Supplementary Materials

May 21, 2019

1 PDB kunitz results

PDB ID	Chain ID	Resolution	Entity ID
1AAP	A	1.50	1
1AAP	B	1.50	1
1B0C	D	2.80	1
1B0C	E	2.80	1
1B0C	A	2.80	1
1B0C	B	2.80	1
1B0C	C	2.80	1
1BHC	C	2.70	1
1BHC	H	2.70	1
1BHC	D	2.70	1
1BHC	I	2.70	1
1BHC	E	2.70	1
1BHC	J	2.70	1
1BHC	A	2.70	1
1BHC	F	2.70	1
1BHC	B	2.70	1
1BHC	G	2.70	1
1BIK	A	2.50	1
1BPI	A	1.09	1
1BRC	I	2.50	2
1BTH	P	2.30	3
1BTH	Q	2.30	3
1BUN	B	2.45	2
1BZ5	A	2.58	1
1BZ5	B	2.58	1

PDB ID	Chain ID	Resolution	Entity ID
1BZ5	C	2.58	1
1BZ5	D	2.58	1
1BZ5	E	2.58	1
1BZX	I	2.10	2
1CA0	D	2.10	4
1CA0	I	2.10	4
1CBW	I	2.60	4
1CBW	D	2.60	4
1CO7	I	1.90	2
1D0D	B	1.62	2
1DTX	A	2.20	1
1EAW	D	2.93	2
1EAW	B	2.93	2
1F5R	I	1.65	2
1F7Z	I	1.55	2
1FY8	I	1.70	2
1KNT	A	1.60	1
1KTH	A	0.95	1
1MTN	D	2.80	4
1MTN	H	2.80	4
1TAW	B	1.80	2
1TFX	C	2.60	2
1TFX	D	2.60	2
1TPA	I	1.90	2
1Y62	C	2.45	1
1Y62	D	2.45	1
1Y62	E	2.45	1
1Y62	A	2.45	1
1Y62	F	2.45	1
1Y62	B	2.45	1
1YC0	I	2.60	2
1YKT	B	1.70	2
1ZJD	B	2.60	2
1ZR0	B	1.80	2
1ZR0	D	1.80	2
2FTL	I	1.62	2
2HEX	D	2.10	1
2HEX	E	2.10	1

PDB ID	Chain ID	Resolution	Entity ID
2HEX	A	2.10	1
2HEX	B	2.10	1
2HEX	C	2.10	1
2IJO	I	2.30	3
2KAI	I	2.50	3
2KNT	A	1.20	1
2ODY	E	2.35	3
2ODY	F	2.35	3
2PTC	I	1.90	2
2R9P	F	1.40	2
2R9P	G	1.40	2
2R9P	I	1.40	2
2R9P	E	1.40	2
2RA3	C	1.46	2
2RA3	I	1.46	2
2TGP	I	1.90	2
2TPI	I	2.10	2
3BTK	I	1.85	2
3BYB	C	1.63	1
3BYB	A	1.63	1
3BYB	B	1.63	1
3D65	I	1.64	1
3FP6	I	1.49	2
3FP7	J	1.46	3
3FP8	I	1.46	2
3GYM	I	2.80	2
3GYM	J	2.80	2
3L33	G	2.48	2
3L33	H	2.48	2
3L33	E	2.48	2
3L33	F	2.48	2
3LDI	C	2.20	1
3LDI	D	2.20	1
3LDI	E	2.20	1
3LDI	A	2.20	1
3LDI	B	2.20	1
3LDJ	C	1.70	1
3LDJ	A	1.70	1

PDB ID	Chain ID	Resolution	Entity ID
3LDJ	B	1.70	1
3LDM	D	2.60	1
3LDM	E	2.60	1
3LDM	A	2.60	1
3LDM	B	2.60	1
3LDM	C	2.60	1
3M7Q	B	1.70	2
3OFW	A	2.50	1
3T62	E	2.00	2
3T62	F	2.00	2
3T62	D	2.00	2
3TGI	I	1.80	2
3TGJ	I	2.20	2
3TGK	I	1.70	2
3TPI	I	1.90	2
3U1J	E	1.80	3
3UIR	C	2.78	2
3UIR	D	2.78	2
3WNY	A	1.30	1
3WNY	G	1.30	1
3WNY	B	1.30	1
3WNY	H	1.30	1
3WNY	C	1.30	1
3WNY	I	1.30	1
3WNY	E	1.30	1
3WNY	F	1.30	1
4BNR	J	2.00	2
4BNR	I	2.00	2
4BQD	A	2.48	1
4BQD	B	2.48	1
4DG4	F	1.40	2
4DG4	H	1.40	2
4DG4	C	1.40	2
4DG4	E	1.40	2
4DTG	K	1.80	3
4ISL	B	2.29	1
4ISN	B	2.45	1
4ISO	B	2.01	2

PDB ID	Chain ID	Resolution	Entity ID
4NTW	B	2.07	2
4NTX	B	2.27	2
4NTY	B	2.65	2
4PTI	A	1.50	1
4U30	W	2.50	2
4U30	X	2.50	2
4U30	Y	2.50	2
4U30	Z	2.50	2
4U32	X	1.65	1
4WWY	C	1.70	2
4WWY	I	1.70	2
4WXV	I	2.10	2
4WXV	C	2.10	2
4Y0Y	I	1.25	2
5EZD	A	2.10	1
5EZD	B	2.10	1
5JBT	Y	1.40	3
5NMV	K	1.65	3
5NX1	C	1.85	4
5NX1	D	1.85	3
5PTI	A	1.00	1
5YV7	A	2.39	1
5YVU	I	2.49	3
5YW1	I	2.60	3
5ZJ3	C	1.88	1
5ZJ3	A	1.88	1
5ZJ3	B	1.88	1
6F1F	D	1.72	1
6F1F	E	1.72	1
6F1F	A	1.72	1
6F1F	B	1.72	1
6F1F	C	1.72	1
6PTI	A	1.70	1
9PTI	A	1.22	1

Table S1: PDB kunitz

2 Common list PDBe fold result

PDB Id	CHAIN
3ldi	E
1brc	I
1bpi	A
3tgi	I
3ldj	C
1yc0	I
4u30	Z
3tgj	I
1f7z	I
3tgk	I
4u32	X
3ldm	E
1knt	A
4dtg	K
1fy8	I
3fp6	I
3m7q	B
3fp7	J
3fp8	I
1d0d	B
3d65	I
1kth	A
1bth	Q
2tpi	I
1zjd	B
1co7	I
2knt	A
1aap	B
5nmv	K
1bun	B
3uir	D
4pti	A
4ntw	B
5zj3	C
4ntx	B
3tpi	I

PDB Id	CHAIN
2ra3	I
4nty	B
1ykt	B
4y0y	I
1mtn	H
1taw	B
3gym	J
2r9p	I
2kai	I
1ca0	I
1bhc	J
4wxv	I
4wwy	I
9pti	A
6f1f	E
3u1j	E
2hex	E
3ofw	A
1dtx	A
5pti	A
4isl	B
1bz5	E
4dg4	H
4bqd	B
1cbw	I
4isn	B
3btk	I
4iso	B
1bzx	I
1tfx	D
5nx1	D
4bnr	J
3byb	C
2ijo	I
1f5r	I
1zr0	D
1tpa	I
1eaw	D

PDB Id	CHAIN
3t62	F
3l33	H
2ptc	I
6pti	A
2tgp	I
5yw1	I
2ftl	I
1b0c	E
5yvu	I
3wny	I
1y62	F
5yv7	A

Table S2:
PDB ids listed both in the PDB search and PDBe fold alignment, just a chain was retained for each structure

3 Python script:

3.1 Get fasta from a list of PDB ids

```
#!/usr/bin/python
import sys

def get_list_fasta(lid,fasta):
    f=open(fasta)
    c=0
    for line in f:
        line=line.rstrip()
        if line[0]=='>':
            tid=line.split(' ')[0][1:]

        if lid.get(tid,False)==1:

#you go here only if the statement is true and so c=1 and it will
#print the value. In the case the id is not present
#in line it will return false and it will pass in the
#else statement

            c=1
        else:
            c=0
        if c==1 :
            print(line)


if __name__=="__main__":
    fid=sys.argv[1]
    fasta= sys.argv[2]
    lid=dict([(i,True) for i in open(fid).read().split('\n')])

    get_list_fasta(lid,fasta)
```

3.2 Sort PDB ids for their resolution

```
#!/usr/bin/env python
import sys
def get_dic(filename):
    d={}
    f=open(filename)
    for line in f:
        v= line.rstrip().split()
        #print(v)
        d[v[0]]=float(v[-1])

    return d
#the key is the identifier and the value is the resolution

#sort on the basis of the resolution

def sort_cluster(clist,d):
    tlist=[]
    #temporary list that will contain
    #the identifiers with the value of resolution
    for pid in clist:
        v=d.get(pid, float ('inf'))
        tlist.append([v,pid])
    tlist.sort()
    return tlist #this is just clustering line by line

if __name__=="__main__":
    f1=sys.argv[1] # list common clusters
    f2=sys.argv[2] #list ids and resolutions
    d=get_dic(f2)
    f=open(f1)
    for line in f:
        lid=line.rstrip().split()
        slid=sort_cluster(lid,d)
        s=''

        print(len(slid), ''.join([i[1] + ':' + str(i[0])

        for i in slid ]))
```

4 Protein set for multiple structural alignment:

PDB id	Chain
5pti	A
1aap	B
3tgi	I
3byb	C
4isn	B
3m7q	B
4ntw	B
1kth	A
4bnr	J
4bqd	B
1zr0	D
1bun	B
1y62	F
5yv7	A
1dtx	A
1tfx	D
4u32	X
3fp7	J

Table S3: PDB ids of the 18 chosen kunitz structure for multiple sequence alignment.

Submission form	List of entries	Source
Multiple	5pti:A 1aap:B 3tgi:I 3byb:C 4isn:B 3m7q:B 4ntw:B 1kth:A 4bnr:J 4bqd:B 1zr0:D 1bun:B 1y62:F 5yv7:A 1dtx:A 1tfx:D 4u32:X 3fp7:J	List of PDB codes

Table S4: Multiple structural alignment parameters

5 Multiple structure alignment:

```
>PDB:5pti:A STRUCTURE OF BOVINE PANCREATIC TRYPSIN INHIBITOR.
-rpdfcleppytgpckARIIRYFYNAKAGLCQTFVYGgCRA-KRNNFKSAEDCMRTCgga

>PDB:1aap:B X-RAY CRYSTAL STRUCTURE OF THE PROTEASE INHIBITOR
-vrevcseqaetgpcrAMISRWYFDVTEGKCAPFFYGgCGG-NRNNFDTEEYCMAVCg--

>PDB:3tgi:I WILD-TYPE RAT ANIONIC TRYPSIN COMPLEXED WITH BOVIN
-rpdfcleppytgpckARIIRYFYNAKAGLCQTFVYGgCRA-KRNNFKSAEDCMRTCg--

>PDB:3byb:C CRYSTAL STRUCTURE OF TEXTILININ-1, A KUNITZ-TYPE S
drpdfcelpadtgpcrVRFPSFYYNPDEKKCLEFIYGgCEG-NANNFITKEECESTCa--

>PDB:4isn:B CRYSTAL STRUCTURE OF MATRIPTASE IN COMPLEX WITH IT
qtedyclasnkvgrcrGSFPRWYYDPTEQICKSFVYGgCLG-NKNNYLREEECILACrgv

>PDB:3m7q:B CRYSTAL STRUCTURE OF RECOMBINANT KUNITZ TYPE SERIN
aeasicsepkkvgrckGYFPRFYFDSETGKCTPFIYGgCGG-NGNNFETLHQCRAICral

>PDB:4ntw:B STRUCTURE OF ACID-SENSING ION CHANNEL IN COMPLEX W
rpafcyedppffqkcgAFVDSYYFNRSRITCVHFFYG-QCDvNQNHFTTMSECNRVChg-

>PDB:1kth:A THE ANISOTROPIC REFINEMENT OF KUNITZ TYPE DOMAIN C
-etdicklpkdegtcrDFILKWYYDPNTKSCARFWYGgCGG-NENKFGSQKECEKVCapv

>PDB:4bnr:J EXTREMELY STABLE COMPLEX OF CRAYFISH TRYPSIN WITH
-rpdfcleppytgpckARIIRYFYNAKAGLCQTFVYGgCRA-KRNNFKSAEDCMRTCgga

>PDB:4bqd:B KD1 OF HUMAN TFPI IN COMPLEX WITH A SYNTHETIC PEPT
lmhsfcafkaddgpckAIMKRFFFNIFTRQCEEFIYGgCEG-NQNRFESLEECKKMCtrd

>PDB:1zr0:D CRYSTAL STRUCTURE OF KUNITZ DOMAIN 1 OF TISSUE FAC
nnaeicllpldygpcrALLLRYYYDRYTQSCRQFLYGgCEG-NANNFYTWEACDDACwri

>PDB:1bun:B STRUCTURE OF BETA2-BUNGAROTOXIN: POTASSIUM CHANNEL
krhpdcdkppdtkicqTVVRAFYYKPSAKRCVQFRYG-GCNgNGNHFKSDHLCRCECley

>PDB:1y62:F A 2.4 CRYSTAL STRUCTURE OF CONKUNITZIN-S1, A NOVEL
-rpslcdlpadsgsgtKAEKRIYYNSARKQCLRFDYTgQGG-NENNFRRTYDCQRTCl--
```

```
>PDB:5yv7:A RACEMIC X-RAY STRUCTURE OF CALCICLUDINE
qppwyckepvrigsckKQFSSFYFKWTAKKCLPFLFSgCGG-NANRFQTIGECRKKClgk

>PDB:1dtx:A CRYSTAL STRUCTURE OF ALPHA-DENDROTOXIN FROM THE GR
prrklcilhrnpgrcyDKIPAFYYNQKKKQCERFDWSgCGG-NSNRFKTIEECRRTCig-

>PDB:1tfx:D COMPLEX OF THE SECOND KUNITZ DOMAIN OF TISSUE FACT
-kpdfcfleedpgicrGYITRYFYNNQTKQCERFKYGgCLG-NMNNFETLEECKNICedg

>PDB:4u32:X HUMAN MESOTRYPSIN COMPLEXED WITH HAI-2 KUNITZ DOMA
--hdfclvskvvgrcrASMPRWWYNVTDGSCQLFVYGgCDG-NSNNYLTKEECLKKC---

>PDB:3fp7:J ANIONIC TRYPSIN VARIANT S195A IN COMPLEX WITH BOVI
----------------ARIIRYFYNAKAGLCQTFVYGgCRA-KRNNFKSAEDCMRTCgga
```

6 Hints Positive set

Uniprot id	E-value full sequence	E-value domain
Q868Z9	6.4e-190	5e-20
O76840	1.2e-172	4.6e-20
Q02445	9.6e-69	7.5e-25
Q28864	8.4e-68	1.4e-25
O54819	1.2e-66	6.2e-25
P19761	1.4e-64	1.9e-24
P84875	2.2e-63	2.1e-22
Q03610	1.1e-62	1.6e-18
P83606	6.9e-62	3.6e-22
Q7YRQ8	2e-58	1.7e-26
O35536	5.2e-50	8.1e-24
P86733	1.3e-47	1.1e-24
Q8WPI2	4.8e-46	9e-26
W4VSH9	8.8e-46	4.6e-25
Q6T269	2.2e-45	1.2e-23
Q8WPI3	2.5e-45	2.7e-25
Q9WU03	2e-44	1.5e-24
Q9R097	1.1e-40	4.1e-22
B2BS84	1.2e-39	1.6e-20
P04365	1.7e-38	4.1e-21
P83609	1.3e-36	8.1e-20
Q60559	3e-36	1.9e-19
P62756	6.9e-36	6.3e-19
P62757	6.9e-36	6.3e-19
P00978	7.7e-36	1.1e-19
Q62577	4e-35	1.2e-18
Q08E66	5.9e-35	1.3e-22
P02760	1.4e-34	2.2e-18
Q07456	1.5e-34	4.6e-18
P04366	1.8e-34	5.2e-19
Q8TEU8	6.1e-34	1.4e-21
Q64240	6.1e-34	8.1e-18
Q7TQN3	2e-33	2.3e-22
Q6NUX0	6.2e-29	1.3e-19
Q90WA0	1.7e-28	1.9e-28

Uniprot id	E-value full sequence	E-value domain
Q6ITB5	3.9e-28	4.5e-28
Q6ITB4	5e-28	5.7e-28
Q6ITB6	5e-28	5.7e-28
Q6ITB9	7.3e-28	8.4e-28
P0DMW7	1.4e-27	1.5e-27
Q90W98	1.7e-27	1.9e-27
B7S4N9	1.9e-27	2.2e-27
P0DMW6	2.1e-27	2.2e-27
P00975	2.1e-27	2.3e-27
Q9TWG0	2.9e-27	3.3e-27
P10280	2.9e-27	3.2e-27
P00985	3.7e-27	3.9e-27
P0DMJ6	5.5e-27	5.8e-27
Q7LZS8	5.5e-27	5.8e-27
Q90W99	5.8e-27	6.5e-27
C0HJU7	6.1e-27	6.4e-27
C0HK72	6.1e-27	6.4e-27
Q6ITB7	6.2e-27	7e-27
P0DN08	6.5e-27	9e-27
P29216	6.7e-27	7.8e-27
B5KL39	7.2e-27	8.2e-27
C0HK74	7.5e-27	7.9e-27
P0DN09	8e-27	1.1e-26
P0DN06	8.1e-27	1.1e-26
P00982	8.6e-27	9.1e-27
P36992	9.9e-27	1e-18
Q6ITB0	9.9e-27	1.1e-26
Q6ITB1	9.9e-27	1.1e-26
B5L5R0	1e-26	1.2e-26
B5G6G6	1e-26	1.2e-26
C1IC50	1.1e-26	1.2e-26
F8J2F3	1.2e-26	1.4e-26
P00984	1.3e-26	1.3e-26
Q7LZE3	1.3e-26	1.3e-26
P0DMJ5	1.3e-26	1.6e-26
P00986	1.3e-26	1.4e-26
Q2ES47	1.3e-26	1.5e-26

Uniprot id	E-value full sequence	E-value domain
B5KF96	1.4e-26	1.5e-26
Q9TWF8	1.5e-26	1.9e-26
B6RLX2	1.6e-26	1.7e-26
P00981	1.7e-26	2.1e-26
Q2ES46	1.9e-26	2.3e-26
F8J2F5	2e-26	2.3e-26
Q6T6T5	2.2e-26	2.5e-26
B5L5R7	2.5e-26	2.8e-26
B5KL37	2.6e-26	2.9e-26
C1IC52	2.6e-26	3e-26
P81129	2.6e-26	2.8e-26
P0DN11	2.7e-26	3.7e-26
P04815	2.8e-26	3.9e-26
Q9TWF9	3.1e-26	3.5e-26
P0DN13	3.5e-26	4.7e-26
B5KL40	3.7e-26	4.2e-26
B5KF95	3.8e-26	4.3e-26
B5KL36	3.8e-26	4.3e-26
P0DN12	4e-26	5.3e-26
C1IC51	4.4e-26	5.1e-26
C0HK73	4.8e-26	5e-26
P86862	4.9e-26	5.3e-26
P0DN19	5e-26	6.5e-26
P12023	5.2e-26	1.2e-25
P00991	5.6e-26	6.5e-26
C0HJF4	5.8e-26	6.1e-26
Q8T3S7	6.3e-26	7.6e-26
H2A0P0	6.3e-26	5.4e-15
Q95241	6.5e-26	1.6e-25
C0HJF3	7e-26	7.4e-26
P08592	7.4e-26	1.6e-25
P53601	7.4e-26	1.6e-25
Q5IS80	7.4e-26	1.6e-25
P00993	7.8e-26	7.8e-26
P0DN10	8.4e-26	1.2e-25
B5KL38	8.5e-26	9.6e-26
Q6ITC0	8.8e-26	1e-25

Uniprot id	E-value full sequence	E-value domain
B5KL35	9.5e-26	1.1e-25
P00992	1e-25	1.2e-25
A8Y7N9	1.1e-25	1.2e-25
P81547	1.2e-25	1.2e-25
B5L5M7	1.2e-25	1.4e-25
Q6T6S5	1.2e-25	1.4e-25
B5L5R4	1.3e-25	1.5e-25
C0HJU6	1.4e-25	1.5e-25
Q6ITB3	1.4e-25	1.6e-25
B5KL33	1.5e-25	1.6e-25
P0DL86	1.6e-25	1.6e-25
F8J2F4	1.6e-25	1.8e-25
P79307	1.6e-25	1.6e-25
Q60495	1.7e-25	4.4e-25
A6MFL4	1.9e-25	2.1e-25
P0DN14	2.3e-25	3.2e-25
P0DN17	2.3e-25	3.2e-25
B5L5R1	2.5e-25	2.8e-25
Q96NZ8	2.7e-25	3.2e-19
P00994	3.1e-25	3.3e-25
B2G331	3.2e-25	3.9e-25
A6MGY1	3.3e-25	3.7e-25
E7FL11	3.5e-25	3.9e-25
A6MFL3	4.3e-25	4.9e-25
Q6ITC1	4.5e-25	5.1e-25
P25660	4.6e-25	5.2e-25
P0DN20	4.7e-25	6.3e-25
B4ESA3	4.7e-25	5.4e-25
P0DN07	5.3e-25	7.4e-25
B5KL30	5.4e-25	6.2e-25
A6MGX9	5.5e-25	6.2e-25
A6MFL1	5.6e-25	6.3e-25
A5X2X1	5.6e-25	5.6e-25
E7FL13	5.6e-25	6.4e-25
P0DKL8	5.9e-25	6.5e-25
P0DMJ4	6.4e-25	6.7e-25
P24541	6.7e-25	7.4e-25

Uniprot id	E-value full sequence	E-value domain
E7FL12	6.7e-25	7.6e-25
Q6ITB8	7e-25	7.9e-25
B1B5I8	7.1e-25	9.1e-25
P00979	7.2e-25	8.3e-25
C1IC53	8e-25	1e-24
B5KL31	8.6e-25	9.7e-25
B2KTG2	8.6e-25	9.8e-25
A8Y7N7	8.9e-25	1e-24
B5L5Q8	9.5e-25	1.1e-24
B2KTG1	1e-24	1.2e-24
A8Y7N6	1e-24	1.2e-24
B5KF94	1.1e-24	1.2e-24
P0DMJ2	1.3e-24	1.4e-24
P0DN18	1.3e-24	1.9e-24
Q1RPT0	1.4e-24	1.6e-24
P00976	1.4e-24	1.7e-24
Q8AY45	1.4e-24	1.6e-24
Q8AY41	1.7e-24	1.9e-24
Q75S50	1.9e-24	2.2e-24
Q8AY44	1.9e-24	2.2e-24
Q06481	1.9e-24	1.9e-24
E5AJX3	2e-24	2.3e-24
P15943	2.3e-24	2.3e-24
F6ULY1	2.5e-24	2.5e-24
Q6ITB2	2.6e-24	2.9e-24
P81902	3.5e-24	3.7e-24
A6MFL2	3.6e-24	4e-24
P20229	3.8e-24	4e-24
B5KL32	4.4e-24	4.9e-24
A0A1Z0YU59	4.4e-24	4.7e-24
H2A0N1	4.4e-24	2.9e-13
B5KL34	4.6e-24	5.2e-24
F8J2F6	4.8e-24	5.4e-24
Q8AY43	5.2e-24	6.2e-24
B5KL41	5.5e-24	6.2e-24
Q29428	6.6e-24	1e-23
H6VC06	8e-24	8.9e-24

Uniprot id	E-value full sequence	E-value domain
P0DMJ3	9e-24	9.4e-24
A8Y7P1	9e-24	1e-23
B4ESA4	1e-23	1.2e-23
A8Y7P5	1.1e-23	1.2e-23
P0DJ50	1.1e-23	1.4e-23
P00990	1.1e-23	1.2e-23
Q2ES48	1.1e-23	1.3e-23
I2G9B4	1.2e-23	1.4e-23
A8Y7N5	1.3e-23	1.4e-23
P0DMX0	1.6e-23	1.6e-23
B5KL27	1.7e-23	1.9e-23
P81548	1.7e-23	1.7e-23
P0C8W3	1.8e-23	2.2e-23
P81162	2e-23	2.4e-23
A8Y7P4	2.1e-23	2.3e-23
Q8AY42	2.1e-23	2.5e-23
Q3UW55	2.4e-23	2.4e-23
A7X3V4	2.4e-23	3.1e-23
A8Y7P6	2.5e-23	2.8e-23
Q90W96	2.6e-23	3e-23
Q7T2Q6	2.7e-23	3e-23
B5KL29	2.7e-23	3.1e-23
A8Y7P0	3.3e-23	3.7e-23
Q2ES50	3.3e-23	3.7e-23
D4A2Z2	3.6e-23	3.6e-23
P26228	3.7e-23	4.4e-23
A8Y7N4	3.9e-23	5e-23
H6VC05	3.9e-23	5e-23
Q28201	4.4e-23	8.4e-23
A8Y7N8	5.2e-23	6.5e-23
B2KTG3	5.2e-23	5.9e-23
A8Y7P3	5.4e-23	6e-23
O93279	5.4e-23	1.2e-22
P0DN15	6.5e-23	7.9e-23
Q0PL65	6.7e-23	7.2e-23
O62845	8.2e-23	1.4e-22
A8Y7P2	8.9e-23	1e-22

Uniprot id	E-value full sequence	E-value domain
A7X3V7	1e-22	1.3e-22
B5L5Q1	1.3e-22	1.5e-22
P15989	1.3e-22	3e-22
B5KL28	1.3e-22	1.5e-22
P0DJ46	1.8e-22	2.1e-22
Q5ZPJ7	2.1e-22	2.4e-22
D8KY58	2.6e-22	3.1e-22
P82966	3.1e-22	3.5e-22
H2A0N5	3.5e-22	5.1e-12
P82968	3.9e-22	3.9e-22
Q29100	4.1e-22	4.1e-22
Q90W97	4.9e-22	5.5e-22
Q9DA01	5.1e-22	5.1e-22
P0DN16	5.6e-22	6.8e-22
P0DJ45	7.1e-22	8.3e-22
O95925	9.3e-22	9.3e-22
P86964	9.8e-22	1.1e-11
B5L5R6	1e-21	1.2e-21
Q4KUS1	1.1e-21	1.1e-21
P0DJ48	1.3e-21	1.4e-21
P49223	1.3e-21	1.8e-21
H2A0M2	2e-21	3.8e-21
Q9BDL1	2.2e-21	2.2e-21
P0C5J5	2.3e-21	1.9e-19
P16044	2.6e-21	2.9e-21
Q8R0S6	2.6e-21	1.2e-19
P86959	2.9e-21	6e-21
P0DJ77	3.1e-21	3.7e-21
P68425	3.1e-21	3.7e-21
P16344	3.2e-21	3.4e-21
P19859	3.7e-21	3.9e-21
B2ZBB6	4.2e-21	4.9e-21
Q29143	4.3e-21	6.2e-21
P0DJ66	4.9e-21	6.7e-21
P0DJ49	5e-21	5.5e-21
B5L5Q6	5.1e-21	5.8e-21
P00983	7.9e-21	8.2e-21

Uniprot id	E-value full sequence	E-value domain
P0DJ76	8e-21	9.3e-21
D2Y489	9.6e-21	1e-20
D2Y491	1.3e-20	1.4e-20
P0DJ47	1.8e-20	2.1e-20
D2Y490	2.1e-20	2.3e-20
P0CY85	2.8e-20	3.3e-20
Q1RPS9	3.5e-20	4.2e-20
D2Y488	3.7e-20	4.4e-20
P11424	4.9e-20	5.4e-20
D2Y2Q9	7.5e-20	8.8e-20
P0DMJ1	7.8e-20	9.2e-20
D2Y2Q2	8.1e-20	9.5e-20
D2Y2G1	9.4e-20	1.1e-19
P0DJ84	9.4e-20	1.1e-19
D2Y2Q7	9.5e-20	1.1e-19
Q1RPS8	1.3e-19	1.6e-19
Q589G4	1.4e-19	1.5e-19
P10832	1.5e-19	1.7e-19
D2Y2Q8	1.6e-19	1.9e-19
P26227	1.6e-19	1.7e-19
Q8T0W4	1.9e-19	2.4e-19
P0DJ82	2.2e-19	2.6e-19
Q9EPX2	2.2e-19	2.2e-19
P10831	2.3e-19	2.5e-19
O95428	2.4e-19	2.4e-19
P81906	2.6e-19	2.9e-19
D2Y2Q1	3.1e-19	3.6e-19
Q9W728	3.8e-19	4.5e-19
H2A0N9	4.5e-19	4.5e-19
B4ESA2	4.6e-19	5.1e-19
D2Y2Q5	4.8e-19	5.6e-19
Q75S49	4.9e-19	5.7e-19
Q8AY46	5.1e-19	6e-19
P00987	7.3e-19	8.6e-19
B2ZBC0	7.5e-19	8.8e-19
P0DJ85	1.3e-18	1.6e-18
B5L5Q3	1.4e-18	1.6e-18

Uniprot id	E-value full sequence	E-value domain
P0DJ80	2.1e-18	2.5e-18
B2ZBB8	2.1e-18	2.5e-18
P0DJ75	2.1e-18	2.5e-18
P0DJ78	2.1e-18	2.5e-18
P0DJ79	2.5e-18	2.9e-18
P07481	3.2e-18	3.6e-18
Q9D263	3.5e-18	4.9e-18
D2Y2C2	3.6e-18	4.2e-18
P0DJ70	3.6e-18	4.2e-18
D2Y2F6	3.7e-18	4.4e-18
P0DJ72	3.7e-18	4.4e-18
D2Y2F4	4.2e-18	4.9e-18
P0DJ67	4.2e-18	4.9e-18
P0DJ69	4.3e-18	4.9e-18
P0DJ74	4.3e-18	4.9e-18
D2Y2F8	4.7e-18	5.6e-18
D2Y2F5	4.8e-18	5.6e-18
P0DJ71	4.8e-18	5.6e-18
Q2UY11	7.5e-18	1.4e-17
B2ZBB9	8.4e-18	1e-17
D2Y2Q6	1.1e-17	1.3e-17
Q2UY09	1.2e-17	2.1e-17
O73683	1.4e-17	3e-17
Q02388	1.5e-17	2.4e-17
D2Y2F3	1.7e-17	2e-17
P0DJ64	1.7e-17	2e-17
D2Y2F7	1.9e-17	2.5e-17
P0DJ73	1.9e-17	2.5e-17
Q63870	2.1e-17	3.2e-17
D2Y2F9	2.8e-17	3.4e-17
P0DJ65	2.8e-17	3.4e-17
Q6UDR6	1.3e-16	1.8e-16
P0DJ81	1.8e-16	2.1e-16
D2Y2G0	3.7e-16	4.5e-16
D2Y2G2	8.3e-16	9.8e-16
P0DJ68	8.3e-16	9.8e-16
P0DMW9	1.4e-15	1.4e-15

Uniprot id	**E-value full sequence**	**E-value domain**
C0LNR2	9.5e-15	1.2e-14
Q8IUA0	1.1e-14	1.1e-14
P0DMW8	1.4e-14	1.5e-14
Q2ES49	7.2e-14	8.3e-14
Q7Z1K3	1.7e-13	4.6e-13
P26226	2.1e-11	2.8e-11
Q9BQY6	3.2e-11	3.2e-11
P0CAR0	3.2e-11	3.3e-11
P0CH75	6.2e-11	6.8e-11
Q11101	4.8e-08	1.8e-07
P86963	3.3e-07	3.3e-07
O62247	2.7e-05	2.7e-05
D3GGZ8	0.00033	0.00033

Table S5: Hints of hmmsearch for the positive set

6.1 Computing the performance

```
#!/usr/bin/env python
import sys
import numpy as np
#filename=allsets.txt
def prettymatrix(m):
    for y in range(0,len(m)):
        print (m[y])

def conf_mat(filename,th,sp=-2,cp=-1):
    cm=[[0.0,0.0] , [0.0,0.0]]
    f=open(filename)
    for line in f:
        v=line.rstrip().split()
        if int(v[cp])==1: i = 1
        if int(v[cp])==0: i = 0
        if float(v[sp]) < th:
            j=1
        else:
            j=0
        cm[i][j]=cm[i][j]+1
    prettymatrix(cm)
    return cm

#1.1 are truepositive (i=1 stays for positive class , j=1
#they were above the trhashold e so they are classified
#as positive by our method)

#0.0 true negative (i=0 stays for negativa class  and j=0
#they were under the trhashold e so they are classified
#as negative by our method

#1.0 are false negative , since i=1 stays for positive
#class but j=0  means they are classidied as negative
#by our method.

# 0.1 are false positive since they belong to is i=0
#that means negative, but they are above the threshold
#since our methods classified them as positive.
```

```
def print_performance(cm):
    tp=cm[1][1] #true positive
    tn=cm[0][0] #true negative
    fn=cm[1][0] #false negative
    fp=cm[0][1] #false positive

    acc=(tn+tp)/(tp+tn+fn+fp) #accuracy
    d=np.sqrt((tn+ fp)*(tn+ fn)* (tp+ fn)*(tp+ fp))
    #denominator of mc
    mc=(tn*tp -fp*fn) /d #matthews correlation coefficient
    n=sum(cm[0]) +sum(cm[1])
    fpr= fn/(fn +tp) #false positive rate
    tpr= tp/ (tp+ fn) #true positive rate


    print ('THR=',th, 'Q2=', acc,'MCC',mc, 'FPR=',fpr,'TPR', tpr)


if __name__== '__main__' :
    filename=sys.argv[1]
    th=float(sys.argv[2])
    sp=-2
    if len(sys.argv)>3: sp=int(sys.argv[3])-1
    cm=conf_mat(filename,th,sp)
    #print(cm)
    print_performance(cm)
```

7 Bash syntax :

7.1 Removing the redundancy

```
blastclust -i common_list.fasta -l 0.95 -s99 -o list_common_c96_i99.txt
```

7.2 Building the model

```
hmmbuild model.hmm alignment.fasta
```

7.3 HMMsearc: positive set

```
hmmsearch -o positive.search --noali model.hmm positive_set.fasta
```

7.4 HMMsearc: negative set

```
hmmsearch --max -E 100000 --domE 100000000 -o negative.search --noali
model.hmm negativeset.fasta
```

8 Results confusion matrices

The confusion matrices at different threshold :

- Threshold=1

$$CM_1 = \begin{bmatrix} [400918.0 & 148875.0 \\ 0.0 & 344.0 \end{bmatrix}$$

$Q2 = 0.7293855894077294$

$MCC = 0.04100108346618696$
$FPR = 0.0\ TPR = 1.0$

- Threshold = 0.1

$$CM_{0.1} = \begin{bmatrix} 530697.0 & 19096.0 \\ 0.0 & 344.0 \end{bmatrix}$$

$Q2 = 0.9652886462826532$

$MCC = 0.13069374545031118$

$FPR = 0.0$, $TPR = 1.0$

- Threshold = 0.01

$$CM_{0.01} = \begin{bmatrix} 548076.0 & 1717.0 \\ 0.0 & 344.0 \end{bmatrix}$$

$Q2 = 0.9968789592410617$

$MCC = 0.40790686532395704$

$FPR = 0.0$, $TPR = 1.0$

- Threshold = 0.001

$$CM_{1\times10^{-2}} = \begin{bmatrix} 549635.0 & 158.0 \\ 0.0 & 344.0 \end{bmatrix}$$

$Q2 = 0.9997127988119323$

$MCC = 0.8276847430117906$

$FPR = 0.0$, $TPR = 1.0$

- Threshold = 0.0001

$$CM_{1\times10^{-4}} = \begin{bmatrix} 549781.0 & 12.0 \\ 0.0 & 344.0 \end{bmatrix}$$

$Q2 = 0.9999781872515392$

$MCC = 0.9829908667715128$

$FPR = 0.0$, $TPR = 1.0$

- Threshold = 0.00001

$$CM_{1\times10^{-5}} = \begin{bmatrix} 549790.0 & 3.0 \\ 0.0 & 344.0 \end{bmatrix}$$

$Q2 = 0.9999945468128848$

$MCC = 0.9956651331822753$

$FPR = 0.0$, $TPR = 1.0$

- Threshold = 0.000001

$$CM_{1\times10^{-6}} = \begin{bmatrix} 549791.0 & 2.0 \\ 0.0 & 344.0 \end{bmatrix}$$

$Q2 = 0.9999963645419232$

$MCC = 0.997103824320116$

$FPR = 0.0$, $TPR = 1.0$

- Threshold = 0.0000001

$$CM_{1\times10^{-7}} = \begin{bmatrix} 549791.0 & 2.0 \\ 1.0 & 344.0 \end{bmatrix}$$

$Q2 = 0.9999945468128848$

$MCC = 0.9956441852517098$

$FPR = 0.0029069767441860465$, $TPR = 0.997093023255814$

- Threshold = 0.00000001

$$CM_{1\times10^{-8}} = \begin{bmatrix} 549791.0 & 2.0 \\ 2.0 & 342.0 \end{bmatrix}$$

$Q2 = 0.9999927290838464$

$MCC = 0.9941824087788812$

$FPR = 0.005813953488372093$, $TPR = 0.9941860465116279$

- Threshold = 0.000000001

$$CM_{1\times10^{-9}} = \begin{bmatrix} 549792.0 & 1.0 \\ 2.0 & 342.0 \end{bmatrix}$$

$Q2 = 0.9999945468128848$

$MCC = 0.9956315155642677$

$FPR = 0.005813953488372093$, $TPR = 0.9941860465116279$

9 Roc_Curve.py

```
# roc curve
import matplotlib.pyplot as plt
import sys
import numpy as np
from sklearn.metrics import roc_curve, auc

def get_data(dataf):
        with open(dataf) as f:
                label = []
                e_val = []
                for line in f:
                        label.append(float(line.split()[3]))
                        e_val.append(-1*float(line.split()[2]))
        return label,e_val

def roc2(fpr,tpr,roc_auc):
        plt.figure()
        plt.plot(fpr, tpr, color='darkorange', lw=1,


        label='ROC curve (area = %0.2f)' % roc_auc)
        plt.plot([0, 1], [0, 1], color='navy', lw=1, linestyle='--')
        plt.xlim([0.0, 1.0])
        plt.ylim([0.0, 1.05])
        plt.xlabel('False Positive Rate')
        plt.ylabel('True Positive Rate')
        plt.title('Receiver operating characteristic')
        plt.legend(loc="lower right")
        plt.show()

def roc1(fpr,tpr,roc_auc):
        fig, ax = plt.subplots(1,1, figsize=(10,5))
        ax.plot(fpr, tpr)
        #~ ax.plot(x,x, "--")
        ax.set_xlim([0,1])
        ax.set_ylim([0,1])
```

```
        ax.set_title("ROC Curve", fontsize=14)
        ax.set_ylabel('TPR', fontsize=12)
        ax.set_xlabel('FPR', fontsize=12)
        ax.grid()
        ax.legend(["AUC=%.5f"%roc_auc])
        plt.show()


if __name__ == '__main__':
        ''' usage: python roc_curve.py labeled_data.txt
         labeled data comprises both pos and neg data,

         each with its label '''
        labeled_data = sys.argv[1]
        # The input data is split into labels

        #and e-values
        label,e_val = get_data(labeled_data)
        #this function takes in input the labels

        #and the e-val
        # and returns a list of FPRs and TPRs
        fpr, tpr, thresholds = roc_curve(label, e_val, pos_label=1)
        # Another function that computes the area under the curve
        roc_auc = auc(fpr, tpr)
        print roc_auc

        roc2(fpr,tpr,roc_auc)
        roc1(fpr,tpr,roc_auc)
```