Supporting Information

A sequence space search engine for computational protein design to modulate molecular functionality

Ayush Malik[1,§], Anupam Banerjee[2,§], Abantika Pal[1], Pralay Mitra[1,*]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, West Bengal 721302, India

[2]School of Medical Science and Technology, Indian Institute of Technology Kharagpur, West Bengal 721302, India

[§]Contributed equally

*Correspondence to: pralay@cse.iitkgp.ac.in

Table S1. Summary of the designed sequences for all the proteins from the data set

PDB ID[§]	Seq_sim (%)	SS_sim*	Final_FoldX (kcal/mol)	C-Score**	TM-score*	RMSD (Å)	ρ (Temp_decay, FoldX_conv)
1A2PA	56.48	0.88	15.31	1.66	0.99	0.34	0.87
1ABAA	47.13	0.91	-9.62	0.3	0.93	0.97	0.75
1BKRA	36.11	0.93	-12.97	1.28	0.98	0.54	0.82
1DBWA	30.08	0.9	-20.42	0.85	0.87	1.81	0.83
1EW4A	28.3	0.89	-5.22	1.2	0.96	0.84	0.90
1F46A	40.29	0.88	-0.84	1.67	0.95	1.05	0.88
1GBSA	51.35	0.86	33.49	1.56	1.00	0.3	0.74
1GUTA	31.34	0.69	5.76	0.24	0.92	0.92	0.63
1HZTA	34.64	0.82	1.06	0.78	0.98	0.7	0.84
1I2TA	27.87	0.69	-15.52	0.18	0.96	0.56	0.58
1IDPA	31.97	0.8	13.01	0.91	0.97	0.77	0.84
1IUJA	22.55	0.84	8.22	0.86	0.95	0.99	0.89
1JB3A	27.56	0.79	15.69	1.5	0.98	0.58	0.83
1JF8A	33.85	0.9	2.82	0.88	0.94	1.26	0.79
1KMTA	53.62	0.77	7.18	1.79	0.94	1.14	0.83
1KNGA	30.56	0.88	2.96	1.01	0.99	0.54	0.65
1KQ1A	21.67	0.68	-4.51	0.12	0.96	0.56	0.93
1M9ZA	34.29	0.8	36.32	1.12	0.90	1.51	0.89
1MF7A	35.05	0.81	-9.07	1.19	0.97	1.15	0.86
1MG4A	40.59	0.89	-0.99	1.2	0.92	1.51	0.89
1NXMA	32.99	0.8	24.44	0.74	0.90	2.1	0.84
1O7IA	28.7	0.86	-1.68	0.86	0.98	0.52	0.81
1OAIA	22.03	0.81	-11.04	-0.85	0.82	1.75	0.85
1OH0A	30.4	0.86	-7.02	0.69	0.95	1.08	0.83
1OK0A	18.92	0.59	5.21	-2.54	0.46	3.67	0.75
1QHQA	33.09	0.81	20.74	0.98	0.99	0.41	0.73
1R26A	24.78	0.93	3.01	0.43	0.90	1.53	0.81
1R6JA	30.49	0.95	-3.42	0.24	0.92	1.01	0.75
1SHUX	32.04	0.91	-22.95	1.16	0.99	0.62	0.77
1T3YA	24.43	0.77	-6.6	0.68	0.94	1.22	0.88
1TQGA	21.9	0.91	-31.22	0.16	0.97	0.74	0.85
1TUKA	22.39	0.78	4.43	0.74	0.96	0.57	0.71
1UCSA	48.44	0.67	-5.66	1.04	0.97	0.56	0.56
1URRA	40.21	0.89	-2.95	1.23	0.97	0.75	0.74
1UTGA	18.57	0.67	-2.62	-0.7	0.95	0.76	0.83
1V5IB	26.32	0.7	-4.22	0.12	0.83	1.92	0.73

1VH5A	32.61	0.95	8.99	1.2	0.99	0.48	0.84
1VKKA	29.93	0.88	-8.33	1	0.94	1.59	0.81
1VZIA	48	0.8	29.26	1.52	0.99	0.43	0.90
1WLUA	34.19	0.94	-3.99	1.3	0.97	0.67	0.81
1X6ZA	42.86	0.87	17.75	0.32	0.96	1.13	0.69
1XTEA	23.28	0.92	-13.79	0.28	0.90	1.52	0.91
1ZHVA	25.37	0.8	12.66	1.08	0.98	0.59	0.85
1ZKEA	16.05	0.9	-19.68	-0.76	0.84	1.88	0.78
1ZZKA	28.75	0.89	-10.21	0.41	0.87	1.67	0.76
2ANXA	30.14	0.8	-7.05	0.89	0.95	0.67	0.87
2BWFA	32.47	0.84	-19.54	0.7	0.90	1.51	0.85
2C9QA	38.24	0.77	-9.3	0.62	0.92	1.32	0.70
2CARA	42.86	0.85	9.62	1.59	0.98	0.77	0.80
2CMPA	23.21	0.53	12.33	-1.12	0.80	1.79	0.68
2CVIA	22.89	0.64	-8.74	0.62	0.93	1.04	0.78
2D3DA	30.12	0.77	-8.51	0.61	0.92	0.98	0.78
2ERBA	31.71	0.95	2.99	1.12	0.99	0.43	0.90
2F01A	47.93	0.85	20.85	0.96	0.97	0.96	0.92
2GMYA	13.61	0.78	-5.94	0.07	0.99	0.34	0.88
2J2JA	38.46	0.77	20.82	1.1	0.99	0.55	0.83
2J5YA	24.59	0.79	-12.95	-1.58	0.85	1.36	0.77
2J8BA	29.49	0.73	5.43	-0.49	0.92	1.04	0.87
2O9SA	35.82	0.87	7.54	0.23	0.86	1.57	0.69
2P5KA	19.05	0.79	-3.07	-1.63	0.73	1.85	0.81
2PTHA	52.33	0.9	-0.22	1.49	0.99	0.77	0.71
2PV2A	29.13	0.86	-8.26	1.19	0.97	0.66	0.70
2QCPX	38.75	0.93	-12.88	1.08	0.94	1.23	0.85
2V0UA	36.99	0.88	-6.46	1.05	0.94	0.82	0.86
2V1QA	36.67	0.83	-8.83	0.43	0.94	0.87	0.80
2VMHA	40.14	0.85	27.7	1.47	0.99	0.59	0.72
2VPBA	26.32	0.68	19.11	-2.5	0.34	4.35	0.81
2VZCA	40.94	0.87	-24.33	0.81	0.96	0.88	0.75
2WLVA	42.36	0.87	13.29	0.97	0.90	1.79	0.87
2ZXYA	37.21	0.92	-0.87	0.95	0.99	0.37	0.72
3CTGA	37.04	0.87	-24.19	1.15	0.97	0.83	0.80
3E9TA	33.93	0.89	21.17	0.93	0.98	0.62	0.81
3FEAA	50.6	0.83	-16.27	1.35	0.98	0.5	0.77
3FILA	49.09	0.85	-9.52	1.1	0.94	0.63	0.71
3G21A	31.17	0.78	-4.59	-0.22	0.86	1.94	0.87
3VUBA	21.78	0.66	6.02	0.53	0.97	0.7	0.90
3I4OA	26.47	0.74	-3.98	0.98	0.92	1.22	0.79

§Fifth character in PDB ID indicates the chain ID;

*Value varies from 0 to 1;

**Value varies from -5 to 2

Table S2. Conservation of residues in designed proteins

Dataset	Residue Conservation			Residue Conservation % (Intra-group)		
	All	Buried	Hydrophobic in Buried	All	Buried	Hydrophobic in Buried
Entire	34.1%	48.7%	59.4%	68.4%	77.7%	93.4%
Representative	28.7%	40.8%	52.8%	68.3%	76.5%	97.2%

Table S3. Conservation of individual residues in designed proteins

Residue	Entire (%)	Buried (%)	Exposed (%)
Alanine	41.82	66.67	19.92
Cysteine	22.14	15.87	32.00
Aspartate	22.69	22.22	21.41
Glutamate	19.52	8.00	21.58
Phenylalanine	38.59	44.81	16.22
Glycine	71.99	77.00	75.00
Histidine	12.37	12.12	12.77
Isoleucine	44.94	52.91	25.26
Lysine	20.85	25.00	20.56
Leucine	52.46	64.20	27.22
Methionine	18.10	30.51	10.13
Aspargine	15.21	26.32	16.19
Proline	71.74	90.48	65.64
Glutamine	13.07	20.69	11.88
Arginine	13.30	11.76	15.79
Serine	18.71	20.93	17.75
Threonine	18.00	33.33	10.53
Valine	46.71	62.36	17.24
Tryptophan	33.90	37.21	10.53
Tyrosine	27.35	24.42	25.58

Table S4. Summary of MD simulation results on the representative dataset. We consider chain A for all the proteins.

PDB ID	Protein Length	SS_sim (%)	RMSD (Å)	TM_Score	RMSD (Avg.)		RG (Avg.)		ρ(RMSF$_N$, RMSF$_D$)
					N*	D**	N*	D**	
1R26	113	93	1.53	0.90	0.37	0.29	1.14	1.08	0.66
1XTE	116	92	1.52	0.90	0.29	0.35	1.26	1.21	0.66
1ZZK	80	89	1.67	0.87	0.16	0.31	1.04	1.04	0.53
2V0U	146	88	0.82	0.94	0.33	0.2	1.23	1.22	0.86
3I40	68	74	1.22	0.92	0.19	0.24	0.93	0.9	0.82

*N denotes native protein

*D denotes designed protein

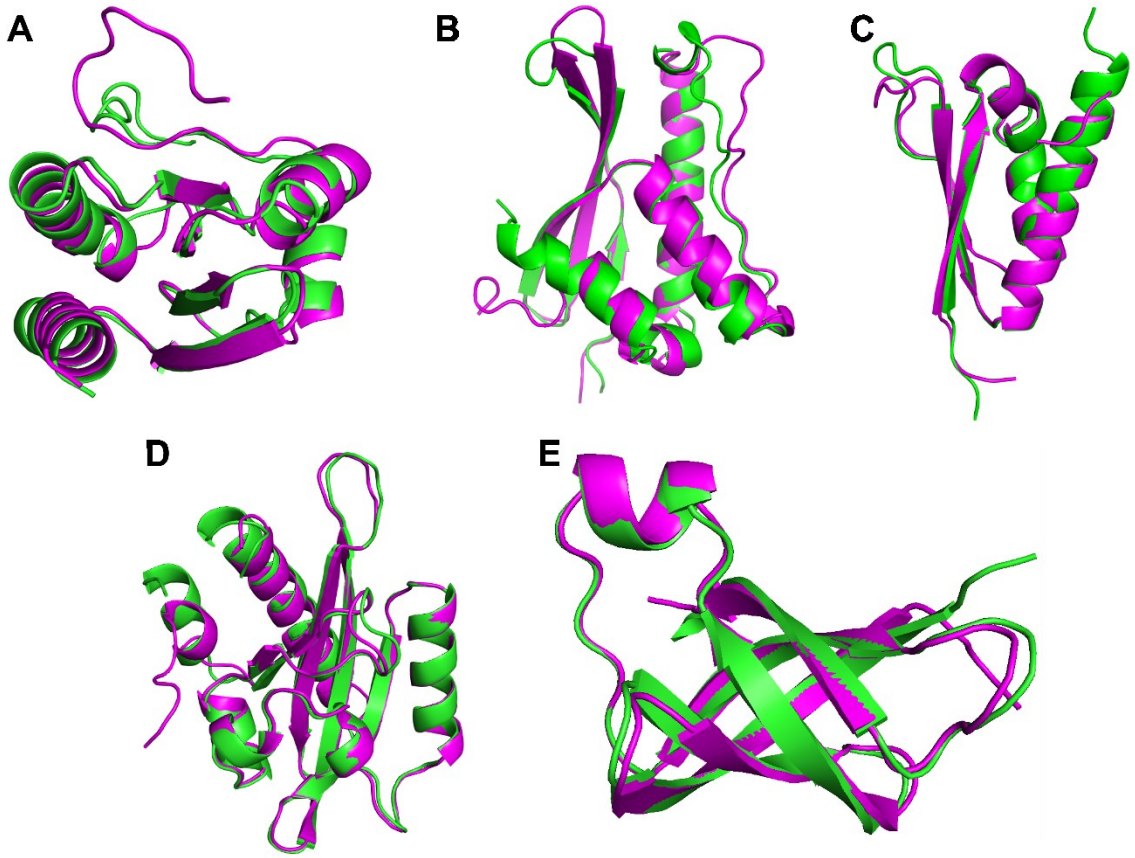

Figure S1. RoseTTAFold predicted structures (in green) of proteins in the representative dataset aligned with target scaffolds (in magenta) bearing PDB IDs- (A) 1R26, chain A (RMSD=1.19 Å), (B) 1XTE chain A (RMSD=0.89 Å), (C) 1ZZK, chain A (RMSD=0.64 Å), (D) 2V0U, chain A (RMSD=0.68 Å), and (E) 3I4O, chain A (RMSD=0.43 Å).

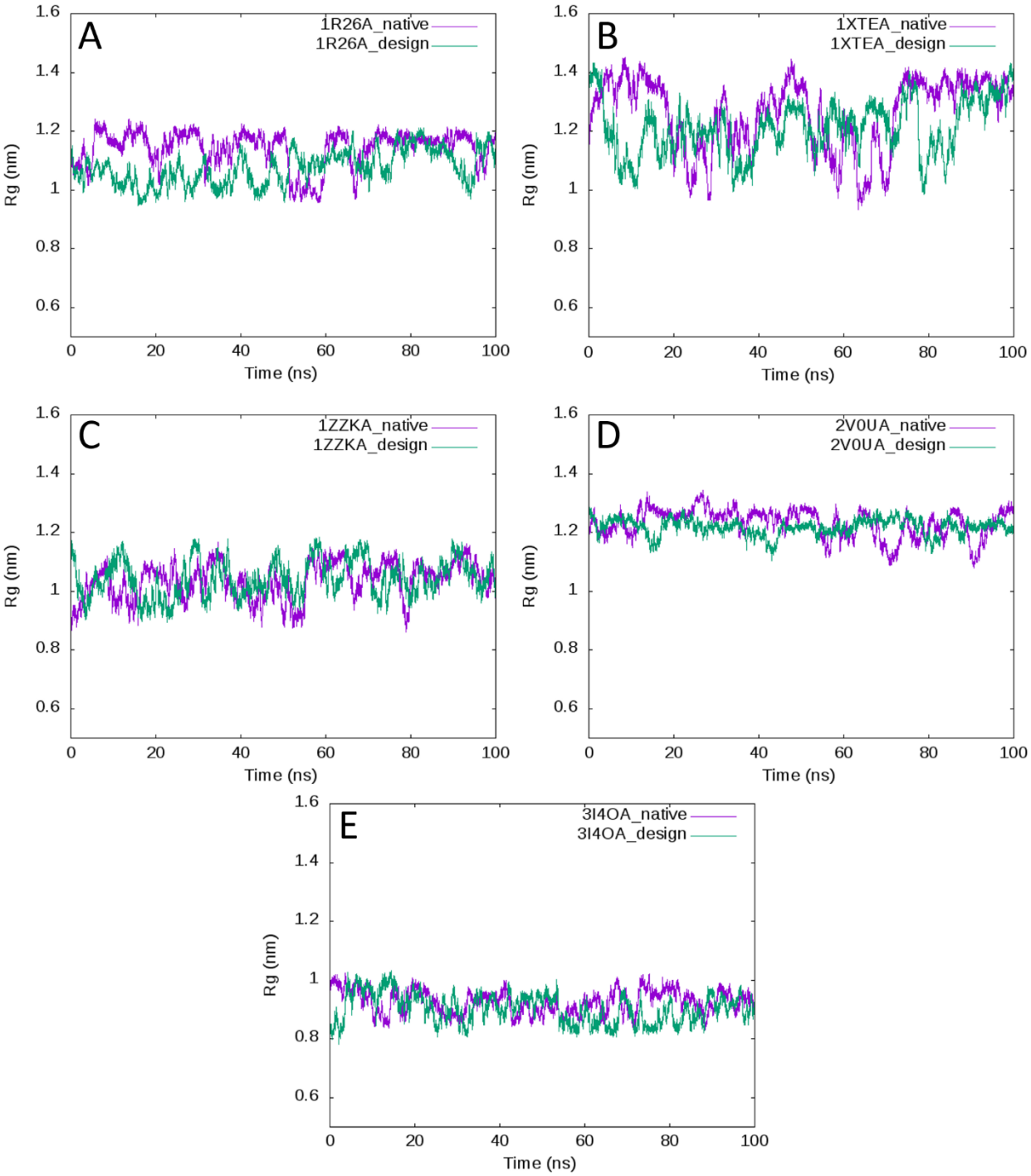

Figure S2. Radius of gyration plots of the design predicted structures (in lime green) and the target scaffolds (in magenta) across 100ns of MD simulation for the proteins bearing PDB IDs- (A) 1R26, chain A, (B) 1XTE chain A, (C) 1ZZK, chain A, (D) 2V0U, chain A, and (E) 3I4O, chain A.

Table S5. FoldX energy ($\Delta G_{folding}$) analysis of ClustENMD generated conformations for each native and designed protein in the representative dataset.

PDB IDs	FoldX energy $\Delta G_{folding}$ (kcal/mol)			
	Native Protein		Designed Protein	
	Mean ± SD	Minimum	Mean ± SD	Minimum
1R26	66.52 ± 7.24	48.79	56.34 ± 6.65	42.58
1XTE	52.71 ± 6.20	38.92	50.72 ± 7.06	31.13
1ZZK	39.92 ± 4.94	31.39	43.26 ± 5.58	28.70
2V0U	77.54 ± 9.55	56.56	60.05 ± 9.20	41.20
3I4O	30.83 ± 5.51	16.26	30.49 ± 4.99	17.76

Table S6. COFACTOR predicted GO Molecular Function (confidence score greater than 0.5) of the target scaffold (bearing native amino acid sequence) and of the design predicted structure (bearing designed sequence). Similar molecular function is color coded in green.

PDB ID:1R26 Chain A					
Native			**Design**		
GO Index	**C-Score**	**Molecular Function**	**GO Index**	**C-Score**	**Molecular Function**
GO:0003824	0.99	catalytic activity	GO:0003824	1	catalytic activity
GO:0016491	0.85	oxidoreductase activity	GO:0015035	0.96	protein disulfide oxidoreductase activity
GO:0016667	0.81	oxidoreductase activity, acting on a sulfur group of donors	GO:0016671	0.78	oxidoreductase activity, acting on a sulfur group of donors, disulfide as acceptor
GO:0015036	0.79	disulfide oxidoreductase activity	GO:0003756	0.54	protein disulfide isomerase activity
GO:0015035	0.78	protein disulfide oxidoreductase activity			
GO:0016671	0.69	oxidoreductase activity, acting on a sulfur group of donors, disulfide as acceptor			
GO:1901363	0.52	heterocyclic compound binding			
GO:0097159	0.52	organic cyclic compound binding			
PDB ID:1XTE Chain A					
Native			**Design**		

GO Index	C-Score	Molecular Function	GO Index	C-Score	Molecular Function
GO:0035091	0.69	phosphatidylinositol binding	GO:0005543	0.79	phospholipid binding
			GO:0035091	0.78	phosphatidylinositol binding
			GO:0003824	0.61	catalytic activity
			GO:0016740	0.6	transferase activity
			GO:0016773	0.59	phosphotransferase activity, alcohol group as acceptor
			GO:0016301	0.59	kinase activity
			GO:0098772	0.58	molecular function regulator
			GO:0004672	0.57	protein kinase activity
			GO:0016247	0.55	channel regulator activity

PDB ID:1ZZK Chain A					
Native			**Design**		
GO Index	**C-Score**	**Molecular Function**	**GO Index**	**C-Score**	**Molecular Function**
GO:1901363	0.95	heterocyclic compound binding	GO:0003676	0.93	nucleic acid binding
GO:0097159	0.95	organic cyclic compound binding	GO:0003723	0.92	RNA binding
GO:0003676	0.93	nucleic acid binding	GO:0044822	0.85	poly(A) RNA binding
GO:0003723	0.92	RNA binding	GO:0003729	0.64	mRNA binding
GO:0044822	0.86	poly(A) RNA binding	GO:0003700	0.6	transcription factor activity, sequence-specific DNA binding
GO:0003729	0.64	mRNA binding	GO:0000981	0.59	RNA polymerase II transcription factor activity, sequence-specific DNA binding
			GO:0001228	0.58	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding
			GO:0001205	0.54	transcriptional activator activity, RNA polymerase II distal enhancer sequence-specific binding

PDB ID:2V0U Chain A					
Native			**Design**		
GO Index	**C-Score**	**Molecular Function**	**GO Index**	**C-Score**	**Molecular Function**
GO:0016740	0.67	transferase activity	GO:0004672	0.85	protein kinase activity
GO:1901363	0.56	heterocyclic compound binding	GO:0038023	0.79	signaling receptor activity
GO:0097159	0.56	organic cyclic compound binding	GO:0000155	0.78	phosphorelay sensor kinase activity
GO:0004672	0.56	protein kinase activity	GO:0032553	0.71	ribonucleotide binding
GO:0004674	0.55	protein serine/threonine kinase activity	GO:0004674	0.71	protein serine/threonine kinase activity
GO:0005249	0.51	voltage-gated potassium channel activity	GO:0005524	0.7	ATP binding
			GO:0005249	0.62	voltage-gated potassium channel activity

PDB ID:3I4O Chain A					
Native			**Design**		
GO Index	**C-Score**	**Molecular Function**	**GO Index**	**C-Score**	**Molecular Function**
GO:1901363	0.97	heterocyclic compound binding	GO:0003743	0.88	translation initiation factor activity
GO:0097159	0.97	organic cyclic compound binding	GO:0043022	0.87	ribosome binding
GO:0003723	0.95	RNA binding	GO:0019843	0.64	rRNA binding
GO:0008135	0.89	translation factor activity, RNA binding			
GO:0003743	0.88	translation initiation factor activity			
GO:0043021	0.85	ribonucleoprotein complex binding			
GO:0043022	0.84	ribosome binding			
GO:0019843	0.83	rRNA binding			

Table S7. COFACTOR predicted binding sites (confidence score greater than 0.35) of the target scaffold (bearing native amino acid sequence) and of the design predicted structure (bearing designed sequence). Similar binding site color coded in green.

PDB ID- 1ZZK Chain A																
	1		2							3	4		5	6	7	
N	7	8	0	1	2	3	4	7	8	1	0	2	0	8	5	
D	7	8	0	1	2	3	4			1	0	2	0			

PDB ID- 2V0U Chain A																												
	1		2	3	4	5				6				7				8		9				10	11			
N	6	8	5	8	9	0	1	3	4		3		6	7	0	2	6	8	0	1	2	2	4	6	9	0	1	3
D	6	8			9	0	1	3	4	0		4		7	0						2	2	4	6	9	0	1	3

Residues renumbered from index one. The first row for individual protein entries indicates the preceding digits of the residue index, while the following digits are mentioned in the next two rows. N stands for the target scaffold bearing native amino acid sequence and D stands for the design predicted structure bearing the design sequence.