylase/hydrolase-like	58
59	c.61	PRTase-like	27
60	c.66	S-adenosyl-L-methionine-dependent methyltransferases	115
61	c.67	PLP-dependent transferase-like	77
62	c.68	Nucleotide-diphospho-sugar transferases	37
63	c.69	alpha/beta-Hydrolases	91
64	c.72	Ribokinase-like	35
65	c.94	Periplasmic binding protein-like II	71
66	d.3	Cysteine proteinases	47
67	d.14	Ribosomal protein S5 domain 2-like	41
68	d.15	beta-Grasp (ubiquitin-like)	124
69	d.17	Cystatin-like	94
70	d.26	FKBP-like	27
71	d.32	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	34
72	d.38	Thioesterase/thiol ester dehydrase-isomerase	56

73	d.52	Alpha-lytic protease prodomain-like	26
74	d.58	Ferredoxin-like	347
75	d.79	Bacillus chorismate mutase-like	41
76	d.81	FwdE/GAPDH domain-like	35
77	d.92	Zincin-like	41
78	d.104	Class II aaRS and biotin synthetases	28
79	d.108	Acyl-CoA N-acyltransferases (Nat)	75
80	d.110	Profilin-like	51
81	d.113	Nudix	33
82	d.129	TBP-like	55
83	d.142	ATP-grasp	30
84	d.144	Protein kinase-like (PK-like)	56
85	d.153	Ntn hydrolase-like	39
86	d.157	Metallo-hydrolase/oxidoreductase	30
87	d.169	C-type lectin-like	41
88	e.3	beta-lactamase/transpeptidase-like	29
89	f.23	Single transmembrane helix	44
90	g.3	Knottins (small inhibitors, toxins, lectins)	107
91	g.18	Complement control module/SCR domain	28
92	g.37	beta-beta-alpha zinc fingers	38
93	g.39	Glucocorticoid receptor-like (DNA-binding domain)	58
94	g.41	Rubredoxin-like	52
95	g.44	RING/U-box	28
Total			**6,791**

Table 3. Summary of the 194 fold classes in the $F194_{new}$ set.

Index	Fold identifier	Fold name	No. of Sequences
1	a.1	Globin-like	51
2	a.2	Long alpha-hairpin	41
3	a.140	LEM/SAP HeH motif	15
4	a.3	Cytochrome c	38
5	a.4	DNA/RNA-binding 3-helical bundle	335
6	a.5	RuvA C-terminal domain-like	45
7	a.6	Putative DNA-binding domain	14
8	a.7	Spectrin repeat-like	46
9	a.8	immunoglobulin/albumin-binding domain-like	23
10	a.22	Histone-fold	21
11	a.24	Four-helical up-and-down bundle	73
12	a.25	Ferritin-like	56
13	a.26	4-helical cytokines	30
14	a.28	Acyl carrier protein-like	13
15	a.29	Bromodomain-like	39
16	a.35	lambda repressor-like DNA-binding domains	31
17	a.39	EF Hand-like	67
18	a.40	CH domain-like	12
19	a.43	Ribbon-helix-helix	16
20	a.45	GST C-terminal domain-like	29
21	a.47	STAT-like	11
22	a.60	SAM domain-like	77
23	a.74	Cyclin-like	23
24	a.80	post-AAA+ oligomerization domain-like	13
25	a.77	DEATH domain	17
26	a.211	HD-domain/PDEase-like	21
27	a.100	6-phosphogluconate dehydrogenase C-terminal domain-like	25
28	a.102	alpha/alpha toroid	45
29	a.104	Cytochrome P450	22
30	a.118	alpha-alpha superhelix	110
31	a.121	Tetracyclin repressor-like, C-terminal domain	34
32	a.123	Nuclear receptor ligand-binding domain	18
33	a.128	Terpenoid synthases	11
34	a.132	Heme oxygenase-like	16
35	a.137	Non-globular all-alpha subunits of globular proteins	13
36	a.138	Multiheme cytochromes	22

37	b.1	Immunoglobulin-like beta-sandwich	401
38	b.2	Common fold of diphtheria toxin/transcription factors/cytochrome f	49
39	b.3	Prealbumin-like	23
40	b.6	Cupredoxin-like	49
41	b.7	C2 domain-like	30
42	b.121	Nucleoplasmin-like/VP (viral coat and capsid proteins)	62
43	b.11	gamma-Crystallin-like	14
44	b.18	Galactose-binding domain-like	61
45	b.22	TNF-like	14
46	b.26	SMAD/FHA domain	18
47	b.29	Concanavalin A-like lectins/glucanases	63
48	b.30	Supersandwich	29
49	b.33	ISP domain	15
50	b.34	SH3-like barrel	137
51	b.35	GroES-like	20
52	b.36	PDZ domain-like	66
53	b.38	Sm-like fold	24
54	b.40	OB-fold	158
55	b.42	beta-Trefoil	49
56	b.43	Reductase/isomerase/elongation factor common domain	43
57	b.44	Elongation factor/aminomethyltransferase common domain	11
58	b.45	Split barrel-like	38
59	b.106	Phage tail proteins	13
60	b.47	Trypsin-like serine proteases	50
61	b.49	Domain of alpha and beta subunits of F1 ATP synthase-like	15
62	b.50	Acid proteases	18
63	b.52	Double psi beta-barrel	23
64	b.55	PH domain-like barrel	78
65	b.60	Lipocalins	41
66	b.61	Streptavidin-like	18
67	b.68	6-bladed beta-propeller	29
68	b.69	7-bladed beta-propeller	38
69	b.71	Glycosyl hydrolase domain	38
70	b.80	Single-stranded right-handed beta-helix	30
71	b.81	Single-stranded left-handed beta-helix	20
72	b.82	Double-stranded beta-helix	98
73	b.84	Barrel-sandwich hybrid	24
74	b.85	beta-clip	25

75	b.92	Composite domain of metallo-dependent hydrolases	28
76	b.122	PUA domain-like	32
77	c.1	TIM beta/alpha-barrel	373
78	c.2	NAD(P)-binding Rossmann-fold domains	220
79	c.3	FAD/NAD(P)-binding domain	73
80	c.6	7-stranded beta/alpha barrel	20
81	c.8	The "swivelling" beta/beta/alpha domain	20
82	c.10	Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)	29
83	c.14	ClpP/crotonase	34
84	c.58	Aminoacid dehydrogenase-like, N-terminal domain	14
85	c.23	Flavodoxin-like	150
86	c.25	Ferredoxin reductase-like, C-terminal NADP-linked domain	13
87	c.26	Adenine nucleotide alpha hydrolase-like	73
88	c.120	PIN domain-like	11
89	c.116	alpha/beta knot	19
90	c.30	PreATP-grasp domain	16
91	c.31	DHS-like NAD/FAD-binding domain	19
92	c.108	HAD-like	75
93	c.124	NagB/RpiA/CoA transferase-like	30
94	c.36	Thiamin diphosphate-binding fold (THDP-binding)	30
95	c.37	P-loop containing nucleoside triphosphate hydrolases	252
96	c.43	CoA-dependent acyltransferases	13
97	c.45	(Phosphotyrosine protein) phosphatases II	24
98	c.46	Rhodanese/Cell cycle control phosphatase	21
99	c.47	Thioredoxin fold	135
100	c.51	Anticodon-binding domain-like	26
101	c.52	Restriction endonuclease-like	46
102	c.55	Ribonuclease H-like motif	130
103	c.56	Phosphorylase/hydrolase-like	58
104	c.61	PRTase-like	27
105	c.62	vWA-like	14
106	c.66	S-adenosyl-L-methionine-dependent methyltransferases	115
107	c.67	PLP-dependent transferase-like	77
108	c.68	Nucleotide-diphospho-sugar transferases	37
109	c.69	alpha/beta-Hydrolases	91
110	c.71	Dihydrofolate reductase-like	15

111	c.72	Ribokinase-like	35
112	c.78	ATC-like	13
113	c.79	Tryptophan synthase beta subunit-like PLP-dependent enzymes	15
114	c.87	UDP-Glycosyltransferase/glycogen phosphorylase	18
115	c.92	Chelatase-like	22
116	c.93	Periplasmic binding protein-like I	19
117	c.94	Periplasmic binding protein-like II	71
118	c.95	Thiolase-like	24
119	c.97	Cytidine deaminase-like	19
120	d.2	Lysozyme-like	21
121	d.3	Cysteine proteinases	47
122	d.9	IL8-like	15
123	d.13	HIT-like	13
124	d.198	Secretion chaperone-like	20
125	d.14	Ribosomal protein S5 domain 2-like	41
126	d.15	beta-Grasp (ubiquitin-like)	124
127	d.16	FAD-linked reductases, C-terminal domain	24
128	d.17	Cystatin-like	94
129	d.19	MHC antigen-recognition domain	15
130	d.20	UBC-like	21
131	d.21	Diaminopimelate epimerase-like	13
132	d.24	Pili subunits	12
133	d.26	FKBP-like	27
134	d.32	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	34
135	d.211	beta-hairpin-alpha-hairpin repeat	19
136	d.37	CBS-domain pair	26
137	d.38	Thioesterase/thiol ester dehydrase-isomerase	56
138	d.41	alpha/beta-Hammerhead	22
139	d.50	dsRBD-like	22
140	d.51	Eukaryotic type KH-domain (KH-domain type I)	24
141	d.52	Alpha-lytic protease prodomain-like	26
142	d.218	Nucleotidyltransferase	23
143	d.54	Enolase N-terminal domain-like	16
144	d.58	Ferredoxin-like	347
145	d.190	Chorismate lyase-like	13
146	d.185	LuxS/MPP-like metallohydrolase	21
147	d.68	IF3-like	20
148	d.74	DCoH-like	13
149	d.79	Bacillus chorismate mutase-like	41
150	d.80	Tautomerase/MIF	16
151	d.81	FwdE/GAPDH domain-like	35

152	d.87	CO dehydrogenase flavoprotein C-domain-like	18
153	d.92	Zincin-like	41
154	d.93	SH2-like	25
155	d.95	Homing endonuclease-like	13
156	d.96	T-fold	14
157	d.101	Ribonuclease PH domain 2-like	12
158	d.104	Class II aaRS and biotin synthetases	28
159	d.108	Acyl-CoA N-acyltransferases (Nat)	75
160	d.109	Gelsolin-like	18
161	d.110	Profilin-like	51
162	d.113	Nudix	33
163	d.122	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase	16
164	d.126	Pentein, beta/alpha-propeller	13
165	d.129	TBP-like	55
166	d.131	DNA clamp	22
167	d.136	Phospholipase D/nuclease	11
168	d.142	ATP-grasp	30
169	d.144	Protein kinase-like (PK-like)	56
170	d.145	FAD-binding/transporter-associated domain-like	22
171	d.153	Ntn hydrolase-like	39
172	d.157	Metallo-hydrolase/oxidoreductase	30
173	d.159	Metallo-dependent phosphatases	24
174	d.162	LDH C-terminal domain-like	16
175	d.166	ADP-ribosylation	17
176	d.169	C-type lectin-like	41
177	e.1	Serpins	13
178	e.8	DNA/RNA polymerases	23
179	e.3	beta-lactamase/transpeptidase-like	29
180	f.1	Toxins' membrane translocation domains	15
181	f.23	Single transmembrane helix	44
182	f.4	Transmembrane beta-barrels	22
183	g.3	Knottins (small inhibitors, toxins, lectins)	107
184	g.7	Snake toxin-like	17
185	g.9	Defensin-like	14
186	g.68	Kazal-type serine protease inhibitors	11
187	g.17	Cystine-knot cytokines	14
188	g.18	Complement control module/SCR domain	28
189	g.24	TNF receptor-like	15
190	g.37	beta-beta-alpha zinc fingers	38
191	g.39	Glucocorticoid receptor-like (DNA-binding domain)	58
192	g.41	Rubredoxin-like	52

193	g.44	RING/U-box	28
194	g.50	FYVE/PHD zinc finger	15
Total			**8,525**