Supporting Information

Estimating Change in Foldability due to Multi-point Deletions in Protein Structures

Anupam Banerjee[1], Amit Kumar[2], Kushal Kanti Ghosh[3], Pralay Mitra[2]*

[1]Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, West Bengal 721302, India

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

[3]Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India

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

Table of Contents

Dataset Preparation

While preparing the 4350 unlabeled MPD entries corresponding to deletion in the loop region, we consult the distribution of the deletion ratio (ratio between the deletion length to the total length of the loop in consideration) (varying from 0.1 to 0.8), deletion length (varying from 2 to 15) and the terminal (N-terminal, C-terminal, and middle region) where the deletion takes place in the 87 positive MPD instances (Figure 1B-1D). Although we consult the distribution of the positive instances, we additionally ensure that we have unlabeled MPD entries corresponding to deletion ratios or deletion lengths, which have a minimum or no representation in the positive dataset. Towards that direction, we allocate an equal probability to each observation (for deletion length, deletion ratio, and deletion terminal) and then, while considering the distribution of each such observation from the positive samples, prepare a cumulative distribution for the entire range of values for that particular attribute. While preparing an unlabeled MPD instance in the loop region, we generate three random numbers (lying between 0 to 2) and based on where those numbers lie on the cumulatative distributions select a deletion ratio, deletion length and deletion terminal specification. Next, we consider a deletion with the same specifications from a randomly chosen loop in a randomly chosen protein from the curated dataset. The distribution of the deletion length, deletion ratio, and deletion terminal for the loop region in the positive and unlabeled samples are shown in Figure 1F-1H. For the 66 positive non-loop MPD instances, we adhere to the same procedure to prepare 3300 unlabeled MPD instances. In the case of the unlabeled non-loop MPD instances, the distribution of only the loop lengths (varying from 2 to 23) in the positive deletion instances are considered, and a single random number is generated to chose a deletion length from the cumulative distribution. The deletion with the same deletion length specification is considered at a randomly chosen residue starting index (from where the MPD will start) in a randomly chosen protein from the curated protein dataset. The distribution of deletion lengths for the non-loop region in the positive and unlabeled samples is also provided in Figure 1E and Figure 1I, respectively. Although we list both the *ProteinL* and *ProteinSm* for the 153 positive MPD instances in Supplementary Table S1, we use only the corresponding *ProteinL*s to encode the deletion effects on foldability in the PU learning classifier. Deletion details corresponding to the unlabeled instances can be found in the metadata files in the PROFOUND_database folder in the supporting data segment in the webserver.

Table S1. Details of the 153 (87 Loop indicated by L and 66 Non-loop indicated by NL in the ID field) *ProteinL*s (colored green) and *ProteinSm*s considered for the database

ID	PDB	Chain	Organism	Protein	SCOP	Cov
L1	1M3C	A	MUS MUSCULUS	PROTO-ONCOGENE C-CRK	b	1.00
	1M3B	A	MUS MUSCULUS	PROTO-ONCOGENE C-CRK	b	1.00
L2	1O1M	A	HOMO SAPIENS	HEMOGLOBIN ALPHA CHAIN	a	1.00
	1O1J	A	HOMO SAPIENS	HEMOGLOBIN ALPHA CHAIN	a	1.00
L3	1XV8	B	HOMO SAPIENS	ALPHA-AMYLASE	b	1.00
	1JXK	A	HOMO SAPIENS	ALPHA-AMYLASE, SALIVARY	b	1.00
L4	2EE5	A	HOMO SAPIENS	RHO GTPASE ACTIVATING PROTEIN 5 VARIANT	a	1.00
	2EE4	A	HOMO SAPIENS	RHO GTPASE ACTIVATING PROTEIN 5 VARIANT	a	1.00
L5	2ESW	A	MUS MUSCULUS	RHO GUANINE NUCLEOTIDE EXCHANGE FACTOR 7	b	1.00
	2G6F	X	RATTUS NORVEGICUS	RHO GUANINE NUCLEOTIDE EXCHANGE FACTOR 7	b	0.98
L6	2GBJ	B	HOMO SAPIENS	UBIQUITIN	d	0.99
	1S1Q	D	HOMO SAPIENS	UBIQUITIN	d	1.00
L7	2GBJ	B	HOMO SAPIENS	UBIQUITIN	d	0.98
	4MDK	F	HOMO SAPIENS	UBIQUITIN	d	0.97
L8	2GQG	B	HOMO SAPIENS	PROTO-ONCOGENE TYROSINE-PROTEIN KINASE ABL1	d	1.00
	2HIW	B	HOMO SAPIENS	PROTO-ONCOGENE TYROSINE-PROTEIN KINASE ABL1	d	1.00
L9	2MD5	A	MUS MUSCULUS	TRANSCRIPTION FACTOR ETV6	a	1.00
	2LF8	A	MUS MUSCULUS	TRANSCRIPTION FACTOR ETV6	x	0.75
L10	2UWA	A	TROPAEOLUM MAJUS	CELLULASE	b	0.99
	2VH9	B	TROPAEOLUM MAJUS	CELLULASE	b	1.00
L11	2YRY	A	HOMO SAPIENS	COMPND 3 MEMBER 6	b	1.00
	2D9Y	A	HOMO SAPIENS	COMPND 3 FAMILY A MEMBER 6	b	0.95
L12	2YT1	A	MUS MUSCULUS	COMPND 3 PRECURSOR PROTEIN-BINDING FAMILY B MEMBER 2	x	1.00
	2YT0	A	MUS MUSCULUS	COMPND 3 PRECURSOR PROTEIN-BINDING FAMILY B MEMBER 2	x	1.00
L13	2Z24	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	0.99
	2Z2B	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	1.00
L14	2E25	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	0.99
	2Z2B	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	1.00
L15	1XGE	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	0.99
	2Z2B	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	1.00
L16	2Z28	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	0.99
	2Z2B	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	1.00

L17	2Z25	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	0.99
	2Z2B	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	1.00
L18	2Z29	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	0.99
	2Z2B	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	1.00
L19	2Z2A	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	0.99
	2Z2B	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	1.00
L20	2Z26	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	0.99
	2Z2B	A	ESCHERICHIA COLI	DIHYDROOROTASE	c	1.00
L21	1KNB	A	HUMAN ADENOVIRUS 5	ADENOVIRUS TYPE 5 FIBER PROTEIN	b	1.00
	4ATZ	A	UNIDENTIFIED ADENOVIRUS	FIBER PROTEIN	x	1.00
L22	2VX9	A	HALOBACTERIUM SALINARUM R1	DODECIN	d	1.00
	4B2H	A	HALOBACTERIUM SALINARUM	DODECIN	d	0.98
L23	1PWK	A	RATTUS NORVEGICUS	DYNEIN LIGHT CHAIN-2	d	1.00
	4D07	A	HOMO SAPIENS	DYNEIN LIGHT CHAIN 2, CYTOPLASMIC	x	0.97
L24	4G79	A	CAENORHABDITIS ELEGANS	SPINDLE ASSEMBLY ABNORMAL PROTEIN 6	x	0.99
	4GFA	C	CAENORHABDITIS ELEGANS	SPINDLE ASSEMBLY ABNORMAL PROTEIN 6	x	0.85
L25	2N69	A	PENTADIPLANDRA BRAZZEANA	DEFENSIN-LIKE PROTEIN	x	0.98
	4HE7	A	PENTADIPLANDRA BRAZZEANA	DEFENSIN-LIKE PROTEIN	g	1.00
L26	4NUP	C	MUS MUSCULUS	N-CADHERIN EC1-2	x	1.00
	2QVI	A	MUS MUSCULUS	CADHERIN-2	b	1.00
L27	3IWI	B	ESCHERICHIA COLI	BETA-LACTAMASE	e	1.00
	4OLG	B	ESCHERICHIA COLI	BETA-LACTAMASE	e	1.00
L28	4F01	B	ESCHERICHIA COLI	CHAPERONE PROTEIN DNAK	a	0.97
	4R5I	A	ESCHERICHIA COLI	CHAPERONE PROTEIN DNAK	x	1.00
L29	4R5L	D	ESCHERICHIA COLI	CHAPERONE PROTEIN DNAK	a	0.98
	4R5J	D	ESCHERICHIA COLI	CHAPERONE PROTEIN DNAK	a	1.00
L30	2X6W	A	ENTEROBACTERIA PHAGE HK620	TAIL SPIKE PROTEIN	x	1.00
	4XQI	A	ENTEROBACTERIA PHAGE HK620	TAIL SPIKE PROTEIN	x	0.99
L31	1IP2	A	HOMO SAPIENS	LYSOZYME C	d	1.00
	1DI4	A	HOMO SAPIENS	LYSOZYME C	d	1.00
L32	1LHL	A	HOMO SAPIENS	HUMAN LYSOZYME	d	1.00
	1DI4	A	HOMO SAPIENS	LYSOZYME C	d	1.00
L33	1FRS	B	ENTEROBACTERIA PHAGE FR	BACTERIOPHAGE FR CAPSID	d	1.00
	1FR5	C	ENTEROBACTERIA PHAGE FR	BACTERIOPHAGE FR CAPSID	d	1.00
L34	3LI0	A	METHANOTHERMOBACTER THERMAUTOTROPHICUS	OROTIDINE 5'-PHOSPHATE DECARBOXYLASE	c	0.96
	1LOS	D	METHANOTHERMO	OROTIDINE MONOPHOSPHATE	c	1.00

			BACTER	DECARBOXYLASE		
L35	4FX8	B	METHANOTHERMOBACTER THERMAUTOTROPHICUS STR.	OROTIDINE 5'-PHOSPHATE DECARBOXYLASE	c	0.99
	1LOS	D	METHANOTHERMOBACTER	OROTIDINE MONOPHOSPHATE DECARBOXYLASE	c	0.98
L36	1LZ5	A	HOMO SAPIENS	HUMAN LYSOZYME	d	1.00
	1B7S	A	HOMO SAPIENS	LYSOZYME	d	1.00
L37	1LZ6	A	HOMO SAPIENS	HUMAN LYSOZYME	d	1.00
	1OUH	A	HOMO SAPIENS	LYSOZYME	d	1.00
L38	1SYG	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L39	1SYE	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L40	1F2Z	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L41	1SYC	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L42	1KDC	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L43	1F2M	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L44	1KDB	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L45	1KAA	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L46	1SNC	A	STAPHYLOCOCCUS AUREUS	THERMONUCLEASE PRECURSOR	b	1.00
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L47	1KDA	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00
L48	1KAB	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.99
	1SND	B	STAPHYLOCOCCUS	STAPHYLOCOCCAL NUCLEASE DIMER	b	1.00

			AUREUS			
L49	2UY9	A	BACILLUS SUBTILIS	OXALATE DECARBOXYLASE OXDC	b	1.00
	2UYA	A	BACILLUS SUBTILIS	OXALATE DECARBOXYLASE OXDC	b	1.00
L50	3LKY	A	GRIFFITHSIA	GRIFFITHSIN	x	1.00
	2GTY	A	GRIFFITHSIA	GRIFFITHSIN	b	1.00
L51	3LL0	A	GRIFFITHSIA	GRIFFITHSIN	x	1.00
	3LKY	A	GRIFFITHSIA	GRIFFITHSIN	x	1.00
L52	3KRA	C	MENTHA X PIPERITA	GERANYL DIPHOSPHATE SYNTHASE SMALL SUBUNIT	x	0.95
	3OAC	C	MENTHA X PIPERITA	GERANYL DIPHOSPHATE SYNTHASE SMALL SUBUNIT	x	1.00
L53	2NPP	C	HOMO SAPIENS	COMPND 18 ALPHA ISOFORM	d	1.00
	3P71	C	HOMO SAPIENS	COMPND 10 ALPHA ISOFORM	d	1.00
L54	3V0G	C	CIONA INTESTINALIS	VOLTAGE-SENSOR CONTAINING PHOSPHATASE	x	1.00
	3V0J	B	CIONA INTESTINALIS	VOLTAGE-SENSOR CONTAINING PHOSPHATASE	x	0.99
L55	4K03	B	DROSOPHILA MELANOGASTER	CRYPTOCHROME-1	x	0.99
	4JZY	B	DROSOPHILA MELANOGASTER	CRYPTOCHROME-1	x	0.99
L56	4NXT	A	HOMO SAPIENS	MITOCHONDRIAL DYNAMIC PROTEIN MID51	x	0.98
	4NXX	A	HOMO SAPIENS	MITOCHONDRIAL DYNAMIC PROTEIN MID51	x	1.00
L57	4OYC	B	SALMONELLA TYPHIMURIUM	LIPOPROTEIN PRGK	x	1.00
	4OYC	A	SALMONELLA TYPHIMURIUM	LIPOPROTEIN PRGK	x	0.99
L58	4QK2	A	HOMO SAPIENS	CARBONIC ANHYDRASE 2	x	1.00
	4QK3	A	HOMO SAPIENS	CARBONIC ANHYDRASE 2	x	1.00
L59	2HKK	A	HOMO SAPIENS	CARBONIC ANHYDRASE 2	b	1.00
	4QK3	A	HOMO SAPIENS	CARBONIC ANHYDRASE 2	x	1.00
L60	5B3Y	A	HOMO SAPIENS, ESCHERICHIA COLI K-12	COMPND 3 MALTOSE-BINDING PERIPLASMIC PROTEIN	x	1.00
	5B3X	A	HOMO SAPIENS, ESCHERICHIA COLI K-12	COMPND 3 MALTOSE-BINDING PERIPLASMIC PROTEIN	x	0.99
L61	4LRM	D	HOMO SAPIENS	EPIDERMAL GROWTH FACTOR RECEPTOR	x	1.00
	5CAV	A	HOMO SAPIENS	EPIDERMAL GROWTH FACTOR RECEPTOR	x	0.88
L62	5CVD	A	HOMO SAPIENS	N-TERMINAL XAA-PRO-LYS N-METHYLTRANSFERASE 1	x	0.98
	5E2B	B	HOMO SAPIENS	N-TERMINAL XAA-PRO-LYS N-METHYLTRANSFERASE 1	x	1.00
L63	5E8E	H	HOMO SAPIENS	THROMBIN HEAVY CHAIN	x	0.99
	3GIC	B	HOMO SAPIENS	THROMBIN HEAVY CHAIN	x	1.00
L64	5EJW	A	MUS MUSCULUS	CHROMOBOX PROTEIN HOMOLOG 7	x	0.86
	4X3K	A	MUS MUSCULUS	CHROMOBOX PROTEIN HOMOLOG 7	x	1.00
L65	5GQM	A	BOMBYX MORI	POLYHEDRIN	x	1.00

			CYPOVIRUS 1			
	5GQJ	A	BOMBYX MORI CYPOVIRUS 1	POLYHEDRIN	x	1.00
L66	5GQI	A	BOMBYX MORI CYPOVIRUS 1	POLYHEDRIN	x	1.00
	5GQN	A	BOMBYX MORI CYPOVIRUS 1	POLYHEDRIN	x	1.00
L67	5EXY	A	BOMBYX MORI CYTOPLASMIC POLYHEDROSIS VIRUS	POLYHEDRIN	x	1.00
	5GQN	A	BOMBYX MORI CYPOVIRUS 1	POLYHEDRIN	x	1.00
L68	1HZB	A	BACILLUS CALDOLYTICUS	COLD SHOCK PROTEIN CSPB	b	0.98
	5JX4	B	BACILLUS CALDOLYTICUS	COLD SHOCK PROTEIN CSPB	x	0.95
L69	5JX8	B	VACCINIA VIRUS (STRAIN WESTERN RESERVE)	URACIL-DNA GLYCOSYLASE	x	0.99
	4IRB	B	VACCINIA VIRUS ANKARA	URACIL-DNA GLYCOSYLASE	x	1.00
L70	5K6B	F	HUMAN RESPIRATORY SYNCYTIAL VIRUS A	FUSION GLYCOPROTEIN F0	x	1.00
	5K6C	F	HUMAN RESPIRATORY SYNCYTIAL VIRUS A (STRAIN A2)	FUSION GLYCOPROTEIN F0	x	0.99
L71	5KC6	B	HOMO SAPIENS	CEREBELLIN-1	x	0.94
	5KWR	A	RATTUS NORVEGICUS	CEREBELLIN-1	x	0.97
L72	1LZ5	A	HOMO SAPIENS	HUMAN LYSOZYME	d	1.00
	5LVK	B	HOMO SAPIENS	LYSOZYME C	x	1.00
L73	1LZ6	A	HOMO SAPIENS	HUMAN LYSOZYME	d	1.00
	5LVK	B	HOMO SAPIENS	LYSOZYME C	x	1.00
L74	1LMT	A	HOMO SAPIENS	HUMAN LYSOZYME	d	1.00
	5LVK	B	HOMO SAPIENS	LYSOZYME C	x	1.00
L75	5LVK	B	HOMO SAPIENS	LYSOZYME C	x	1.00
	1DI4	A	HOMO SAPIENS	LYSOZYME C	d	1.00
L76	5MXN	f	VIBRIO CHOLERAE	TYPE VI SECRETION PROTEIN	x	1.00
	5OJQ	G	VIBRIO CHOLERAE	VIPA	x	1.00
L77	5BOZ	C	RICINUS COMMUNIS	RICIN	x	0.68
	5SV3	D	RICINUS COMMUNIS	RICIN	x	1.00
L78	5THF	C	INFLUENZA A VIRUS	HEMAGGLUTININ HA1 CHAIN	x	0.85
	5UMN	A	INFLUENZA A VIRUS	HEMAGGLUTININ	x	1.00
L79	5VG3	C	BACILLUS SUBTILIS	OXALATE DECARBOXYLASE	x	1.00
	2UYA	A	BACILLUS SUBTILIS	OXALATE DECARBOXYLASE OXDC	b	1.00

L80	1K5M	B	HUMAN RHINOVIRUS 14	CHIMERA OF HRV14 COAT PROTEIN VP2 (P1B) AND THE V3 LOOP OF HIV-1 GP120	x	1.00
	5W3M	C	HUMAN RHINOVIRUS 14	VIRAL PROTEIN 2	x	1.00
L81	5W57	A	PARACOCCUS DENITRIFICANS (STRAIN PD 1222)	PERIPLASMIC SOLUTE BINDING PROTEIN	x	1.00
	5KZJ	A	PARACOCCUS DENITRIFICANS (STRAIN PD 1222)	PERIPLASMIC SOLUTE BINDING PROTEIN	x	1.00
L82	4RCA	A	HOMO SAPIENS	RECEPTOR-TYPE TYROSINE-PROTEIN PHOSPHATASE DELTA	x	0.69
	2YD7	A	HOMO SAPIENS	PTPRD PROTEIN	b	1.00
L83	3LL0	A	GRIFFITHSIA	GRIFFITHSIN	x	1.00
	2GTY	A	GRIFFITHSIA	GRIFFITHSIN	b	1.00
L84	3V0E	A	CIONA INTESTINALIS	VOLTAGE-SENSOR CONTAINING PHOSPHATASE	x	1.00
	3V0J	B	CIONA INTESTINALIS	VOLTAGE-SENSOR CONTAINING PHOSPHATASE	x	0.98
L85	4GNQ	A	RATTUS NORVEGICUS	PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC [GTP]	c	1.00
	4GMM	A	RATTUS NORVEGICUS	PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC [GTP]	c	1.00
L86	1E21	A	HOMO SAPIENS	RIBONUCLEASE 1	d	0.98
	4KXH	D	HOMO SAPIENS	RIBONUCLEASE PANCREATIC	d	0.94
L87	5LWF	B	BACILLUS LICHENIFORMIS	BETA-LACTAMASE	x	1.00
	4BLM	A	BACILLUS LICHENIFORMIS	BETA-LACTAMASE	e	1.00
NL1	1D1M	B	ENTEROBACTERIA PHAGE LAMBDA	LAMBDA CRO REPRESSOR	a	1.00
	1D1L	A	ENTEROBACTERIA PHAGE LAMBDA	LAMBDA CRO REPRESSOR	a	0.98
NL2	2DFF	A	THERMOCOCCUS KODAKARENSIS	RIBONUCLEASE HII	c	1.00
	2DFE	A	THERMOCOCCUS KODAKARENSIS	RIBONUCLEASE HII	c	0.99
NL3	2DFH	A	THERMOCOCCUS KODAKARENSIS	RIBONUCLEASE HII	c	0.99
	2DFF	A	THERMOCOCCUS KODAKARENSIS	RIBONUCLEASE HII	c	1.00
NL4	2DFH	A	THERMOCOCCUS KODAKARENSIS	RIBONUCLEASE HII	c	1.00
	2DFE	A	THERMOCOCCUS KODAKARENSIS	RIBONUCLEASE HII	c	1.00
NL5	2DMF	A	HOMO SAPIENS	RING FINGER PROTEIN 25	d	1.00
	2DAY	A	HOMO SAPIENS	RING FINGER PROTEIN 25	d	1.00
NL6	2GBK	D	HOMO SAPIENS	UBIQUITIN	d	1.00
	3EEC	A	HOMO SAPIENS	UBIQUITIN	d	1.00
NL7	2KJK	A	LISTERIA INNOCUA	LIN2157 PROTEIN	x	0.94
	3I1E	B	LISTERIA INNOCUA	LIN2157 PROTEIN	x	1.00
NL8	3Q29	C	ESCHERICHIA COLI, HOMO	CHIMERIC PROTEIN	c	0.99

			SAPIENS			
	3Q26	A	ESCHERICHIA COLI, HOMO SAPIENS	CHIMERIC PROTEIN	x	0.92
NL9	2KY9	A	BACILLUS SUBTILIS	UNCHARACTERIZED PROTEIN YDHK	x	0.95
	4FIB	C	BACILLUS SUBTILIS SUBSP. SUBTILIS	UNCHARACTERIZED PROTEIN YDHK	x	1.00
NL10	4O4F	A	ENTAMOEBA HISTOLYTICA	INOSITOL HEXAKISPHOSPHATE KINASE	x	1.00
	4O4B	B	ENTAMOEBA HISTOLYTICA	INOSITOL HEXAKISPHOSPHATE KINASE	x	0.97
NL11	104L	B	ENTEROBACTERIA PHAGE T4	T4 LYSOZYME	d	1.00
	1L68	A	ENTEROBACTERIA PHAGE T4	LYSOZYME	d	1.00
NL12	104L	B	ENTEROBACTERIA PHAGE T4	T4 LYSOZYME	d	1.00
	171L	A	ENTEROBACTERIA PHAGE T4	T4 LYSOZYME	d	1.00
NL13	3IFK	A	RATTUS NORVEGICUS	CALMODULIN	x	1.00
	1AHR	A	GALLUS GALLUS	CALMODULIN	a	0.58
NL14	1C7P	A	HOMO SAPIENS	LYSOZYME	d	0.98
	1DI4	A	HOMO SAPIENS	LYSOZYME C	d	1.00
NL15	1FRS	A	ENTEROBACTERIA PHAGE FR	BACTERIOPHAGE FR CAPSID	d	1.00
	1FR5	C	ENTEROBACTERIA PHAGE FR	BACTERIOPHAGE FR CAPSID	d	1.00
NL16	1IFG	A	ESCHERICHIA COLI	ECOTIN	b	1.00
	1AZZ	C	ESCHERICHIA COLI	ECOTIN	b	1.00
NL17	1MB8	A	HOMO SAPIENS	PLECTIN	a	0.95
	4Q59	A	HOMO SAPIENS	PLECTIN	a	1.00
NL18	1SRA	A	HOMO SAPIENS	SPARC	a	1.00
	1NUB	B	HOMO SAPIENS	BASEMENT MEMBRANE PROTEIN BM-40	a	0.63
NL19	1RJ7	H	HOMO SAPIENS	ECTODYSPLASIN A	b	0.99
	1RJ8	G	HOMO SAPIENS	ECTODYSPLASIN-A ISOFORM EDA-A2	b	0.99
NL20	1S16	A	ESCHERICHIA COLI	TOPOISOMERASE IV SUBUNIT B	d	0.53
	1S14	B	ESCHERICHIA COLI	TOPOISOMERASE IV SUBUNIT B	d	1.00
NL21	1STA	A	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE	b	0.77
	2F3W	A	STAPHYLOCOCCUS AUREUS	THERMONUCLEASE	b	0.95
NL22	209L	A	ENTEROBACTERIA PHAGE T4	T4 LYSOZYME	d	1.00
	210L	A	ENTEROBACTERIA PHAGE T4	T4 LYSOZYME	d	1.00
NL23	2CME	B	HUMAN SARS CORONAVIRUS	HYPOTHETICAL PROTEIN 5	b	0.99
	2CME	H	HUMAN SARS CORONAVIRUS	HYPOTHETICAL PROTEIN 5	b	1.00
NL24	1M7V	A	BACILLUS SUBTILIS	NITRIC OXIDE SYNTHASE	d	1.00
	2FC1	A	BACILLUS SUBTILIS	NITRIC OXIDE SYNTHASE	d	1.00

NL25	3AFI	E	BRADYRHIZOBIUM JAPONICUM	HALOALKANE DEHALOGENASE	c	0.98
	3A2L	B	BRADYRHIZOBIUM JAPONICUM	HALOALKANE DEHALOGENASE	c	0.99
NL26	3AAK	A	HOMO SAPIENS	PROGRAMMED CELL DEATH PROTEIN 6	a	1.00
	3AAJ	A	HOMO SAPIENS	PROGRAMMED CELL DEATH PROTEIN 6	a	0.99
NL27	2ZRS	C	HOMO SAPIENS	PROGRAMMED CELL DEATH PROTEIN 6	a	1.00
	3AAJ	A	HOMO SAPIENS	PROGRAMMED CELL DEATH PROTEIN 6	a	0.96
NL28	3QDO	A	RATTUS NORVEGICUS	POTASSIUM CHANNEL 3 CHIMERA	b	0.95
	3QE1	A	RATTUS NORVEGICUS	POTASSIUM CHANNEL 3 CHIMERA	b	1.00
NL29	4E89	A	XENOTROPIC MULV-RELATED VIRUS	RNASE H	c	1.00
	3V1Q	A	XENOTROPIC MULV-RELATED VIRUS VP35	REVERSE TRANSCRIPTASE/RIBONUCLEASE H P80	c	0.93
NL30	4AMJ	B	MELEAGRIS GALLOPAVO	BETA-1 ADRENERGIC RECEPTOR	x	0.96
	2YCX	A	MELEAGRIS GALLOPAVO	BETA-1 ADRENERGIC RECEPTOR	x	1.00
NL31	2X5X	A	PAUCIMONAS LEMOIGNEI	PHB DEPOLYMERASE PHAZ7	x	1.00
	4BVL	D	PAUCIMONAS LEMOIGNEI	PHB DEPOLYMERASE PHAZ7	x	1.00
NL32	4LZB	I	VACCINIA VIRUS	URACIL-DNA GLYCOSYLASE	x	1.00
	4IRB	B	VACCINIA VIRUS ANKARA	URACIL-DNA GLYCOSYLASE	x	1.00
NL33	4GU5	A	DROSOPHILA MELANOGASTER	CRYPTOCHROME-1	x	1.00
	4JZY	B	DROSOPHILA MELANOGASTER	CRYPTOCHROME-1	x	0.99
NL34	3SJZ	A	SULFOLOBUS SOLFATARICUS P2	TRANSLATION INITIATION FACTOR 2 SUBUNIT GAMMA	b	1.00
	4M53	A	SULFOLOBUS SOLFATARICUS	TRANSLATION INITIATION FACTOR 2 SUBUNIT GAMMA	b	1.00
NL35	3RBB	A	HIV-1 M	PROTEIN NEF	d	1.00
	4ORZ	B	HIV-1 M	PROTEIN NEF	d	0.99
NL36	1SH6	A	MUS MUSCULUS	PLECTIN 1	a	1.00
	4Q57	B	MUS MUSCULUS	PLECTIN	a	0.92
NL37	5L6T	B	HOMO SAPIENS	CARBONIC ANHYDRASE 2	x	1.00
	4QK3	A	HOMO SAPIENS	CARBONIC ANHYDRASE 2	x	1.00
NL38	5APY	C	SACCHAROMYCES CEREVISIAE	GENERAL CONTROL PROTEIN GCN4	x	1.00
	5APW	B	SACCHAROMYCES CEREVISIAE	GENERAL CONTROL PROTEIN GCN4	x	0.91
NL39	5AZ8	A	ESCHERICHIA COLI (STRAIN K12)	RECEPTOR SUBUNIT TOM20 HOMOLOG	x	1.00
	5AZ6	B	ESCHERICHIA COLI (STRAIN K12)	RECEPTOR SUBUNIT TOM20 HOMOLOG	x	1.00
NL40	4WMC	G	KLEBSIELLA PNEUMONIAE	BETA-LACTAMASE	e	1.00

	5FDH	A	SERRATIA MARCESCENS	BETA-LACTAMASE	x	0.98
NL41	3AJP	A	HOMO SAPIENS	FERRITIN HEAVY CHAIN	a	0.97
	5GN8	A	HOMO SAPIENS	FERRITIN HEAVY CHAIN	x	0.99
NL42	3AJQ	A	HOMO SAPIENS	FERRITIN HEAVY CHAIN	a	0.97
	5GN8	A	HOMO SAPIENS	FERRITIN HEAVY CHAIN	x	0.99
NL43	3AJP	A	HOMO SAPIENS	FERRITIN HEAVY CHAIN	a	0.86
	5GN8	B	HOMO SAPIENS	FERRITIN HEAVY CHAIN	x	0.99
NL44	4XHS	A	HOMO SAPIENS	MALTOSE-BINDING PERIPLASMIC PROTEIN,NACHT, LRR AND PYD DOMAINS-CONTAINING PROTEIN 12	x	1.00
	5H7N	B	HOMO SAPIENS	NLRP12-PYD WITH MBP TAG	x	0.99
NL45	1X23	D	HOMO SAPIENS	UBIQUITIN-CONJUGATING ENZYME E2 D3	d	1.00
	5IFR	A	HOMO SAPIENS	UBIQUITIN-CONJUGATING ENZYME E2 D3	x	0.99
NL46	1HKX	E	MUS MUSCULUS	II ALPHA CHAIN	d	0.98
	5IG3	A	HOMO SAPIENS	ALPHA	x	0.92
NL47	5IG6	A	HOMO SAPIENS	BROMODOMAIN-CONTAINING PROTEIN 2	x	0.99
	2DVV	A	HOMO SAPIENS	BROMODOMAIN-CONTAINING PROTEIN 2	a	0.99
NL48	5JBH	7	PYROCOCCUS ABYSSI GE5	AIF2-GAMMA	x	1.00
	3SJZ	A	SULFOLOBUS SOLFATARICUS P2	TRANSLATION INITIATION FACTOR 2 SUBUNIT GAMMA	b	0.99
NL49	5JBH	7	PYROCOCCUS ABYSSI GE5	AIF2-GAMMA	x	1.00
	3PEN	A	SULFOLOBUS SOLFATARICUS	TRANSLATION INITIATION FACTOR 2 SUBUNIT GAMMA	b	0.99
NL50	5JVG	M	DEINOCOCCUS RADIODURANS R1	50S RIBOSOMAL PROTEIN L19	x	0.96
	5DM6	M	DEINOCOCCUS RADIODURANS	50S RIBOSOMAL PROTEIN L19	x	1.00
NL51	5JVS	A	DROSOPHILA MELANOGASTER	ASSOCIATED PROTEIN RP/EB FAMILY MEMBER 1	x	1.00
	5JVR	F	DROSOPHILA MELANOGASTER	HUMAN EB1	x	0.77
NL52	5JVU	B	MULTIPLE ORGANISMS	ASSOCIATED PROTEIN RP/EB FAMILY MEMBER 1	x	0.93
	5JVR	C	MULTIPLE ORGANISMS	HUMAN EB1	x	0.85
NL53	5KDO	A	RATTUS NORVEGICUS	GUANINE NUCLEOTIDE-BINDING PROTEIN G(I) SUBUNIT ALPHA-1	x	0.92
	1GIA	A	RATTUS NORVEGICUS	G PROTEIN GI ALPHA 1	a	0.99
NL54	2X0L	A	HOMO SAPIENS	LYSINE-SPECIFIC HISTONE DEMETHYLASE 1	x	1.00
	5LGU	A	HOMO SAPIENS	LYSINE-SPECIFIC HISTONE DEMETHYLASE 1A	x	1.00
NL55	5LGN	A	HOMO SAPIENS	LYSINE-SPECIFIC HISTONE DEMETHYLASE 1A	x	1.00
	5LGU	A	HOMO SAPIENS	LYSINE-SPECIFIC HISTONE DEMETHYLASE 1A	x	1.00
NL56	3MHY	B	AZOSPIRILLUM BRASILENSE	PII-LIKE PROTEIN PZ	d	1.00
	5OVO	B	AZOSPIRILLUM	NITROGEN REGULATORY PROTEIN P-II 1	x	1.00

			BRASILENSE			
NL57	1IFS	A	RICINUS COMMUNIS	RICIN	d	0.68
	5SV3	D	RICINUS COMMUNIS	RICIN	x	0.99
NL58	104L	A	ENTEROBACTERIA PHAGE T4	T4 LYSOZYME	d	1.00
	5VNR	A	ENTEROBACTERIA PHAGE T4	ENDOLYSIN	x	0.99
NL59	209L	A	ENTEROBACTERIA PHAGE T4	T4 LYSOZYME	d	1.00
	5VNR	A	ENTEROBACTERIA PHAGE T4	ENDOLYSIN	x	0.99
NL60	5W57	B	PARACOCCUS DENITRIFICANS (STRAIN PD 1222)	PERIPLASMIC SOLUTE BINDING PROTEIN	x	1.00
	5KZJ	A	PARACOCCUS DENITRIFICANS (STRAIN PD 1222)	PERIPLASMIC SOLUTE BINDING PROTEIN	x	1.00
NL61	5H7Q	A	HOMO SAPIENS	MNDA PYD DOMAIN WITH MBP TAG	x	1.00
	5WQ6	D	HOMO SAPIENS	MBP TAGGED HMNDA-PYD	x	1.00
NL62	5Y32	A	MUS MUSCULUS	RECEPTOR-TYPE TYROSINE-PROTEIN PHOSPHATASE DELTA	x	0.99
	2YD7	A	HOMO SAPIENS	PTPRD PROTEIN	b	1.00
NL63	5YI5	X	HOMO SAPIENS	FERRITIN HEAVY CHAIN	x	0.97
	5GN8	A	HOMO SAPIENS	FERRITIN HEAVY CHAIN	x	0.99
NL64	5YI5	X	HOMO SAPIENS	FERRITIN HEAVY CHAIN	x	0.86
	5GN8	B	HOMO SAPIENS	FERRITIN HEAVY CHAIN	x	0.99
NL65	1JI1	A	THERMOACTINOMYCES VULGARIS	ALPHA-AMYLASE I	b	1.00
	5Z0U	A	THERMOACTINOMYCES VULGARIS	NEOPULLULANASE 1	x	1.00
NL66	1ORC	A	ENTEROBACTERIA PHAGE LAMBDA	CRO REPRESSOR INSERTION MUTANT K56-[DGEVK]	a	1.00
	6CRO	A	ENTEROBACTERIA PHAGE LAMBDA	LAMBDA CRO REPRESSOR	a	0.98

Table S2. Deletion details for the 87 MPD instances in the loop region (Amino acid indices start from 1. For actual starting index consult pdb files in the database provided in the supporting data segment in the webserver.)

#	PDB_ChainID	Start_index_loop	End_index_loop	Start_index_deletion	End_index_deletion	Deletion_terminal
1	1M3C_A	1	4	3	4	N
2	1O1M_A	139	147	142	143	M
3	1XV8_B	303	316	305	309	M
4	2EE5_A	1	26	8	17	N
5	2ESW_A	1	6	5	6	N
6	2GBJ_B	7	18	10	17	M
7	2GBJ_B	8	19	11	18	M
8	2GQG_B	1	16	6	8	N
9	2MD5_A	1	11	5	10	N
10	2UWA_A	116	127	122	126	M
11	2YRY_A	1	21	8	18	N
12	2YT1_A	36	70	42	50	M
13	2Z24_A	99	113	102	111	M
14	2E25_A	99	113	102	111	M
15	1XGE_A	91	113	102	111	M
16	2Z28_A	99	113	102	111	M
17	2Z25_A	99	113	102	111	M
18	2Z29_A	99	113	102	111	M
19	2Z2A_A	99	113	102	111	M
20	2Z26_A	99	113	102	111	M
21	1KNB_A	92	99	94	97	M
22	2VX9_A	45	51	49	50	M
23	1PWK_A	60	74	61	62	M
24	4G79_A	119	126	120	125	M
25	2N69_A	17	21	18	19	M
26	4NUP_C	1	8	3	4	N
27	3IWI_B	207	212	208	210	M
28	4F01_B	1	12	6	7	N
29	4R5L_D	1	12	6	7	N
30	2X6W_A	355	366	357	358	M
31	1IP2_A	47	50	47	48	M
32	1LHL_A	47	50	47	48	M
33	1FRS_B	66	78	70	73	M
34	3LI0_A	177	185	180	183	M
35	4FX8_B	172	180	175	178	M
36	1LZ5_A	61	84	74	77	M
37	1LZ6_A	61	88	74	81	M

38	1SYG_A	106	114	108	113	M
39	1SYE_A	106	114	108	113	M
40	1F2Z_A	106	114	108	113	M
41	1SYC_A	106	114	108	113	M
42	1KDC_A	106	114	108	113	M
43	1F2M_A	106	114	108	113	M
44	1KDB_A	106	114	108	113	M
45	1KAA_A	106	114	108	113	M
46	1SNC_A	106	114	108	113	M
47	1KDA_A	106	114	108	113	M
48	1KAB_A	106	114	108	113	M
49	2UY9_A	151	160	157	158	M
50	3LKY_A	17	21	18	19	M
51	3LL0_A	17	23	20	21	M
52	3KRA_C	81	100	83	92	M
53	2NPP_C	289	308	292	296	C
54	3V0G_C	149	162	151	155	M
55	4K03_B	290	306	295	299	M
56	4NXT_A	104	114	106	110	M
57	4OYC_B	38	46	39	45	M
58	4QK2_A	224	252	225	235	M
59	2HKK_A	224	252	225	235	M
60	5B3Y_A	6	21	6	13	M
61	4LRM_D	75	83	77	79	M
62	5CVD_A	1	5	4	5	N
63	5E8E_H	141	160	145	153	M
64	5EJW_A	1	15	5	9	N
65	5GQM_A	183	194	192	193	M
66	5GQI_A	183	193	191	192	M
67	5EXY_A	183	194	191	193	M
68	1HZB_A	33	44	35	36	M
69	5JX8_B	159	174	171	172	M
70	5K6B_F	74	85	79	80	M
71	5KC6_B	135	140	136	138	C
72	1LZ5_A	61	84	75	78	M
73	1LZ6_A	61	88	75	82	M
74	1LMT_A	61	86	75	80	M
75	5LVK_B	47	50	47	48	M
76	5MXN_f	1	50	28	30	N
77	5BOZ_C	27	48	30	39	M
78	5THF_C	112	123	116	117	M
79	5VG3_C	151	160	157	158	M
80	1K5M_B	149	176	153	167	M
81	5W57_A	82	114	96	108	M
82	4RCA_A	158	171	164	169	M

83	3LL0_A	17	23	18	21	M
84	3V0E_A	145	150	146	150	M
85	4GNQ_A	462	464	462	463	M
86	1E21_A	6	16	9	13	M
87	5LWF_B	162	169	165	166	M

Table S3. Deletion details for the 66 MPD instances in the non-loop region (Amino acid indices start from 1. For actual starting index consult pdb files in the database provided in the supporting data segment in the webserver)

#	PDB_ChainID	Start_index_deletion	End_index_deletion
1	1D1M_B	54	58
2	2DFF_A	201	204
3	2DFH_A	205	212
4	2DFH_A	201	212
5	2DMF_A	123	131
6	2GBK_D	10	16
7	2KJK_A	82	86
8	3Q29_C	370	378
9	2KY9_A	121	123
10	4O4F_A	5	9
11	104L_B	44	45
12	104L_B	46	47
13	3IFK_A	76	77
14	1C7P_A	47	48
15	1FRS_A	70	73
16	1IFG_A	126	128
17	1MB8_A	22	26
18	1SRA_A	61	68
19	1RJ7_H	61	62
20	1S16_A	77	99
21	1STA_A	6	7
22	209L_A	74	77
23	2CME_B	17	18
24	1M7V_A	138	139
25	3AFI_E	131	137
26	3AAK_A	97	98
27	2ZRS_C	95	96
28	3QDO_A	97	98
29	4E89_A	89	99
30	4AMJ_B	211	212
31	2X5X_A	202	208
32	4LZB_I	171	172
33	4GU5_A	292	296
34	3SJZ_A	36	41
35	3RBB_A	87	107
36	1SH6_A	21	25
37	5L6T_B	225	235
38	5APY_C	30	32
39	5AZ8_A	370	371
40	4WMC_G	189	192

41	3AJP_A	130	135
42	3AJQ_A	130	135
43	3AJP_A	135	140
44	4XHS_A	371	372
45	1X23_D	5	8
46	1HKX_E	5	13
47	5IG6_A	5	7
48	5JBH_7	35	39
49	5JBH_7	31	41
50	5JVG_M	106	110
51	5JVS_A	18	24
52	5JVU_B	15	21
53	5KDO_A	301	304
54	2X0L_A	200	203
55	5LGN_A	200	203
56	3MHY_B	42	54
57	1IFS_A	29	38
58	104L_A	45	46
59	209L_A	75	77
60	5W57_B	96	108
61	5H7Q_A	370	372
62	5Y32_A	164	169
63	5YI5_X	130	135
64	5YI5_X	135	140
65	1JI1_A	363	373
66	1ORC_A	52	56

Table S4. Short description of the features (along with their contributions to classification) used in the PU learning framework.

Sl. No.	Feature Name	Feature Type	Cbn_loop[1]	Cbn_nloop[2]	Cbn_loop (e)[3]	Cbn_nloop (e)[4]
1	end_end_distance	Deletion site specific properties	3.53	9.90	2.86	10.13
2	Mean of weighted contact number(WCN)	Environmental compatibility	0.47	0.56	0.39	0.34
3	Standard deviation of weighted contact number(WCN)	Environmental compatibility	0.09	0.20	0.08	0.14
4	Mean of hydro weighted contact number(WCN)	Environmental compatibility	0.87	1.11	0.83	0.97
5	Standard deviation of hydro weighted contact number(WCN)	Environmental compatibility	0.44	0.47	0.32	0.34
6	Number of salt bridge bond in deleted region	Deletion site specific properties	0.39	0.42	0.23	0.31
7	Number of disulfide bond in deleted region	Deletion site specific properties	0.09	0.10	0.07	0.07
8	Number of hydrogen bond in deleted region	Deletion site specific properties	2.08	11.82	1.83	13.80
9	Number of salt bridge bond between two subunits	Environmental compatibility	0.82	0.89	0.73	0.53
10	Number of disulfide bond between two subunits	Environmental compatibility	0.33	0.26	0.23	0.18
11	Maximum value of phi angles	Deletion site specific properties	2.95	2.95	2.57	2.25
12	Minimum value of phi angles	Deletion site specific properties	2.71	3.48	2.26	2.68
13	Maximum value of psi angles	Deletion site specific properties	2.67	3.99	2.83	3.22
14	Minimum value of psi angles	Deletion site specific properties	3.75	3.13	2.88	2.47
15	Size of N-terminal subunit	Environmental compatibility	9.76	3.51	7.02	2.15

16	Size of C-terminal subunit	Environmental compatibility	15.16	2.62	11.63	1.92
17	Mean of surface area	Deletion site specific properties	17.59	18.26	16.40	11.92
18	Standard deviation of surface area	Deletion site specific properties	0.71	0.63	0.59	0.64
19	avg_loop_propensity	Deletion site specific properties	2.31	0.00	1.59	0.00
20	std_loop_propensity	Deletion site specific properties	0.97	0.00	0.87	0.00
21	ALA Amino acid propensity	Deletion site specific properties	1.74	5.08	1.39	3.62
22	CYS Amino acid propensity	Deletion site specific properties	0.28	0.44	0.21	0.31
23	ASP Amino acid propensity	Deletion site specific properties	1.55	0.97	1.06	0.74
24	GLU Amino acid propensity	Deletion site specific properties	0.93	2.00	0.87	1.40
25	PHE Amino acid propensity	Deletion site specific properties	0.65	0.63	0.67	0.53
26	GLY Amino acid propensity	Deletion site specific properties	3.12	1.05	1.78	1.12
27	HIS Amino acid propensity	Deletion site specific properties	0.51	0.52	0.42	0.39
28	ILE Amino acid propensity	Deletion site specific properties	0.53	0.88	0.47	0.65
29	LYS Amino acid propensity	Deletion site specific properties	0.80	1.27	0.59	0.92
30	LEU Amino acid propensity	Deletion site specific properties	4.42	1.67	2.61	1.18
31	MET Amino acid propensity	Deletion site specific properties	0.35	0.43	0.31	0.30
32	ASN Amino acid propensity	Deletion site specific properties	6.56	1.06	8.37	0.96

33	PRO Amino acid propensity	Deletion site specific properties	0.89	1.40	0.75	0.75
34	GLN Amino acid propensity	Deletion site specific properties	0.74	0.71	0.59	0.55
35	ARG Amino acid propensity	Deletion site specific properties	0.99	0.89	0.77	0.66
36	SER Amino acid propensity	Deletion site specific properties	1.50	0.98	1.42	0.75
37	THR Amino acid propensity	Deletion site specific properties	0.85	0.87	0.66	0.64
38	VAL Amino acid propensity	Deletion site specific properties	0.90	1.32	0.79	1.44
39	TRP Amino acid propensity	Deletion site specific properties	0.23	0.36	0.21	0.31
40	TYR Amino acid propensity	Deletion site specific properties	0.57	0.79	0.51	0.50
41	Total_foldx	Fold level attributes	4.22	12.40	4.16	8.26
42	Normalized gap count in deletion stretch	Evolutionary information	0.00	0.00	4.60	4.38
43	Normalized gap count in entire MSA	Evolutionary information	0.00	0.00	3.98	1.56
44	Category match count in deletion stretch	Evolutionary information	0.00	0.00	2.39	7.63
45	Support gap count in deletion stretch	Evolutionary information	0.00	0.00	2.46	2.28
46	Number of structural homologs	Evolutionary information	0.00	0.00	1.79	4.09

Contribution to PU learning-based classification of MPDs in the 1. Loop region, 2. Non-loop region, 3. Loop region (with evolutionary features) and 4. Non-loop region (with evolutionary features) in percentage

Table S5. Cross-validation (on only positive instances) based foldability classification due to MPDs in the loop and non-loop region for the entire database

	MPD at a loop region		**MPD at a non-loop region**	
	Positive: 87 Unlabeled: 4350		Positive: 66 Unlabeled: 3300	
	Recall (%)	**train positive (%)**	**Recall (%)**	**train positive (%)**
5-fold CV	80.1	14.0	83.3	19.6
10-fold CV	81.4	13.5	83.5	19.1

Table S6. Performance evaluation of PROFOUND on the Single Point Deletion database

Testing Protocol	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**	**Accuracy**
LOOCV on SPD database*	131	0	30	1	100%	99.2%	99.4%
PROFOUND tested on SPD database	103	9	21	29	92.0%	78.0%	76.5%

*Random Forest classifier trained on an augmented classification database

Table S7. Cross-validation (on only positive instances) based foldability classification due to MPDs in the loop and non-loop region with and without considering evolutionary features

	MPD at a loop region		**MPD at a non-loop region**	
	Positive: 76 Unlabeled: 3060		Positive: 55 Unlabeled: 2640	
	Recall (%)	**train positive (%)**	**Recall (%)**	**train positive (%)**
	Without evolutionary features			
5-fold CV	77.2	13.2	83.6	19.7
10-fold CV	79.3	12.4	86.9	19.0
	With evolutionary features			
5-fold CV	80.0	12.1	88.0	19.8
10-fold CV	80.4	11.3	89.9	19.2

Table S8. Details of the MPD in the 10 proteins whose effect are analyzed by Molecular Dynamics simulations

Sl.No	PDBID_ChainID	Molecule	Length	Deletion_at	Deletion_pos*
1	4BOL_B	AMPDH2	243	Loop	2-16
2	4Y8F_A	Triosephosphate Isomerase	251	Loop	69-79
3	4ZZ5_A	Glucanase/chitosanase	132	Loop	76-83
4	4P9C_A	Deoxycytidylate deaminase	136	Loop	45-49
5	1AFC_A	Acidic fibroblast growth factor	127	Loop	8-9
6	4XGQ_A	Ribonuclease VapC30	132	NonLoop	115-129
7	5UMH_A	Catechol 1,2-dioxygenase	307	NonLoop	42-52
8	4A5M_A	Uncharacterized HTH-type transcriptional regulator YYBR	96	NonLoop	49-56
9	3GUD_A	Neck appendage protein	119	NonLoop	31-35
10	2SAM_A	SIV protease	99	NonLoop	92-93

*Deletion_pos is the range of residues considered for deletion. The residues in the corresponding PDB files are renumbered and the first residue is assigned index 1.

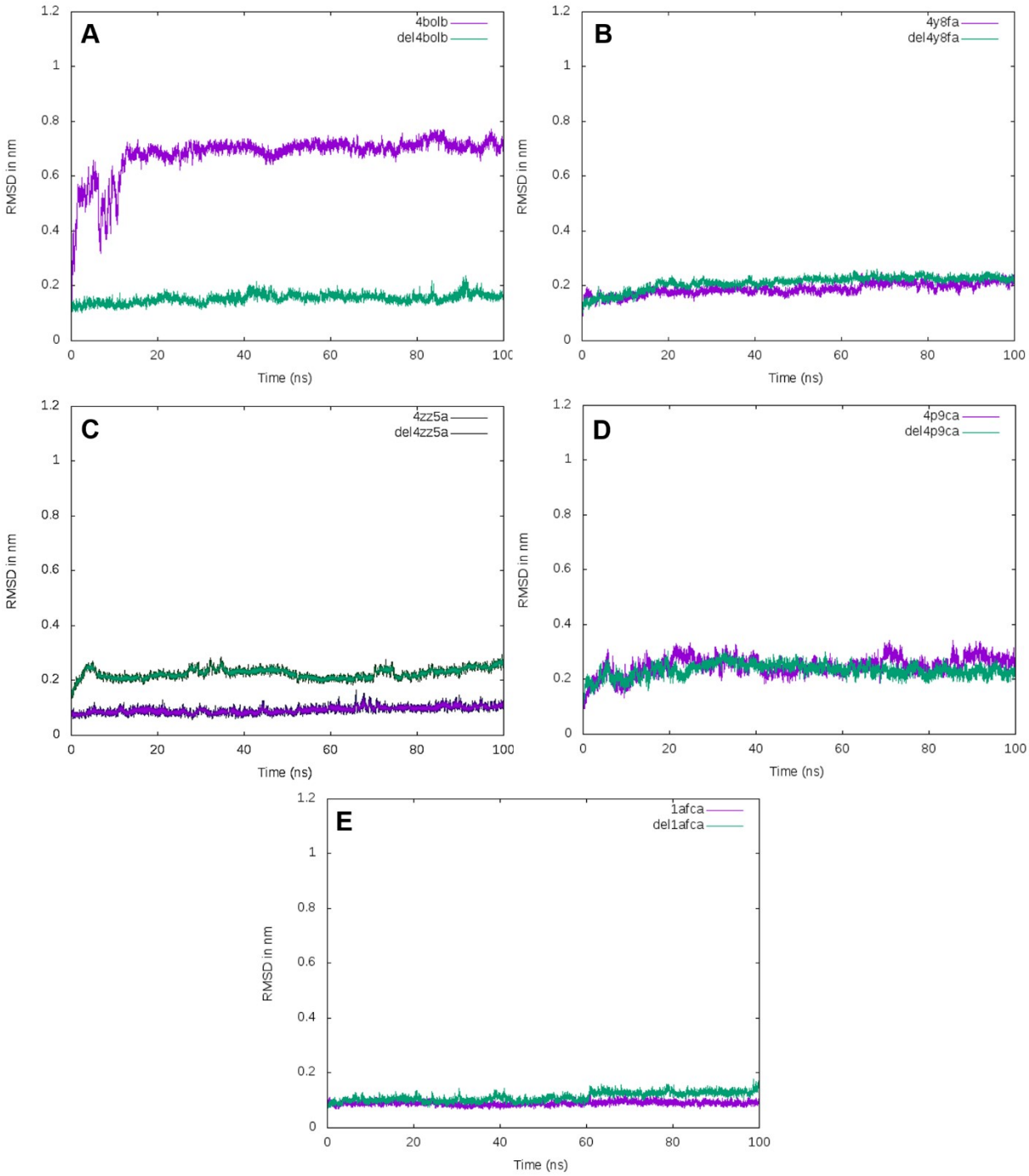

Figure S1. Root mean square deviation plots of the protein in its original conformation (in magenta) and of the protein subject to MPD in residue stretches in the loop region (in sea green) for PDB IDs **(A)** 4BOL Chain B, **(B)** 4Y8F Chain A, **(C)** 4ZZ5 Chain A, **(D)** 4P9C Chain A and **(E)** 1AFC Chain A during 100ns of Molecular Dynamics simulation.

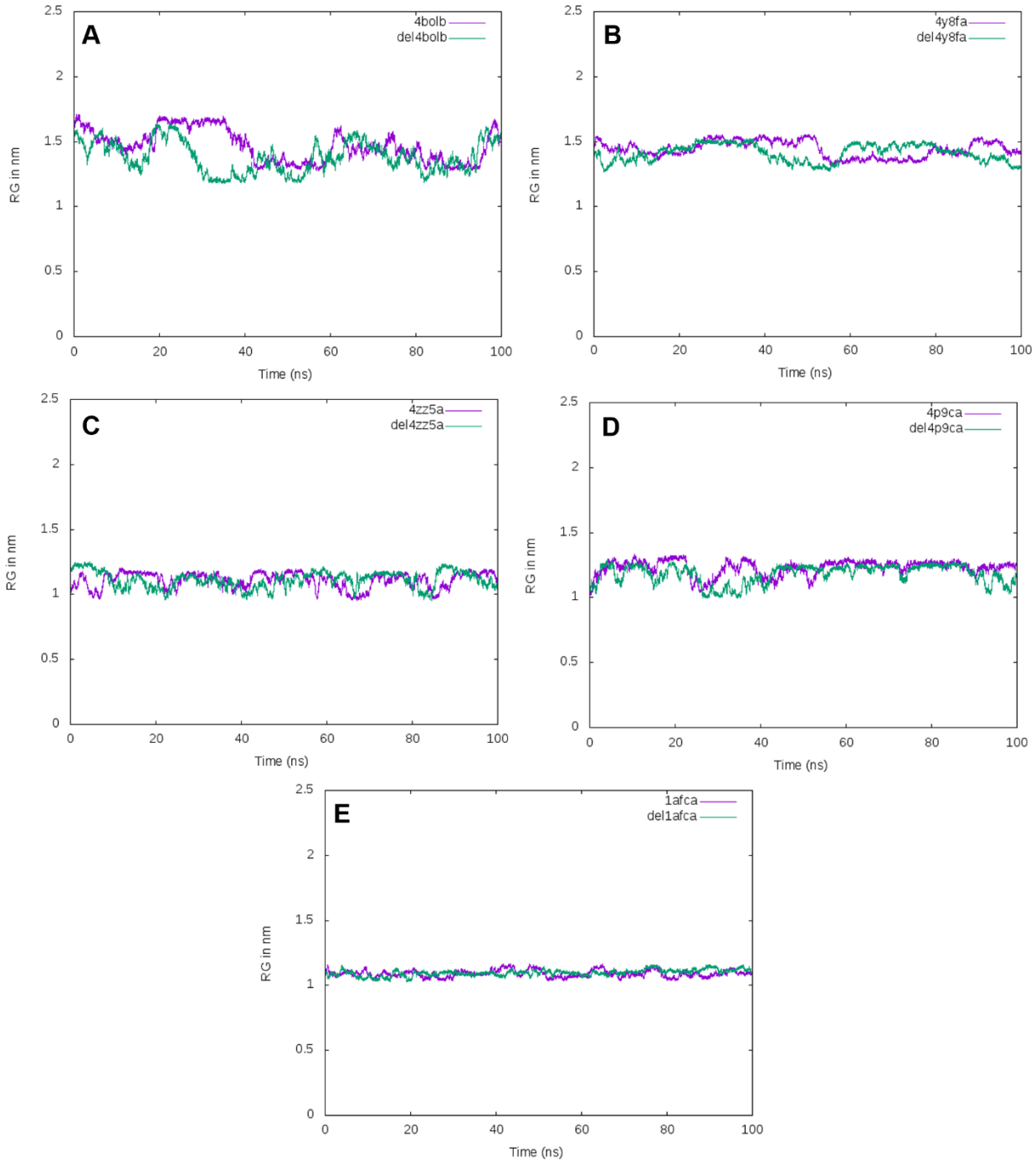

Figure S2. Radius of gyration plots of the protein in its original conformation (in magenta) and of the protein subject to MPD in residue stretches in the loop region (in sea green) for PDB IDs **(A)** 4BOL Chain B, **(B)** 4Y8F Chain A, **(C)** 4ZZ5 Chain A, **(D)** 4P9C Chain A and **(E)** 1AFC Chain A during 100ns of Molecular Dynamics simulation.

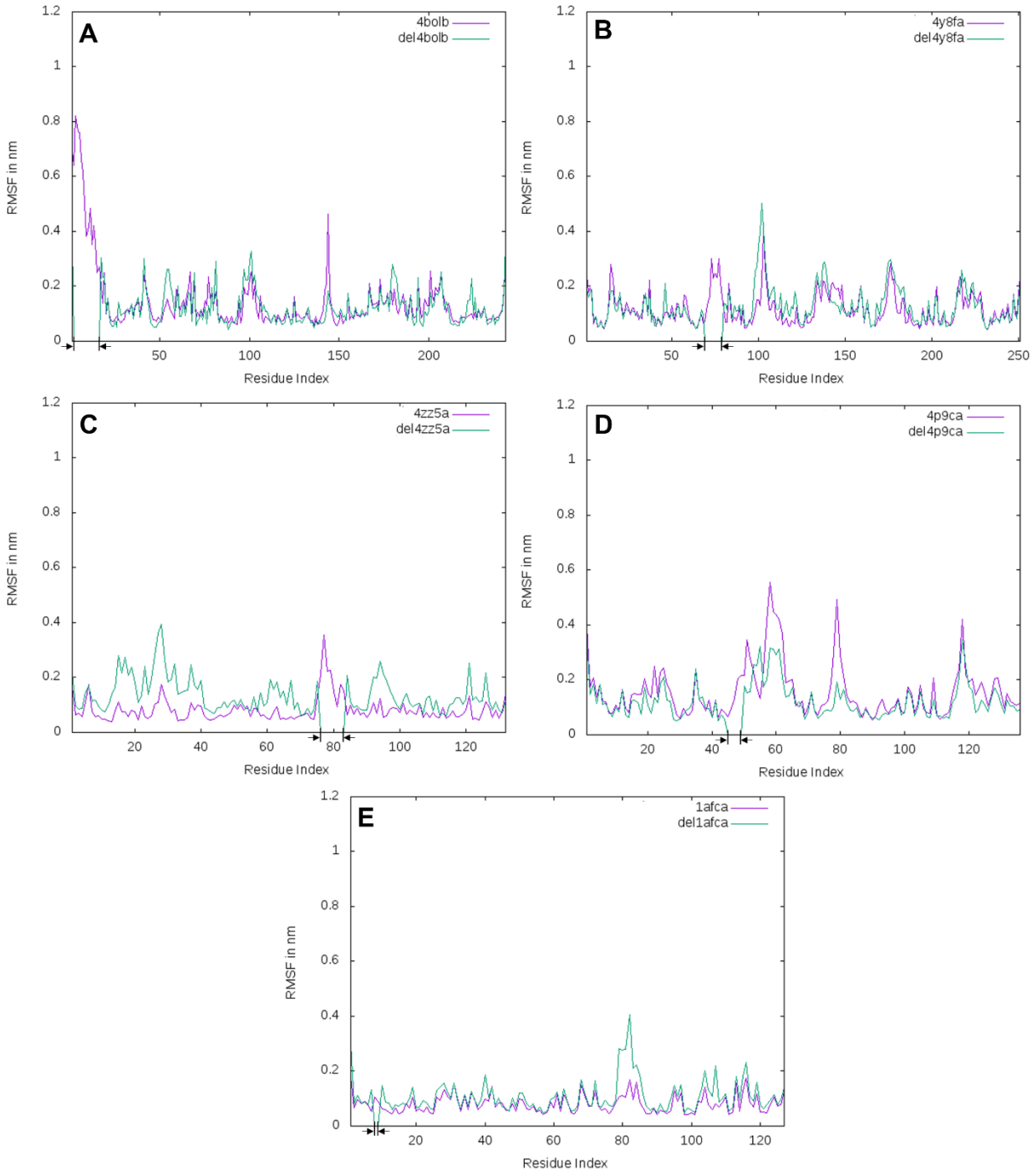

Figure S3. Root mean square fluctuation of each residue of the protein in its original conformation (in magenta) and of the protein subject to MPD in residue stretches in the loop region (in sea green) for PDB IDs **(A)** 4BOL Chain B, **(B)** 4Y8F Chain A, **(C)** 4ZZ5 Chain A, **(D)** 4P9C Chain A and **(E)** 1AFC Chain A during 100ns of Molecular Dynamics simulation. The deleted residue stretch is shown in the X-axis (with the help of arrows).

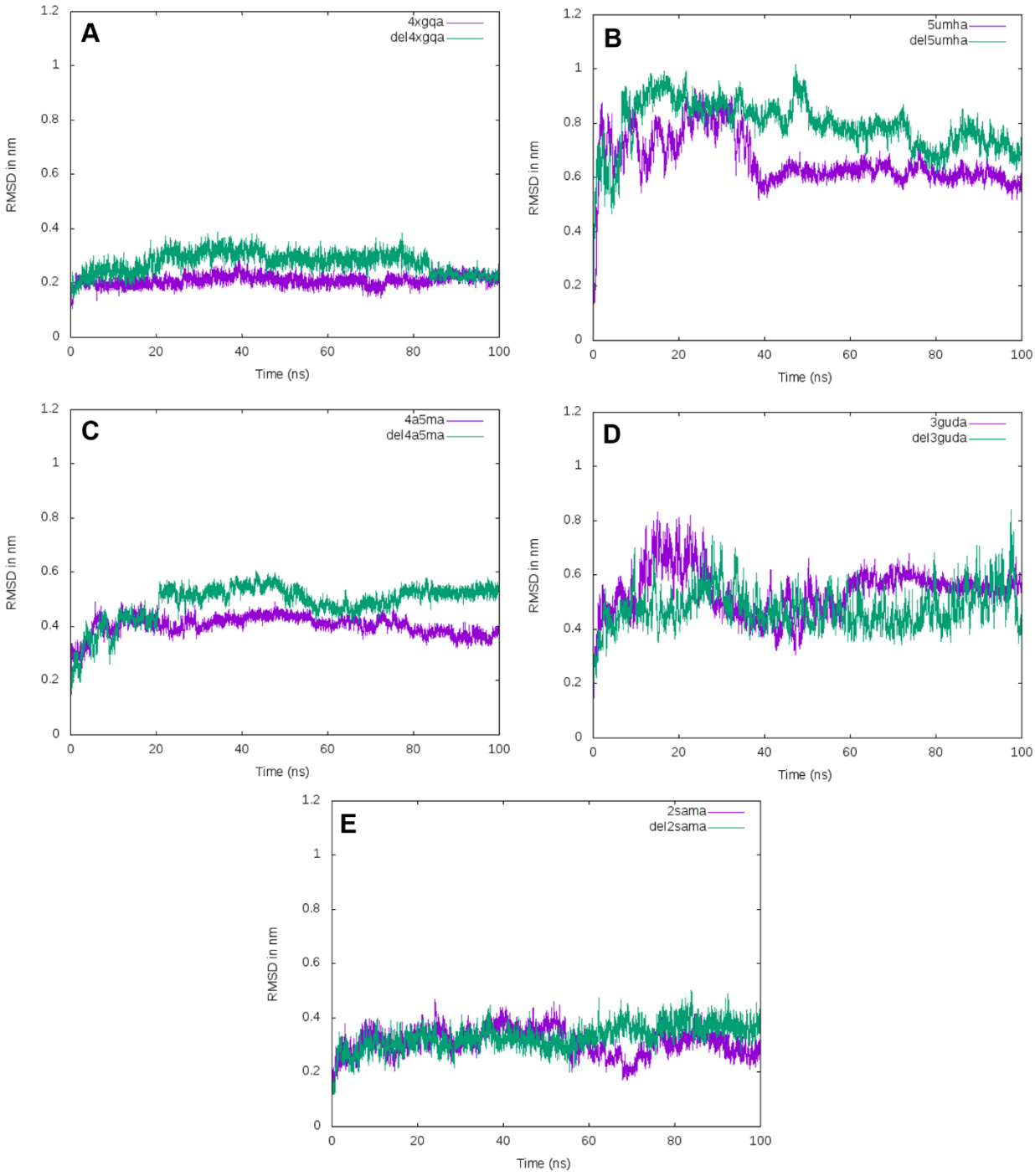

Figure S4. Root mean square deviation plots of the protein in its original conformation (in magenta) and of the protein subject to MPD in residue stretches in the non-loop region (in sea green) for PDB IDs **(A)** 4XGQ Chain A, **(B)** 5UMH Chain A, **(C)** 4A5M Chain A, **(D)** 3GUD Chain A and **(E)** 2SAM Chain A during 100ns of Molecular Dynamics simulation.

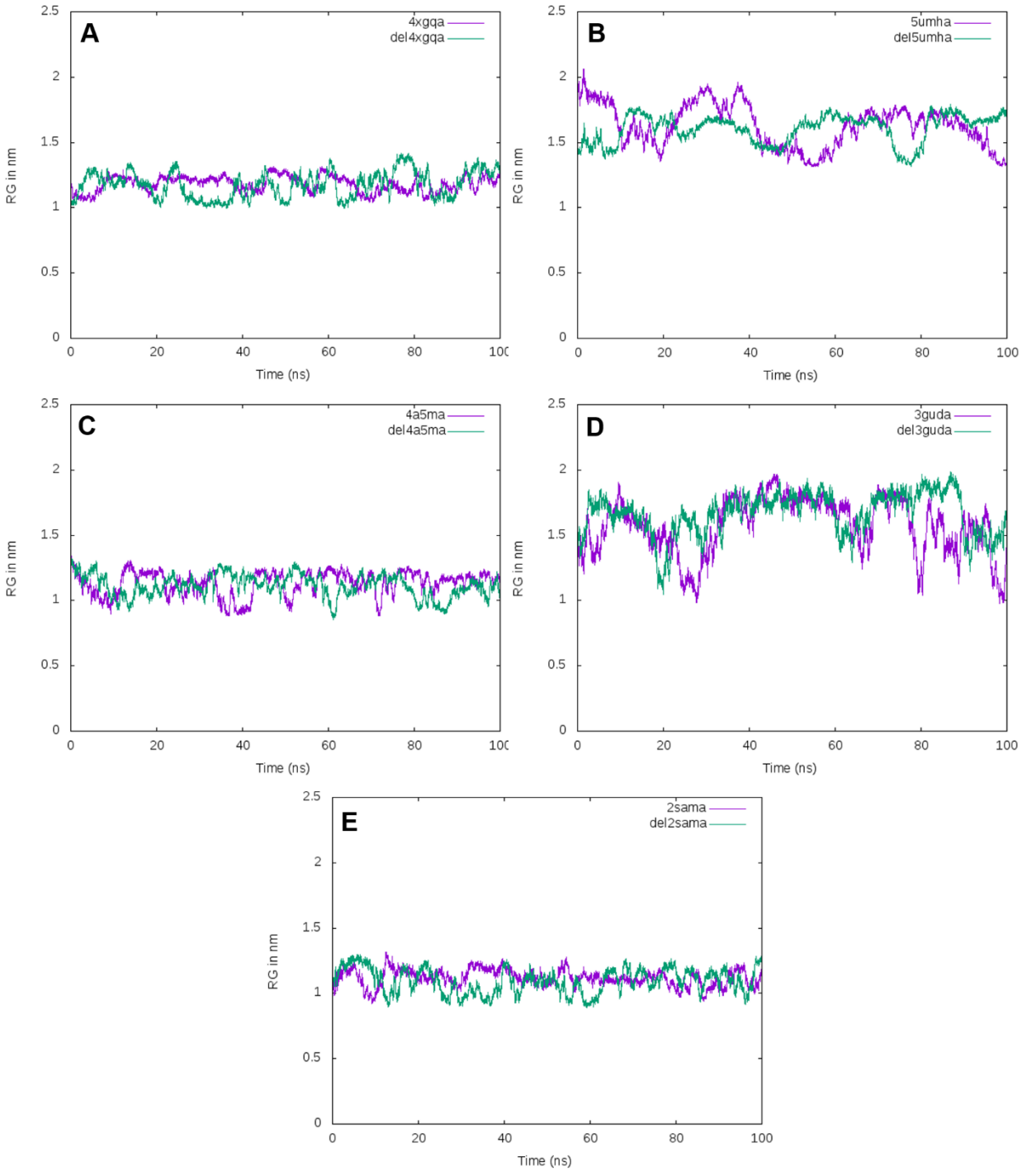

Figure S5. Radius of gyration plots of the protein in its original conformation (in magenta) and of the protein subject to MPD in residue stretches in the non-loop region (in sea green) for PDB IDs **(A)** 4XGQ Chain A, **(B)** 5UMH Chain A, **(C)** 4A5M Chain A, **(D)** 3GUD Chain A and **(E)** 2SAM Chain A during 100ns of Molecular Dynamics simulation.

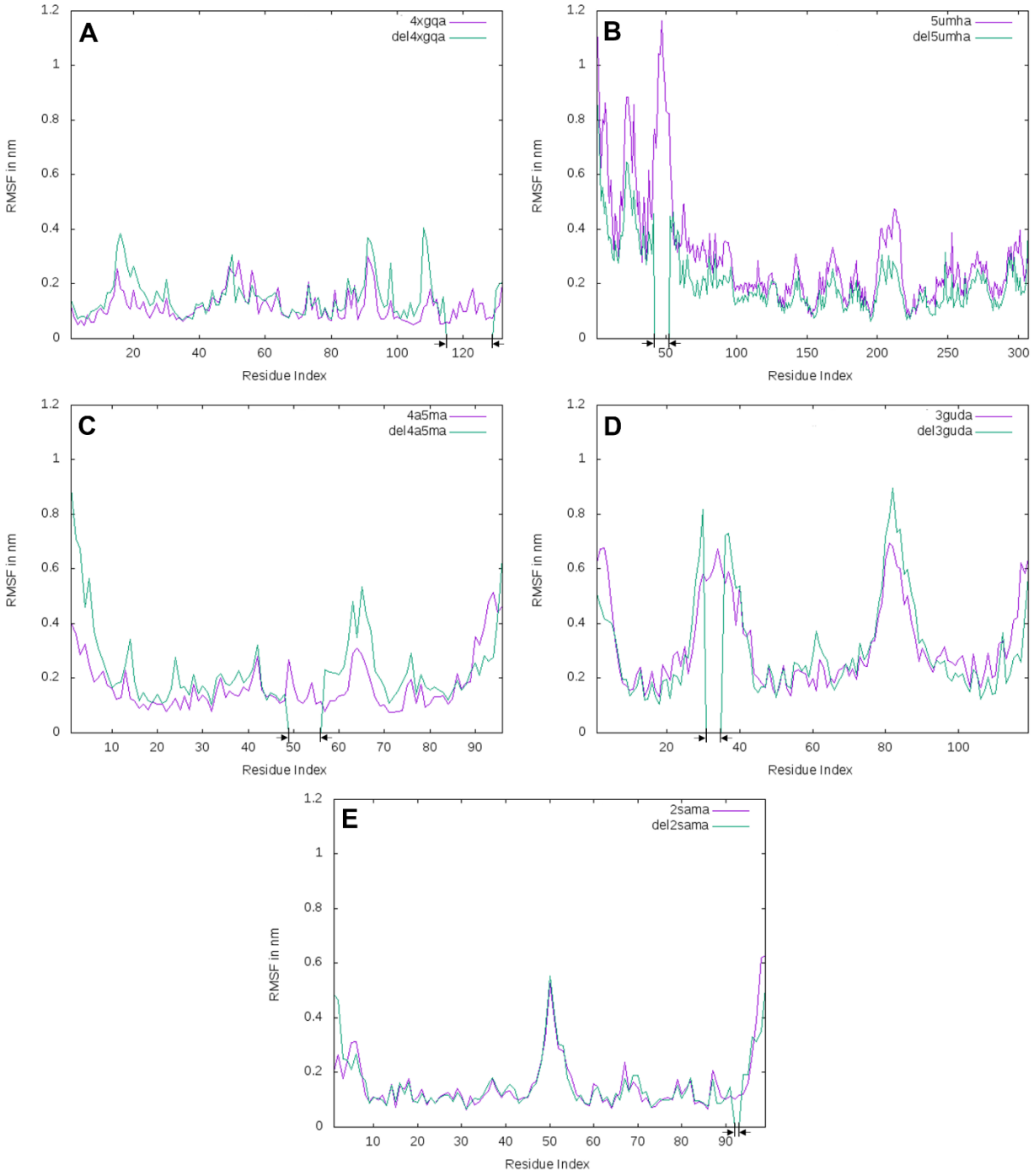

Figure S6. Root mean square fluctuation of each residue of the protein in its original conformation (in magenta) and of the protein subject to MPD in residue stretches in the non-loop region (in sea green) for PDB IDs **(A)** 4XGQ Chain A, **(B)** 5UMH Chain A, **(C)** 4A5M Chain A, **(D)** 3GUD Chain A and **(E)** 2SAM Chain A during 100ns of Molecular Dynamics simulation. The deleted residue stretch is shown in the X-axis (with the help of arrows).

Table S9. Summary of the MD simulation results for the 10 proteins in their original conformations and for the same proteins subject to MPD

PDB	$RMSD_{native}^{max}$ (nm)	$RMSD_{native}^{avg}$ (nm)	$RMSD_{del}^{max}$ (nm)	$RMSD_{del}^{avg}$ (nm)	Rg_{native}^{max} (nm)	Rg_{native}^{avg} (nm)	Rg_{del}^{max} (nm)	Rg_{del}^{avg} (nm)	RMSF_Corr*
4BOL_B	0.77	0.68	0.24	0.15	1.71	1.47	1.64	1.38	0.72
4Y8F_A	0.25	0.19	0.26	0.21	1.56	1.44	1.53	1.41	0.77
4ZZ5_A	0.16	0.09	0.30	0.23	1.20	1.11	1.25	1.11	0.51
4P9C_A	0.34	0.25	0.31	0.23	1.33	1.23	1.30	1.18	0.82
1AFC_A	0.13	0.09	0.18	0.11	1.17	1.09	1.16	1.10	0.78
4XGQ_A	0.30	0.21	0.39	0.27	1.32	1.18	1.42	1.18	0.67
5UMH_A	0.93	0.66	1.02	0.79	2.06	1.63	1.79	1.60	0.94
4A5M_A	0.49	0.40	0.60	0.49	1.34	1.13	1.32	1.11	0.69
3GUD_A	0.83	0.54	0.84	0.48	1.97	1.57	1.99	1.66	0.86
2SAM_A	0.47	0.31	0.50	0.33	1.32	1.12	1.30	1.09	0.86

*Pearson correlation coefficient between the RMSF of corresponding residues in the orginal conformation and in the protein subject to MPD

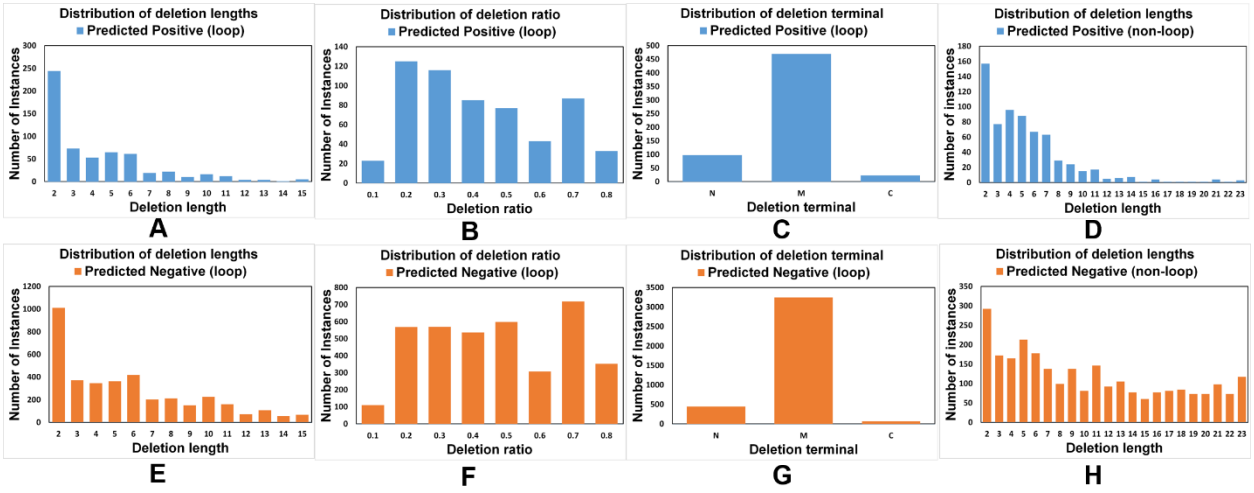

Figure S7. The distribution of **(A,E)** length of deletion in the loop region, **(B,F)** loop length to deletion length ratio, **(C,G)** terminal in which the deletion (in the loop) takes place and **(D,H)** length of deletion in the non-loop region in the unlabeled but predicted positive and unlabeled but predicted negative MPD instances respectively.

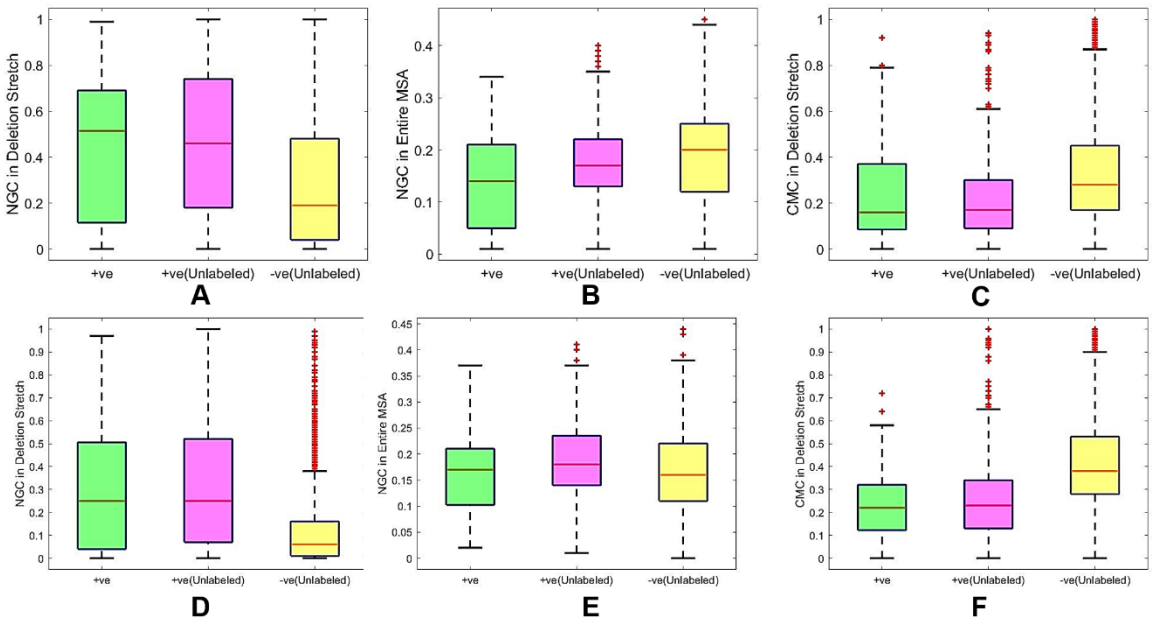

Figure S8. Box plots showing the distribution in positive , unlabeled but predicted positive and unlabeled but predicted negative MPD instances of **(A,D)** NGC of deletion stretch, and **(B,E)** NGC of entire MSA, **(C,F)** CMC of deletion stretch in the loop and non-loop region respectively.