1 Proteins with kunitz domain, PDB (resolution lower than 3.5 length 40-80 pfam pf00014)

protein	chain	leng hth	resolution	protein	chain	lenghth	resolution
3BTM	I	58	1.8	4TPI	I	58	2.2
3BTQ	I	58	1.9	4U30	Z	59	2.5
3BTT	I	58	1.9	4U32	X	55	1.65
3BTW	I	58	2.05	4WWY	I	58	1.7
3BYB	C	59	1.63	4WXV	I	55	2.1
3D65	I	57	1.64	4Y0Y	I	58	1.25
3FP6	I	58	1.49	4Y0Z	I	58	1.37
3FP7	J	43	1.46	4Y10	I	58	1.37
3FP8	I	58	1.46	4Y11	I	58	1.3
3GYM	J	58	2.8	5JB4	C	58	1.99
3L33	H	52	2.48	5JB5	C	58	1.6
3LDI	E	58	2.2	5JB6	C	58	1.9
3LDJ	C	58	1.7	5JB7	C	58	1.9
3LDM	E	58	2.6	5M4V	A	57	1.06
3M7Q	B	61	1.7	5NX1	D	56	1.853
3OFW	A	60	2.5	5NX3	D	56	2.296
3OTJ	I	58	2.15, 1.6	5PTI	A	58	1
3P92	E	58	1.5992	5XX2	B	58	1.12
3P95	E	58	1.2991	5XX3	B	58	1.12
3T62	F	54	2	5XX4	B	58	1.67
3TGI	I	65	1.8	5XX5	B	58	1.38
3TGJ	I	65	2.2	5XX6	B	58	1.31
3TGK	I	65	1.7	5XX7	D	58	1.38
3TPI	I	58	1.9	5XX8	B	58	1.3
3U1J	E	51	1.8	5YV7	A	60	2.395
3U1J	E	58	1.8	5YVU	I	55	2.491
3UIR	D	59	2.777	5YVU	I	55	2.491
3UOU	B	55	2	5YW1	I	54	2.6
3WNY	I	63	1.3	5YW1	I	55	2.6
4BQD	B	79	2.48	5ZJ3	C	59	1.88
4DG4	H	58	1.4	6BX8	H	80	1.98
4DTG	K	66	1.8	6F1F	E	58	1.716
4ISL	B	60	2.29	6HAR	E	80	1.497
4ISN	B	62	2.45	6KZF	A	60	2.52
4ISO	B	60	2.01	6KZF	A	60	2.52
4NTW	B	60	2.07	6PTI	A	58	1.7
4NTX	B	60	2.27	6Q61	A	61	1.3
4NTY	B	60	2.65	6Q6C	B	61	1.3
4PTI	A	58	1.5	6YHY	B	59	1.55
9PTI	A	58	1.22	7PTI	A	58	1.6
8PTI	A	58	1.8				

2 Clustering list (Common list between PDB and PDBeFOLD)

cluster NO. &representative	protein	chain	cluster NO. &representative	protein	chain	cluster NO. &representative	protein	chain
clsuter0 & F5RI	1AAL	A	clsuter0 & F5RI	2HEX	C	clsuter0 & F5RI	4WXV	I
	1AAL	B		2HEX	D		4Y0Y	I
	1B0C	A		2HEX	E		4Y0Z	I
	1B0C	B		2IJO	I		4Y10	I
	1B0C	C		2KAI	I		4Y11	I
	1B0C	D		2PTC	I		5PTI	A
	1B0C	E		2R9P	E		5YVU	I
	1BHC	A		2R9P	F		6F1F	A
	1BHC	B		2R9P	G		6F1F	B
	1BHC	C		2R9P	I		6F1F	C
	1BHC	D		2RA3	C		6F1F	D
	1BHC	E		2RA3	I		6F1F	E
	1BHC	F		2TGP	I		6PTI	A
	1BHC	G		2TPI	I		7PTI	A
	1BHC	H		3BTK	I		8PTI	A
	1BHC	I		3FP6	I		9PTI	A
	1BHC	J		3FP7	J	cluster 1 &1T8LB	1EJM	B
	1BPI	A		3FP8	I		1EJM	D
	1BPT	A		3GYM	I		1EJM	F
	1BTH	P		3GYM	J		1P2I	I
	1BTH	Q		3LDI	A		1P2J	I
	1BTI	A		3LDI	B		1P2K	I
	1BZ5	A		3LDI	C		1P2M	B
	1BZ5	B		3LDI	D		1P2M	D
	1BZ5	C		3LDI	E		1P2N	B
	1BZ5	D		3LDJ	A		1P2N	D
	1BZ5	E		3LDJ	B		1P2O	B
	1BZX	I		3LDJ	C		1P2O	D
	1CBW	D		3LDM	A		1P2Q	B
	1CBW	I		3LDM	B		1P2Q	D
	1D0D	A		3LDM	C		1T7C	B
	1D0D	B		3LDM	D		1T7C	D
	1EAW	B		3LDM	E		1T8L	B
	1EAW	D		3OTJ	I		1T8L	D
	1F5R	I		3P92	E		1T8M	B
	1F7Z	I		3P95	E		1T8M	D
	1FAN	A		3TGI	I		1T8N	B
	1FY8	I		3TGJ	I		1T8N	D
	1MTN	D		3TGK	I		1T8O	B
	1MTN	H		3TPI	I		1T8O	D
	1NAG	A		3U1J	E		3BTD	I
	1TPA	I		4DG4	C		3BTE	I
	1YKT	B		4DG4	E		3BTF	I
	2FI3	I		4DG4	F		3BTG	I
	2FI4	I		4DG4	H		3BTH	I
	2FI5	I		4PTI	A		3BTM	I
	2FTL	I		4TPI	I		3BTQ	I
	2FTM	B		4WWY	C		3BTT	I
	2HEX	A		4WWY	I		3BTW	I
	2HEX	B		4WXV	C			

Table S: Common identifiers and clusterizaiton

Continued Clustering list (Common list between PDB and PDBeFOLD)

cluster NO. &representative	protein	chain	cluster NO. &representative	protein	chain	cluster NO. &representative	protein	chain
cluster 2 & 2ZJXA	2ZJX	A	cluster 6 & 5JB4A	5JB4	A	cluster 17 &	5JB7	A
	2ZJX	B		5JB4	B		5JB7	B
	2ZVX	A		5JB4	C		5JB7	C
	2ZVX	B		5JB5	A	cluster 18 &	1ZR0	B
	5XX2	A		5JB5	B		1ZR0	D
	5XX2	B		5JB5	C	cluster 19 &	1D0D	A
	5XX3	A		5JB6	A		1D0D	B
	5XX3	B		5JB6	B	clsuter 20 &	6YHY	A
	5XX4	A		5JB6	C		6YHY	B
	5XX4	B	cluster 7 & 3WNYA	3WNY	A	cluster 21 &	1YLC	B
	5XX5	A		3WNY	B		1YLD	B
	5XX5	B		3WNY	C	clsuter 22 &	5NX1	C
	5XX6	A		3WNY	E		5NX1	D
	5XX6	B		3WNY	F	clsuter 23 &	5NX3	C
	5XX7	A		3WNY	G		5NX3	D
	5XX7	D		3WNY	H	cluster 24	2KAI	I
	5XX8	A		3WNY	I	cluster 25	6HAR	E
	5XX8	B	cluster 8 & 3M7QB	3M7Q	B	cluster 26	4BQD	B
cluster 3 & 1AAPA	1AAP	A		3OFW	A	cluster 27	1BUN	B
	1AAP	B		3T62	D	cluster 28	1DTX	A
	1BRC	I		3T62	E	cluster 29	1BRB	I
	1CA0	D		3T62	F	cluster 30	2IJO	I
	1CA0	I		3UOU	B	cluster 31	5M4V	A
	1TAW	B	cluster 9 &	6BX8	B	cluster 32	1FAK	I
	1ZJD	B		6BX8	D	cluster 33	4U32	X
	3L33	E		6BX8	F	cluster 34	5YVU	I
	3L33	F		6BX8	H	cluster 35	5YW1	I
	3L33	G	cluster 10 &	1YC0	I	cluster 36	5YW1	I
	3L33	H		4ISL	B	clsuter 37	3U1J	E
cluster 4 & 6Q61A	1Y62	A		4ISN	B			
	1Y62	B		4ISO	B			
	1Y62	C	cluster 11 &	4U30	W			
	1Y62	D		4U30	X			
	1Y62	E		4U30	Y			
	1Y62	F		4U30	Z			
	6Q61	A	cluster 12 &	1TFX	C			
	6Q6C	A		1TFX	D			
	6Q6C	B		4DTG	K			

Table S: Common identifiers and clusterizaiton

3- multiple alignment by mtm-align website

>1AALA.pdb
--------RP-DF-C-LEPPY----T-------GPCKARIIRYFYNAKAGLVQTFVYGG--CR-A--KRNNFKSAEDAMRTCGG---A-
>1AAPA.pdb
--------VR-EV-C-SEQAE----T-------GPCRAMISRWYFDVTEGKCAPFFYGG--CG-G--NRNNFDTEEYCMAVCG------
>1BRBI.pdb
--------------A-GEPPY----T-------GPCKARIIRYFYNAKAGLCQTFVYGG--CR-A--KRNNFKSAEDCMRTA-------
>1BUNB.pdb
R------KRH-PD-C-DKPPD----T-------KICQTVVRAFYYKPSAKRCVQFRYGG---C-N-GNGNHFKSDHLCRCECLE---YR
>1D0DA.pdb
---------YN-RLC--I-KPRDWI-DECDSNE----GGERAYFRNG-KGGCDSFWICPEDHTGAD----YYSSYRDCFNACI------
>1DTXA.pdb
-------PRR-KL-C-ILHRN----P-------GRCYDKIPAFYYNQKKKQCERFDWSG--CG-G--NSNRFKTIEECRRTCIG-----
>1EJMB.pdb
--------RP-DF-C-LEPPY----T-------GPCRLRIIRYFYNAKAGLCQTFVYGG--CR-A--KRNNFKSAEDCLRTCGG---A-
>1FAKI.pdb
--------AP-DF-C-LEPPY----D-------GPCRALHLRYFYNAKAGLCQTFYYGG--CL-A--KRNNFESAEDCMRTC-------
>1G6XA.pdb
--------RP-DF-C-LEPPY----A-------GACRARIIRYFYNAKAGLCQTFVYGG--CR-A--KRNNFKSAEDCLRTCGG---A-
>1KNTA.pdb
---------T-DI-C-KLPKD----E-------GTCRDFILKWYYDPNTKSCARFWYGG--CG-G--NENKFGSQKECEKVCA------
>1TFXC.pdb
--------KP-DF-C-FLEED----P-------GICRGYITRYFYNNQTKQCERFKYGG--CL-G--NMNNFETLEECKNICEDG----
>1Y62A.pdb
--------RP-SL-C-DLPAD----S-------GSGTKAEKRIYYNSARKQCLRFDYTG--QG-G--NENNFRRTYDCQRTCL------
>1YLCB.pdb
--------RP-DF---LEPPY----T-------GPCKARIIRYFYNAKAGL-QTFVYGG--CR-A--KRNNFKSAED-MRT-G------
>1ZR0B.pdb
--PTGN--NA-EI-C-LLPLD----Y-------GPCRALLLRYYYDRYTQSCRQFLYGG--CE-G--NANNFYTWEACDDACWR---IE
>2IJOI.pdb
--------RP-DF-C-LEPPY----T-------GPCKARIIRYFYNAKAGLCQTFVYGG--CR-A--KRNNFKSAEDCMRTCG------
>2KAII.pdb
---------P-DF-C-LEPPY----T-------GPCKARIIRYFYNAKAGLCQTFVYGG--CR-A--KRNNFKSAEDCMRTCGG-----
>2ZJXA.pdb
--------RP-DF-C-LEPPY----T-------GPGKARIIRYFYNAKAGLAQTFVYGG--AR-A--KRNNFKSAEDALRTCGG-A---
>3BYBA.pdb
------KDRP-DF-C-ELPAD----T-------GPCRVRFPSFYYNPDEKKCLEFIYGG--CE-G--NANNFITKEECESTCA------
>3U1JE.pdb
--------RP-DF-C-LEPPY----T-------GPCKARIIRYFYNAKAGLCQTFVYGG--CR-A--KRNNFKSAEDCMRTCG------
>4NTWB.pdb
-------IRP-AF-C-YEDPP----FF------QKCGAFVDSYYFNRSRITCVHFFYGQ---C-D-VNQNHFTTMSECNRVCHG-----
>4U30W.pdb
---------A----CANLPIV----R-------GPCRAFIQLWAFDAVKGKCVLFPYGG--CQ-G--NGNKFYSEKECREYCG------
>4U32X.pdb
-----------HDFC-LVSKV----V-------GRCRASMPRWWYNVTDGSCQLFVYGG--CD-G--NSNNYLTKEECLKKC-------
>5JB4A.pdb
--------RP-AF-C-LEPPY----A-------GPGAAAIIRYFYNAAAGAAQAFVYGG--VR-A--KRNNFASAADALAACAA--A--
>5M4VA.pdb
--------RP-SF-C-NLPVK----P-------GPCKAFFSAFYYSQKTNKCHSFTYGG--CK-G--NANRFSTLEKCRRTCVG-----
>5NX1C.pdb
-----------EV-C-SEQAE----T-------GPCRAMISRWYFDVTEGKCAPFFYGG--CG-G--NRNNFDTEEYCMAVCG------
>5NX3C.pdb
-----------EV-C-SEQAE----T-------GPCRALFFRWYFDVTEGKCAPFVYGG--CG-G--NRNNFDTEEYCMAVC-------
>5YV7A.pdb
-W-----QPP-WY-C-KEPVR----I-------GSCKKQFSSFYFKWTAKKCLPFLFSG--CG-G--NANRFQTIGECRKKCLGK----
>5YVUI.pdb
--------RP-DF-C-LEPPY----T-------GPCKARIIRYFYNAKAGLCQTFVYGG--CR-A--KRNNFKSAEDCMRTC-------
>5YW1I.pdb
--------RP-DF-C-LEPPY----T-------GPCKARMIRYFYNAKAGLCQPFVYGG--CR-A--KRNNFKSSEDCMRTC-------

4- Domain and sequence E-value of all positive ids

Protein	E-value sequence	E-value domain	protein	E-value sequence	E-value domain
A0A6P8HC43	2.70E-78	2.80E-24	B2KTG2	2.90E-22	3.60E-22
Q02445	3.70E-59	3.50E-24	P26228	3.00E-22	3.70E-22
O54819	4.20E-57	5.10E-24	B2G331	4.20E-22	6.10E-22
P83606	1.20E-56	9.80E-23	P0DMX0	4.20E-22	5.50E-22
P10646	1.50E-56	9.90E-23	F6ULY1	4.30E-22	4.30E-22
P19761	1.80E-55	1.70E-22	P0DMJ3	5.00E-22	5.70E-22
P84875	4.70E-55	1.50E-21	A0A3G2FQK2	5.40E-22	5.40E-22
Q03610	7.80E-54	1.70E-17	Q8R0S6	5.90E-22	1.10E-19
Q7YRQ8	5.50E-47	1.20E-23	P10832	6.40E-22	8.30E-22
Q9WU03	3.00E-42	1.70E-24	F8J2F4	6.70E-22	8.10E-22
P86733	1.10E-41	6.60E-23	G9I929	8.40E-22	1.10E-21
P04365	4.90E-40	5.60E-22	P0DJ46	8.70E-22	8.70E-22
W4VSH9	1.50E-38	4.60E-23	P10831	9.10E-22	1.20E-21
Q6T269	3.10E-38	2.80E-21	A6MFL2	1.20E-21	1.50E-21
P62756	3.30E-38	2.70E-21	P16044	1.30E-21	1.50E-21
P62757	3.30E-38	2.70E-21	A5X2X1	1.40E-21	1.40E-21
Q60559	1.20E-36	1.90E-20	P00990	1.70E-21	1.90E-21
Q62577	4.10E-36	5.90E-20	D2Y491	2.10E-21	2.60E-21
Q64240	5.30E-35	1.60E-19	B5KL32	2.20E-21	2.70E-21
Q08E66	7.60E-34	9.60E-24	F8J2F6	2.40E-21	2.90E-21
P83609	9.90E-34	3.00E-20	P0DJ45	2.60E-21	2.60E-21
Q9R097	6.20E-33	1.10E-19	Q28201	2.60E-21	4.80E-21
Q8TEU8	9.10E-31	3.20E-22	A8Y7P3	2.70E-21	3.30E-21
P00974	9.20E-30	1.40E-29	P0DJ48	2.70E-21	3.90E-21
H2A0P0	3.10E-27	1.80E-17	A8Y7N4	2.70E-21	3.40E-21
P29216	3.20E-27	3.90E-27	A8Y7N6	3.00E-21	3.80E-21
P0DMJ6	2.40E-26	2.80E-26	B5KL34	3.10E-21	3.80E-21
Q7LZS8	2.40E-26	2.80E-26	A8Y7N7	3.60E-21	4.60E-21
P10280	3.80E-26	4.80E-26	A7X3V4	4.40E-21	5.50E-21
Q6NUX0	4.30E-26	2.60E-18	A8Y7P0	5.50E-21	6.70E-21
P0DN09	5.90E-26	7.90E-26	Q2ES50	5.50E-21	6.70E-21
Q7LZE3	7.60E-26	8.90E-26	Q6ITB2	6.20E-21	7.50E-21
P0DN08	8.00E-26	1.10E-25	Q9BDL1	6.50E-21	6.50E-21
P0DMW6	8.90E-26	1.10E-25	A8Y7P1	7.10E-21	9.00E-21
P79307	1.10E-25	2.20E-25	P15989	8.80E-21	2.30E-20
Q95241	1.10E-25	2.10E-25	D2Y488	9.00E-21	1.20E-20
P08592	1.20E-25	2.20E-25	I2G9B4	1.50E-20	1.90E-20
Q96NZ8	1.50E-25	3.00E-20	H2A0M2	1.60E-20	3.10E-20
P0DN11	1.50E-25	2.00E-25	Q2ES48	2.90E-20	3.70E-20
C1IC50	1.70E-25	2.10E-25	Q90W96	5.60E-20	6.80E-20
P0DN10	2.70E-25	3.60E-25	P0DN16	6.70E-20	9.00E-20
Q9TWG0	4.30E-25	5.30E-25	B2KTG3	7.10E-20	9.00E-20
Q90WA0	4.60E-25	5.50E-25	P19859	7.20E-20	8.40E-20
B5KL33	2.90E-22	3.50E-22	Q11101	3.70E-10	1.30E-09
D3GGZ8	4.00E-07	4.00E-07	O62247	4.90E-08	1.30E-07

4 Domain and sequence E-value of all positive ids

Protein	E-value sequence	E-value domain	protein	E-value sequence	E-value domain
Q6ITB5	5.90E-25	7.00E-25	B5KL29	7.60E-20	9.20E-20
Q6ITB4	5.90E-25	7.00E-25	P0DQR0	9.70E-20	1.10E-19
P36992	5.90E-25	2.80E-18	B5L5R6	9.70E-20	1.20E-19
Q6ITB9	6.40E-25	7.70E-25	P07481	2.60E-19	3.20E-19
P00980	1.00E-24	1.20E-24	H2A0N9	3.30E-19	3.30E-19
P0DN06	1.10E-24	1.50E-24	P0DJ77	3.50E-19	4.50E-19
P0DN17	1.30E-24	1.70E-24	P68425	3.50E-19	4.50E-19
Q9TWF8	1.70E-24	2.10E-24	P0DJ66	3.50E-19	3.50E-19
P81547	1.90E-24	2.50E-24	P0DJ47	5.20E-19	6.40E-19
B6RLX2	2.20E-24	2.70E-24	G3LH89	5.60E-19	7.10E-19
Q06481	2.70E-24	2.70E-24	B5KL28	6.90E-19	8.40E-19
C0HLB2	3.70E-24	4.60E-24	P0DJ76	8.00E-19	1.00E-18
P0DMJ4	7.20E-24	8.20E-24	Q9EPX2	8.10E-19	8.10E-19
P24541	7.40E-24	8.70E-24	Q90W97	9.00E-19	1.10E-18
P00994	8.00E-24	9.30E-24	Q9DA01	1.30E-18	1.30E-18
C1IC51	8.00E-24	9.90E-24	Q589G4	1.90E-18	2.30E-18
P00979	1.20E-23	1.50E-23	P81906	3.20E-18	3.80E-18
P00993	1.30E-23	1.30E-23	D2Y2Q7	3.30E-18	4.40E-18
B5KF96	1.30E-23	1.60E-23	D2Y2Q2	3.80E-18	5.00E-18
B4ESA3	1.40E-23	1.70E-23	P26227	4.00E-18	4.50E-18
C0HK74	1.40E-23	1.60E-23	Q9D263	4.40E-18	6.20E-18
P25660	1.50E-23	1.80E-23	P0DJ84	5.70E-18	7.50E-18
A0A1Z0YU59	1.50E-23	1.80E-23	D2Y2Q8	6.70E-18	8.80E-18
Q2ES47	1.60E-23	2.00E-23	P0CY85	1.60E-17	2.10E-17
Q6ITB1	2.00E-23	2.40E-23	D2Y2Q5	1.70E-17	2.30E-17
Q2ES46	2.00E-23	2.50E-23	P0DJ82	2.80E-17	3.70E-17
P0DN19	2.70E-23	3.50E-23	Q1RPS8	2.90E-17	3.70E-17
P0DMJ5	3.00E-23	4.40E-23	Q1RPS9	2.90E-17	3.70E-17
P00976	3.80E-23	4.60E-23	P0DMJ1	5.20E-17	6.80E-17
P0DL86	4.10E-23	4.90E-23	Q6UDR6	1.00E-16	1.60E-16
C0HJF4	4.30E-23	4.90E-23	Q75S49	1.50E-16	2.20E-16
P0DN07	4.80E-23	6.40E-23	Q02388	4.40E-16	6.60E-16
B5L5R7	5.40E-23	6.60E-23	P0DJ69	5.40E-16	6.80E-16
A6MGX9	6.50E-23	7.80E-23	P0DJ79	6.30E-16	8.30E-16
A6MFL1	8.70E-23	1.10E-22	B2ZBB9	9.10E-16	1.20E-15
E7FL11	1.10E-22	1.40E-22	D2Y2F9	9.20E-16	1.20E-15
P00992	1.50E-22	1.90E-22	B2ZBB8	9.20E-16	1.20E-15
A6MGY1	1.50E-22	1.80E-22	P0DJ78	9.20E-16	1.20E-15
P81162	1.60E-22	2.30E-22	D2Y2F6	1.20E-15	1.60E-15
P31713	1.70E-22	1.90E-22	D2Y2F8	1.20E-15	1.60E-15
B5L5R4	1.70E-22	2.10E-22	Q63870	1.50E-15	2.30E-15
A0A6B7FA07	1.80E-22	1.80E-22	P0DJ73	1.50E-15	2.10E-15
C1IC53	1.90E-22	2.40E-22	D2Y2F3	7.90E-15	1.00E-14
P20229	2.10E-22	2.50E-22	Q2ES49	8.50E-14	1.20E-13
A8Y7N5	2.30E-22	2.80E-22	D2Y2G0	1.40E-13	2.00E-13
Q29428	2.40E-22	3.80E-22	D2Y2G2	1.90E-13	2.60E-13

protein	E-value sequence	E-value domain	protein	E-value sequence	E-value domain
Q868Z9	7.40E-147	1.30E-18	F8J2F5	4.90E-24	5.90E-24
O76840	1.30E-143	3.30E-19	Q9TWF9	5.00E-24	6.10E-24
Q28864	4.20E-57	1.00E-22	F8J2F3	5.50E-24	6.60E-24
P48307	2.00E-44	1.60E-23	Q90W98	5.90E-24	7.20E-24
Q8WPI2	1.00E-42	7.00E-24	Q6ITB7	6.80E-24	8.20E-24
Q8WPI3	2.00E-42	5.20E-24	O93279	7.70E-24	1.40E-23
O43291	7.80E-41	2.50E-24	Q90W99	1.10E-23	1.30E-23
O35536	1.20E-38	1.20E-21	C0HJU6	1.30E-23	1.50E-23
P02760	6.90E-38	1.20E-21	Q29100	1.30E-23	2.20E-23
P00978	3.70E-37	6.30E-21	B5KL39	1.80E-23	2.20E-23
Q07456	3.30E-36	3.50E-20	Q6ITB0	2.00E-23	2.40E-23
P04366	4.40E-36	6.00E-21	P0DN13	2.10E-23	2.80E-23
B2BS84	1.60E-34	1.30E-20	B5L5R0	2.70E-23	3.20E-23
O43278	1.70E-34	4.70E-20	B5G6G6	2.70E-23	3.20E-23
Q7TQN3	9.70E-32	5.00E-23	B5L5M7	2.90E-23	3.50E-23
P00975	7.50E-28	8.90E-28	P86862	3.00E-23	3.60E-23
P86964	1.10E-26	7.00E-16	P0DN20	3.20E-23	4.30E-23
H2A0N5	1.30E-26	2.90E-16	B5KF95	3.40E-23	4.10E-23
P04815	1.60E-26	2.30E-26	B5KL36	3.40E-23	4.10E-23
P00982	1.60E-26	2.00E-26	P00991	3.70E-23	4.70E-23
P12023	5.60E-26	1.00E-25	B5KF94	4.50E-23	5.40E-23
P00984	7.60E-26	8.90E-26	P0DQQ9	5.20E-23	6.00E-23
P00981	9.60E-26	1.30E-25	P0DN18	5.80E-23	7.80E-23
P53601	1.20E-25	2.20E-25	P0DJ50	6.30E-23	8.10E-23
P05067	1.20E-25	2.20E-25	B5KL40	6.30E-23	7.70E-23
Q5IS80	1.20E-25	2.20E-25	C0HJF3	7.10E-23	8.20E-23
P0DN11	1.50E-25	2.00E-25	Q8AY41	7.90E-23	9.60E-23
P0DMW7	1.90E-25	2.30E-25	P81129	8.10E-23	9.30E-23
Q90WA1	2.20E-25	2.60E-25	B1B5I8	8.40E-23	1.10E-22
Q60495	4.30E-25	8.90E-25	P0DQR1	8.70E-23	1.00E-22
Q6ITB6	5.90E-25	7.00E-25	A8Y7N9	1.10E-22	1.40E-22
C0HJU7	8.60E-25	1.00E-24	P0DKL8	1.50E-22	1.80E-22
C0HK72	8.60E-25	1.00E-24	A6MFL3	1.50E-22	1.80E-22
P0DN14	1.00E-24	1.40E-24	B5KL35	1.90E-22	2.30E-22
H2A0N1	1.20E-24	2.20E-14	Q6ITB8	1.90E-22	2.30E-22
Q6T6T5	1.70E-24	2.20E-24	Q6ITC1	1.90E-22	2.30E-22
P81902	1.80E-24	2.10E-24	P12111	2.00E-22	4.30E-22
P00986	1.80E-24	2.10E-24	B5KL37	2.20E-22	2.60E-22
P15943	2.70E-24	2.70E-24	Q75S50	2.20E-22	2.80E-22
P0DMJ2	3.10E-24	3.70E-24	Q8AY44	2.20E-22	2.80E-22
P0DN12	3.10E-24	4.20E-24	H6VC06	2.30E-22	2.80E-22
C0HK73	3.10E-24	3.60E-24	B2KTG1	2.60E-22	3.20E-22
Q8T3S7	3.20E-24	4.10E-24	Q6T6S5	2.60E-22	3.40E-22
B7S4N9	3.70E-24	4.50E-24	B5L5Q8	2.80E-22	3.40E-22
P00985	4.00E-24	4.50E-24	E7FL12	2.80E-22	3.40E-22
C1IC52	4.90E-24	6.10E-24	Q1RPT0	2.90E-22	3.70E-22

protein	E-value sequence	E-value domain	protein	E-value sequence	E-value domain
P00989	2.90E-22	3.70E-22	D8KY58	1.30E-19	1.70E-19
Q6ITB3	2.90E-22	3.50E-22	B5L5Q1	2.10E-19	2.60E-19
B2KTG2	2.90E-22	3.60E-22	P82966	2.40E-19	2.80E-19
P82968	2.90E-22	2.90E-22	Q0PL65	2.60E-19	3.20E-19
Q8AY45	2.90E-22	3.70E-22	B2ZBB6	3.90E-19	5.10E-19
A6MFL4	3.20E-22	3.90E-22	B5L5Q6	1.10E-18	1.30E-18
A0A6B7FBD3	3.30E-22	3.30E-22	O95428	1.40E-18	1.40E-18
B5KL31	3.40E-22	4.10E-22	P0DJ49	2.20E-18	3.00E-18
P00983	5.80E-22	6.70E-22	Q8T0W4	2.40E-18	3.60E-18
Q6ITC0	5.90E-22	7.20E-22	D2Y2Q9	5.50E-18	7.50E-18
P81658	5.90E-22	7.70E-22	D2Y2G1	5.70E-18	7.50E-18
B5KL38	6.90E-22	8.30E-22	Q2UY11	8.10E-18	1.50E-17
A8Y7P4	8.40E-22	1.00E-21	B4ESA2	2.30E-17	2.90E-17
B5L5R1	1.10E-21	1.30E-21	P16344	3.20E-17	4.20E-17
Q7T2Q6	1.10E-21	1.40E-21	O73683	3.50E-17	7.00E-17
P0C5J5	1.20E-21	2.90E-19	D2Y2Q1	4.10E-17	5.40E-17
B5KL30	1.30E-21	1.60E-21	Q2UY09	5.80E-17	1.10E-16
E5AJX3	1.40E-21	1.80E-21	Q8AY46	1.60E-16	2.30E-16
E7FL13	1.70E-21	2.00E-21	Q9W728	2.20E-16	3.20E-16
B5KL41	1.80E-21	2.20E-21	B5L5Q3	2.50E-16	3.00E-16
D2Y489	1.90E-21	2.30E-21	Q7Z1K3	4.40E-16	1.00E-15
Q5ZPJ7	2.50E-21	3.10E-21	P00987	5.10E-16	7.50E-16
H6VC05	2.70E-21	3.40E-21	P0DJ74	5.40E-16	6.80E-16
A8Y7N8	2.80E-21	3.40E-21	D2Y2Q6	5.50E-16	7.30E-16
P11424	4.20E-21	5.20E-21	P0DJ85	5.80E-16	7.40E-16
A7X3V7	4.50E-21	6.10E-21	P0DJ65	9.20E-16	1.20E-15
Q8AY43	5.20E-21	6.50E-21	D2Y2C2	9.20E-16	1.20E-15
A8Y7P6	5.40E-21	6.70E-21	P0DJ70	9.20E-16	1.20E-15
P86959	5.90E-21	1.20E-20	P0DJ80	9.20E-16	1.20E-15
P0C1X2	6.00E-21	7.50E-21	P0DJ75	9.20E-16	1.20E-15
P81548	6.90E-21	7.90E-21	B2ZBC0	9.20E-16	1.20E-15
D2Y490	7.50E-21	9.30E-21	D2Y2F4	1.10E-15	1.40E-15
P0DN15	7.80E-21	1.00E-20	P0DJ67	1.10E-15	1.40E-15
A8Y7P5	8.40E-21	1.10E-20	P0DJ72	1.20E-15	1.60E-15
Q4KUS1	9.00E-21	9.00E-21	D2Y2F7	1.50E-15	2.10E-15
B4ESA4	9.80E-21	1.20E-20	D2Y2F5	2.20E-15	2.80E-15
Q3UW55	1.50E-20	1.50E-20	P0DJ71	2.20E-15	2.80E-15
D4A2Z2	1.50E-20	1.50E-20	P0DJ81	3.90E-15	5.30E-15
A8Y7P2	1.50E-20	1.90E-20	P0DJ64	7.90E-15	1.00E-14
O95925	1.90E-20	1.90E-20	P0DJ68	1.90E-13	2.60E-13
P0C8W3	2.10E-20	2.90E-20	C0LNR2	6.20E-13	8.80E-13
O62845	2.50E-20	2.50E-20	Q9BQY6	6.20E-13	6.20E-13
P49223	2.50E-20	3.20E-20	Q8IUA0	2.10E-12	2.10E-12
Q8AY42	2.70E-20	3.40E-20	P86963	6.50E-11	6.50E-11
Q29143	5.30E-20	7.60E-20	P26226	8.00E-11	1.30E-10
B5KL27	5.50E-20	6.70E-20			

Calculating the performance of the model

Performance of set_all.res					
Threshold	**ACC**	**MCC**	**CM**	**FPR**	**TPR**
1.00E-06	**0.999989241**	**0.991875446**	**[[557281,0],[6,365]]**	**0.016172507**	**0.983827493**
1.00E-05	0.999985654	0.989210914	[[557279,0],[8,365]]	0.021447721	0.978552279
1.00E-08	0.999985654	0.989108347	[[557281,2],[6,363]]	0.016260163	0.983739837
1.00E-07	0.999985654	0.989108347	[[557281,2],[6,363]]	0.016260163	0.983739837
1.00E-09	0.999985654	0.989067557	[[557282,3],[5,362]]	0.013623978	0.986376022
1.00E-10	0.999985654	0.989033918	[[557283,4],[4,361]]	0.010958904	0.989041096
1.00E-11	0.999983861	0.987647177	[[557283,5],[4,360]]	0.010989011	0.989010989
1.00E-12	0.999982068	0.986258491	[[557283,6],[4,359]]	0.011019284	0.988980716
0.0001	0.999976688	0.982642326	[[557274,0],[13,365]]	0.034391534	0.965608466
1.00E-13	0.999971308	0.977885091	[[557283,12],[4,353]]	0.011204482	0.988795518
1.00E-14	0.999969515	0.976482564	[[557283,13],[4,352]]	0.011235955	0.988764045
1.00E-15	0.999928271	0.943651039	[[557283,36],[4,329]]	0.012012012	0.987987988
1.00E-16	0.999899579	0.920123106	[[557283,52],[4,313]]	0.012618297	0.987381703
1.00E-17	0.999876267	0.900555475	[[557283,65],[4,300]]	0.013157895	0.986842105
1.00E-18	0.999842195	0.871168651	[[557283,84],[4,281]]	0.014035088	0.985964912
1.00E-19	0.999802744	0.83585573	[[557283,106],[4,259]]	0.015209125	0.984790875
0.001	0.999670045	0.815245767	[[557103,0],[184,365]]	0.335154827	0.664845173
0.01	0.996135583	0.379843621	[[555132,0],[2155,365]]	0.85515873	0.14484127
0.1	0.958520726	0.122023665	[[534156,0],[23131,365]]	0.984465441	0.015534559
1	0.528471161	0.027057843	[[294338,0],[262949,365]]	0.998613822	0.001386178

Table S5: calculating the performance of the built HMM model on set_all.res according to ACC, MCC, CM, FPR, TPR in different thresholds and finding the optimal one(the optimal Threshold is bold)

Calculating the performance of the model

Performance of set_all1.res					
Threshold	**ACC**	**MCC**	**MCC**	**FPR**	**TPR**
1.00E-06	**0.999989241**	**0.991853413**	**[[278640,0],[3,182]]**	**0.016216216**	**0.983783784**
1.00E-05	0.999985654	0.989181772	[[278639,0],[4,182]]	0.021505376	0.978494624
0.0001	0.999982068	0.986531585	[[278638,0],[5,182]]	0.026737968	0.973262032
1.00E-08	0.999982068	0.986296104	[[278640,2],[3,180]]	0.016393443	0.983606557
1.00E-07	0.999982068	0.986296104	[[278640,2],[3,180]]	0.016393443	0.983606557
1.00E-12	0.999982068	0.98622067	[[278641,3],[2,179]]	0.011049724	0.988950276
1.00E-11	0.999982068	0.98622067	[[278641,3],[2,179]]	0.011049724	0.988950276
1.00E-10	0.999982068	0.98622067	[[278641,3],[2,179]]	0.011049724	0.988950276
1.00E-09	0.999978481	0.983505717	[[278640,3],[3,179]]	0.016483516	0.983516484
1.00E-14	0.999971308	0.977823864	[[278641,6],[2,176]]	0.011235955	0.988764045
1.00E-13	0.999971308	0.977823864	[[278641,6],[2,176]]	0.011235955	0.988764045
1.00E-15	0.99993903	0.952190616	[[278641,15],[2,167]]	0.01183432	0.98816568
1.00E-16	0.999921097	0.937648077	[[278641,20],[2,162]]	0.012195122	0.987804878
1.00E-17	0.999892406	0.913900222	[[278641,28],[2,154]]	0.012820513	0.987179487
1.00E-18	0.999849368	0.877077152	[[278641,40],[2,142]]	0.013888889	0.986111111
1.00E-19	0.999795571	0.828757374	[[278641,55],[2,127]]	0.015503876	0.984496124
0.001	0.999670044	0.814870705	[[278551,0],[92,182]]	0.335766423	0.664233577
0.01	0.996151708	0.380080882	[[277570,0],[1073,182]]	0.85498008	0.14501992
0.1	0.958522371	0.121861691	[[267078,0],[11565,182]]	0.984506683	0.015493317
1	0.528263248	0.027009605	[[147111,0],[131532,182]]	0.998618218	0.001381782

Table S6: calculating the performance of the built HMM model on set_all1.res according to ACC, MCC, CM, FPR, TPR in different thresholds and finding the optimal one(the optimal Threshold is bold)

Calculating the performance of the model

Performance of set_all2.res					
Threshold	**ACC**	**MCC**	**CM**	**FPR**	**TPR**
1.00E-09	**0.999992827**	**0.994576337**	**[[278642,0],[2,183]]**	**0.010810811**	**0.989189189**
1.00E-08	0.999989241	0.991897361	[[278641,0],[3,183]]	0.016129032	0.983870968
1.00E-07	0.999989241	0.991897361	[[278641,0],[3,183]]	0.016129032	0.983870968
1.00E-06	0.999989241	0.991897361	[[278641,0],[3,183]]	0.016129032	0.983870968
1.00E-10	0.999989241	0.991823917	[[278642,1],[2,182]]	0.010869565	0.989130435
1.00E-05	0.999985654	0.989239899	[[278640,0],[4,183]]	0.021390374	0.978609626
1.00E-11	0.999985654	0.989063861	[[278642,2],[2,181]]	0.010928962	0.989071038
1.00E-12	0.999982068	0.986296104	[[278642,3],[2,180]]	0.010989011	0.989010989
1.00E-04	0.999971308	0.978819532	[[278636,0],[8,183]]	0.041884817	0.958115183
1.00E-13	0.999971308	0.97794598	[[278642,6],[2,177]]	0.011173184	0.988826816
1.00E-14	0.999967722	0.975146765	[[278642,7],[2,176]]	0.011235955	0.988764045
1.00E-15	0.999917512	0.935080969	[[278642,21],[2,162]]	0.012195122	0.987804878
1.00E-16	0.999878061	0.90235729	[[278642,32],[2,151]]	0.013071895	0.986928105
1.00E-17	0.999860128	0.887085075	[[278642,37],[2,146]]	0.013513514	0.986486486
1.00E-18	0.999835023	0.865252508	[[278642,44],[2,139]]	0.014184397	0.985815603
1.00E-19	0.999809918	0.842856195	[[278642,51],[2,132]]	0.014925373	0.985074627
0.001	0.999670046	0.815619294	[[278552,0],[92,183]]	0.334545455	0.665454545
0.01	0.996119458	0.379608072	[[277562,0],[1082,183]]	0.855335968	0.144664032
0.1	0.958519082	0.122185397	[[267078,0],[11566,183]]	0.984424206	0.015575794
1	0.528679073	0.027106066	[[147227,0],[131417,183]]	0.998609422	0.001390578

Table S7: calculating the performance of the built HMM model on set_all2.res according to ACC, MCC, CM, FPR, TPR in different thresholds and finding the optimal one(the optimal Threshold is bold)

2 Useful Linux Commands

2.1 Common identifiers (150 common proteins between two lists)

python compare.py clean_pdb.seq list2 |awk '{print ">"$1;print $2}' >comm_seq.fasta

2.2 Getting the clusterized list

awk '{if (substr($0,1,1)==">") {print ""} else {printf "%s ",substr($3,2,5)}}' cd-hit.cluster |tail -n +2 >cdhit-seq.list

2.3 Extract the pdf file of all of these representative form pdb website

for i in `cat list_pdb.txt`
do
wget -q https://files.rcsb.org/view/$i.pdb done

2.4 Getting the list of representative chains

awk '{print substr($1,1,4),substr($1,5,1)}' ../cdhit-seq.list |sort -u >list_chain.txt

2.5 Extract the information of representative chains about their ATOM and TER part

vi selch.sh
#!/bin/bash
pdbfile=$1
chain=$2
awk -v c=$chain '{if ((substr($0,1,4)=="ATOM" || (substr($0,1,3)=="TER")) && substr($0,22,1)==c) print $0}' $pdbfile
awk '{print "./selch.sh",$1".pdb",$2" >chains/"$1$2".pdb"}' list_chain.txt >run.sh
/bin/bash run.sh

2.6 Extract the fasta file of multipe structure alignment based on representative identifiers chains

wget https://yanglab.nankai.edu.cn/mTM-align/output/mTM018114/seq.fasta -o tm-ali.fasta

2.7 Remove the inital and final gaps in N terminal and C terminal of alignment

awk '{if (substr($0,1,1)==">") {printf "\n%s ",$0} else {printf "%s",$0}}' tm-ali.fasta |awk '{print substr($1,1,6);print substr($2,27,77)}' |tail -n +3 >bpti-kunitz.ali

2.8 Generate a Hidden Markov Model

hmmbuild bpti-kunitz.hmm bpti-kunitz.ali

2.9 Creating two groups of negatives and positives based on the reviewed proteins in uniprot and get just its IDs and save as positive.ids and negative ids

positive.fasta file
https://www.uniprot.org/ pfam: pf00014 reviewed leneth of 40 to *
negative.fasta file
https://www.uniprot.org/ pfam: not pf00014 reviewed
leneth of 40 to *
cat positives.fasta |awk '{if (substr($0,1,1)==">") {split($0,a,"|");print a[2]} }' >positives.ids
cat negatives.fasta |awk '{if (substr($0,1,1)==">") {split($0,a,"|");print a[2]} }' >negatives.ids

2.10 Sorting randomly the two groups of positives and negatives identifiers
sort -R positives.ids >positives.rids sort -R negatives.ids >negatives.rids

2.11 Cut the positives and negatives groups to halve
head -n 182 positives.rids >pos1.ids
tail -n +181 positives.rids >pos2.ids

head -n 278643 negatives.rids >neg1.ids
tail -n +278644 negatives.rids >neg2.ids

2.12 Run the hmmsearch we need a fasta file of list of identifiers
python3 select-seqs.py set_pos1.ids positives.fasta >set_pos1.fasta
python3 select-seqs.py pos2.ids positives.fasta >set_pos2.fasta
python3 select-seqs.py neg1.ids negatives.fasta >set_neg1.fasta
python3 select-seqs.py neg2.ids negatives.fasta >set_neg2.fasta

2.13 Running the hidden markov model on the fasta file of identifiers
hmmsearch -Z 1 --noali --max --tblout set_pos1.out bpti-kunitz.hmm set_pos1.fasta
hmmsearch -Z 1 --noali --max --tblout set_pos2.out bpti-kunitz.hmm set_pos2.fasta
hmmsearch -Z 1 --noali --max --tblout set_neg1.out bpti-kunitz.hmm set_neg1.fasta
hmmsearch -Z 1 --noali --max --tblout set_neg2.out bpti-kunitz.hmm set_neg2.fasta

grep -v “^#” set_pos1.out |awk ‘{print $1,$8,1}’ >set_pos1.res
grep -v “^#” set_pos2.out |awk ‘{print $1,$8,1}’ >set_pos2.res
grep -v “^#” set_neg1.out |awk ‘{print $1,$8,1}’ >set_neg1.res
grep -v “^#” set_neg2.out |awk ‘{print $1,$8,1}’ >set_neg2.res

2.14 Add info about the e value
comm -23 <(sort neg1.ids) <(awk '{print $1}' set_neg1.res |sort) |awk '{print $1,10.0,0}' >set_neg1.add
comm -23 <(sort neg2.ids) <(awk '{print $1}' set_neg2.res |sort) |awk '{print $1,10.0,0}' >set_neg2.add
comm -23 <(sort pos1.ids) <(awk '{print $1}' set_pos1.res |sort) |awk '{print $1,10.0,0}' >set_pos1.add
comm -23 <(sort pos2.ids) <(awk '{print $1}' set_pos2.res |sort) |awk '{print $1,10.0,0}' >set_pos2.add

2.15 Categorize the groups and applying performance on them

cat set_neg1.res set_neg1.add set_pos1.res >set_all1.res
cat set_neg2.res set_neg2.add set_pos2.res >set_all2.res

2.16 Calculating the performance of set_all1.res
python3 performance.py set_all1.res |sort -nrk 6 >set_all1.out

2.19 Calculating the performance of set_all2.res
python3 performance.py set_all2.res |sort -nrk 6>set_all2.out

2.20 Cross-validation (apply optiaml threshold on set_all1.res)
python3 performance.py set_all1.res 1e-09 awk '{p=0;if ($2<1e-9) {p=1};print $1,p, $3}' set_all1.res
python3 performance.py set_all2.res 1e-06 awk '{p=0;if ($2<1e-6) {p=1};print $1,p, $3}' set_all2.res

2.21 finding Wrong prediction of set1 & set2

- awk '{if ($2<=1e-9 && $3==0 || ($2>1e-9 && $3==1))print $0}' set_all1.res

P0DV04 9.6e-23 0
P0DV06 2.8e-22 0
P17726 2.2e-10 0
Q11101 1.3e-09 1
O62247 1.3e-07 1
D3GGZ8 4e-07 1

- awk '{if ($2<=1e-6 && $3==0 || ($2>1e-6 && $3==1))print $0}' set_all2.res

P0DV05 6.7e-25 0
P0DV03 2.6e-23 0
P36235 8.3e-09 0

Overall Wrong predictions

awk '{p=0;if ($2<1e-6) {p=1};print $1,p, $3}' set_all2.res |> set_all22.res
awk '{p=0;if ($2<1e-9) {p=1};print $1,p, $3}' set_all1.res > set_all11.res
cat set_all11.res set_all22.res > set_all.res
awk '{if ($2!=$3) print $0}' set_all.res > wrong_pred.res

P0DV04 1 0
P0DV06 1 0
P17726 1 0
Q11101 0 1
O62247 0 1
D3GGZ8 0 1
P0DV05 1 0
P0DV03 1 0
P36235 1 0

Calculating the overall performance:
cat set_all1.res set_all2.res >set_overall.res
python3 performance.py set_overall.res |sort -nrk 6 >overall_performance.res

Select-seqs.py

```
#!usr/bin/env python3

import sys

def get_ids(fileids):
    d={}
    listid=open(fileids).read().rstrip().split('\n')
    d=dict([(i,True) for i in listid])
    return d

def get_sequences(fileseq,listid):
    c=0
    f=open(fileseq)
    for line in f:
        if line.find('>')==0:
            pid=line.split('|')[1]
            if pid in listid:
                c=1
                print(">"+pid)
                continue
            else:
                c=0
        if c==1: print (line.rstrip())
    return

if __name__=='__main__':
    fileids=sys.argv[1]
    fileseq=sys.argv[2]
    listid=get_ids(fileids)
    get_sequences(fileseq,listid)
```

Compare .py

```
#!usr/bin/env python3
import sys
def get_dic(filename):
    d={}
    f=open(filename)
    for line in f:
        v=line.rstrip().split()
        d[v[0]]=d.get(v[0],[])
        d[v[0]].append(line)
    return d
def get_common(d1,d2):
    s1=set(list(d1.keys()))
    s2=set(list(d2.keys()))
    c=list(s1.intersection(s2))
    for i in c:
        print(d1[i][0].rstrip())

    #print(len(list(c))

if __name__ == "__main__":
    file1=sys.argv[1]
    file2=sys.argv[2]
    d1=get_dic(file1)
    d2=get_dic(file2)
    get_common(d1,d2)
```

```
#!/usr/bin/env python3

from multiprocessing.context import ForkProcess
import sys
import numpy as np

def confusion(m):
    for k in range(0,len(m)):
        print (m[k])

def get_preds(filename, sp=-2, rc=-1):
    pred_list = []
    f = open(filename)
    for line in f:
        v = line.rstrip().split()
        pred_list.append([v[0], float(v[sp]), int(v[rc])])
    return pred_list


def get_cm(data, th=0.1):
    cm = [[0, 0], [0, 0]]
    for i in data:
        rc = i[-1]
        #if (i[-2] <=th):
        if (i[-2] <= th): #when evaluating all the sets
            pc = 1
        else:
            pc = 0
        cm[pc][rc] = cm[pc][rc] + 1
        #confusion(cm)
    return cm


def calculate_performance(cm):
    n = float(cm[0][0] + cm[1][1] + cm[0][1] + cm[1][0])
    d = np.sqrt((cm[0][0] + cm[0][1]) * (cm[0][0] + cm[1][0]) * (cm[1][1] + cm[1][0]) * (cm[1][1] + cm[0][1]))
    mcc = (cm[0][0] * cm[1][1] - cm[0][1] * cm[1][0]) / d
    acc = (cm[0][0] + cm[1][1]) / n
    fpr= (cm[1][0])/(cm[1][0]+cm[1][1])
    tpr= (cm[1][1])/ (cm[1][1]+ cm[1][0])
    return acc, mcc,fpr, tpr

def opt_th(pred_list):
    for i in range(20):
        cm = get_cm(pred_list, 10 ** -i)
        acc, mcc,fpr,tpr = calculate_performance(cm)
        print('TH:', 10 ** -i, 'ACC:', acc, 'MCC:', mcc,'CM',cm, 'FPR',fpr,'TPR',tpr)

if __name__ == "__main__":
    predfile = sys.argv[1]
```

```
pred_list = get_preds(predfile)
if len(sys.argv) > 2:
    th = float(sys.argv[2])
    cm = get_cm(pred_list, th)
    acc, mcc,fpr,tpr = calculate_performance(cm)
    print('TH:', th, 'ACC:', acc, 'MCC:', mcc,'CM:',cm, 'FPR:',fpr,'TPR:',tpr)
else:
    opt_th(pred_list)
```