SESSION10. PRACTICE

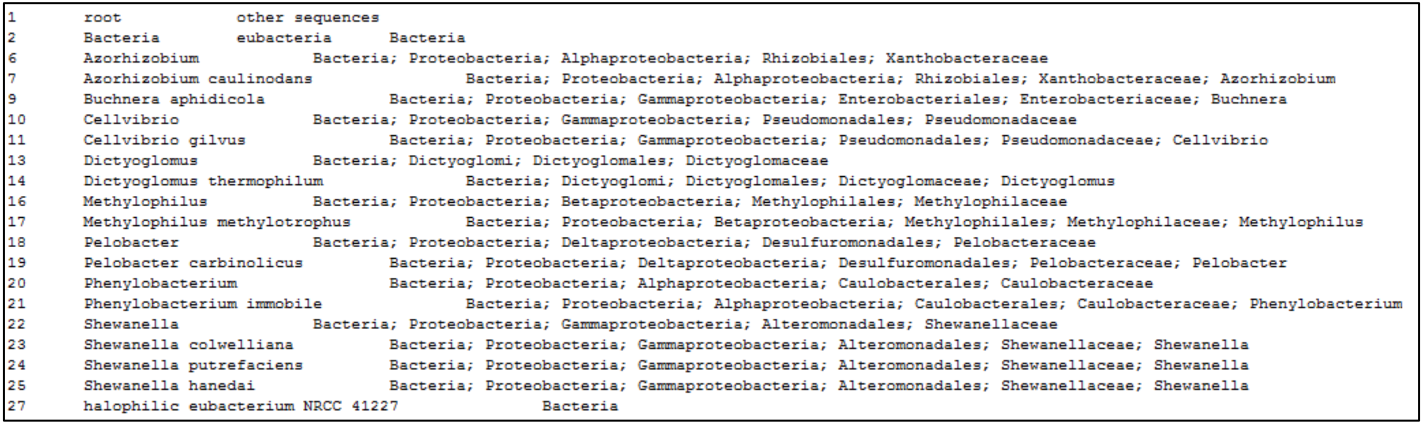
Royal blood and order In the sequence universe

Basic Shell Commands

```
$ cd [User_Folder]
$ mkdir Session10
$ cd Session10
```

Unix commands

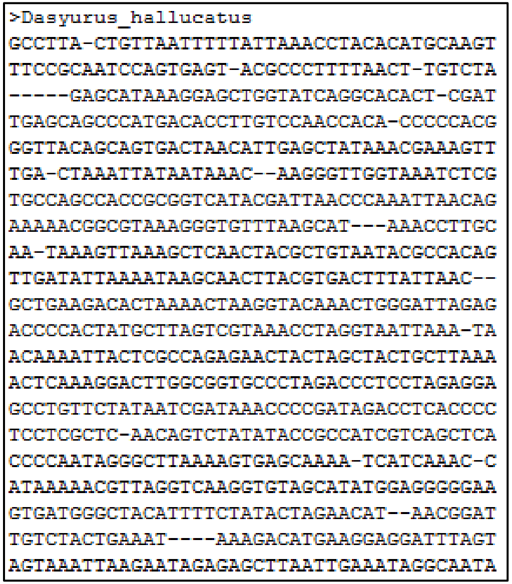
```
** 반드시 여러분의 working directory에서 아래 명령어를 수행하세요!
$ ln -s /home/biguser/tutor/Session10/tax.txt .
$ ln -s /home/biguser/tutor/Session10/mito.fa .

$ less tax.txt
```



```
1       root            other sequences
2       Bacteria        eubacteria      Bacteria
6       Azorhizobium            Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Xanthobacteraceae
7       Azorhizobium caulinodans                Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Xanthobacteraceae; Azorhizobium
9       Buchnera aphidicola             Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Buchnera
10      Cellvibrio              Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae
11      Cellvibrio gilvus               Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Cellvibrio
13      Dictyoglomus            Bacteria; Dictyoglomi; Dictyoglomales; Dictyoglomaceae
14      Dictyoglomus thermophilum               Bacteria; Dictyoglomi; Dictyoglomales; Dictyoglomaceae; Dictyoglomus
16      Methylophilus           Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae
17      Methylophilus methylotrophus            Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; Methylophilus
18      Pelobacter              Bacteria; Proteobacteria; Deltaproteobacteria; Desulfuromonadales; Pelobacteraceae
19      Pelobacter carbinolicus         Bacteria; Proteobacteria; Deltaproteobacteria; Desulfuromonadales; Pelobacteraceae; Pelobacter
20      Phenylobacterium                Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae
21      Phenylobacterium immobile               Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Phenylobacterium
22      Shewanella              Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales; Shewanellaceae
23      Shewanella colwelliana          Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales; Shewanellaceae; Shewanella
24      Shewanella putrefaciens         Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales; Shewanellaceae; Shewanella
25      Shewanella hanedai              Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales; Shewanellaceae; Shewanella
27      halophilic eubacterium NRCC 41227               Bacteria
```

Unix commands

```
$ less mito.fa
```



```
>Dasyurus_hallucatus
GCCTTA-CTGTTAATTTTTATTAAACCTACACATGCAAGT
TTCCGCAATCCAGTGAGT-ACGCCCTTTTAACT-TGTCTA
-----GAGCATAAAGGAGCTGGTATCAGGCACACT-CGAT
TGAGCAGCCCATGACACCTTGTCCAACCACA-CCCCCACG
GGTTACAGCAGTGACTAACATTGAGCTATAAACGAAAGTT
TGA-CTAAATTATAATAAAC--AAGGGTTGGTAAATCTCG
TGCCAGCCACCGCGGTCATACGATTAACCCAAATTAACAG
AAAAACGGCGTAAAGGGTGTTTAAGCAT---AAACCTTGC
AA-TAAAGTTAAAGCTCAACTACGCTGTAATACGCCACAG
TTGATATTAAAATAAGCAACTTACGTGACTTTATTAAC--
GCTGAAGACACTAAAACTAAGGTACAAACTGGGATTAGAG
ACCCCACTATGCTTAGTCGTAAACCTAGGTAATTAAA-TA
ACAAAATTACTCGCCAGAGAACTACTAGCTACTGCTTAAA
ACTCAAAGGACTTGGCGGTGCCCTAGACCCTCCTAGAGGA
GCCTGTTCTATAATCGATAAACCCCGATAGACCTCACCCC
TCCTCGCTC-AACAGTCTATATACCGCCATCGTCAGCTCA
CCCCAATAGGGCTTAAAAGTGAGCAAAA-TCATCAAAC-C
ATAAAAACGTTAGGTCAAGGTGTAGCATATGGAGGGGGAA
GTGATGGGCTACATTTTCTATACTAGAACAT--AACGGAT
TGTCTACTGAAAT----AAAGACATGAAGGAGGATTTAGT
AGTAAATTAAGAATAGAGAGCTTAATTGAAATAGGCAATA
```

• • •

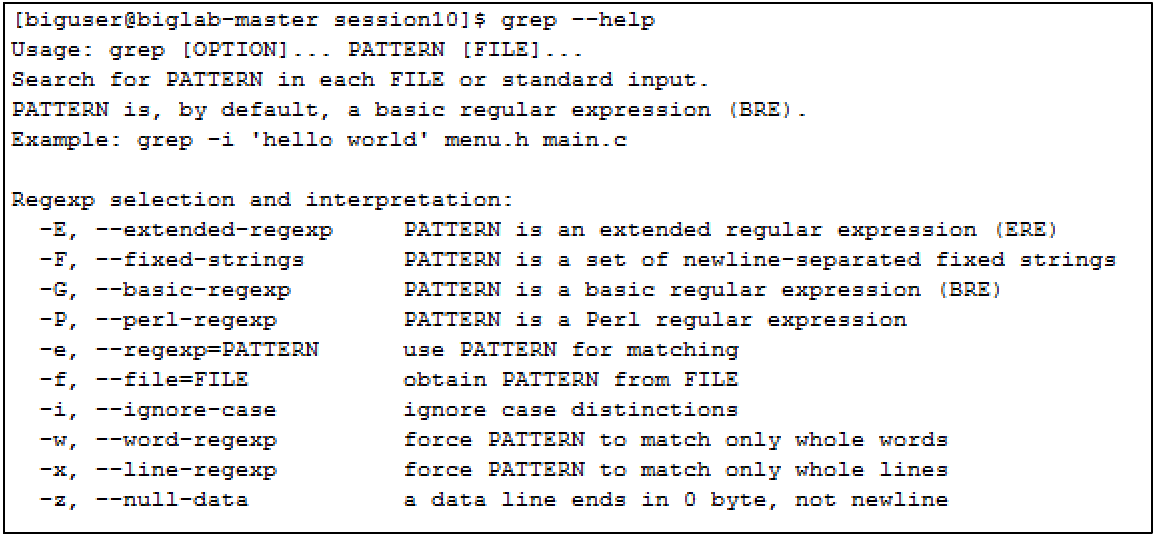
```
>Phascogale_tapoatafa
GCCTTA-CTGTTAATTTTTATTAGACCTACACATGCAAGT
TTCCGCTACCCAGTGAGT-ATGCCCTTTTAGCT-TTCTTA
-----GAGTATAAAGGAGTAGGTATCAGGCACACTTCTGT
GAAGTAGCCCATGACACCTAGTTTGACCACA-CCCCCACG
GGCTACAGCAGTGACTAACATTGAGCTATAAACGAAAGTT
TGA-CTAAATCATAATAAA---AAGGGTTGGTAAATTTCG
TGCCAGCCACCGCGGTCATACGATTAACCCGAATTAACAG
AAGAACGGCGTAAAGTGTGTTTAAGCAATAAGAATTTTCC
AAATAAGGTTAAAGATCAACTAAGCTGTAATACGCTCAGG
TTGATGTTAAAATACGCAACTTACGTGACTTTACCCCCT-
GCTGAAGACATTAAAGCTAAGGTACAAACTGGGATTAGAG
ACCCCACTATGCTTAGCCGTAAACCGAGGTAGTTATA-TA
ACAAGACTATCCGCCAGAGAACTACGAGCCACTGCTTAAA
ACTCAAAGGACTTGGCGGTGCCCTAGACCCTCCTAGAGGA
GCCTGTTCTGTAATCGATAAACCCCGATATACCTCACCTC
TCCTGGCT--GTCAGTCTATATACCGCCATCGTCAGCTCA
CCCCAATAGGGTACAAAAGTGAGCAAGA-TCATGAAAC-C
ATAAAAACGTTAGGTCAAGGTGTAGCATATGGAAAGGGAA
GTAATGGGCTACATTTTCTATATTAGAACAT--AACGGAT
```

Unix commands

```
$ cat tax.txt
$ cat mito.fa

## 아래 명령어는 실제로 실행하지는 마세요. 단순 예시입니다!
$ cat tax.txt mito.fa > taxmito.txt



*cat = concatenate
*cat file1 file2 fil3 … > merged_file
```

Unix commands

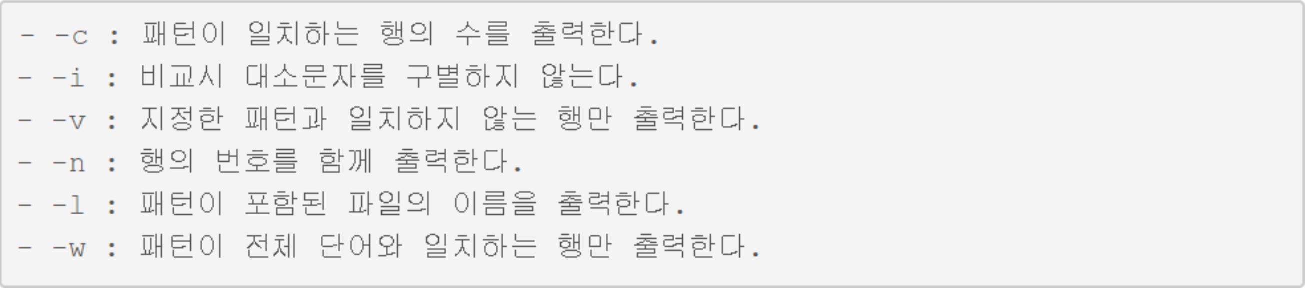
```
$ grep --help
```



```
[biguser@biglab-master session10]$ grep --help
Usage: grep [OPTION]... PATTERN [FILE]...
Search for PATTERN in each FILE or standard input.
PATTERN is, by default, a basic regular expression (BRE).
Example: grep -i 'hello world' menu.h main.c

Regexp selection and interpretation:
  -E, --extended-regexp     PATTERN is an extended regular expression (ERE)
  -F, --fixed-strings       PATTERN is a set of newline-separated fixed strings
  -G, --basic-regexp        PATTERN is a basic regular expression (BRE)
  -P, --perl-regexp         PATTERN is a Perl regular expression
  -e, --regexp=PATTERN      use PATTERN for matching
  -f, --file=FILE           obtain PATTERN from FILE
  -i, --ignore-case         ignore case distinctions
  -w, --word-regexp         force PATTERN to match only whole words
  -x, --line-regexp         force PATTERN to match only whole lines
  -z, --null-data           a data line ends in 0 byte, not newline
```

Unix commands

```
$ grep --help
```

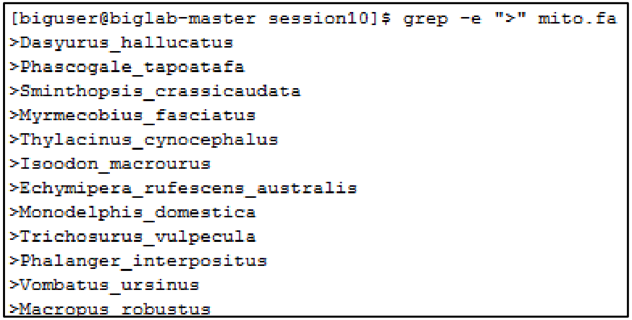


```
- -c : 패턴이 일치하는 행의 수를 출력한다.
- -i : 비교시 대소문자를 구별하지 않는다.
- -v : 지정한 패턴과 일치하지 않는 행만 출력한다.
- -n : 행의 번호를 함께 출력한다.
- -l : 패턴이 포함된 파일의 이름을 출력한다.
- -w : 패턴이 전체 단어와 일치하는 행만 출력한다.
```

http://www.incodom.kr/Linux/기본명령어/grep

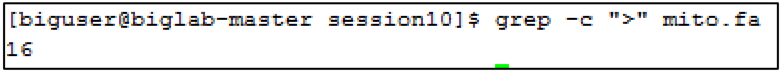
Unix commands

```
$ grep –e “>” mito.fa
##-e, --regexp=PATTERN, use PATTERN for matching
```



```
[biguser@biglab-master session10]$ grep -e ">" mito.fa
>Dasyurus_hallucatus
>Phascogale_tapoatafa
>Sminthopsis_crassicaudata
>Myrmecobius_fasciatus
>Thylacinus_cynocephalus
>Isoodon_macrourus
>Echymipera_rufescens_australis
>Monodelphis_domestica
>Trichosurus_vulpecula
>Phalanger_interpositus
>Vombatus_ursinus
>Macropus_robustus
```

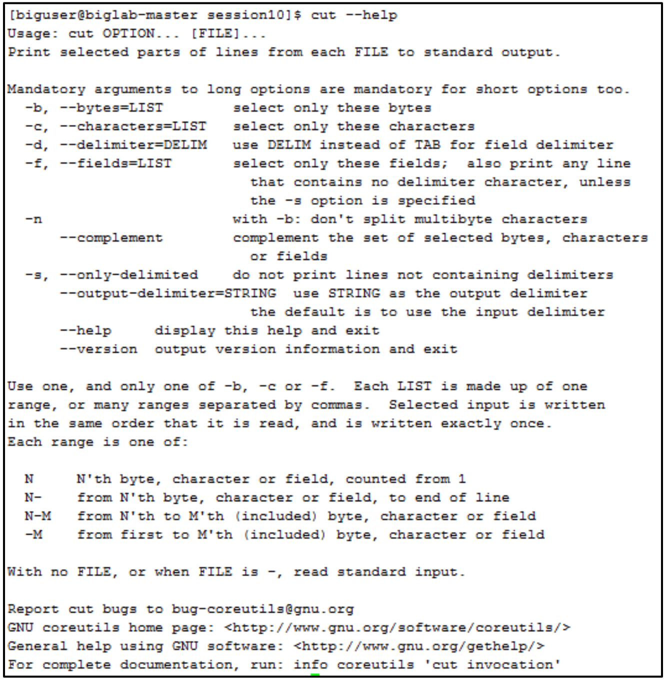
```
$ grep –c “>” mito.fa
## -c, --count, print only a count of matching lines
per FILE
```



```
[biguser@biglab-master session10]$ grep -c ">" mito.fa
16
```

Unix commands

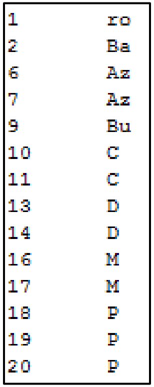
```
$ cut --help
```



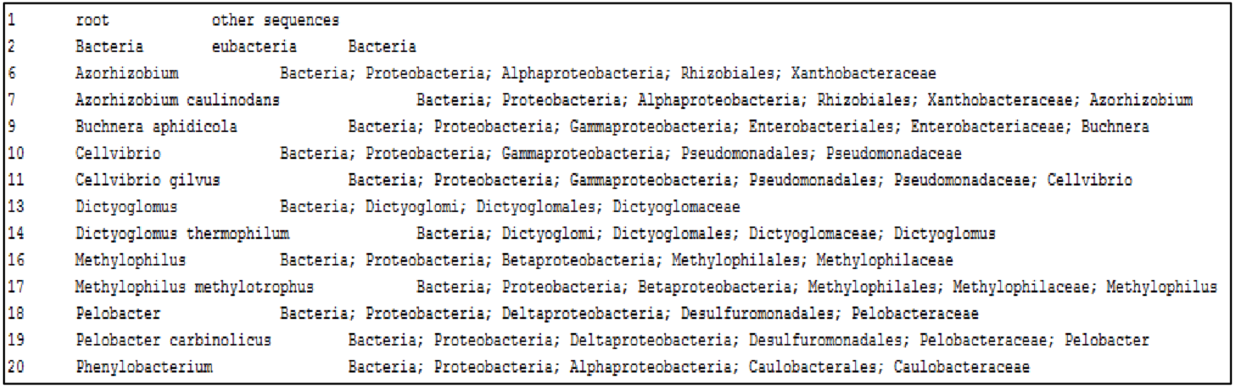
```
[biguser@biglab-master session10]$ cut --help
Usage: cut OPTION... [FILE]...
Print selected parts of lines from each FILE to standard output.

Mandatory arguments to long options are mandatory for short options too.
  -b, --bytes=LIST        select only these bytes
  -c, --characters=LIST   select only these characters
  -d, --delimiter=DELIM   use DELIM instead of TAB for field delimiter
  -f, --fields=LIST       select only these fields;  also print any line
                            that contains no delimiter character, unless
                            the -s option is specified
  -n                      with -b: don't split multibyte characters
      --complement        complement the set of selected bytes, characters
                            or fields
  -s, --only-delimited    do not print lines not containing delimiters
      --output-delimiter=STRING  use STRING as the output delimiter
                            the default is to use the input delimiter
      --help     display this help and exit
      --version  output version information and exit

Use one, and only one of -b, -c or -f.  Each LIST is made up of one
range, or many ranges separated by commas.  Selected input is written
in the same order that it is read, and is written exactly once.
Each range is one of:

  N     N'th byte, character or field, counted from 1
  N-    from N'th byte, character or field, to end of line
  N-M   from N'th to M'th (included) byte, character or field
  -M    from first to M'th (included) byte, character or field

With no FILE, or when FILE is -, read standard input.

Report cut bugs to bug-coreutils@gnu.org
GNU coreutils home page: <http://www.gnu.org/software/coreutils/>
General help using GNU software: <http://www.gnu.org/gethelp/>
For complete documentation, run: info coreutils 'cut invocation'
```

Unix commands

```
$ cut -f 1-4 tax.txt
##-f, --fields=LIST          select only these fields
```

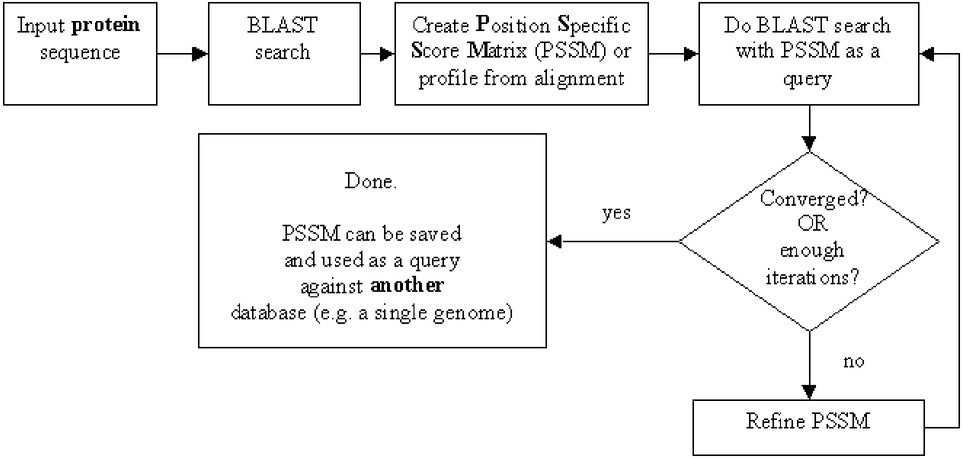


```
1       ro
2       Ba
6       Az
7       Az
9       Bu
10      C
11      C
13      D
14      D
16      M
17      M
18      P
19      P
20      P
```



```
1       root            other sequences
2       Bacteria        eubacteria      Bacteria
6       Azorhizobium            Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Xanthobacteraceae
7       Azorhizobium caulinodans                Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Xanthobacteraceae; Azorhizobium
9       Buchnera aphidicola             Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Buchnera
10      Cellvibrio              Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae
11      Cellvibrio gilvus               Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Cellvibrio
13      Dictyoglomus            Bacteria; Dictyoglomi; Dictyoglomales; Dictyoglomaceae
14      Dictyoglomus thermophilum               Bacteria; Dictyoglomi; Dictyoglomales; Dictyoglomaceae; Dictyoglomus
16      Methylophilus           Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae
17      Methylophilus methylotrophus            Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; Methylophilus
18      Pelobacter              Bacteria; Proteobacteria; Deltaproteobacteria; Desulfuromonadales; Pelobacteraceae
19      Pelobacter carbinolicus         Bacteria; Proteobacteria; Deltaproteobacteria; Desulfuromonadales; Pelobacteraceae; Pelobacter
20      Phenylobacterium                Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae
```

PSI-BLAST workflow



Input **protein** sequence

BLAST search

Create **P**osition **S**pecific **S**core **M**atrix (PSSM) or profile from alignment

Do BLAST search with PSSM as a query

Converged?
OR
enough iterations?

yes

Done.

PSSM can be saved
and used as a query
against **another**
database (e.g. a single genome)

no

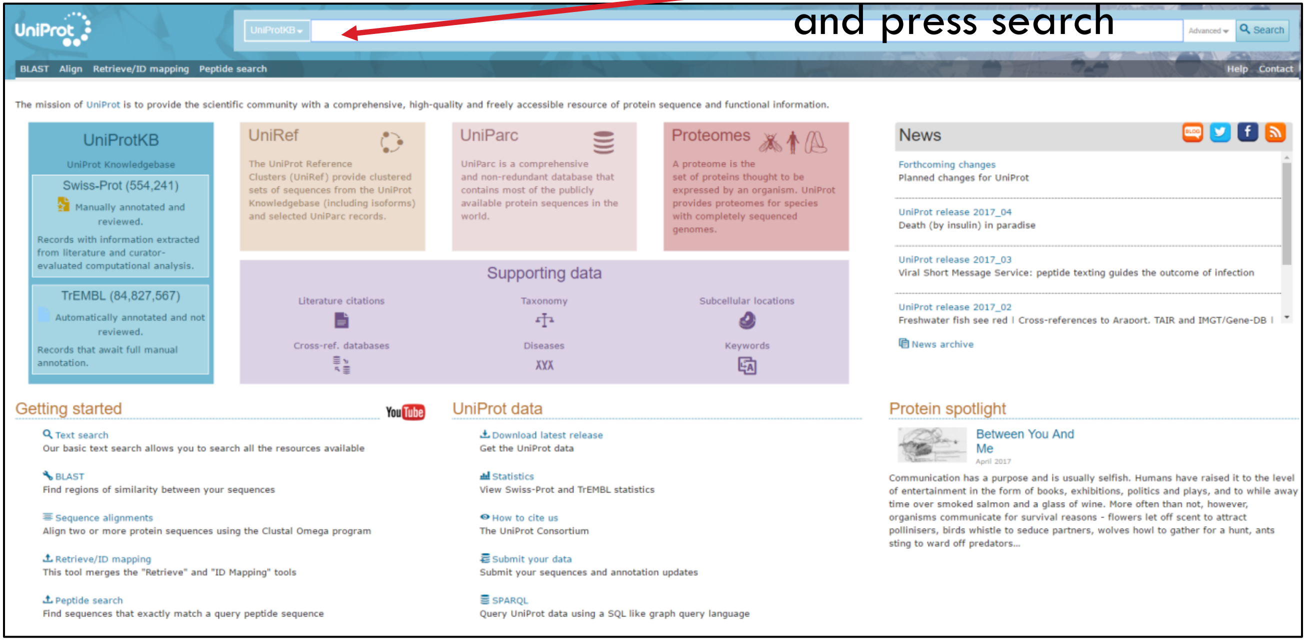
Refine PSSM

Iterate

© Olga Zhaxybayeva

Get sequence from UniProt

- Input: ABL1 (ABL1_HUMAN) – SH domain



Type “ABL1_HUMAN” and press search

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (554,241)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (84,827,567)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
Keywords

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2017_04
Death (by insulin) in paradise

UniProt release 2017_03
Viral Short Message Service: peptide texting guides the outcome of infection

UniProt release 2017_02
Freshwater fish see red | Cross-references to Araport, TAIR and IMGT/Gene-DB |

News archive

Getting started

Text search
Our basic text search allows you to search all the resources available

BLAST
Find regions of similarity between your sequences

Sequence alignments
Align two or more protein sequences using the Clustal Omega program

Retrieve/ID mapping
This tool merges the "Retrieve" and "ID Mapping" tools

Peptide search
Find sequences that exactly match a query peptide sequence

UniProt data

Download latest release
Get the UniProt data

Statistics
View Swiss-Prot and TrEMBL statistics

How to cite us
The UniProt Consortium

Submit your data
Submit your sequences and annotation updates

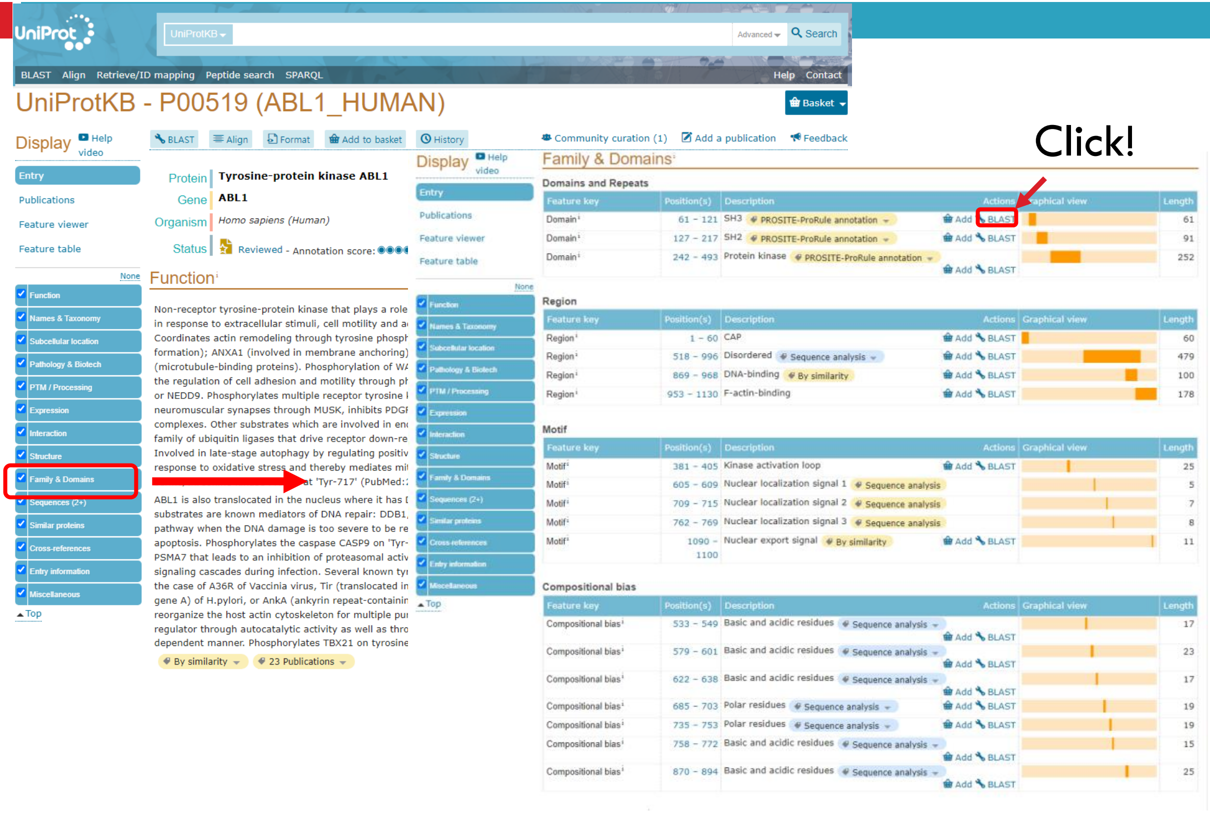
SPARQL
Query UniProt data using a SQL like graph query language

Protein spotlight

Between You And Me
April 2017

Communication has a purpose and is usually selfish. Humans have raised it to the level of entertainment in the form of books, exhibitions, politics and plays, and to while away time over smoked salmon and a glass of wine. More often than not, however, organisms communicate for survival reasons - flowers let off scent to attract pollinisers, birds whistle to seduce partners, wolves howl to gather for a hunt, ants sting to ward off predators...

Get sequence from UniProt



Click!

UniProtKB - P00519 (ABL1_HUMAN)

Protein Tyrosine-protein kinase ABL1
Gene ABL1
Organism *Homo sapiens (Human)*
Status Reviewed - Annotation score:

Function

Non-receptor tyrosine-protein kinase that plays a role
in response to extracellular stimuli, cell motility and a
Coordinates actin remodeling through tyrosine phosph
formation); ANXA1 (involved in membrane anchoring)
(microtubule-binding proteins). Phosphorylation of WA
the regulation of cell adhesion and motility through ph
or NEDD9. Phosphorylates multiple receptor tyrosine k
neuromuscular synapses through MUSK, inhibits PDGF
complexes. Other substrates which are involved in end
family of ubiquitin ligases that drive receptor down-re
Involved in late-stage autophagy by regulating positiv
response to oxidative stress and thereby mediates mi
at 'Tyr-717' (PubMed:

ABL1 is also translocated in the nucleus where it has
substrates are known mediators of DNA repair: DDB1,
pathway when the DNA damage is too severe to be re
apoptosis. Phosphorylates the caspase CASP9 on 'Tyr-
PSMA7 that leads to an inhibition of proteasomal activ
signaling cascades during infection. Several known ty
the case of A36R of Vaccinia virus, Tir (translocated in
gene A) of H.pylori, or AnkA (ankyrin repeat-containin
reorganize the host actin cytoskeleton for multiple pu
regulator through autocatalytic activity as well as thro
dependent manner. Phosphorylates TBX21 on tyrosine

By similarity 23 Publications

Family & Domains

Domains and Repeats

Feature key	Position(s)	Description	Actions	Graphical view	Length
Domain	61 – 121	SH3 PROSITE-ProRule annotation	Add BLAST		61
Domain	127 – 217	SH2 PROSITE-ProRule annotation	Add BLAST		91
Domain	242 – 493	Protein kinase PROSITE-ProRule annotation	Add BLAST		252

Region

Feature key	Position(s)	Description	Actions	Graphical view	Length
Region	1 – 60	CAP	Add BLAST		60
Region	518 – 996	Disordered Sequence analysis	Add BLAST		479
Region	869 – 968	DNA-binding By similarity	Add BLAST		100
Region	953 – 1130	F-actin-binding	Add BLAST		178

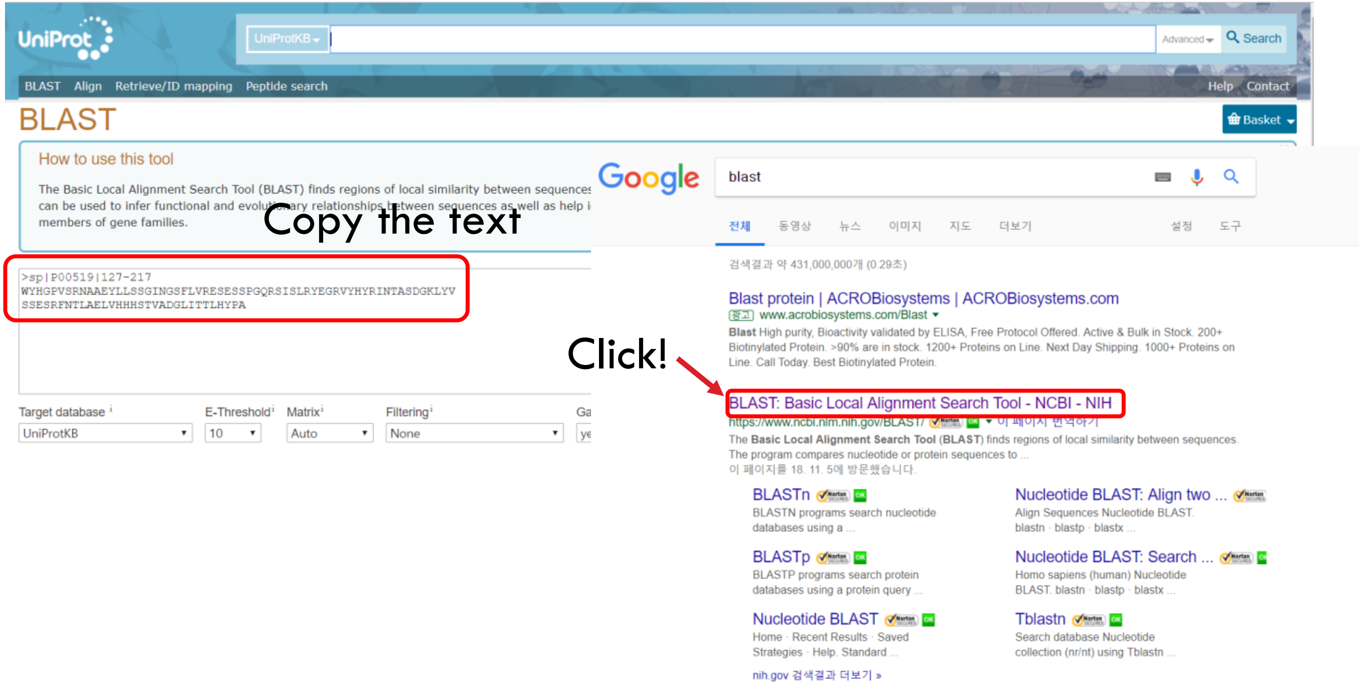
Motif

Feature key	Position(s)	Description	Actions	Graphical view	Length
Motif	381 – 405	Kinase activation loop	Add BLAST		25
Motif	605 – 609	Nuclear localization signal 1 Sequence analysis			5
Motif	709 – 715	Nuclear localization signal 2 Sequence analysis			7
Motif	762 – 769	Nuclear localization signal 3 Sequence analysis			8
Motif	1090 – 1100	Nuclear export signal By similarity	Add BLAST		11

Compositional bias

Feature key	Position(s)	Description	Actions	Graphical view	Length
Compositional bias	533 – 549	Basic and acidic residues Sequence analysis	Add BLAST		17
Compositional bias	579 – 601	Basic and acidic residues Sequence analysis	Add BLAST		23
Compositional bias	622 – 638	Basic and acidic residues Sequence analysis	Add BLAST		17
Compositional bias	685 – 703	Polar residues Sequence analysis	Add BLAST		19
Compositional bias	735 – 753	Polar residues Sequence analysis	Add BLAST		19
Compositional bias	758 – 772	Basic and acidic residues Sequence analysis	Add BLAST		15
Compositional bias	870 – 894	Basic and acidic residues Sequence analysis	Add BLAST		25

Get sequence from UniProt



BLAST

How to use this tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences can be used to infer functional and evolutionary relationships between sequences as well as help i members of gene families.

Copy the text

```
>sp|P00519|127-217
WYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISLRYEGRVYHYRINTASDGKLYV
SSESRFNTLAELVHHHSTVADGLITTLHYPA
```

Target database: UniProtKB
E-Threshold: 10
Matrix: Auto
Filtering: None
Ga: ye

blast

검색결과 약 431,000,000개 (0.29초)

Blast protein | ACROBiosystems | ACROBiosystems.com
광고 www.acrobiosystems.com/Blast
Blast High purity, Bioactivity validated by ELISA, Free Protocol Offered. Active & Bulk in Stock. 200+ Biotinylated Protein. >90% are in stock. 1200+ Proteins on Line. Next Day Shipping. 1000+ Proteins on Line. Call Today. Best Biotinylated Protein.

Click!

BLAST: Basic Local Alignment Search Tool - NCBI - NIH
https://www.ncbi.nlm.nih.gov/BLAST/ 이 페이지 번역하기
The **Basic Local Alignment Search Tool** (**BLAST**) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to ...
이 페이지를 18. 11. 5에 방문했습니다.

BLASTn
BLASTN programs search nucleotide databases using a ...

Nucleotide BLAST: Align two ...
Align Sequences Nucleotide BLAST. blastn · blastp · blastx ...

BLASTp
BLASTP programs search protein databases using a protein query ...

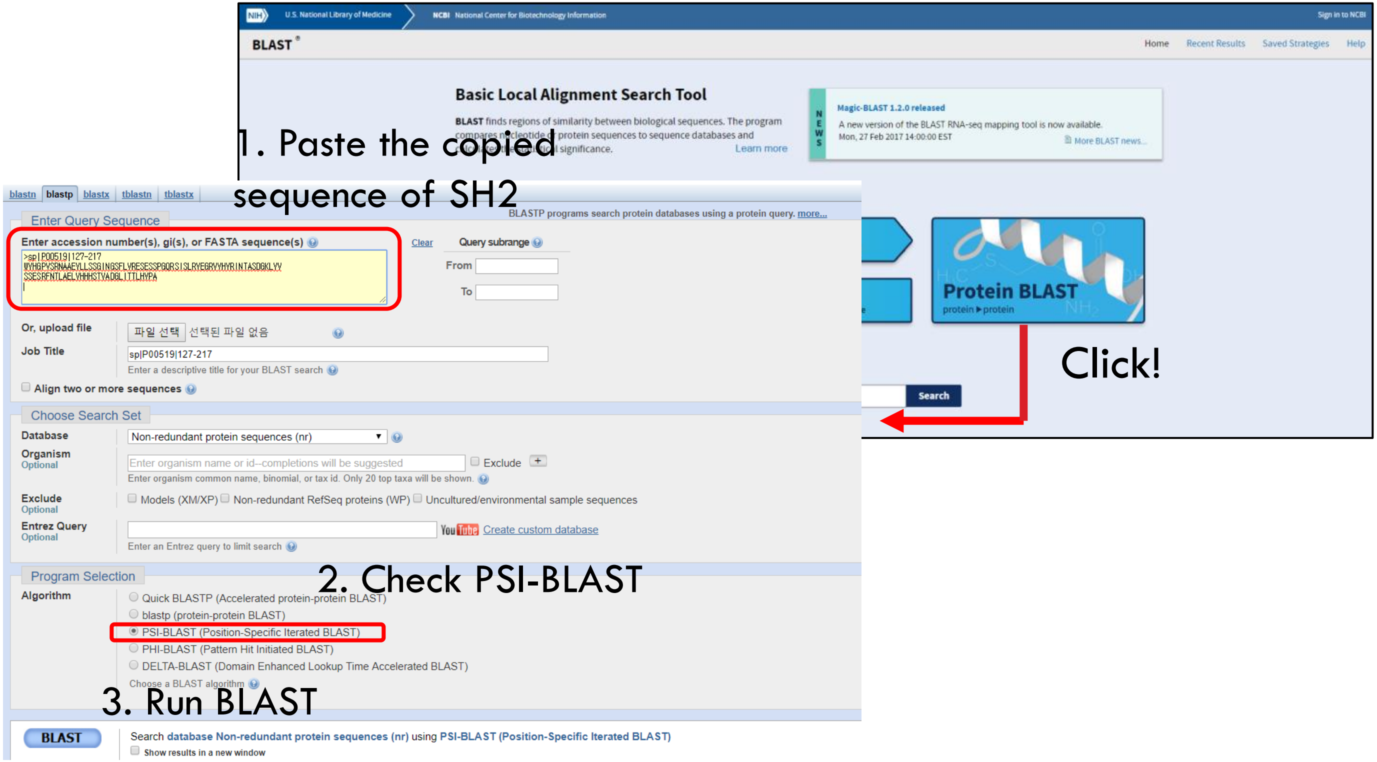
Nucleotide BLAST: Search ...
Homo sapiens (human) Nucleotide BLAST. blastn · blastp · blastx ...

Nucleotide BLAST
Home · Recent Results · Saved Strategies · Help. Standard ...

Tblastn
Search database Nucleotide collection (nr/nt) using Tblastn ...

nih.gov 검색결과 더보기 »

PSI-BLAST



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and significance. Learn more

NEWS
Magic-BLAST 1.2.0 released
A new version of the BLAST RNA-seq mapping tool is now available.
Mon, 27 Feb 2017 14:00:00 EST
More BLAST news...

1. Paste the copied sequence of SH2

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

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

```
>sp|P00519|127-217
WYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISLRYEGRVYHYRINTASDGKLYV
SSESRFNTLAELVHHHSTVADGLITTLHYPA
```

Clear

Query subrange
From
To

Or, upload file: 파일 선택 선택된 파일 없음

Job Title: sp|P00519|127-217
Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Click!

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested ☐ Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Entrez Query Optional: Create custom database
Enter an Entrez query to limit search

Program Selection

2. Check PSI-BLAST

Algorithm
- ○ Quick BLASTP (Accelerated protein-protein BLAST)
- ○ blastp (protein-protein BLAST)
- ● PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)
- ○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

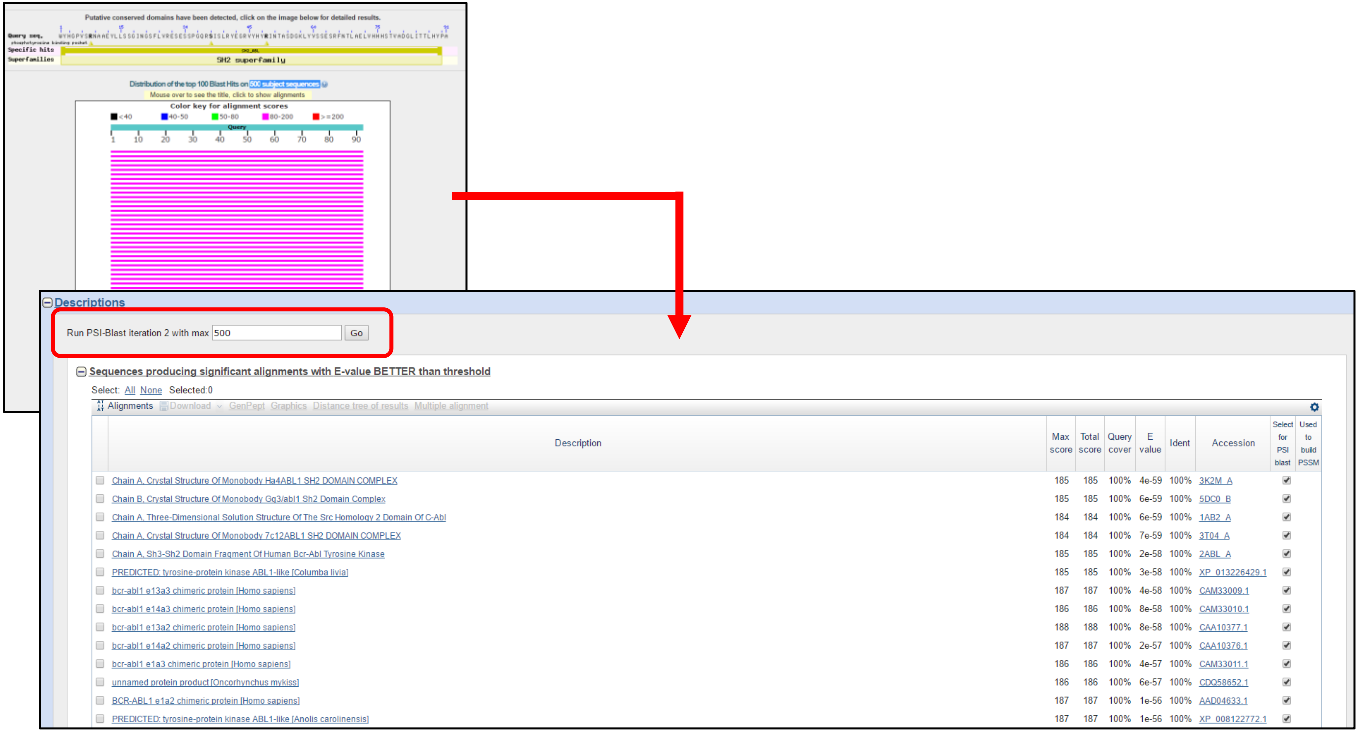
Choose a BLAST algorithm

3. Run BLAST

Search database Non-redundant protein sequences (nr) using PSI-BLAST (Position-Specific Iterated BLAST)

☐ Show results in a new window

PSI-BLAST



Putative conserved domains have been detected, click on the image below for detailed results.

Query seq. WYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISLRYEGRVYHYRINTASDGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPA

Distribution of the top 100 Blast Hits on 500 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores

Descriptions

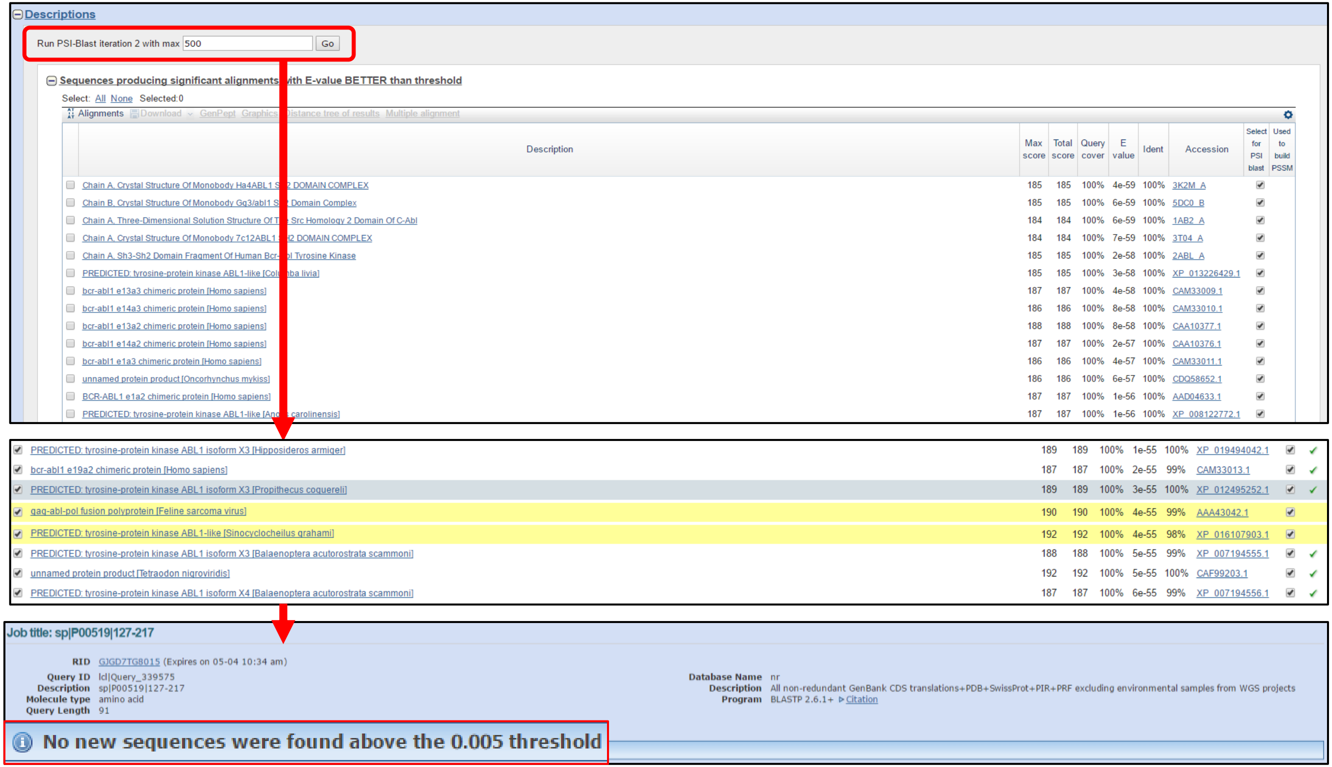
Run PSI-Blast iteration 2 with max 500

Sequences producing significant alignments with E-value BETTER than threshold

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
☐	Chain A, Crystal Structure Of Monobody Ha4ABL1 SH2 DOMAIN COMPLEX	185	185	100%	4e-59	100%	3K2M_A	☑	
☐	Chain B, Crystal Structure Of Monobody Gg3/abl1 Sh2 Domain Complex	185	185	100%	6e-59	100%	5DC0_B	☑	
☐	Chain A, Three-Dimensional Solution Structure Of The Src Homology 2 Domain Of C-Abl	184	184	100%	6e-59	100%	1AB2_A	☑	
☐	Chain A, Crystal Structure Of Monobody 7c12ABL1 SH2 DOMAIN COMPLEX	184	184	100%	7e-59	100%	3T04_A	☑	
☐	Chain A, Sh3-Sh2 Domain Fragment Of Human Bcr-Abl Tyrosine Kinase	185	185	100%	2e-58	100%	2ABL_A	☑	
☐	PREDICTED: tyrosine-protein kinase ABL1-like [Columba livia]	185	185	100%	3e-58	100%	XP_013226429.1	☑	
☐	bcr-abl1 e13a3 chimeric protein [Homo sapiens]	187	187	100%	4e-58	100%	CAM33009.1	☑	
☐	bcr-abl1 e14a3 chimeric protein [Homo sapiens]	186	186	100%	8e-58	100%	CAM33010.1	☑	
☐	bcr-abl1 e13a2 chimeric protein [Homo sapiens]	188	188	100%	8e-58	100%	CAA10377.1	☑	
☐	bcr-abl1 e14a2 chimeric protein [Homo sapiens]	187	187	100%	2e-57	100%	CAA10376.1	☑	
☐	bcr-abl1 e1a3 chimeric protein [Homo sapiens]	186	186	100%	4e-57	100%	CAM33011.1	☑	
☐	unnamed protein product [Oncorhynchus mykiss]	186	186	100%	6e-57	100%	CDQ58652.1	☑	
☐	BCR-ABL1 e1a2 chimeric protein [Homo sapiens]	187	187	100%	1e-56	100%	AAD04633.1	☑	
☐	PREDICTED: tyrosine-protein kinase ABL1-like [Anolis carolinensis]	187	187	100%	1e-56	100%	XP_008122772.1	☑	

PSI-BLAST



Run PSI-Blast iteration 2 with max 500 Go

Sequences producing significant alignments ith E-value BETTER than threshold

☐	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
☐	Chain A, Crystal Structure Of Monobody Ha4ABL1 S 2 DOMAIN COMPLEX	185	185	100%	4e-59	100%	3K2M_A	☑	
☐	Chain B, Crystal Structure Of Monobody Gg3/abl1 S 2 Domain Complex	185	185	100%	6e-59	100%	5DC0_B	☑	
☐	Chain A, Three-Dimensional Solution Structure Of T e Src Homology 2 Domain Of C-Abl	184	184	100%	6e-59	100%	1AB2_A	☑	
☐	Chain A, Crystal Structure Of Monobody 7c12ABL1 H2 DOMAIN COMPLEX	184	184	100%	7e-59	100%	3T04_A	☑	
☐	Chain A, Sh3-Sh2 Domain Fragment Of Human Bcr- bl Tyrosine Kinase	185	185	100%	2e-58	100%	2ABL_A	☑	
☐	PREDICTED: tyrosine-protein kinase ABL1-like [Col mba livia]	185	185	100%	3e-58	100%	XP_013226429.1	☑	
☐	bcr-abl1 e13a3 chimeric protein [Homo sapiens]	187	187	100%	4e-58	100%	CAM33009.1	☑	
☐	bcr-abl1 e14a3 chimeric protein [Homo sapiens]	186	186	100%	8e-58	100%	CAM33010.1	☑	
☐	bcr-abl1 e13a2 chimeric protein [Homo sapiens]	188	188	100%	8e-58	100%	CAA10377.1	☑	
☐	bcr-abl1 e14a2 chimeric protein [Homo sapiens]	187	187	100%	2e-57	100%	CAA10376.1	☑	
☐	bcr-abl1 e1a3 chimeric protein [Homo sapiens]	186	186	100%	4e-57	100%	CAM33011.1	☑	
☐	unnamed protein product [Oncorhynchus mykiss]	186	186	100%	6e-57	100%	CDQ58652.1	☑	
☐	BCR-ABL1 e1a2 chimeric protein [Homo sapiens]	187	187	100%	1e-56	100%	AAD04633.1	☑	
☐	PREDICTED: tyrosine-protein kinase ABL1-like [An carolinensis]	187	187	100%	1e-56	100%	XP_008122772.1	☑	
☑	PREDICTED: tyrosine-protein kinase ABL1 isoform X3 [Hipposideros armiger]	189	189	100%	1e-55	100%	XP_019494042.1	☑	✓
☑	bcr-abl1 e19a2 chimeric protein [Homo sapiens]	187	187	100%	2e-55	99%	CAM33013.1	☑	✓
☑	PREDICTED: tyrosine-protein kinase ABL1 isoform X3 [Propithecus coquereli]	189	189	100%	3e-55	100%	XP_012495252.1	☑	✓
☑	gag-abl-pol fusion polyprotein [Feline sarcoma virus]	190	190	100%	4e-55	99%	AAA43042.1	☑	
☑	PREDICTED: tyrosine-protein kinase ABL1-like [Sinocyclocheilus grahami]	192	192	100%	4e-55	98%	XP_016107903.1	☑	
☑	PREDICTED: tyrosine-protein kinase ABL1 isoform X3 [Balaenoptera acutorostrata scammoni]	188	188	100%	5e-55	99%	XP_007194555.1	☑	✓
☑	unnamed protein product [Tetraodon nigroviridis]	192	192	100%	5e-55	100%	CAF99203.1	☑	✓
☑	PREDICTED: tyrosine-protein kinase ABL1 isoform X4 [Balaenoptera acutorostrata scammoni]	187	187	100%	6e-55	99%	XP_007194556.1	☑	✓

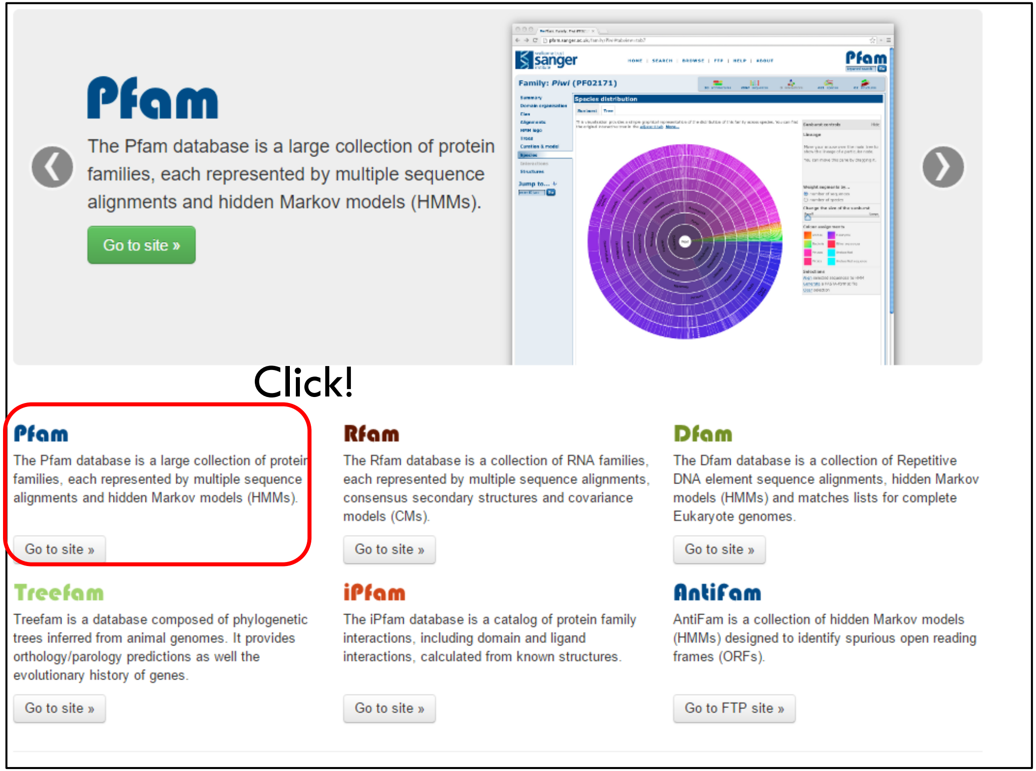
Job title: sp|P00519|127-217

RID GJGD7TG8015 (Expires on 05-04 10:34 am)
Query ID lcl|Query_339575
Description sp|P00519|127-217
Molecule type amino acid
Query Length 91

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.6.1+ ▷ Citation

No new sequences were found above the 0.005 threshold

Data download (xfam.org)



Pfam

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).

Go to site »

Click!

Pfam

The Pfam database is a large collection of protei families, each represented by multiple sequence alignments and hidden Markov models (HMMs).

Go to site »

Rfam

The Rfam database is a collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models (CMs).

Go to site »

Dfam

The Dfam database is a collection of Repetitive DNA element sequence alignments, hidden Markov models (HMMs) and matches lists for complete Eukaryote genomes.

Go to site »

Treefam

Treefam is a database composed of phylogenetic trees inferred from animal genomes. It provides orthology/parology predictions as well the evolutionary history of genes.

Go to site »

iPfam

The iPfam database is a catalog of protein family interactions, including domain and ligand interactions, calculated from known structures.

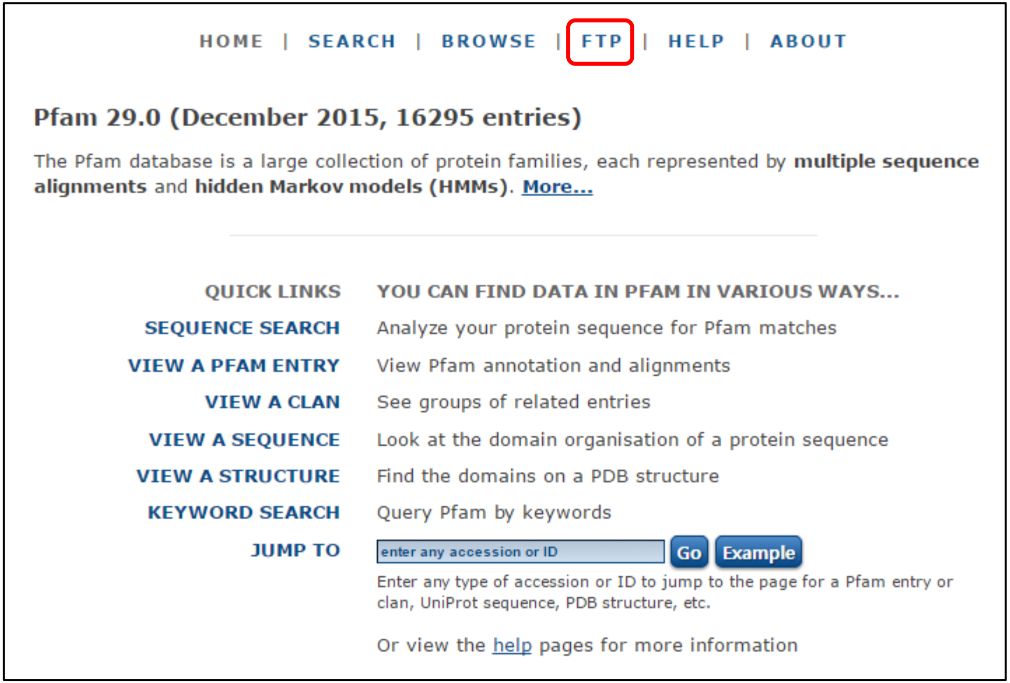
Go to site »

AntiFam

AntiFam is a collection of hidden Markov models (HMMs) designed to identify spurious open reading frames (ORFs).

Go to FTP site »

Data download (xfam.org)



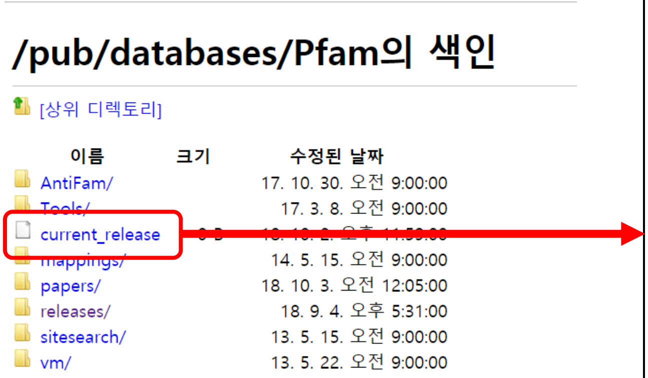
Pfam 29.0 (December 2015, 16295 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/

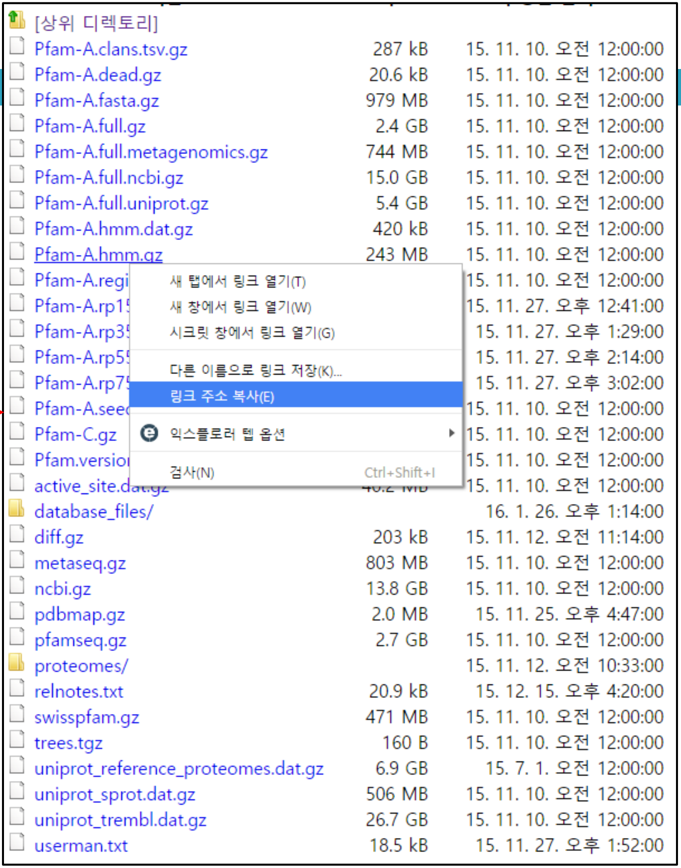
Data download



/pub/databases/Pfam의 색인

[상위 디렉토리]

이름	크기	수정된 날짜
AntiFam/		17. 10. 30. 오전 9:00:00
Tools/		17. 3. 8. 오전 9:00:00
current_release	0 B	18. 10. 2. 오후 11:59:00
mappings/		14. 5. 15. 오전 9:00:00
papers/		18. 10. 3. 오전 12:05:00
releases/		18. 9. 4. 오후 5:31:00
sitesearch/		13. 5. 15. 오전 9:00:00
vm/		13. 5. 22. 오전 9:00:00



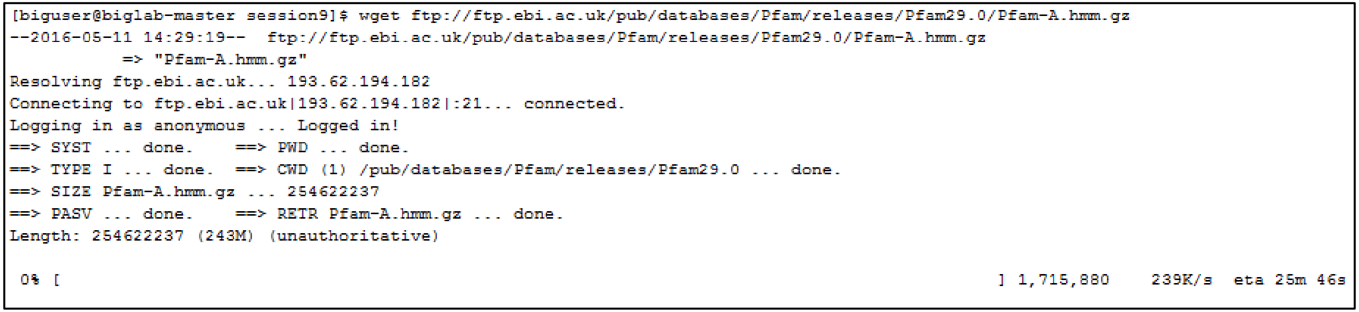
[상위 디렉토리]

Pfam-A.clans.tsv.gz	287 kB	15. 11. 10. 오전 12:00:00
Pfam-A.dead.gz	20.6 kB	15. 11. 10. 오전 12:00:00
Pfam-A.fasta.gz	979 MB	15. 11. 10. 오전 12:00:00
Pfam-A.full.gz	2.4 GB	15. 11. 10. 오전 12:00:00
Pfam-A.full.metagenomics.gz	744 MB	15. 11. 10. 오전 12:00:00
Pfam-A.full.ncbi.gz	15.0 GB	15. 11. 10. 오전 12:00:00
Pfam-A.full.uniprot.gz	5.4 GB	15. 11. 10. 오전 12:00:00
Pfam-A.hmm.dat.gz	420 kB	15. 11. 10. 오전 12:00:00
Pfam-A.hmm.gz	243 MB	15. 11. 10. 오전 12:00:00
Pfam-A.regi		15. 11. 10. 오전 12:00:00
Pfam-A.rp1		15. 11. 27. 오후 12:41:00
Pfam-A.rp3		15. 11. 27. 오후 1:29:00
Pfam-A.rp5		15. 11. 27. 오후 2:14:00
Pfam-A.rp7		15. 11. 27. 오후 3:02:00
Pfam-A.see		15. 11. 10. 오전 12:00:00
Pfam-C.gz		15. 11. 10. 오전 12:00:00
Pfam.versio		15. 11. 10. 오전 12:00:00
active_site.dat.gz	40.2 MB	15. 11. 10. 오전 12:00:00
database_files/		16. 1. 26. 오후 1:14:00
diff.gz	203 kB	15. 11. 12. 오전 11:14:00
metaseq.gz	803 MB	15. 11. 10. 오전 12:00:00
ncbi.gz	13.8 GB	15. 11. 10. 오전 12:00:00
pdbmap.gz	2.0 MB	15. 11. 25. 오후 4:47:00
pfamseq.gz	2.7 GB	15. 11. 10. 오전 12:00:00
proteomes/		15. 11. 12. 오전 10:33:00
relnotes.txt	20.9 kB	15. 12. 15. 오후 4:20:00
swisspfam.gz	471 MB	15. 11. 10. 오전 12:00:00
trees.tgz	160 B	15. 11. 10. 오전 12:00:00
uniprot_reference_proteomes.dat.gz	6.9 GB	15. 7. 1. 오전 12:00:00
uniprot_sprot.dat.gz	506 MB	15. 11. 10. 오전 12:00:00
uniprot_trembl.dat.gz	26.7 GB	15. 11. 10. 오전 12:00:00
userman.txt	18.5 kB	15. 11. 27. 오후 1:52:00

Data download

```
$ wget [copy the URL link]
```

→ 실제로 실행하지는 말고 아래 link 명령어를 사용하세요

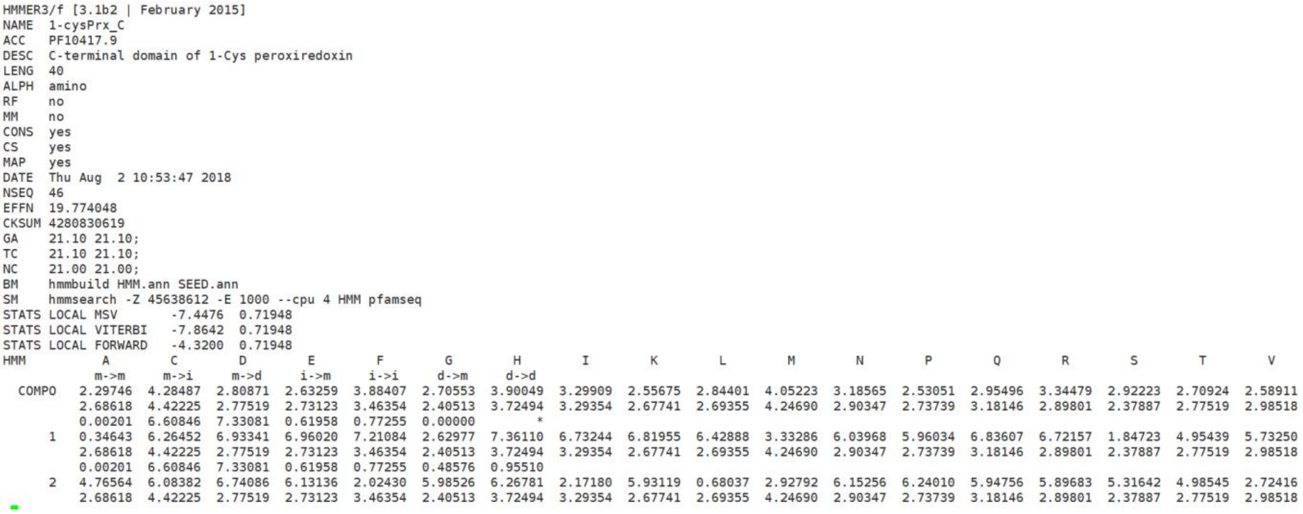


```
[biguser@biglab-master session9]$ wget ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam29.0/Pfam-A.hmm.gz
--2016-05-11 14:29:19--  ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam29.0/Pfam-A.hmm.gz
           => "Pfam-A.hmm.gz"
Resolving ftp.ebi.ac.uk... 193.62.194.182
Connecting to ftp.ebi.ac.uk|193.62.194.182|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /pub/databases/Pfam/releases/Pfam29.0 ... done.
==> SIZE Pfam-A.hmm.gz ... 254622237
==> PASV ... done.    ==> RETR Pfam-A.hmm.gz ... done.
Length: 254622237 (243M) (unauthoritative)

 0% [                                                                                                ] 1,715,880    239K/s  eta 25m 46s
```

```
$ ln -s /home/biguser/tutor/Session10/Pfam-A.hmm .
```

Build Pfam-A

```
$ less Pfam-A.hmm
```

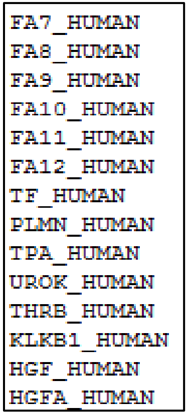


```
HMMER3/f [3.1b2 | February 2015]
NAME  1-cysPrx_C
ACC   PF10417.9
DESC  C-terminal domain of 1-Cys peroxiredoxin
LENG  40
ALPH  amino
RF    no
MM    no
CONS  yes
CS    yes
MAP   yes
DATE  Thu Aug  2 10:53:47 2018
NSEQ  46
EFFN  19.774048
CKSUM 4280830619
GA    21.10 21.10;
TC    21.10 21.10;
NC    21.00 21.00;
BM    hmmbuild HMM.ann SEED.ann
SM    hmmsearch -Z 45638612 -E 1000 --cpu 4 HMM pfamseq
STATS LOCAL MSV       -7.4476  0.71948
STATS LOCAL VITERBI   -7.8642  0.71948
STATS LOCAL FORWARD   -4.3200  0.71948
HMM          A        C        D        E        F        G        H        I        K        L        M        N        P        Q        R        S        T        V
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.29746  4.28487  2.80871  2.63259  3.88407  2.70553  3.90049  3.29909  2.55675  2.84401  4.05223  3.18565  2.53051  2.95496  3.34479  2.92223  2.70924  2.58911
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518
          0.00201  6.60846  7.33081  0.61958  0.77255  0.00000        *
      1   0.34643  6.26452  6.93341  6.96020  7.21084  2.62977  7.36110  6.73244  6.81955  6.42888  3.33286  6.03968  5.96034  6.83607  6.72157  1.84723  4.95439  5.73250
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518
          0.00201  6.60846  7.33081  0.61958  0.77255  0.48576  0.95510
      2   4.76564  6.08382  6.74086  6.13136  2.02430  5.98526  6.26781  2.17180  5.93119  0.68037  2.92792  6.15256  6.24010  5.94756  5.89683  5.31642  4.98545  2.72416
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518
```

Get sequence from swissprot

```
$ cp /home/biguser/tutor/Session10/bc_seqid.txt .
```

```
$ less bc_seqid.txt
```



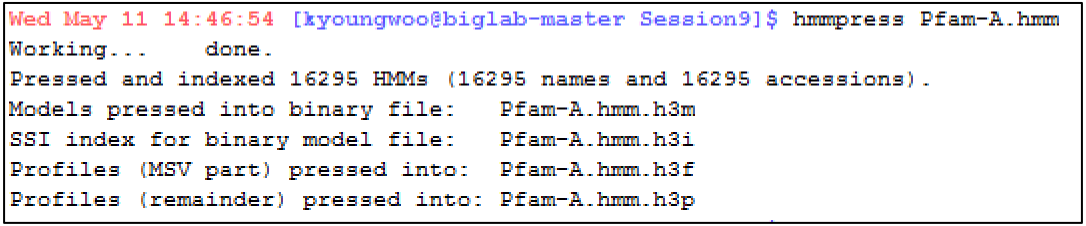
```
FA7_HUMAN
FA8_HUMAN
FA9_HUMAN
FA10_HUMAN
FA11_HUMAN
FA12_HUMAN
TF_HUMAN
PLMN_HUMAN
TPA_HUMAN
UROK_HUMAN
THRB_HUMAN
KLKB1_HUMAN
HGF_HUMAN
HGFA_HUMAN
```

```
$ ln -s /home/biguser/tutor/Session6/swissprot* .

$ blastdbcmd -entry_batch bc_seqid.txt -db swissprot -long_seqids > clotting.fa
```

hmmscan from Pfam-A.hmm

HMM profile indexing

```
$ hmmpress Pfam-A.hmm
```



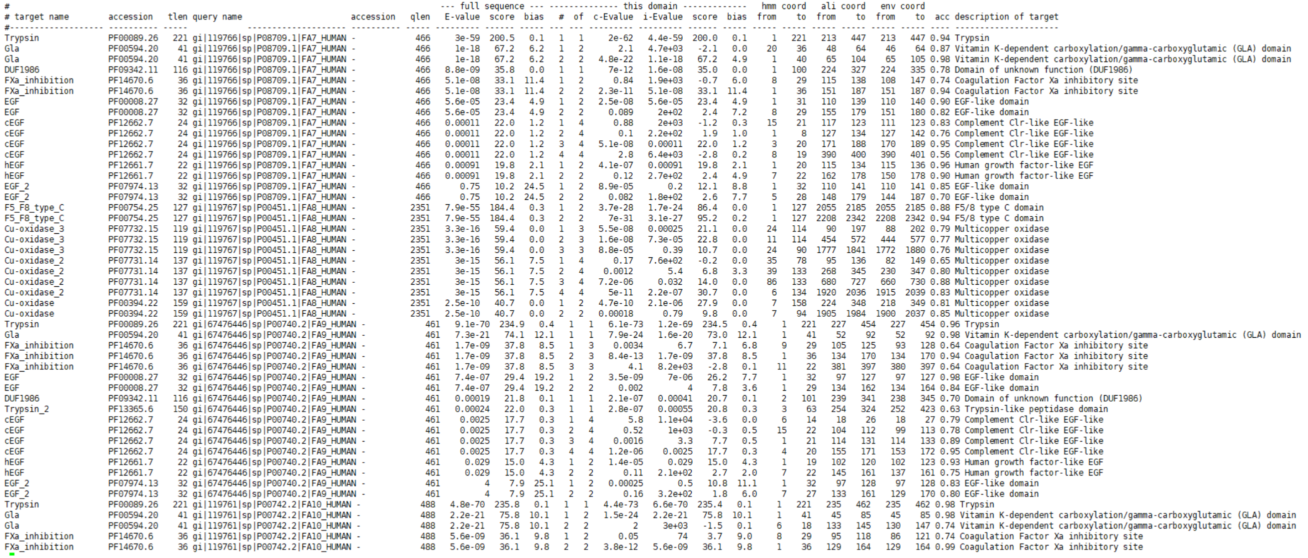
```
Wed May 11 14:46:54 [kyoungwoo@biglab-master Session9]$ hmmpress Pfam-A.hmm
Working...    done.
Pressed and indexed 16295 HMMs (16295 names and 16295 accessions).
Models pressed into binary file:   Pfam-A.hmm.h3m
SSI index for binary model file:   Pfam-A.hmm.h3i
Profiles (MSV part) pressed into:  Pfam-A.hmm.h3f
Profiles (remainder) pressed into: Pfam-A.hmm.h3p
```

Searching HMM profile with a query sequence

```
$ hmmscan --domtblout clotting.tab Pfam-A.hmm clotting.fa
```

Result of hmmscan

```
$ less clotting.tab
```



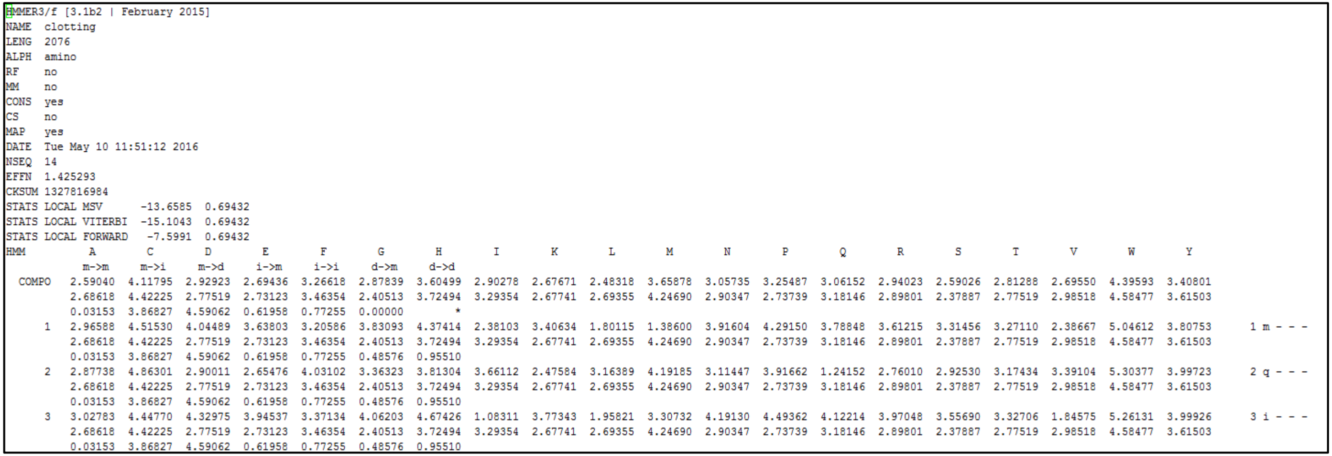
```
#                                                                                      --- full sequence --- -------------- this domain -------------   hmm coord   ali coord   env coord
# target name        accession   tlen query name                      accession   qlen   E-value  score  bias   #  of  c-Evalue  i-Evalue  score  bias  from    to  from    to  from    to  acc description of target
#------------------- ---------- -----            -------------------- ---------- ----- --------- ------ ----- --- --- --------- --------- ------ ----- ----- ----- ----- ----- ----- ----- ---- ---------------------
Trypsin              PF00089.26   221 gi|119766|sp|P08709.1|FA7_HUMAN -            466     3e-59  200.5   0.1   1   1     2e-62   4.4e-59  200.0   0.1     1   221   213   447   213   447 0.94 Trypsin
Gla                  PF00594.20    41 gi|119766|sp|P08709.1|FA7_HUMAN -            466     1e-18   67.2   6.2   1   2       2.1   4.7e+03   -2.1   0.0    20    36    48    64    46    64 0.87 Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain
Gla                  PF00594.20    41 gi|119766|sp|P08709.1|FA7_HUMAN -            466     1e-18   67.2   6.2   2   2   4.8e-22   1.1e-18   67.2   4.9     1    40    65   104    65   105 0.98 Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain
DUF1986              PF09342.11   116 gi|119766|sp|P08709.1|FA7_HUMAN -            466   8.8e-09   35.8   0.0   1   1     7e-12   1.6e-08   35.0   0.0     1   100   224   327   224   335 0.78 Domain of unknown function (DUF1986)
FXa_inhibition       PF14670.6     36 gi|119766|sp|P08709.1|FA7_HUMAN -            466   5.1e-08   33.1  11.4   1   2      0.84   1.9e+03   -0.7   6.0     8    29   115   138   108   147 0.74 Coagulation Factor Xa inhibitory site
FXa_inhibition       PF14670.6     36 gi|119766|sp|P08709.1|FA7_HUMAN -            466   5.1e-08   33.1  11.4   2   2   2.3e-11   5.1e-08   33.1  11.4     1    36   151   187   151   187 0.94 Coagulation Factor Xa inhibitory site
EGF                  PF00008.27    32 gi|119766|sp|P08709.1|FA7_HUMAN -            466   5.6e-05   23.4   4.9   1   2   2.5e-08   5.6e-05   23.4   4.9     1    31   110   139   110   140 0.90 EGF-like domain
EGF                  PF00008.27    32 gi|119766|sp|P08709.1|FA7_HUMAN -            466   5.6e-05   23.4   4.9   2   2     0.089     2e+02    2.4   7.2     8    29   155   179   151   180 0.82 EGF-like domain
cEGF                 PF12662.7     24 gi|119766|sp|P08709.1|FA7_HUMAN -            466   0.00011   22.0   1.2   1   4      0.88     2e+03   -1.2   0.3    15    21   117   123   111   123 0.83 Complement Clr-like EGF-like
cEGF                 PF12662.7     24 gi|119766|sp|P08709.1|FA7_HUMAN -            466   0.00011   22.0   1.2   2   4       0.1   2.2e+02    1.9   1.0     1     8   127   134   127   142 0.76 Complement Clr-like EGF-like
cEGF                 PF12662.7     24 gi|119766|sp|P08709.1|FA7_HUMAN -            466   0.00011   22.0   1.2   3   4   5.1e-08   0.00011   22.0   1.2     3    20   171   188   170   189 0.95 Complement Clr-like EGF-like
cEGF                 PF12662.7     24 gi|119766|sp|P08709.1|FA7_HUMAN -            466   0.00011   22.0   1.2   4   4       2.8   6.4e+03   -2.8   0.2     8    19   390   400   390   401 0.56 Complement Clr-like EGF-like
hEGF                 PF12661.7     22 gi|119766|sp|P08709.1|FA7_HUMAN -            466   0.00091   19.8   2.1   1   2   4.1e-07   0.00091   19.8   2.1     1    20   115   134   115   136 0.96 Human growth factor-like EGF
hEGF                 PF12661.7     22 gi|119766|sp|P08709.1|FA7_HUMAN -            466   0.00091   19.8   2.1   2   2      0.12   2.7e+02    2.4   4.9     7    22   162   178   150   178 0.90 Human growth factor-like EGF
EGF_2                PF07974.13    32 gi|119766|sp|P08709.1|FA7_HUMAN -            466      0.75   10.2  24.5   1   2   8.9e-05       0.2   12.1   8.8     1    32   110   141   110   141 0.85 EGF-like domain
EGF_2                PF07974.13    32 gi|119766|sp|P08709.1|FA7_HUMAN -            466      0.75   10.2  24.5   2   2     0.082   1.8e+02    2.6   7.7     5    28   148   179   144   187 0.70 EGF-like domain
F5_F8_type_C         PF00754.25   127 gi|119767|sp|P00451.1|FA8_HUMAN -           2351   7.9e-55  184.4   0.3   1   2   3.7e-28   1.7e-24   86.4   0.0     1   127  2055  2185  2055  2185 0.88 F5/8 type C domain
F5_F8_type_C         PF00754.25   127 gi|119767|sp|P00451.1|FA8_HUMAN -           2351   7.9e-55  184.4   0.3   2   2     7e-31   3.1e-27   95.2   0.2     1   127  2208  2342  2208  2342 0.94 F5/8 type C domain
Cu-oxidase_3         PF07732.15   119 gi|119767|sp|P00451.1|FA8_HUMAN -           2351   3.3e-16   59.4   0.0   1   3   5.5e-08   0.00025   21.1   0.0    24   114    90   197    88   202 0.79 Multicopper oxidase
Cu-oxidase_3         PF07732.15   119 gi|119767|sp|P00451.1|FA8_HUMAN -           2351   3.3e-16   59.4   0.0   2   3   1.6e-08   7.3e-05   22.8   0.0    11   114   454   572   444   577 0.77 Multicopper oxidase
Cu-oxidase_3         PF07732.15   119 gi|119767|sp|P00451.1|FA8_HUMAN -           2351   3.3e-16   59.4   0.0   3   3   8.8e-05      0.39   10.7   0.0    24    90  1777  1841  1772  1880 0.76 Multicopper oxidase
Cu-oxidase_2         PF07731.14   137 gi|119767|sp|P00451.1|FA8_HUMAN -           2351     3e-15   56.1   7.5   1   4      0.17   7.6e+02   -0.2   0.0    35    78    95   136    82   149 0.65 Multicopper oxidase
Cu-oxidase_2         PF07731.14   137 gi|119767|sp|P00451.1|FA8_HUMAN -           2351     3e-15   56.1   7.5   2   4    0.0012       5.4    6.8   3.3    39   133   268   345   230   347 0.80 Multicopper oxidase
Cu-oxidase_2         PF07731.14   137 gi|119767|sp|P00451.1|FA8_HUMAN -           2351     3e-15   56.1   7.5   3   4   7.2e-06     0.032   14.0   0.0    86   133   680   727   660   730 0.88 Multicopper oxidase
Cu-oxidase_2         PF07731.14   137 gi|119767|sp|P00451.1|FA8_HUMAN -           2351     3e-15   56.1   7.5   4   4     5e-11   2.2e-07   30.7   0.0     6   134  1920  2036  1915  2039 0.83 Multicopper oxidase
Cu-oxidase           PF00394.22   159 gi|119767|sp|P00451.1|FA8_HUMAN -           2351   2.5e-10   40.7   0.0   1   2   4.7e-10   2.1e-06   27.9   0.0     7   158   224   348   218   349 0.81 Multicopper oxidase
Cu-oxidase           PF00394.22   159 gi|119767|sp|P00451.1|FA8_HUMAN -           2351   2.5e-10   40.7   0.0   2   2   0.00018      0.79    9.8   0.0     7    94  1905  1984  1900  2037 0.85 Multicopper oxidase
Trypsin              PF00089.26   221 gi|67476446|sp|P00740.2|FA9_HUMAN -            461   9.1e-70  234.9   0.4   1   1   6.1e-73   1.2e-69  234.5   0.4     1   221   227   454   227   454 0.96 Trypsin
Gla                  PF00594.20    41 gi|67476446|sp|P00740.2|FA9_HUMAN -            461   7.3e-21   74.1  12.1   1   1   7.9e-24   1.6e-20   73.0  12.1     1    41    52    92    52    92 0.98 Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain
FXa_inhibition       PF14670.6     36 gi|67476446|sp|P00740.2|FA9_HUMAN -            461   1.7e-09   37.8   8.5   1   3    0.0034       6.7    7.1   6.8     9    29   105   125    93   128 0.64 Coagulation Factor Xa inhibitory site
FXa_inhibition       PF14670.6     36 gi|67476446|sp|P00740.2|FA9_HUMAN -            461   1.7e-09   37.8   8.5   2   3   8.4e-13   1.7e-09   37.8   8.5     1    36   134   170   134   170 0.94 Coagulation Factor Xa inhibitory site
FXa_inhibition       PF14670.6     36 gi|67476446|sp|P00740.2|FA9_HUMAN -            461   1.7e-09   37.8   8.5   3   3       4.1   8.2e+03   -2.8   0.1    11    22   381   397   380   397 0.64 Coagulation Factor Xa inhibitory site
EGF                  PF00008.27    32 gi|67476446|sp|P00740.2|FA9_HUMAN -            461   7.4e-07   29.4  19.2   1   2   3.5e-09     7e-06   26.2   7.7     1    32    97   127    97   127 0.98 EGF-like domain
EGF                  PF00008.27    32 gi|67476446|sp|P00740.2|FA9_HUMAN -            461   7.4e-07   29.4  19.2   2   2     0.002         4    7.8   3.6     1    29   134   162   134   164 0.84 EGF-like domain
DUF1986              PF09342.11   116 gi|67476446|sp|P00740.2|FA9_HUMAN -            461   0.00019   21.8   0.1   1   1   2.1e-07   0.00041   20.7   0.1     2   101   239   341   238   345 0.70 Domain of unknown function (DUF1986)
Trypsin_2            PF13365.6    150 gi|67476446|sp|P00740.2|FA9_HUMAN -            461   0.00024   22.0   0.3   1   1   2.8e-07   0.00055   20.8   0.3     3    63   254   324   252   423 0.63 Trypsin-like peptidase domain
cEGF                 PF12662.7     24 gi|67476446|sp|P00740.2|FA9_HUMAN -            461    0.0025   17.7   0.3   1   4       5.8   1.1e+04   -3.6   0.0     6    14    18    26    18    27 0.79 Complement Clr-like EGF-like
cEGF                 PF12662.7     24 gi|67476446|sp|P00740.2|FA9_HUMAN -            461    0.0025   17.7   0.3   2   4      0.52     1e+03   -0.3   0.5    15    22   104   112    99   113 0.78 Complement Clr-like EGF-like
cEGF                 PF12662.7     24 gi|67476446|sp|P00740.2|FA9_HUMAN -            461    0.0025   17.7   0.3   3   4    0.0016       3.3    7.7   0.5     1    21   114   131   114   133 0.89 Complement Clr-like EGF-like
cEGF                 PF12662.7     24 gi|67476446|sp|P00740.2|FA9_HUMAN -            461    0.0025   17.7   0.3   4   4   1.2e-06    0.0025   17.7   0.3     4    20   155   171   153   172 0.95 Complement Clr-like EGF-like
hEGF                 PF12661.7     22 gi|67476446|sp|P00740.2|FA9_HUMAN -            461     0.029   15.0   4.3   1   2   1.4e-05     0.029   15.0   4.3     1    19   102   120   102   123 0.93 Human growth factor-like EGF
hEGF                 PF12661.7     22 gi|67476446|sp|P00740.2|FA9_HUMAN -            461     0.029   15.0   4.3   2   2      0.11   2.1e+02    2.7   2.0     7    22   145   161   137   161 0.75 Human growth factor-like EGF
EGF_2                PF07974.13    32 gi|67476446|sp|P00740.2|FA9_HUMAN -            461         4    7.9  25.1   1   2   0.00025       0.5   10.8  11.1     1    32    97   128    97   128 0.83 EGF-like domain
EGF_2                PF07974.13    32 gi|67476446|sp|P00740.2|FA9_HUMAN -            461         4    7.9  25.1   2   2      0.16   3.2e+02    1.8   6.0     7    27   133   161   129   170 0.80 EGF-like domain
Trypsin              PF00089.26   221 gi|119761|sp|P00742.2|FA10_HUMAN -            488   4.8e-70  235.8   0.1   1   1   4.4e-73   6.6e-70  235.4   0.1     1   221   235   462   235   462 0.98 Trypsin
Gla                  PF00594.20    41 gi|119761|sp|P00742.2|FA10_HUMAN -            488   2.2e-21   75.8  10.1   1   2   1.5e-24   2.2e-21   75.8  10.1     1    41    45    85    45    85 0.98 Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain
Gla                  PF00594.20    41 gi|119761|sp|P00742.2|FA10_HUMAN -            488   2.2e-21   75.8  10.1   2   2         2     3e+03   -1.5   0.1     6    18   133   145   130   147 0.74 Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain
FXa_inhibition       PF14670.6     36 gi|119761|sp|P00742.2|FA10_HUMAN -            488   5.6e-09   36.1   9.8   1   2      0.05        74    3.7   9.0     8    29    95   118    86   121 0.74 Coagulation Factor Xa inhibitory site
FXa_inhibition       PF14670.6     36 gi|119761|sp|P00742.2|FA10_HUMAN -            488   5.6e-09   36.1   9.8   2   2   3.8e-12   5.6e-09   36.1   9.8     1    36   129   164   129   164 0.99 Coagulation Factor Xa inhibitory site
```

Searching with custome HMM profiles

Building HMM profile

```
clustalw2 clotting.fa
hmmbuild clotting.hmm clotting.aln
```

`clotting.hmm`

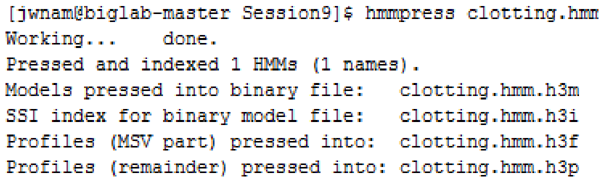


```
HMMER3/f [3.1b2 | February 2015]
NAME  clotting
LENG  2076
ALPH  amino
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Tue May 10 11:51:12 2016
NSEQ  14
EFFN  1.425293
CKSUM 1327816984
STATS LOCAL MSV      -13.6585  0.69432
STATS LOCAL VITERBI  -15.1043  0.69432
STATS LOCAL FORWARD   -7.5991  0.69432
HMM          A        C        D        E        F        G        H        I        K        L        M        N        P        Q        R        S        T        V        W        Y   
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.59040  4.11795  2.92923  2.69436  3.26618  2.87839  3.60499  2.90278  2.67671  2.48318  3.65878  3.05735  3.25487  3.06152  2.94023  2.59026  2.81288  2.69550  4.39593  3.40801
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03153  3.86827  4.59062  0.61958  0.77255  0.00000        *
      1   2.96588  4.51530  4.04489  3.63803  3.20586  3.83093  4.37414  2.38103  3.40634  1.80115  1.38600  3.91604  4.29150  3.78848  3.61215  3.31456  3.27110  2.38667  5.04612  3.80753      1 m - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03153  3.86827  4.59062  0.61958  0.77255  0.48576  0.95510
      2   2.87738  4.86301  2.90011  2.65476  4.03102  3.36323  3.81304  3.66112  2.47584  3.16389  4.19185  3.11447  3.91662  1.24152  2.76010  2.92530  3.17434  3.39104  5.30377  3.99723      2 q - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03153  3.86827  4.59062  0.61958  0.77255  0.48576  0.95510
      3   3.02783  4.44770  4.32975  3.94537  3.37134  4.06203  4.67426  1.08311  3.77343  1.95821  3.30732  4.19130  4.49362  4.12214  3.97048  3.55690  3.32706  1.84575  5.26131  3.99926      3 i - - -
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.03153  3.86827  4.59062  0.61958  0.77255  0.48576  0.95510
```

Searching with custome HMM profiles

HMM profile indexing

```
hmmpress clotting.hmm
```

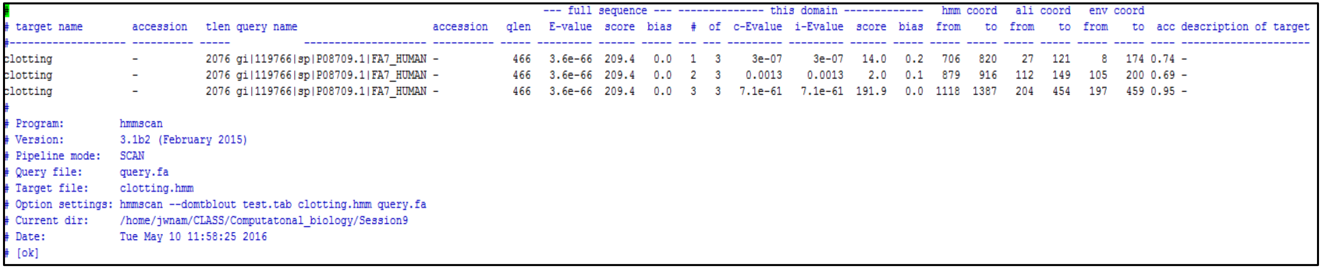


```
[jwnam@biglab-master Session9]$ hmmpress clotting.hmm
Working...    done.
Pressed and indexed 1 HMMs (1 names).
Models pressed into binary file:   clotting.hmm.h3m
SSI index for binary model file:   clotting.hmm.h3i
Profiles (MSV part) pressed into:  clotting.hmm.h3f
Profiles (remainder) pressed into: clotting.hmm.h3p
```

Searching HMM profile with a query sequence

```
hmmscan --domtblout query.tab clotting.hmm query.fa
```

- query.tab

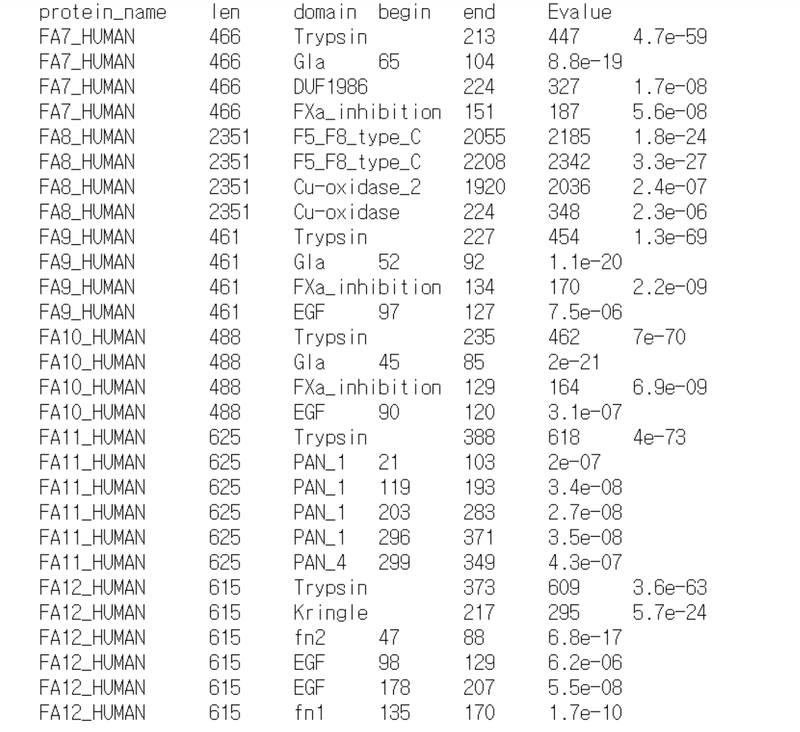


```
#                                                                            --- full sequence --- -------------- this domain -------------   hmm coord   ali coord   env coord
# target name        accession   tlen query name           accession   qlen   E-value  score  bias   #  of  c-Evalue  i-Evalue  score  bias  from    to  from    to  from    to  acc description of target
#------------------- ---------- -----          -------------------- ---------- ----- --------- ------ ----- --- --- --------- --------- ------ ----- ----- ----- ----- ----- ----- ----- ---- ---------------------
clotting             -           2076 gi|119766|sp|P08709.1|FA7_HUMAN -            466   3.6e-66  209.4   0.0   1   3     3e-07     3e-07   14.0   0.2   706   820    27   121     8   174 0.74 -
clotting             -           2076 gi|119766|sp|P08709.1|FA7_HUMAN -            466   3.6e-66  209.4   0.0   2   3    0.0013    0.0013    2.0   0.1   879   916   112   149   105   200 0.69 -
clotting             -           2076 gi|119766|sp|P08709.1|FA7_HUMAN -            466   3.6e-66  209.4   0.0   3   3   7.1e-61   7.1e-61  191.9   0.0  1118  1387   204   454   197   459 0.95 -
#
# Program:         hmmscan
# Version:         3.1b2 (February 2015)
# Pipeline mode:   SCAN
# Query file:      query.fa
# Target file:     clotting.hmm
# Option settings: hmmscan --domtblout test.tab clotting.hmm query.fa
# Current dir:     /home/jwnam/CLASS/Computatonal_biology/Session9
# Date:            Tue May 10 11:58:25 2016
# [ok]
```

Code 1 parse_hmmscan.ipynb

```
import re
import sys

print 'protein_name₩tlen₩tdomain₩tbegin₩tend₩tEvalue'

for line in open("clotting.tab", 'r'):
    if not re.search('^₩#', line): # avoid all lines beginning
                      # with the '#' character
        col = re.split(' +', line)
        domname = col[0]
        protname = col[3]
        protname = re.sub('.*₩|', '', protname)
        length = col[5]
        evalue = float(col[12])
        begin = col[17]
        end = col[18]
        if evalue < 1e-5:
            print protname, '₩t', length, '₩t', domname,
            print '₩t', begin, '₩t', end, '₩t', str(evalue)
            pass
    else:
        pass
```

Code 1 parse_hmmscan.ipynb



```
protein_name    len     domain  begin   end     Evalue
FA7_HUMAN       466     Trypsin         213     447     4.7e-59
FA7_HUMAN       466     Gla     65      104     8.8e-19
FA7_HUMAN       466     DUF1986         224     327     1.7e-08
FA7_HUMAN       466     FXa_inhibition  151     187     5.6e-08
FA8_HUMAN       2351    F5_F8_type_C    2055    2185    1.8e-24
FA8_HUMAN       2351    F5_F8_type_C    2208    2342    3.3e-27
FA8_HUMAN       2351    Cu-oxidase_2    1920    2036    2.4e-07
FA8_HUMAN       2351    Cu-oxidase      224     348     2.3e-06
FA9_HUMAN       461     Trypsin         227     454     1.3e-69
FA9_HUMAN       461     Gla     52      92      1.1e-20
FA9_HUMAN       461     FXa_inhibition  134     170     2.2e-09
FA9_HUMAN       461     EGF     97      127     7.5e-06
FA10_HUMAN      488     Trypsin         235     462     7e-70
FA10_HUMAN      488     Gla     45      85      2e-21
FA10_HUMAN      488     FXa_inhibition  129     164     6.9e-09
FA10_HUMAN      488     EGF     90      120     3.1e-07
FA11_HUMAN      625     Trypsin         388     618     4e-73
FA11_HUMAN      625     PAN_1   21      103     2e-07
FA11_HUMAN      625     PAN_1   119     193     3.4e-08
FA11_HUMAN      625     PAN_1   203     283     2.7e-08
FA11_HUMAN      625     PAN_1   296     371     3.5e-08
FA11_HUMAN      625     PAN_4   299     349     4.3e-07
FA12_HUMAN      615     Trypsin         373     609     3.6e-63
FA12_HUMAN      615     Kringle         217     295     5.7e-24
FA12_HUMAN      615     fn2     47      88      6.8e-17
FA12_HUMAN      615     EGF     98      129     6.2e-06
FA12_HUMAN      615     EGF     178     207     5.5e-08
FA12_HUMAN      615     fn1     135     170     1.7e-10
```

Exercise

- Merged the BCR_HUMAN.fa, ABL1_HUMAN.fa, BCR_ABL1_fusion_HUMAN.fa into single file using “cat” command. Then run “hmmscan” to search for the similar proteins (domains) of query proteins. After obtaining the output file named as “bcr_abl1.tab”, run the “parse_hmmscan.py” to get summary of the hmmscan search.

아래 명령어를 실행해서 BCR_HUMAN.fa, ABL1_HUMAN.fa, BCR_ABL1_fusion_HUMAN.fa 파일들을 여러분들의 directory에 복사해 가세요

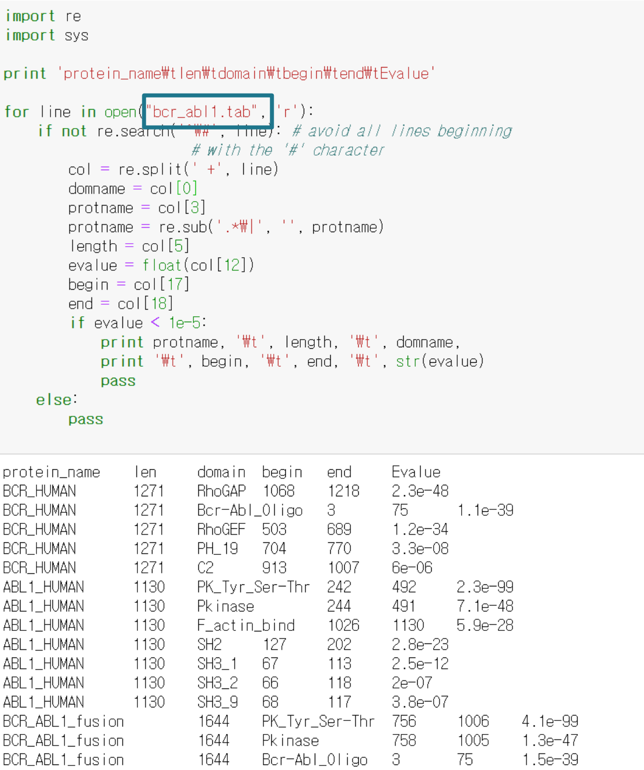
```
$ cp /home/biguser/tutor/Session10/bcr_abl_sequences/* .
```

Exercise results

```
Step 1) cat BCR_HUMAN.fa ABL1_HUMAN.fa BCR_ABL1_fusion_HUMAN.fa >
merged.fa

Step2 ) hmmscan --domtblout bcr_abl1.tab Pfam-A.hmm merged.fa
```

Exercise results

3. Code 수정과 실행



```
import re
import sys

print 'protein_name₩tlen₩tdomain₩tbegin₩tend₩tEvalue'

for line in open("bcr_abl1.tab", 'r'):
    if not re.search('₩#', line): # avoid all lines beginning
                                   # with the '#' character
        col = re.split(' +', line)
        domname = col[0]
        protname = col[3]
        protname = re.sub('.*₩|', '', protname)
        length = col[5]
        evalue = float(col[12])
        begin = col[17]
        end = col[18]
        if evalue < 1e-5:
            print protname, '₩t', length, '₩t', domname,
            print '₩t', begin, '₩t', end, '₩t', str(evalue)
            pass
    else:
        pass
```

```
protein_name    len     domain  begin   end     Evalue
BCR_HUMAN       1271    RhoGAP  1068    1218    2.3e-48
BCR_HUMAN       1271    Bcr-Abl_Oligo   3       75      1.1e-39
BCR_HUMAN       1271    RhoGEF  503     689     1.2e-34
BCR_HUMAN       1271    PH_19   704     770     3.3e-08
BCR_HUMAN       1271    C2      913     1007    6e-06
ABL1_HUMAN      1130    PK_Tyr_Ser-Thr  242     492     2.3e-99
ABL1_HUMAN      1130    Pkinase 244     491     7.1e-48
ABL1_HUMAN      1130    F_actin_bind    1026    1130    5.9e-28
ABL1_HUMAN      1130    SH2     127     202     2.8e-23
ABL1_HUMAN      1130    SH3_1   67      113     2.5e-12
ABL1_HUMAN      1130    SH3_2   66      118     2e-07
ABL1_HUMAN      1130    SH3_9   68      117     3.8e-07
BCR_ABL1_fusion         1644    PK_Tyr_Ser-Thr  756     1006    4.1e-99
BCR_ABL1_fusion         1644    Pkinase 758     1005    1.3e-47
BCR_ABL1_fusion         1644    Bcr-Abl_Oligo   3       75      1.5e-39
```