

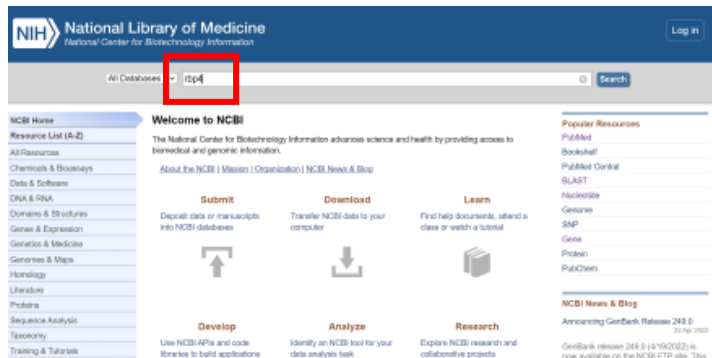


Semester II 2021/2022

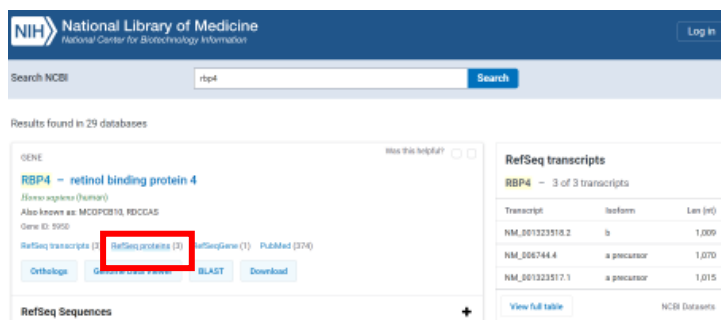
Subject : Bioinformatics I (SCSB2103)
Section : 01 – Dr Haslina Hashim
Topic : Lab 05 – Advanced Database Searching

Name : Gui Yu Xuan
Matric ID : A20EC0039

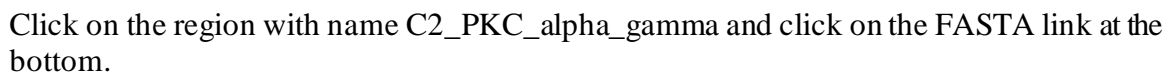
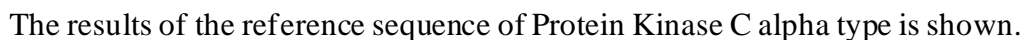
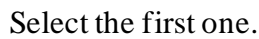
- 1) Create an artificial protein sequence consisting of human RBP4 followed by the C2 domain of human protein kinase C α . An example of this is shown in Web Document 5.5 (available in your e-Learning site). Enter this combined sequence into a PSI-BLAST search
 - i) In general, are multiple domains always detected by the PSI-BLAST program?
 - ii) Do any naturally occurring proteins have both lipocalin and C2 domains?



Go to NCBI website and type rbp4 on the search bar.



Then, click on the RefSeq proteins.





You will get the results.

Here is RBP4, obtained via NCBI Gene then RefSeq:

```
>NP_006735.2 retinol-binding protein 4 isoform a precursor [Homo sapiens]
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLNNWDVCADMVGTFTDTEDPAKFVKMYWGVASFQKGNDDHWIVD TDYDTYAVQYSCRL
LNL DGT CADSY SFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
```

Here is the entire PKCA protein: C2 domain

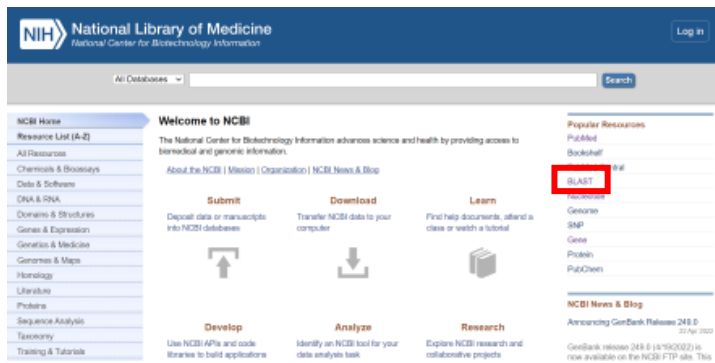
```
>NP_002728.2 protein kinase C alpha type [Homo sapiens]
MADVFPGNDS TASQDVANRFARKGALRQKNVHEVKDHKFIARFFKQPTFC SHCTDFIWGF GKQGFQCQVC
CFV VHKRCHEFVTFSCPGADKGPDTDDPRS KHKFKIHTYGSPTFCDHCGSLLYGLIHQGMKCDTCDMNVH
KQCVINVP SLCGMDHTEKRGRIYLKAEVADEKLHVTVRDAKNLI PMDPNGLSDPYVKLKLIPDPKNESKQ
KTKTIRSTLNPQWNESFTFKLKPSDKDRRLSVEIWDWDRTTRNDFMGSLSFGVSELMKMPASGWYKLLNQ
EEGEYYNVPIPEGDEEGNME LRQKFEKAKLGPAGNKVISPSEDRKQPSNNLDRVKLTD FNFMLVLGKGSF
GKVMLADRKGTEELYAIKILK KDVIQDDVECTMVEKRVLALLDKPPFLTQLHSCFQTVDRLYFVMEYV
NGGDIMYHIQQVGKFKEPQAVFYAAEISIGLFFLHKRGIIYRDLKLDNVMLDSEGH IKIADF GMCKEHMM
DGVTTTRTFCTGTPDYIAPEIIAYQPYGKSVDDWAYGVLLYEMLAGQPPFDGEDEDEL FQSIMEHNVSYPKS
LSKEAVSVCKGLMTKHPAKRLGCGPEGERDVREHAFRRIDWEKLENREIQPPFKPKVCGKGAENFDKFF
TRGQPVLT PPDQLVIANIDQSDFE GFSYVNPQFVHPILQSAV
```

Here is the C2 domain of PKCA, obtained by clicking the C2 “region” link from the NCBI Protein page (http://www.ncbi.nlm.nih.gov/protein/NP_002728.1):

```
>NP_002728.2:159-289 protein kinase C alpha type [Homo sapiens]
RGR IYLKAEVADEKLHVTVRDAKNLI PMDPNGLSDPYVKLKLIPDPKNESKQKTKTIRSTLNPQWNESFT
FKLKPSDKDRRLSVEIWDWDRTTRNDFMGSLSFGVSELMKMPASGWYKLLNQEEGEYYNVP
```

By combining the RBP4 reference sequence and the C2 domain of PKCA obtained by clicking C2 “region” link. Here is a chimeric protein:

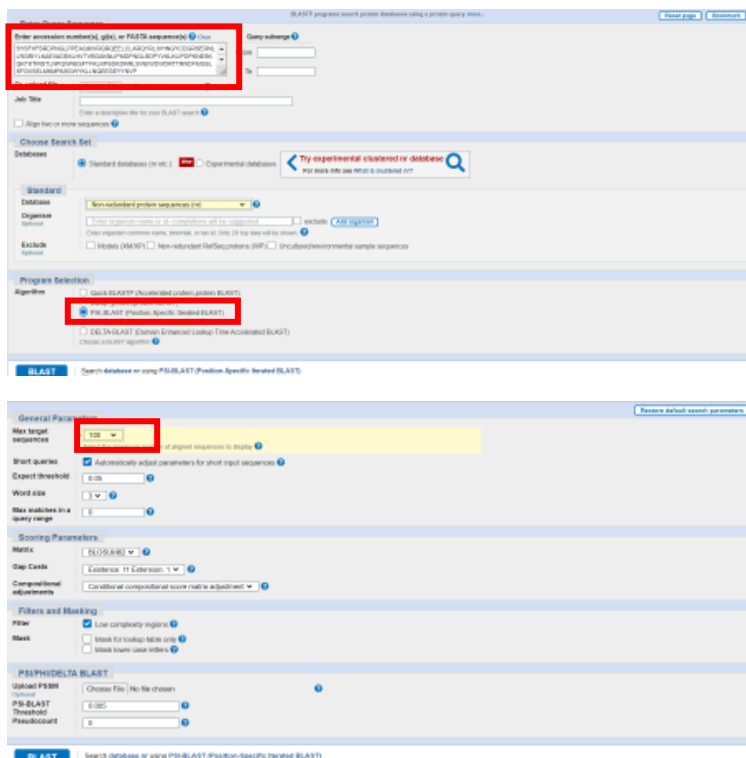
```
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLNNWDVCADMVGTFTDTEDPAKFVKMYWGVASFQKGNDDHWIVD TDYDTYAVQY
SCRL LNL DGT CADSY SFVFSRDPNGLPPEAQKI VRQRQEELCLARQYRLIVHNGYCDGRSERNLLRGR
IYLKAEVADEKLHVTVRDAKNLI PMDPNGLSDPYVKLKLIPDPKNESKQKTKTIRSTLNPQWNESFTF
KLKPSDKDRRLSVEIWDWDRTTRNDFMGSLSFGVSELMKMPASGWYKLLNQEEGEYYNVP
```



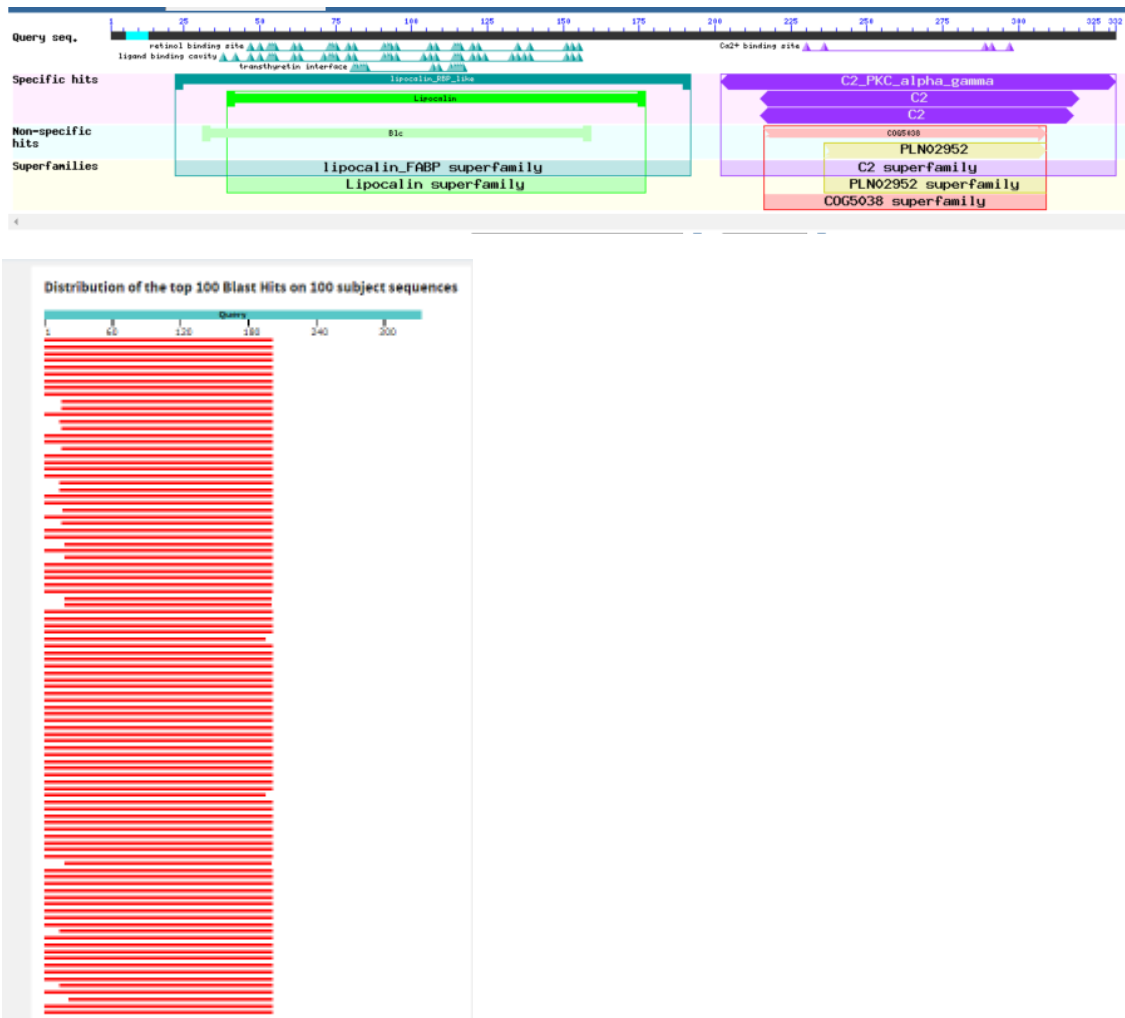
Click on the BLAST on the NCBI Home page.



Click on the Protein BLAST.



Copy the chimeric protein and paste it on the query sequence. Change the algorithm to the PSI-BLAST under program selection and change the max target sequences to 100 under general parameters. Then, click on the BLAST button.

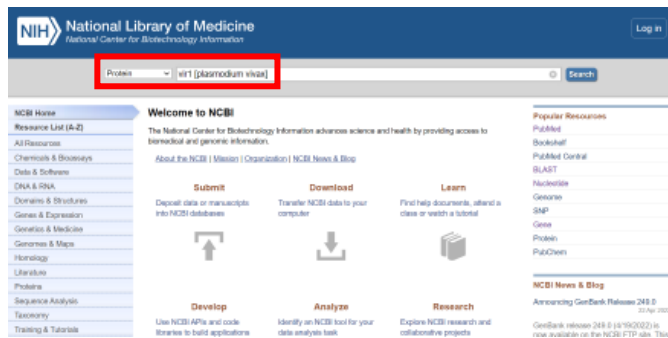


Above are the results. From the graphical summary, we can see that there are some matches to either RBP4 or PKC but not both. The RBP4 portion of the protein is larger so it accumulates higher scores (and lower E values) and those results are listed first. By inspection, there are no proteins with both domains.

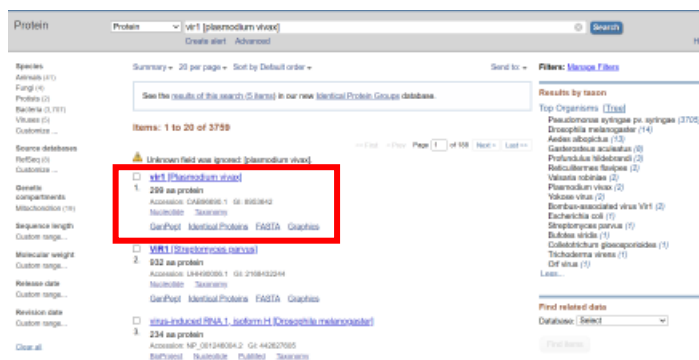
2) The purpose of this problem is to compare BLASTP to DELTA-BLAST

The malarium parasite *Plasmodium vivax* has a multigene family called *vir* that is specific to that organism (del Portillo et al., 2001). There are 600–1000 copies of these genes, and they may have a role in causing chronic infection through antigenic variation.

- i) Select *vir1* and perform a BLASTP search of the nonredundant protein database (restricting the species to *Plasmodium vivax*).
- ii) Then perform a DELTA-BLAST search with the same entry. For each search, approximately how many proteins have an E value less than 1×10^{-10} ?



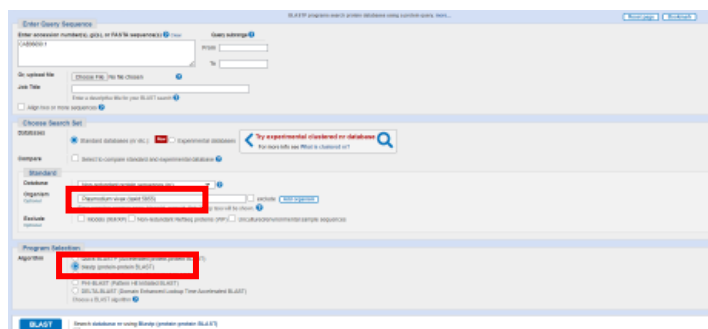
On the NCBI Home page, select protein and type vir1 [plasmodium vivax] on the search bar.



Select the first one.



Click on the Run BLAST.



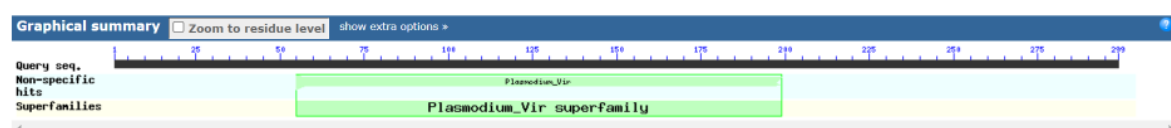
You will be redirected to this page. Type Plasmodium vivax on the organism bar and choose blastp under program selection. Click on the BLAST button and the results is shown as below.



From the results, we can see that there are many matches with significant E values.

The image shows the BLAST search interface. The 'Program Selection' section is highlighted, showing the selection of the DELTA-BLAST algorithm. The 'BLAST' button is also visible.

Click on the edit search at the top of the page and change the algorithm to DELTA-BLAST. Click on the BLAST button and the result is shown below.



3) Are there globins in fungi?

- Perform a PSI-BLAST search using human beta globin (NP_000509) as a query, restricting the output to sequences from fungi (taxid:4751) in the nr database.
- What is the approximate range of lengths of fungal proteins having globin domains? What non-globin domains are often present in fungal globins? Does the presence of these unrelated domains lead to corruption? Why or why not?
- In the first iteration there are several hits (with the E values below the 0.005 threshold). After several more iterations there are many dozens of hits including flavohemoproteins that include a globin domain. These fungal proteins have globin domains that are more related to bacterial than vertebrate orthologs. Most of the fungal flavohemoproteins and are quite long (over 400 amino acids and sometimes about 1000 amino acids long), having multiple domains. However, only the globin domain is used for the continued PSI-BLAST iterations.

Enter the NP_000509 on the query sequence and type fungi on the organism bar. Change the program selection to the PSI-BLAST.

Description	Accession	Score
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	48.9
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	47.8
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	47.8
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	47.8
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	47.8
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	47.8
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	47.8
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	47.8
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	47.8
hemoglobin subunit beta (HBB) [Homo sapiens]	NP_000509.1	47.8

The description results of the PSI-BLAST sequences with E value BETTER than threshold.

Table 1. The 100 most common words in the corpus (frequency)									
Rank	Word	Frequency	Percentage	Log-probability	Log-probability	Log-probability	Log-probability	Log-probability	Log-probability
1	the	10000	10.00%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	and	8000	8.00%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	of	7000	7.00%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	is	6000	6.00%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	in	5000	5.00%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	to	4000	4.00%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	a	3500	3.50%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	on	3000	3.00%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	for	2500	2.50%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	with	2000	2.00%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
11	at	1800	1.80%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	by	1600	1.60%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
13	as	1400	1.40%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
14	from	1200	1.20%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	but	1000	1.00%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
16	not	900	0.90%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
17	so	800	0.80%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
18	that	700	0.70%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
19	which	600	0.60%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	when	500	0.50%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	if	450	0.45%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
22	and	400	0.40%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
23	of	350	0.35%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
24	is	300	0.30%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25	in	250	0.25%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
26	to	200	0.20%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
27	a	180	0.18%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
28	on	160	0.16%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
29	for	140	0.14%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
30	with	120	0.12%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
31	at	100	0.10%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
32	by	90	0.09%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
33	as	80	0.08%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
34	from	70	0.07%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
35	but	60	0.06%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
36	not	50	0.05%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	so	45	0.045%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
38	that	40	0.04%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
39	which	35	0.035%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
40	when	30	0.03%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
41	if	25	0.025%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
42	and	20	0.02%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
43	of	18	0.018%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
44	is	16	0.016%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
45	in	14	0.014%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
46	to	12	0.012%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
47	a	10	0.01%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
48	on	9	0.009%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
49	for	8	0.008%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
50	with	7	0.007%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
51	at	6	0.006%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
52	by	5	0.005%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
53	as	4	0.004%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
54	from	3	0.003%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
55	but	2	0.002%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
56	not	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
57	so	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
58	that	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
59	which	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
60	when	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
61	if	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
62	and	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
63	of	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
64	is	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
65	in	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
66	to	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
67	a	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
68	on	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
69	for	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
70	with	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
71	at	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
72	by	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
73	as	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
74	from	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
75	but	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
76	not	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
77	so	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
78	that	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
79	which	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
80	when	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
81	if	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
82	and	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
83	of	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
84	is	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
85	in	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
86	to	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
87	a	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
88	on	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
89	for	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
90	with	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
91	at	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
92	by	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
93	as	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
94	from	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
95	but	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
96	not	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
97	so	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
98	that	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
99	which	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	when	1	0.001%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

The description results of PSI-BLAST sequences with E value WORSE than threshold.

[illegible][illegible]

To see the length of these matches, inspect the pairwise outputs. You can also reformat to view the results as a table.

Most fungal proteins have lengths of 250 to 450 amino acids. Some are more than that.

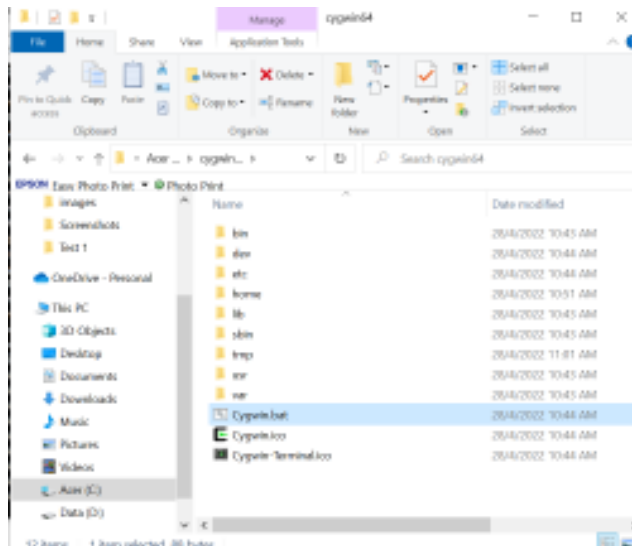
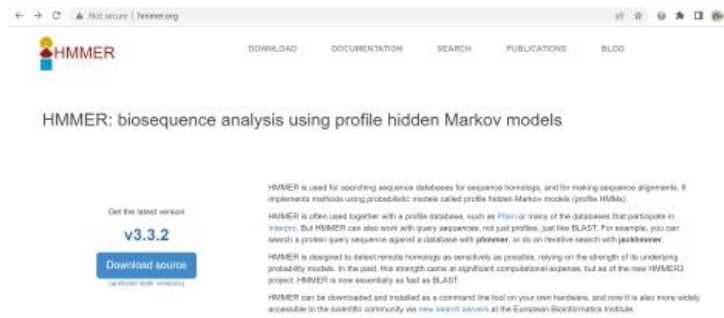
S globin are often present in fungal proteins. Presence of unrelated domains can lead to corruption which results in the high scores for homologous to the unrelated domains. This corruption can then lead high scores for the assignment and statistical significance to biological-incorrect relationship.

4) Perform HMMER searches

- i) First make two different HMMs. You can obtain sets of vertebrate globin and bacterial/fungal/vertebrate globin sequences as web documents 5.6 and 5.7. The multiple sequence alignments that we use as input to HMMER are in these documents.
- ii) When the profile HMM was built from a multiple sequence alignment of vertebrate alpha and beta globins and used to search the human RefSeq database, there were many database matches, including myoglobin that we could not detect with BLASTP. In contrast, when an alignment of bacterial and fungal globins was used to generate a profile HMM, the output consisted of one result with a non-significant expect value. Combining several human globins with the bacterial and fungal globins in a multiple sequence alignment resulted in the creation of an HMM that readily detected human globins. Thus, the profile HMM is a model that is sensitive to the choice of sequences that are used as input for the multiple sequence alignment.
- iii) The full results of the HMMER searches for (1) vertebrate, (2) bacterial plus fungal, and (3) bacterial plus fungal plus vertebrate globins are shown in web documents 5.8, 5.9, and 5.10. The HMM match to human myoglobin had a higher score and lower E value in search (3) than in (1).
- iv) HMMer searches are run locally. This search was run against all human RefSeq proteins. You can download NCBI databases such as RefSeq by visiting the file transfer protocol (FTP) site from the home page of NCBI or going directly to ► <http://www.ncbi.nlm.nih.gov/Ftp/> (WebLink 5.1). Place the downloaded database into the same directory as your input sequences for HMMER.

Follow the YouTube link, <https://www.youtube.com/watch?v=MBtbgZ7OmNM> install Cygwin and Hmmer.





Then, open Cygwin.bat to run the command.

Type:

- i. `cd C:/cygwin64`
- ii. `ls`
- iii. `cd home`
- iv. `ls`
- v. `cd user`
- vi. `ls`
- vii. `tar xvzf hmmer-3.3.2.tar.gz`
- viii. `ls`
- ix. `cd hmmer-3.3.2` (move to new directory)
- x. `./configure`
- xi. `Make`
- xii. `make check`

[illegible]

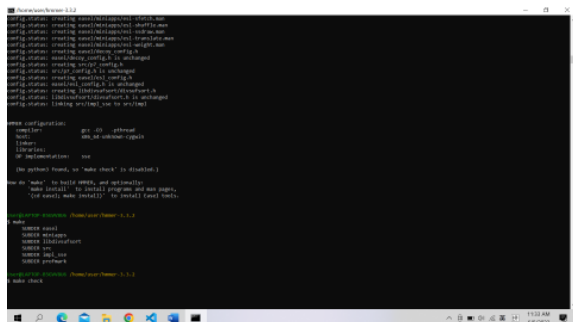
```

root@kali:~/kernsec# ./kernsec.py --help
usage: kernsec.py [-h] [-v] [-c CONFIG] [-s SCRIPTS_PATH] [-t TEST_PATH] [-m MONITOR_PATH] [-e EXECUTABLE_PATH] [-f FILTER_PATH] [-i INCLUDE_PATH] [-o OUTPUT_PATH] [-p PROFILE_PATH] [-r REPORT_PATH] [-u UPDATE_PATH] [-d DEBUG_PATH] [-l LOG_PATH] [-x XMONITOR_PATH] [-y YMONITOR_PATH] [-z ZMONITOR_PATH] [-a AMONITOR_PATH] [-b BMONITOR_PATH] [-c CONFIG] [-s SCRIPTS_PATH] [-t TEST_PATH] [-m MONITOR_PATH] [-e EXECUTABLE_PATH] [-f FILTER_PATH] [-i INCLUDE_PATH] [-o OUTPUT_PATH] [-p PROFILE_PATH] [-r REPORT_PATH] [-u UPDATE_PATH] [-d DEBUG_PATH] [-l LOG_PATH] [-x XMONITOR_PATH] [-y YMONITOR_PATH] [-z ZMONITOR_PATH] [-a AMONITOR_PATH] [-b BMONITOR_PATH]

options:
  -h, --help            show this help message and exit
  -v, --verbose          verbose mode
  -c CONFIG, --config CONFIG
                        configuration file
  -s SCRIPTS_PATH, --scripts SCRIPTS_PATH
                        scripts path
  -t TEST_PATH, --test TEST_PATH
                        test path
  -m MONITOR_PATH, --monitor MONITOR_PATH
                        monitor path
  -e EXECUTABLE_PATH, --executable EXECUTABLE_PATH
                        executable path
  -f FILTER_PATH, --filter FILTER_PATH
                        filter path
  -i INCLUDE_PATH, --include INCLUDE_PATH
                        include path
  -o OUTPUT_PATH, --output OUTPUT_PATH
                        output path
  -p PROFILE_PATH, --profile PROFILE_PATH
                        profile path
  -r REPORT_PATH, --report REPORT_PATH
                        report path
  -u UPDATE_PATH, --update UPDATE_PATH
                        update path
  -d DEBUG_PATH, --debug DEBUG_PATH
                        debug path
  -l LOG_PATH, --log LOG_PATH
                        log path
  -x XMONITOR_PATH, --xmonitor XMONITOR_PATH
                        xmonitor path
  -y YMONITOR_PATH, --ymonitor YMONITOR_PATH
                        ymonitor path
  -z ZMONITOR_PATH, --zmonitor ZMONITOR_PATH
                        zmonitor path
  -a AMONITOR_PATH, --amonitor AMONITOR_PATH
                        amonitor path
  -b BMONITOR_PATH, --bmonitor BMONITOR_PATH
                        bmonitor path

```

[illegible][illegible]



- 20

iii) What do you observe? Are there any nonviral sequences detected in the PSI-BLAST search?

iv) Did you expect to find any?

[illegible]

Go to https://www.ncbi.nlm.nih.gov/protein/NP_057849.4 and click on Run BLAST.

This output shows the lineage and E values (not shown) for different taxonomic groups:

Taxonomy Report for BLASTP:

[illegible][illegible]

Taxonomy Report for PSI-BLAST:

[illegible]

BLAST program search protein database using custom query, seq1

Enter Query Sequence

Enter accession number(s), gis, or FASTA sequence(s) 

seq1_seq1.fasta

Query settings 

Input:

To:

Do you want to:

☒ Choose the "Yes" checkbox 

seq1_seq1.fasta

☐ Align more or more sequences 

Choose Search Set

Database:  Selected databases on NCBI  Experimental database

 Try experimental database or database 

For more info see What is Custom set?

Standard

Database: 

Organism:  ☐ include

Exclude: ☐ Organism across taxa: Interact, interact-2. Only 200k seqs will be chosen

☐ repeats (200k+); ☐ Non-scientific (only policy pages) (200k); ☐ Unclassified/uncharacterized sample sequences

Program Selection:

Algorithms

- ☐ Quick BLAST (Accelerated protein-protein BLAST)
 - ☐ Search protein-protein (BLAST)
 - ☐ PIR-BLAST (Protein Specific Repeat BLAST)
 - ☐ BLAST (Protein vs. Nucleotide) (NT)
- ☒ BLAST (Standard Database Lookup (Non-Accelerated) BLAST)
 - ☐ Protein vs. NT (NT) algorithm

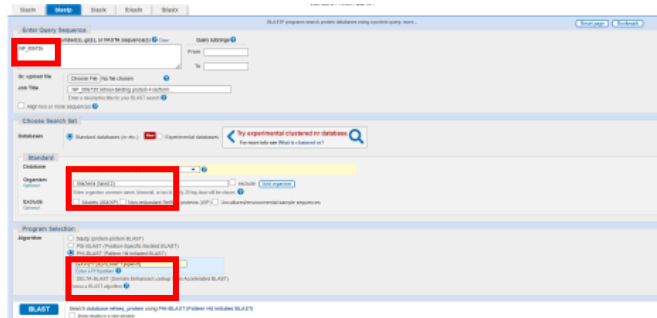
 Search database by using BLAST-BLAST (Standard Database Lookup (Non-Accelerated) BLAST)

Taxonomy Report for DELTA-BLAST:

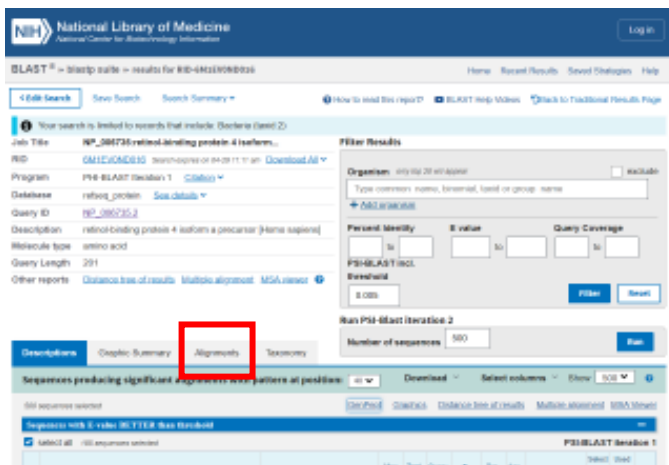
[illegible]

Based on the outputs, there are more hits with DELTA-BLAST than BLASTP and PSI-BLAST. There are both viral and non-viral detected in the PSI search. This is expected because the PSSM-based approach is far more sensitive than that of standard BLAST.

- 6) Explore PHI-BLAST using human RBP4 (NP_006735) as a query, restricting the output to bacteria and the RefSeq database.
 - i) Use the PHI pattern GXW[YF]X[VILMAFY]A[RKH]. Perform this search, and save the results.
 - ii) Then repeat the search using the PHI pattern GXW[YF][EA][IVLM].
 - iii) How do the results differ?
 - iv) Select one protein that appears as a bacterial protein in a pairwise alignment with the human RBP4 query; what are the E values, and why do they differ?



Go to NCBI website to enter the details. After entering the accession number, change the database to the reference proteins and organism to bacteria under Standard section. Then, select PHI-BLAST algorithm and paste the PHI pattern under Program Selection.



After click on the BLAST button, you will come to this page. Then, click on the alignment.

[illegible]

For every database match, the results include asterisks indicating where this pattern matches. Above are the top two results.

[illegible]

Next, click on edit search, change to the second PHI pattern which is shorter and more restrictive.

The screenshot shows the NCBI BLAST search results for the query 'NP_065735-retinol-binding protein 4 isoform...'. The top navigation bar includes links for 'BLAST', 'BLAST page - results for R06-6ML2V0X002R', 'Home', 'Recent Results', and 'Saved Searches'. Below the navigation bar, there are links for 'Edit Search', 'Save Search', 'Search Summary', 'How to read this report?', 'BLAST help pages', and 'Back to Traditional Results Page'. The main content area is divided into two columns. The left column displays the search details: Job Title, RID, Program, Database, Query ID, Description, Molecule type, Query Length, and Other reports. The right column displays the 'Filter Results' section, which includes a search bar for 'Organism', a 'Type comments' field, a 'Filter results' button, and a 'Run' button. The 'Filter Results' section also includes a 'Percent identity' and 'E value' filter, a 'Query coverage' filter, and a 'PSI-BLAST incl.' filter. The 'Run' button is highlighted in blue. The bottom of the page shows the 'Alignment view' and 'Parents' tabs, with a 'Download' button and a 'Run' button.

BLAST[®] - BLAST page - results for R06-6ML2V0X002R

Home Recent Results Saved Searches

[Edit Search](#) [Save Search](#) [Search Summary](#) [How to read this report?](#) [BLAST help pages](#) [Back to Traditional Results Page](#)

1 Your search is limited to records that include GeneID (taxid 3)

Job Title NP_065735-retinol-binding protein 4 isoform...

RID 6ML2V0X002R Report options 01/24/2011 10:59 AM [Download All](#)

Program PSI-BLAST iteration 1 [Details](#)

Database refseq_protein [See details](#)

Query ID NP_065735.2

Description retinol-binding protein 4 isoform (Homo sapiens)

Molecule type amino acid

Query Length 291

Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#)

Filter Results

Organism: only top 20 will appear ☐ include all

Type comments: name, taxonid, taxid or group name

[Add organism](#)

Percent identity: to E value: to Query coverage: to

PSI-BLAST incl.

Threshold:

Run PSI-Blast iteration 2

Number of sequences:

Descriptions [Graphic Summary](#) **[Alignments](#)** Taxonomy

Alignment view: [Parents](#) [Reset defaults](#) [Download](#)

0/0 sequences selected

[Download](#) [Reset Defaults](#) [Send](#) [Clipboard](#) [Descriptions](#)

lipocalin family protein (Lipocalin aquaporin)

Click on the alignment. The results are shown.

[Download ▾](#)
[GenBank Genbank](#)

[▼ Next ▲ Previous](#)
[→ Descriptions](#)

Apocalin family protein [Pseudoalteromonas luteoviolacea]

Sequence ID: [WP_023408096.1](#) Length: 179 Number of Matches: 1

Range 1: 33 to 148 [GenBank](#) [Download](#) [▼ Next Results ▲ Previous Results](#)

Score	Expect	Method	Identical	Positives	Score	
38.130(11.1)	4e-04	Compositional matrix adjusted	23/79(29%)	45/79(57%)	3.79(1.9)	
Query 31	100.0	100.0	100.0	100.0	100.0	
Subject 31	100.0	100.0	100.0	100.0	100.0	
Query 32	100.0	100.0	100.0	100.0	100.0	
Subject 32	100.0	100.0	100.0	100.0	100.0	

If we instead do a PHI-BLAST search without using a PHI pattern, only *Pseudoalteromonas luteoviolacea*] is significant.

Download
GenBank
GenBank
Next
Previous
Descriptions

Apocalin family protein [Leisingera aquemixtae]

Sequence ID: [WP_14188816.1](#) Length: 189 Number of Matches: 1

Range 1: 34 to 149 [GenBank](#) [Download](#) [View Next Results](#) [Previous Results](#)

Score	Expect	Method	Identical	Positives	Score	Expect
76.430(0.00)	6e-05	Simple	50/188(27%)	80/188(43%)	28.188(1.5)	
Query 1: AF040574.1 (Pseudomonas fluorescens) [100%]						
1	100.0	100.0	100.0	100.0	100.0	
Subject 1: AF040574.1 (Pseudomonas fluorescens) [100%]						
1	100.0	100.0	100.0	100.0	100.0	
Query 2: AF040574.1 (Pseudomonas fluorescens) [100%]						
2	100.0	100.0	100.0	100.0	100.0	
Subject 2: AF040574.1 (Pseudomonas fluorescens) [100%]						
2	100.0	100.0	100.0	100.0	100.0	
Query 3: AF040574.1 (Pseudomonas fluorescens) [100%]						
3	100.0	100.0	100.0	100.0	100.0	
Subject 3: AF040574.1 (Pseudomonas fluorescens) [100%]						
3	100.0	100.0	100.0	100.0	100.0	
Query 4: AF040574.1 (Pseudomonas fluorescens) [100%]						
4	100.0	100.0	100.0	100.0	100.0	
Subject 4: AF040574.1 (Pseudomonas fluorescens) [100%]						
4	100.0	100.0	100.0	100.0	100.0	
Query 5: AF040574.1 (Pseudomonas fluorescens) [100%]						
5	100.0	100.0	100.0	100.0	100.0	
Subject 5: AF040574.1 (Pseudomonas fluorescens) [100%]						
5	100.0	100.0	100.0	100.0	100.0	
Query 6: AF040574.1 (Pseudomonas fluorescens) [100%]						
6	100.0	100.0	100.0	100.0	100.0	
Subject 6: AF040574.1 (Pseudomonas fluorescens) [100%]						
6	100.0	100.0	100.0	100.0	100.0	
Query 7: AF040574.1 (Pseudomonas fluorescens) [100%]						
7	100.0	100.0	100.0	100.0	100.0	
Subject 7: AF040574.1 (Pseudomonas fluorescens) [100%]						
7	100.0	100.0	100.0	100.0	100.0	
Query 8: AF040574.1 (Pseudomonas fluorescens) [100%]						
8	100.0	100.0	100.0	100.0	100.0	
Subject 8: AF040574.1 (Pseudomonas fluorescens) [100%]						
8	100.0	100.0	100.0	100.0	100.0	
Query 9: AF040574.1 (Pseudomonas fluorescens) [100%]						
9	100.0	100.0	100.0	100.0	100.0	
Subject 9: AF040574.1 (Pseudomonas fluorescens) [100%]						
9	100.0	100.0	100.0	100.0	100.0	
Query 10: AF040574.1 (Pseudomonas fluorescens) [100%]						
10	100.0	100.0	100.0	100.0	100.0	
Subject 10: AF040574.1 (Pseudomonas fluorescens) [100%]						
10	100.0	100.0	100.0	100.0	100.0	
Query 11: AF040574.1 (Pseudomonas fluorescens) [100%]						
11	100.0	100.0	100.0	100.0	100.0	
Subject 11: AF040574.1 (Pseudomonas fluorescens) [100%]						
11	100.0	100.0	100.0	100.0	100.0	
Query 12: AF040574.1 (Pseudomonas fluorescens) [100%]						
12	100.0	100.0	100.0	100.0	100.0	
Subject 12: AF040574.1 (Pseudomonas fluorescens) [100%]						
12	100.0	100.0	100.0	100.0	100.0	
Query 13: AF040574.1 (Pseudomonas fluorescens) [100%]						
13	100.0	100.0	100.0	100.0	100.0	
Subject 13: AF040574.1 (Pseudomonas fluorescens) [100%]						
13	100.0	100.0	100.0	100.0	100.0	
Query 14: AF040574.1 (Pseudomonas fluorescens) [100%]						
14	100.0	100.0	100.0	100.0	100.0	
Subject 14: AF040574.1 (Pseudomonas fluorescens) [100%]						
14	100.0	100.0	100.0	100.0	100.0	
Query 15: AF040574.1 (Pseudomonas fluorescens) [100%]						
15	100.0	100.0	100.0	100.0	100.0	
Subject 15: AF040574.1 (Pseudomonas fluorescens) [100%]						
15	100.0	100.0	100.0	100.0	100.0	
Query 16: AF040574.1 (Pseudomonas fluorescens) [100%]						
16	100.0	100.0	100.0	100.0	100.0	
Subject 16: AF040574.1 (Pseudomonas fluorescens) [100%]						
16	100.0	100.0	100.0	100.0	100.0	
Query 17: AF040574.1 (Pseudomonas fluorescens) [100%]						
17	100.0	100.0	100.0	100.0	100.0	
Subject 17: AF040574.1 (Pseudomonas fluorescens) [100%]						
17	100.0	100.0	100.0	100.0	100.0	
Query 18: AF040574.1 (Pseudomonas fluorescens) [100%]						
18	100.0	100.0	100.0	100.0	100.0	
Subject 18: AF040574.1 (Pseudomonas fluorescens) [100%]						
18	100.0	100.0	100.0	100.0	100.0	
Query 19: AF040574.1 (Pseudomonas fluorescens) [100%]						
19	100.0	100.0	100.0	100.0	100.0	
Subject 19: AF040574.1 (Pseudomonas fluorescens) [100%]						
19	100.0	100.0	100.0	100.0	100.0	
Query 20: AF040574.1 (Pseudomonas fluorescens) [100%]						
20	100.0	100.0	100.0	100.0	100.0	
Subject 20: AF040574.1 (Pseudomonas fluorescens) [100%]						
20	100.0	100.0	100.0	100.0	100.0	
Query 21: AF040574.1 (Pseudomonas fluorescens) [100%]						
21	100.0	100.0	100.0	100.0	100.0	
Subject 21: AF040574.1 (Pseudomonas fluorescens) [100%]						
21	100.0	100.0	100.0	100.0	100.0	
Query 22: AF040574.1 (Pseudomonas fluorescens) [100%]						
22	100.0	100.0	100.0	100.0	100.0	
Subject 22: AF040574.1 (Pseudomonas fluorescens) [100%]						
22	100.0	100.0	100.0	100.0	100.0	
Query 23: AF040574.1 (Pseudomonas fluorescens) [100%]						
23	100.0	100.0	100.0	100.0	100.0	
Subject 23: AF040574.1 (Pseudomonas fluorescens) [100%]						
23	100.0	100.0	100.0	100.0	100.0	
Query 24: AF040574.1 (Pseudomonas fluorescens) [100%]						
24	100.0	100.0	100.0	100.0	100.0	
Subject 24: AF040574.1 (Pseudomonas fluorescens) [100%]						
24	100.0	100.0	100.0	100.0	100.0	
Query 25: AF040574.1 (Pseudomonas fluorescens) [100%]						
25	100.0	100.0	100.0	100.0	100.0	
Subject 25: AF040574.1 (Pseudomonas fluorescens) [100%]						
25	100.0	100.0	100.0	100.0	100.0	
Query 26: AF040574.1 (Pseudomonas fluorescens) [100%]						
26	100.0	100.0	100.0	100.0	100.0	
Subject 26: AF040574.1 (Pseudomonas fluorescens) [100%]						
26	100.0	100.0	100.0	100.0	100.0	
Query 27: AF040574.1 (Pseudomonas fluorescens) [100%]						