9.2 BLAST

9.2.1 Protein Sequence Alignments

Amino Acid Sequence Alignment

SARS-CoV-2 is the virus that causes COVID-19. The SARS-CoV-2 genome codes for a surface glycoprotein, or the spike protein, which allows SARS-CoV-2 to bind to human cells. The portion of the protein that allows it to bind to the cells is the receptor binding domain (RBD).

Mutations in the viral genome lead to new virus variants, including the Alpha, Delta, and Omicron variants of SARS-CoV-2. When new variants are identified, their genomes are sequenced to identify differences from earlier variants.

To determine how similar the RBDs of the Alpha, Delta, and Omicron variants are to the initial SARS-CoV-2 RBD sequence, BLAST can be used align the sequences. As a query sequence the RBD from the original SARS-CoV-2 genome will be used. The exact locations of the RBDs in surface glycoproteins in the variants are not available, therefore the full surface glycoprotein sequences can be aligned to the RBD of the first identified variant.

The analysis is two-fold, the alignment will identify both the coordinates of the RBDs in the variant surface glycoprotein sequences, and the similarity of the variant RBDs to the original RBD can be determined.

The `blastp` command is used for protein-to-protein alignments and has the same syntax as the `blastn` command, however, the FASTA files provided must contain amino acid sequences.

```
blastp -query QUERY.fa -subject SUBJECT.fa
```

The file `QHR63250.1_wuhan_RBD.fa` contains the RBD from the original SARS-CoV-2 gene and the file `variant_surface_glycoprotein.fa`, contains the sequences of the Alpha, Delta, and Omicron variant surface glycoproteins (sequences were retrieved from GenBank, accessions are contained within the files). To following command will return all alignments between the query and subject sequences with an Expect-value < 1.

```
j:~/Week.9/9.2.BLAST.Applications$ blastp -query QHR63250.1_wuhan_RBD.fa
-subject variant_surface_glycoprotein.fa -evalue 1
```

The output of this command is shown in smaller portions to go over each part individually.

The output is similar to the output returned from `blastn`. In this case there are three alignments, one for each of the variants in the subject file. Notice that under 'Database' it states that there are three sequences and outputs the combined total number of amino acid residues.

```
BLASTP 2.12.0+

Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference for composition-based statistics: Alejandro A. Schaffer,
L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri
I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001),
"Improving the accuracy of PSI-BLAST protein database searches with
composition-based statistics and other refinements", Nucleic Acids
Res. 29:2994-3005.

Database: User specified sequence set (Input:
variant_surface_glycoproteins.fa).
           3 sequences; 3,811 total letters

Query= QHR63250.1 spike glycoprotein RBD [Wuhan seafood market pneumonia
virus]

Length=183
                                                                 Score     E
Sequences producing significant alignments:                     (Bits)  Value

QTX93774.1_alpha surface glycoprotein [Severe acute respi...  381     6e-129
UEQ01935.1_delta surface glycoprotein [Severe acute respi...  380     2e-128
UFP04971.1_omicron surface glycoprotein [Severe acute res...  353     4e-118
```

The next portion of the output shows alignments, in this case there are three alignments—one for each variant. Each alignment starts with a block of information followed by the protein sequence alignment.

```
> QTX93774.1_alpha surface glycoprotein [Severe acute respiratory
syndrome coronavirus 2]
Length=1270

 Score = 381 bits (979),  Expect = 6e-129, Method: Compositional matrix adjust.
 Identities = 182/183 (99%), Positives = 182/183 (99%), Gaps = 0/183 (0%)

Query  1    CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNV  60
            CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNV
Sbjct  333  CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNV  392

Query  61   YADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRL  120
            YADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRL
Sbjct  393  YADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRL  452

Query  121  FRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSF  180
            FRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPT GVGYQPYRVVVLSF
Sbjct  453  FRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSF  512
```

```
Query  181  ELL  183
            ELL
Sbjct  513  ELL  515



> UEQ01935.1_delta surface glycoprotein [Severe acute respiratory
syndrome coronavirus 2]
Length=1271

 Score = 380 bits (976),  Expect = 2e-128, Method: Compositional matrix adjust.
 Identities = 181/183 (99%), Positives = 181/183 (99%), Gaps = 0/183 (0%)

Query  1    CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNV  60
            CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNV
Sbjct  334  CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNV  393

Query  61   YADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRL  120
            YADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNY YRL
Sbjct  394  YADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYRYRL  453

Query  121  FRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSF  180
            FRKSNLKPFERDISTEIYQAGS PCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSF
Sbjct  454  FRKSNLKPFERDISTEIYQAGSKPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSF  513

Query  181  ELL  183
            ELL
Sbjct  514  ELL  516


> UFP04971.1_omicron surface glycoprotein [Severe acute respiratory
syndrome coronavirus 2]
Length=1270

 Score = 353 bits (905),  Expect = 4e-118, Method: Compositional matrix adjust.
 Identities = 168/183 (92%), Positives = 172/183 (94%), Gaps = 0/183 (0%)

Query  1    CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNV  60
            CPF EVFNATRFASVYAWNRKRISNCVADYSVLYN A F TFKCYGVSPTKLNDLCFTNV
Sbjct  333  CPFDEVFNATRFASVYAWNRKRISNCVADYSVLYNLAPFFTFKCYGVSPTKLNDLCFTNV  392

Query  61   YADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRL  120
            YADSFVIRGDEVRQIAPGQTG IADYNYKLPDDFTGCVIAWNSN LDSKV GNYNYLYRL
Sbjct  393  YADSFVIRGDEVRQIAPGQTGNIADYNYKLPDDFTGCVIAWNSNKLDSKVSGNYNYLYRL  452

Query  121  FRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSF  180
            FRKSNLKPFERDISTEIYQAG+ PCNGV GFNCYFPL+SY F+PT GVG+QPYRVVVLSF
Sbjct  453  FRKSNLKPFERDISTEIYQAGNKPCNGVAGFNCYFPLRSYSFRPTYGVGHQPYRVVVLSF  512

Query  181  ELL  183
            ELL
Sbjct  513  ELL  515
```

To compare the differences between blastp and blastn output, view the third alignment, the alignment between the query SARS-CoV-2 RBD sequence (QHR63250.1 spike glycoprotein RBD) and the Omicron variant surface glycoprotein sequence (UFP04971.1_omicron

surface glycoprotein). Much of the information is the same as the blastn results: percent identity is 92%, there are 163 matches, alignment length of 183, and an Expect-value of 4e-118. Protein alignments also have the percent positive identity (see 9.1.1) next to the percent identity, labeled “Positives”.

Protein sequence alignments in BLAST differ from nucleotide alignments in that instead of vertical bars displaying matches, the single letter amino acid representation is displayed between matches instead. Where mismatches are positives a plus sign is shown between the bases in the query and subject sequences. As a reminder, these indicate conservation of similar properties between the mismatched amino acids. For example, view the alignment between query amino acids 121 and 180, and subject amino acids 453 and 512. There is a plus sign between aligned occurrences of: serine (S) and asparagine (N), glutamine (Q) and arginine (R), and tyrosine (Y) and histidine (H).

Finally, there is a section at the end of the output that describes the parameters used in the alignment and statistical calculations. These will not be discussed in detail in this course, however more information can be found here:
https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

```
Lambda      K        H        a         alpha
   0.322    0.139    0.442    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 527813


  Database: User specified sequence set (Input:
variant_surface_glycoproteins.fa).
    Posted date:  Unknown
  Number of letters in database: 3,811
  Number of sequences in database:  3



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 11
Window for multiple hits: 40
```

To view the summary information for the alignment to all three variants, the same command can be run with `-outfmt` set to 7.

```
j:~/Week.9/9.2.BLAST.Applications$ blastp -query QHR63250.1_wuhan_RBD.fa
-subject variant_surface_glycoprotein.fa -evalue 1
# BLASTP 2.12.0+
# Query: QHR63250.1 spike glycoprotein RBD [Wuhan seafood market
pneumonia virus]
# Database: User specified sequence set (Input:
variant_surface_glycoproteins.fa)
# Fields: query acc.ver, subject acc.ver, % identity, alignment length,
mismatches, gap opens, q. start, q. end, s. start,s. end, evalue, bit
score
# 3 hits found
QHR63250.1      QTX93774.1_alpha        99.454  183     1       0
1       183     333     515     6.48e-129       381
QHR63250.1      UEQ01935.1_delta        98.907  183     2       0
1       183     334     516     1.79e-128       380
QHR63250.1      UFP04971.1_omicron      91.803  183     15      0
1       183     333     515     3.52e-118       353
# BLAST processed 1 queries
```

Now the similarity of the RBD of the first sequenced SARS-Cov-2 sample to Alpha, Delta and Omicron variant RBDs can easily be compared. The third column displays the percent identity of the alignment. The alpha variant (1st hit) is over 99% identical, the Delta variant (2nd hit) is over 98% identical, and the Omicron variant (3rd hit) is only ~92% identical. The number of mismatches is in the 5th column, with 1, 2, and 15 mismatches for Alpha, Delta and Omicron, respectively. Thus, there are far more mutations in the receptor binding domain of the Omicron variant than there are in other two SARS-CoV-2 variants.

9.2.2 Sequence Evolution

Synonymous & Non-synonymous Mutations

Codons determine the amino acid sequence of the protein product of a gene. Thus, when there is a mutation in the DNA, it can alter the protein product. When a nucleotide sequence mutation changes the amino acid sequence of a protein it is called a **non-synonymous mutation**. For example, the codon CCU encodes the amino acid proline. If the first C is mutated to an A, the codon will become ACU, which encodes the amino acid threonine.

Multiple codons can code for the same amino acid (see codon table below), thus, nucleotide mutations do not always change the amino acid sequence. These are called **synonymous mutations**. For example, the codon CCU encodes the amino acid proline. If the last U is mutated to an A, the codon will become CCA, which still encodes the amino acid proline. In

fact, if the U was mutated to a G or C, it would still encode for proline (see codon table below).

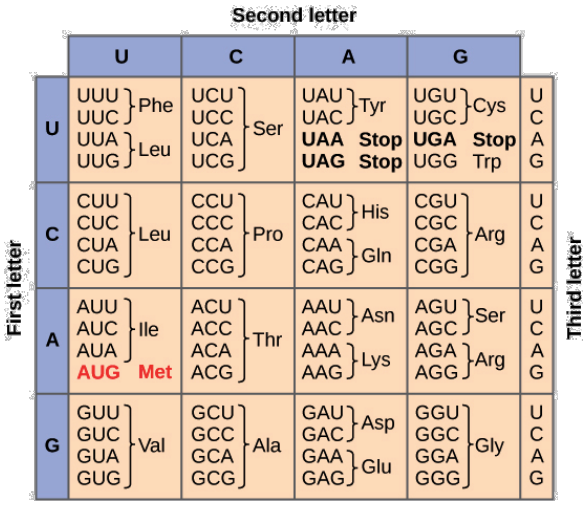


Image source: "The genetic code," by OpenStax College, Biology (CC BY 3.0).

As some nucleotide sequence mutations can have no effect on the amino acid sequence (synonymous mutations), nucleotide sequences generally diverge more rapidly than amino acid sequences. To see this trend, compare the alignment of the coding sequences and amino acid sequences of human and mouse MAP2K1 orthologs.

First, align the human MAP2K1 CDS with the mouse Map2k1 CDS. Note that because there will be only one hit, the BLAST command is piped to `tail -n 4` so that it will output only the last two header lines, the line with information about the hit, and the footer line.

```
j:~/Week.9/9.2.BLAST.Applications$ blastp -query Hs_MAP2K1_CDS.fa -
subject Mm_Map2k1_CDS.fa -outfmt 7 | tail -n 4
# Fields: query acc.ver, subject acc.ver, % identity, alignment length,
mismatches, gap opens, q. start, q. end, s. start,s. end, evalue, bit
score
# 1 hits found
H.sapiens_MAP2K1_CDS    M.musculus_Map2k1_CDS   90.870  1172    107
0       1       1172    1       1172    0.0     1572
# BLAST processed 1 queries
```

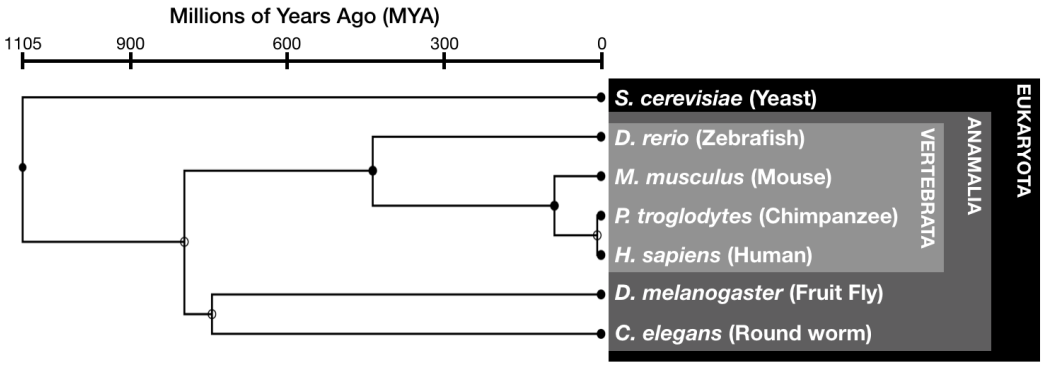
The percent identity of the two CDs is ~91%, with 107 mismatches and 0 gaps. Now align the amino acid sequences of the orthologs. Note that because there will be only one hit, the BLAST command is piped to `tail -n 4` so that it will output only the last two header lines, the line with information about the hit, and the footer line.

```
j:~/Week.9/9.2.BLAST.Applications$ blastp -query Hs_MAP2K1_protein.fa -
subject Mm_Map2k1_protein.fa -outfmt 7 | tail -n 4
# Fields: query acc.ver, subject acc.ver, % identity, alignment length,
mismatches, gap opens, q. start, q. end, s. start,s. end, evalue, bit
score
# 1 hits found
H.sapiens_MAP2K1_protein    M.musculus_Map2k1_protein    98.982  393
4       0       1       393     1       393     0.0     805
# BLAST processed 1 queries
```

The percent identity of the amino acid sequences is ~99%, with 4 mismatches and 0 gaps. This is because the synonymous mutations will affect the coding sequence alignment, but not the amino acid sequence alignment. Only non-synonymous mutations will affect the amino acid sequence alignment.

Conservation of Sequences

A **phylogenetic tree** is a branching diagram displaying the evolutionary relationships between species or other elements. For example, the evolutionary relationships between orthologs. The following phylogenetic tree shows the evolutionary relationships between seven eukaryotes: humans, chimpanzees, mice, zebrafish, fruit flies, round worms, and yeast.



Tree source: http://www.timetree.org/

While humans and chimpanzees diverged relatively recently, between 6 and 7 million years ago (MYA), humans and yeast diverged ~1100 MYA.

Generally, a gene or protein will have higher similarity to orthologs in more closely related species. The chimpanzee orthologs of human genes will be much more similar than fruit fly orthologs of human genes because less time has passed since the divergence of the two species.

An essential gene is a gene that is required for the survival of an organism. When a mutation arises in one of these genes, in order for that mutation to be retained it cannot

affect the structure or function of the gene. Thus, essential genes evolve more slowly than non-essential gene and are therefore more highly conserved.

An example of an essential gene is ACTR3, which is part of protein complex that regulates actin formation. The file `Hs_ACTR3.fa` contains the amino acid sequence of human ACTR3. The file `ACTR3_orthologs.fa` contains the amino acid sequences of mouse (*M. musculus*), zebrafish (*D. rerio*), and round worm (*C. elegans*) orthologs of human ACTR3. Note that by piping the BLAST command to `grep "H sapiens_ACTR3"` only the line with the query name and the lines with information about the hits will be output.

```
j:~/Week.9/9.2.BLAST.Applications$ blastp -query Hs_ACTR3.fa -subject
ACTR3_orthologs.fa -outfmt 7 -evalue 1 | grep "H_sapiens_ACTR3"
# Query: H_sapiens_ACTR3
H_sapiens_ACTR3 M_musculus_Actr3        99.761  418     1       0
1       418     1       418     0.0     868
H_sapiens_ACTR3 D_rerio_actr3   96.954  394     12      0       1
394     1       394     0.0     801
H_sapiens_ACTR3 C_elegans_arx-1 76.123  423     95      3       2
418     3       425     0.0     676
```

Alignment of human ACTR3 with its orthologs in other species shows that the human and mouse amino acid sequences are nearly 100% identical, meaning it is well conserved since the diverge of human and mouse ~90 MYA. As we go further back to zebra fish (diverged from humans ~435 MYA), the percent identity drops to about 97%. Although humans and *C. elegans* diverged ~797 MYA, the ACTR3 sequences are still 76% identical. The protein is extremely well conserved across hundreds of millions of years.

Because essential genes evolve more slowly than non-essential genes, they are generally more well conserved. HEATR1 is a non-essential human gene involved in rRNA processing. The file `Hs_HEATR1.fa` contains the amino acid sequence of human HEATR1. The file `HEATR1_orthologs.fa` contains the amino acid sequences of mouse (*M. musculus*), zebrafish (*D. rerio*), and round worm (*C. elegans*) orthologs of human HEATR1. Note that by piping the BLAST command to `grep "H sapiens_HEATR1"` only the line with the query name and the lines with information about the hits will be output.

```
j:~/Week.9/9.2.BLAST.Applications$ blastp -query Hs_HEATR1.fa -subject
HEATR1_orthologs.fa -outfmt 7 -evalue 1 | grep "H_sapiens_HEATR1"
# Query: H_sapiens_HEATR1
H_sapiens_HEATR1        M_musculus_Heatr1       83.815  2144    346
1       1       2144    1       2143    0.0     3703
H_sapiens_HEATR1        D_rerio_heatr1  53.836  2177    954     20
1       2144    1       2159    0.0     2243
H_sapiens_HEATR1        C_elegans_toe-1 22.398  1076    685     33
1181    2142    606     1645    1.63e-46        174
H_sapiens_HEATR1        C_elegans_toe-1 29.539  369     234     9
2       356     3       359     5.86e-40        152
```

The output shows the same pattern as before, highest similarity to mouse, then zebrafish, then *C. elegans*. However, this gene was much less conserved. The zebrafish protein ortholog is just over 50% identical to the human protein. Furthermore, only two shorter segments of the *C. elegans* ortholog align with the human protein, each of which has less than 30% identity.

9.2.3 BLAST Databases

BLAST Databases

So far, all alignments have used a subject FASTA file with one or a few sequences. Alignments to small sets of subject sequences run very quickly, however, if the set of subject sequences is very large, finding alignments can take a very long time. A **BLAST database** is a FASTA file with a set of associated files that speed up the search time. They are helpful for querying a large set of sequences, like the entire transcriptome or proteome of a species. A database can be created using the command `makeblastdb`:

```
makeblastdb -in SEQUENCES.fa -out DB_NAME -parse_seqids
-dbtype prot/nucl
```

The command requires a FASTA file of sequences (`SEQEUENCES.fa`) with which to create the database, provided after the option `-in`. A database name (`DB_NAME`) must also be provided after the option `-out`. The option `-parse_seqids` should be used as it allows one to later extract individual sequences from the database, and the type of database (`-dbtype`) must be provided: `prot` if `SEQUENCES.fa` contains amino acid sequences or `nucl` if `SEQUENCES.fa` contains nucleic acid sequences.

In the directory for this module is a directory called `BLAST_databases` which contains two FASTA files, one containing the human proteome (`homo_sapiens_proteome.fasta`) and one containing the mouse proteome (`mus_musculus_proteome.fasta`) (sourced from UniProt: https://www.uniprot.org/). These are very large files each containing over 20,000 amino acid sequences.

```
j:~/Week.9/9.2.BLAST.Applications$ cd BLAST_databases
j:~/Week.9/9.2.BLAST.Applications/BLAST_databases$ ls
homo_sapiens_proteome.fasta mus_musculus_proteome.fasta
```

Create a BLAST database for the *Homo sapiens* proteome called `homo_sapiens_proteome`:

```
j:~/Week.9/9.2.BLAST.Applications/BLAST_databases$ makeblastdb -in
homo_sapiens_proteome.fasta -out homo_sapiens_proteome -parse_seqids -
dbtype prot
Building a new DB, current time: 01/06/2022 17:06:29
New DB name:
/home/jovyan/Week.9/9.2.BLAST.Applications/BLAST_databases/homo_sapiens_
proteome
New DB title:  homo_sapiens_proteome.fasta
Sequence type: Protein
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 20588 sequences in 0.580569 seconds.
```

The output of the command is information about the creation of the BLAST database. The database contains 20,588 sequences and is located at the path: `/home/jovyan/Week.9/9.2.BLAST.Applications/BLAST_databases/homo_sapiens_proteome`

Create a BLAST database for the *Mus musculus* proteome called `mus_musculus_proteome`:

```
j:~/Week.9/9.2.BLAST.Applications/BLAST_databases$ makeblastdb -in
mus_musculus_proteome.fasta -out mus_musculus_proteome -parse_seqids -
dbtype prot
Building a new DB, current time: 01/06/2022 17:08:47
New DB name:
/home/jovyan/Week.9/9.2.BLAST.Applications/BLAST_databases/mus_musculus_
proteome
New DB title:  mus_musculus_proteome.fasta
Sequence type: Protein
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 21986 sequences in 0.806985 seconds.
```

This database contains 21,986 sequences and is located at the path: `/home/jovyan/Week.9/9.2.BLAST.Applications/BLAST_databases/mus_musculus_proteome`

To query a database in a BLAST command requires the name of the database, provided via the `-db` option. This is used instead of the `-subject` option.

```
blastp -query QUERY.fa -db DB_NAME
```

Note that the `DB_NAME` should include the path to the database, and the name given to the database when `makeblastdb` was run.

As an example, in the directory `9.2.BLAST.Applications`, the file `Mm_Tp53_protein.fa` contains the amino acid sequence of the mouse Tp53 protein. To identify the most similar protein(s) in the human proteome, BLAST the protein sequence

against the proteome. Note that the database name is `BLAST_databases/homo_sapiens_proteome` as the command is run from the parent directory of the `BLAST_databases` directory.

```
j:~/Week.9/9.2.BLAST.Applications$ blastp -query Mm_Tp53_protein.fa -db
BLAST_databases/homo_sapiens_proteome -outfmt 7 -evalue 1 >
Mm_Tp53_human_proteome.txt
j:~/Week.9/9.2.BLAST.Applications$ cat Mm_Tp53_human_proteome.txt
# BLASTP 2.12.0+
# Query: sp|P02340|P53_MOUSE Cellular tumor antigen p53 OS=Mus musculus
OX=10090 GN=Tp53 PE=1 SV=4
# Database: BLAST_databases/homo_sapiens_proteome
# Fields: query acc.ver, subject acc.ver, % identity, alignment length,
mismatches, gap opens, q. start, q. end, s. start,s. end, evalue, bit
score
# 3 hits found
sp|P02340|P53_MOUSE     P04637  77.354  393     83      4       4
390     1       393     0.0 578
sp|P02340|P53_MOUSE     O15350  50.943  265     121     3       94
349     115     379     4.38e-85     271
sp|P02340|P53_MOUSE     Q9H3D4  48.951  286     138     4       69
347     139     423     7.04e-84     268
# BLAST processed 1 queries
```

There are three hits with an Expect-value of less than 1 in the human proteome for the mouse Tp53 protein. The top hit, with 77 % identity, is a protein with the ID P04637. The ID of the hit can be used to get the full description and sequence of the hit from the database using the command `blastdbcmd`. The `blastdbcmd` requires the ID of a protein and the path to the database and database name

```
blastp -entry SEQUENCE_ID -db DB_NAME
```

Note that the `DB_NAME` should include the path to the database, and the name given to the database when `makeblastdb` was run.

To determine the identity of the human protein with the ID P04637 it can be looked up in the `homo_sapiens_proteome` database. Note that one can only use this command to search a database if the option `-parse_seqids` was used during database creation.

```
j:~/Week.9/9.2.BLAST.Applications$ blastdbcmd -entry P04637 -db
BLAST_databases/homo_sapiens_proteome
>P04637 Cellular tumor antigen p53 OS=Homo sapiens OX=9606 GN=TP53 PE=1
SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPP
VAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQ
LWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSV
VVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEEN
LRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

The description of this sequence in the human proteome reveals that this is the human protein TP53 (`GN=TP53`). This is the direct ortholog of mouse Tp53. Looking up the other two hits with lower percent identity to mouse Tp53 in humans (O15350 ~51% identity & Q9H3D4 ~49% identity) reveals that they are the proteins TP73 and TP63, paralogs of human of TP53.

Function Prediction

In section 9.1 the concept of sequence alignment to predict the function of a protein was mentioned. By identifying similar protein sequences to the sequence of a protein with an unknown function, inference about the role of the protein of unknown function can be made. Proteins in the same species can be searched for, or orthologs in other species can be identified, examining the functions of similar proteins with known functions provides insight into what the protein of unknown function may do in the cell.

For example, the FASTA file `unknown_alpaca_protein.fa` contains the amino acid sequence of an alpaca protein of unknown function. To identify the most similar protein in the mouse proteome, BLAST the unknown protein sequence against `mus_musculus_proteome` and extract the first hit (there are always 5 header lines when using `-outfmt 7`, thus `head -6` will return the header lines and the first hit).

```
j:~/Week.9/9.2.BLAST.Applications$ blastp -query
unknown_alpaca_protein.fa -db BLAST_databases/mus_musculus_proteome
-outfmt 7 -evalue 1 | head -n 6
# BLASTP 2.12.0+
# Query: NP_001372054 length=345
# Database: BLAST_databases/mus_musculus_proteome
# Fields: query acc.ver, subject acc.ver, % identity, alignment length,
mismatches, gap opens, q. start, q. end, s. start,s. end, evalue, bit
score
# 507 hits found
NP_001372054    P51491    86.006  343     48      0       3       345
4       346     0.0     625
```

The top alignment of the unknown alpaca protein has 86% identity with a mouse protein with the ID P51491. Nearly the entire alpaca sequence (length = 345) is included in the alignment (query start = 3, query end = 345).

It is useful to check multiple proteomes for a similar protein, as the ortholog in another species may be more similar. Identify the most similar protein in the human proteome:

```
j:~/Week.9/9.2.BLAST.Applications$ blastp -query
unknown_alpaca_protein.fa -db BLAST_databases/homo_sapiens_proteome
-outfmt 7 -evalue 1 | head -n 6
# BLASTP 2.12.0+
# Query: NP_001372054 length=345
# Database: BLAST_databases/homo_sapiens_proteome
# Fields: query acc.ver, subject acc.ver, % identity, alignment length,
mismatches, gap opens, q. start, q. end, s. start,s. end, evalue, bit
score
# 411 hits found
NP_001372054    P03999    100.000 345     0       0       1       345
4       348     0.0     711
```

The alignment with the most similar human gene covers the entire alpaca protein sequence (query start = 1, query end = 345) and is 100% identical. The function of the human protein with the ID P03999, therefore, will provide a very good prediction for the function of the alpaca protein.

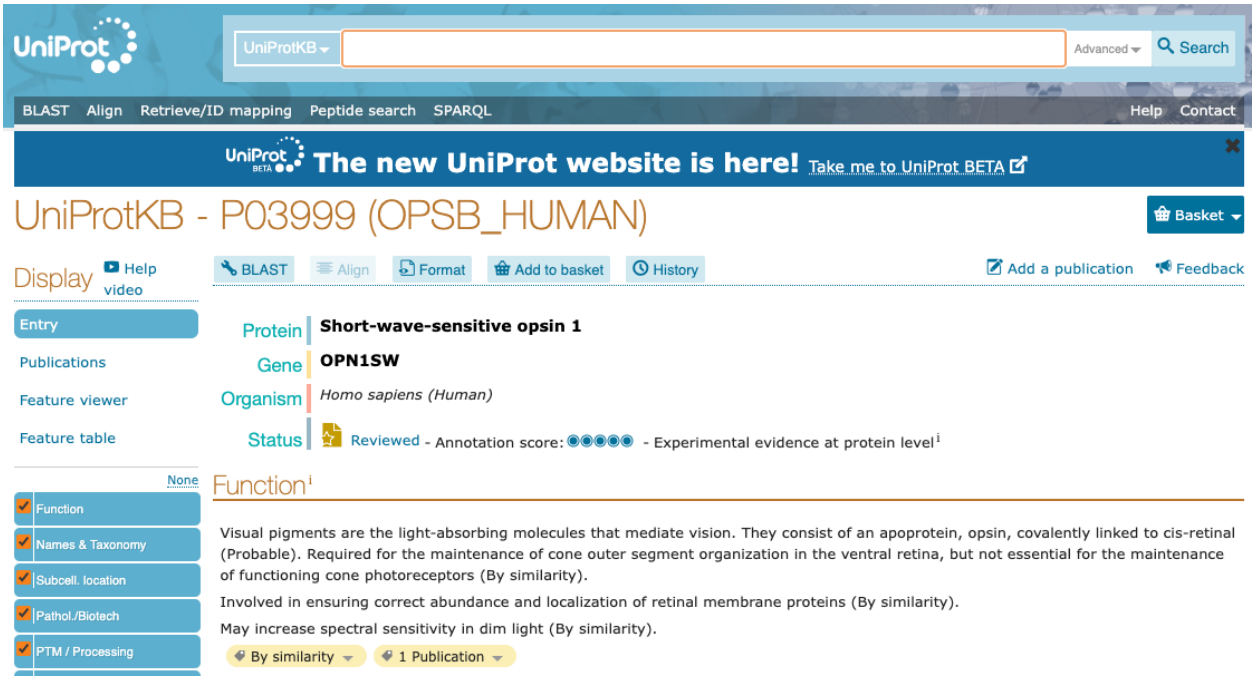
```
j:~/Week.9/9.2.BLAST.Applications$ blastdbcmd -entry P03999 -db
BLAST_databases/homo_sapiens_proteome
>P03999 Short-wave-sensitive opsin 1 OS=Homo sapiens OX=9606 GN=OPN1SW
PE=1 SV=1
MRKMSEEEFYLFKNISSVGPWDGPQYHIAPVWAFYLQAAFMGTVFLIGFPLNAMVLVATLRYKKLRQPLNYI
LVNVSFGGFLLCIFSVFPVFVASCNGYFVFGRHVCALEGFLGTVAGLVTGWSLAFLAFERYIVICKPFGNFR
FSSKHALTVVLATWTIGIGVSIPPFFGWSRFIPEGLQCSCGPDWYTVGTKYRSESYTWFLFIFCFIVPLSLI
CFSYTQLLRALKAVAAQQQESATTQKAEREVSRMVVVMVGSFCVCYVPYAAFAMYMVNNRNHGLDLRLVTIP
SFFSKSACIYNPIIYCFMNKQFQACIMKMVCGKAMTDESDTCSSQKTEVSTVSSTQVGPN
```

Searching this protein ID in the human proteome reveals that the gene is named OPN1SW with a longer description of “Short-wave-sensitive opsin 1”. It is difficult to determine the function of this protein from the name alone.

The proteomes and amino acid sequences used in 9.2 are downloaded from UniProt, an online protein database (https://www.uniprot.org/). The IDs of the proteins in the database (ex. P03999) are UniProt protein IDs.

At the top of the UniProt website homepage there is a search bar with which UniProt IDs can be searched:

This directly opens the page for the protein:



The protein ID is at the top of the page, followed by key information including the gene name, the species, and the protein function. Reading the function information, this protein has a role in vision.

Each protein page on UniProt has several sections with further information about the protein. The menu listing the sections is on the left side of the page:

The “Subcellular location” section provides information on where in the cell the protein is located. In the case of OPN1SW, it is located in the cell membrane and the inner and outer segments of photoreceptors.

The “Pathology & Biotech” section provides information on diseases the protein is involved in, amino acid changes found in these diseases, and links to other databases with more information. For OPN1SW, this section provides details on the involvement of the protein in tritan colour blindness (tritanopia).

Other sections provide information on where the protein is expressed, which proteins it interacts with, the protein structure, the protein domains, the protein sequence, and other details.