Practicum Bioinformatics - answer key

Prof. Kris Laukens - Adrem Data Lab
Department of Computer Science, University of Antwerp

Academic year 2024-2025

Contents

Note about the answers

These are just example solutions. For many of the exercises there will be other correct solutions as well. The main goal of these practicals is to become acquainted with the different databases, techniques and principles of bioinformatics. Try to understand why we are using a specific technique, the different aspects of the outputs of the tools and how they work, rather than learning everything by heart. You will need to be able to interpret related questions to the ones you've solved here while using this text as a quick reference, but you will need to know the major sections and themes of the practicals if you want to be able to retrieve them in a timely fashion.

1 Search & motifs

Learning goals

- Get familiar with sequences searches
- Understand the use of different parameters in sequence searches
  - Use of different matrices
  - Complexity masks
- Characterize domains from transcripts
- Visualize motifs from promotor sequences

Useful resources

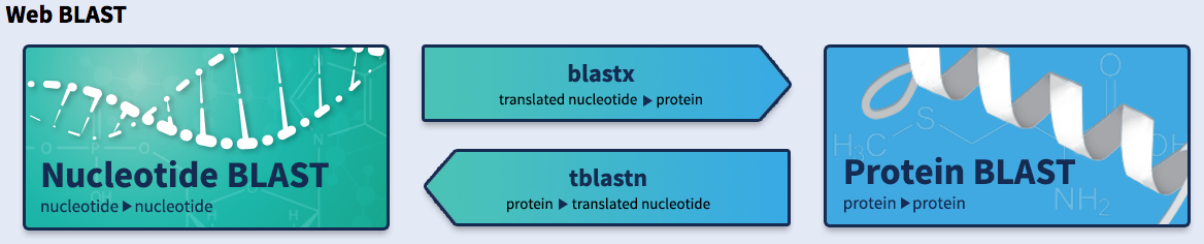
Introduction to BLAST

Section 4: Sequence similarity search

Exercise 4.1: Getting familiar with sequence similarity search: word size

The most frequently used algorithm to perform database searches is the BLAST algorithm. BLAST comes in a number of *"flavours"* depending on the input sequence and the queried database.

The BLAST page of NCBI can be found at `http://blast.ncbi.nlm.nih.gov/Blast.cgi` and provides an overview of the different BLAST flavours available.



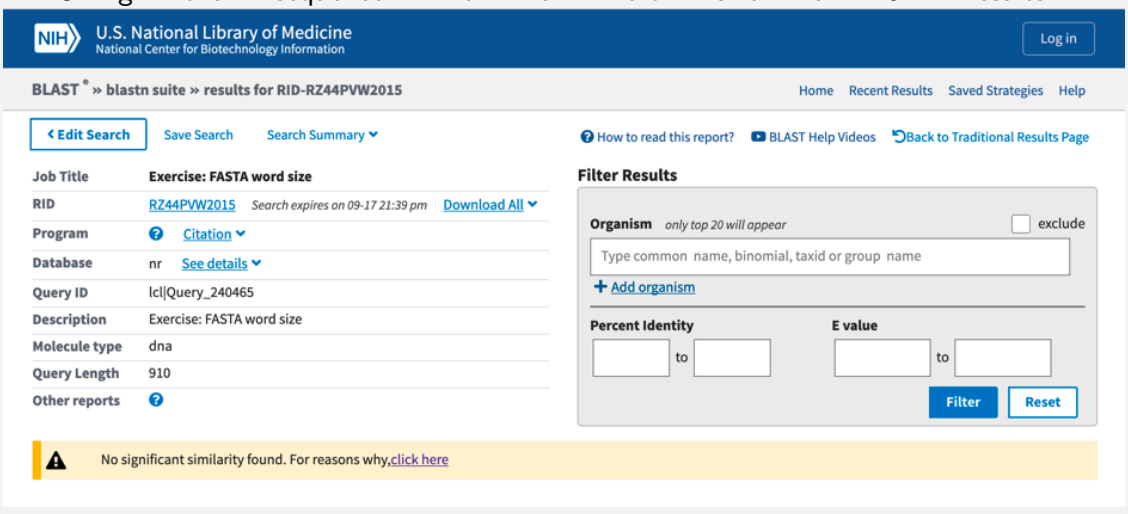
Navigate to `https://blast.ncbi.nlm.nih.gov/Blast.cgi`. This page contains the NCBI BLAST tool for executing BLAST. Run two searches with Nucleotide BLAST using the following unknown nucleotide sequence. Use different word sizes (28 and 64) for each search. Word size can be adjusted by changing the Word size parameter in the drop-down Algorithm Parameters. Make sure to only search in the nr/nt database which contains the nucleotide collection of also other non-human, non-mice species.

```
>Exercise: word size
TTGGCAGATTCCCGCCAGAGCAAAACAGCCGCTAGTCTTAGTCCGAGTCGCCCGCAAAGTTCCTCGATTA
ACTCCGTACCCGGAGCGCCAAACCGGGACTCCTTCGCTAAGCTGCGCGAACCACTTGTGGTTCCGGGACT
CCTTGACGTCCAGACCGTTTCGTTCGAGTGGCTGATCGGTTCGCCGCTCTGGCGCGAATCCGCCGCCGAG
CGGGGTGTTGTCAACCCAGTGGGTGGCCTGAAGAGGAGCTCTACGAGCTGTCTCCGATCGAGGACTACTC
CGGGTCGATGTCGTTGTCGTTCTCTGTCCCTCGTTTCGACGATGTCAAGGCACCCGACGACGAGTGCAAA
GACAAGGACATGACGTTCGCGGCTCCACTGTTCGTCACCGCCGAGTTACATCAACAACAACACCGGTGAG
ATCAAGATTCAGACGGTGTTCATGGGTGACTTCCCGAAGATGACCGAGAAGGGCACGTTCATCATCATCG
GGACCGAGCGTGTGGTGGTCAGCCAGCAGGTGCGGTCGCCCGGGGTGTACTTCGACGTGACCATTGACAA
GTCCACCGACAAGACGCAGCACAGCGTCAAGGTGATCCCGAGCCGCGTCGCGTGGCTCGAGTTTGACGTC
GACAAGCTCGACACCGTCGGCGTGCGCATCGACCGCATACGCCGGCAACCGGTCACCGTGCTGCTCATGG
CGCTGGGCTGGACCAGCGAGCAGATTGACGAGCGGTTCGGGTTCTCCGAGATCATGCTATCGACGCTGGA
GAAGGACAACACCGTCGTCACCGACGAGGCGCTGTTGGACATCTACCTCAAGCTGCGTCCGGGCGAGCCC
CCGACCATAGAGTCAGCGCAGACGCTGTTGGAAAACTAGTTCTTCAAGGAGAAGCGCTACGACCTGGTCC
```

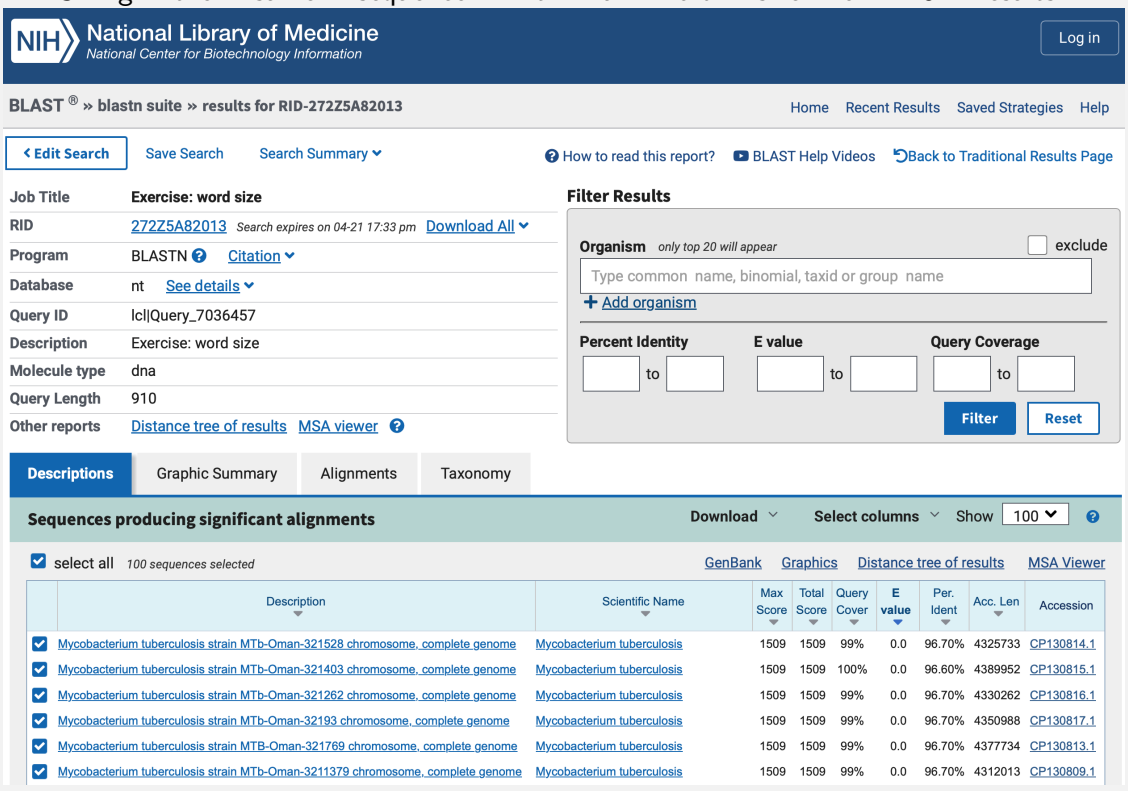
Compare the results of the two searches. Look at the Alignment tab and try to explain why the results are different. Recall how the algorithm depends on word size! What organisms is this sequence from?

As explained in the lectures, there are different BLAST flavours: blastn (DNA → DNA), blastp (protein → protein), ... For this exercise you need Nucleotide blast, so blastn.

BLASTing the sequence with a word size of 64 results in:



BLASTing the same sequence with a word size of 28 results in:



Every +- 30 bases a nucleotide was changed artificially, which corresponds to the sequencing error of a long read sequencer in some cases.

Using a smaller word size allows Blast to still find matching words and overlook words containing a sequencing error. At higher word sizes all words will contain sequencing errors so BLAST will not find matching results.

Do you recall the definition of an E-value (expected value)? It's *the number of hits you can expect to see by chance when searching a database of a particular size*.

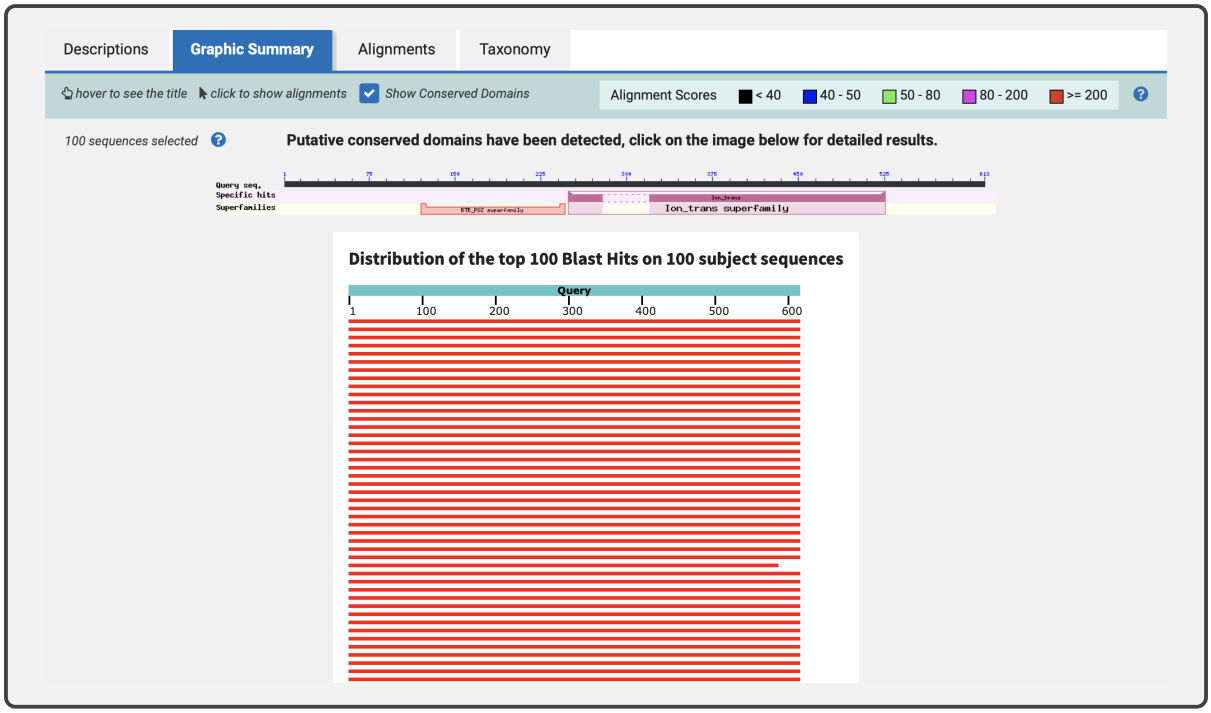
Exercise 4.2: Getting familiar with sequence similarity search: BLAST

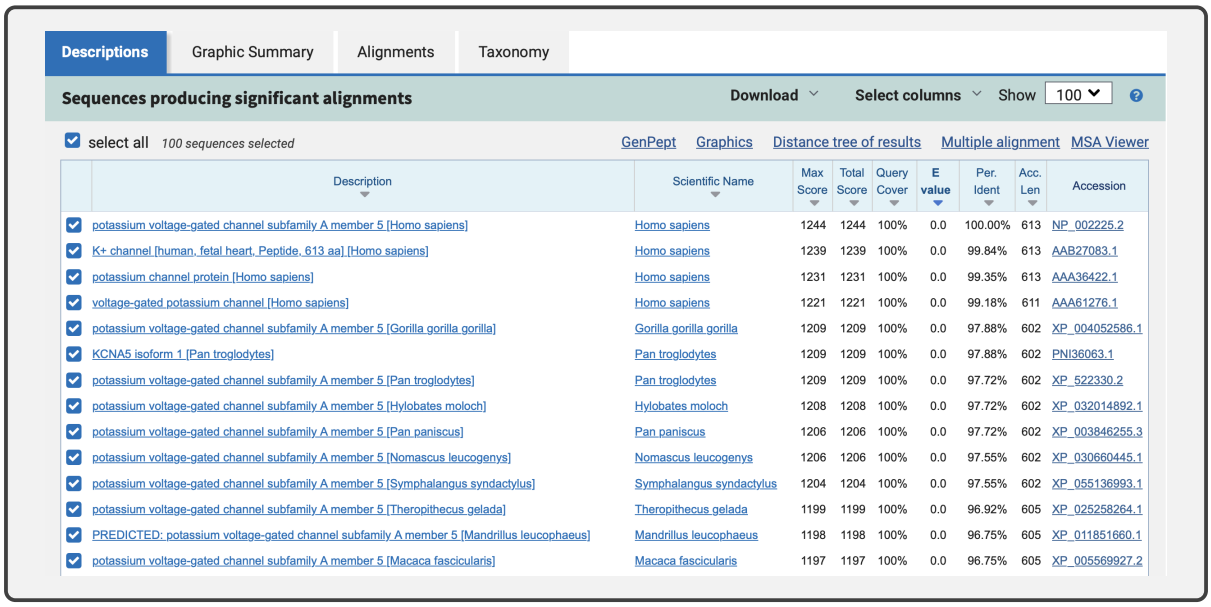
A) BLAST introduction

Go to the BLASTp page and try to find out more information about the following sequence by using the default BLAST settings. BLAST jobs usually require a small amount of time (a few minutes).

```
>Unknown protein sequence
MEIALVPLENGGAMTVRGGDEARAGCGQATGGELQCPPTAGLSDGPKEPAPKGRGAQRDA
DSGVRPLPPLPDPGVRPLPPLPEELPRPRRPPPEDEEEEGDPGLGTVEDQALGTASLHHQ
RVHINISGLRFETQLGTLAQFPNTLLGDPAKRLRYFDPLRNEYFFDRNRPSFDGILYYYQ
SGGRLRRPVNVSLDVFADEIRFYQLGDEAMERFREDEGFIKEEEKPLPRNEFQRQVWLIF
EYPESSGSARAIAIVSVLVILISIITFCLETLPEFRDERELLRHPPAPHQPPAPAPGANG
SGVMAPPSGPTVAPLLPRTLADPFFIVETTCVIWFTFELLVRFFACPSKAGFSRNIMNII
DVVAIFPYFITLGTELAEQQPGGGGGGQNGQQAMSLAILRVIRLVRVFRIFKLSRHSKGL
QILGKTLQASMRELGLLIFFLFIGVILFSSAVYFAEADNQGTHFSSIPDAFWWAVVTMTT
VGYGDMRPITVGGKIVGSLCAIAGVLTIALPVPVIVSNFNYFYHRETDHEEPAVLKEEQG
TQSQGPGLDRGVQRKVSGSRGSFCKAGGTLENADSARRGSCPLEKCNVKAKSNVDLRRSL
YALCLDTSRETDL
```

Take a look at the results windows, what can you learn from the different graphical representations of the results? Try to find out the name of this protein sequence and explain why you think this is the best hit. Given the goal of BLAST (querying a database) and the graphical results you found, do they suggest BLAST relies on a local or global alignment (recall the distinction mentioned in the previous sections)?





Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
potassium voltage-gated channel subfamily A member 5 [Homo sapiens]	Homo sapiens	1244	1244	100%	0.0	100.00%	613	NP_002225.2
K+ channel [human, fetal heart, Peptide, 613 aa] [Homo sapiens]	Homo sapiens	1239	1239	100%	0.0	99.84%	613	AAB27083.1
potassium channel protein [Homo sapiens]	Homo sapiens	1231	1231	100%	0.0	99.35%	613	AAA36422.1
voltage-gated potassium channel [Homo sapiens]	Homo sapiens	1221	1221	100%	0.0	99.18%	611	AAA61276.1
potassium voltage-gated channel subfamily A member 5 [Gorilla gorilla gorilla]	Gorilla gorilla gorilla	1209	1209	100%	0.0	97.88%	602	XP_004052586.1
KCNA5 isoform 1 [Pan troglodytes]	Pan troglodytes	1209	1209	100%	0.0	97.88%	602	PNI36063.1
potassium voltage-gated channel subfamily A member 5 [Pan troglodytes]	Pan troglodytes	1209	1209	100%	0.0	97.72%	602	XP_522330.2
potassium voltage-gated channel subfamily A member 5 [Hylobates moloch]	Hylobates moloch	1208	1208	100%	0.0	97.72%	602	XP_032014892.1
potassium voltage-gated channel subfamily A member 5 [Pan paniscus]	Pan paniscus	1206	1206	100%	0.0	97.72%	602	XP_003846255.3
potassium voltage-gated channel subfamily A member 5 [Nomascus leucogenys]	Nomascus leucogenys	1206	1206	100%	0.0	97.55%	602	XP_030660445.1
potassium voltage-gated channel subfamily A member 5 [Symphalangus syndactylus]	Symphalangus syndactylus	1204	1204	100%	0.0	97.55%	602	XP_055136993.1
potassium voltage-gated channel subfamily A member 5 [Theropithecus gelada]	Theropithecus gelada	1199	1199	100%	0.0	96.92%	605	XP_025258264.1
PREDICTED: potassium voltage-gated channel subfamily A member 5 [Mandrillus leucophaeus]	Mandrillus leucophaeus	1198	1198	100%	0.0	96.75%	605	XP_011851660.1
potassium voltage-gated channel subfamily A member 5 [Macaca fascicularis]	Macaca fascicularis	1197	1197	100%	0.0	96.75%	605	XP_005569927.2

Download GenPept Graphics

potassium voltage-gated channel subfamily A member 5 [Homo sapiens]

Sequence ID: ref|NP_002225.2| **Length:** 613 **Number of Matches:** 1

See 8 more title(s)

Range 1: 1 to 613 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1244 bits(3218)	0.0	Compositional matrix adjust.	613/613(100%)	613/613(100%)	0/613(0%)

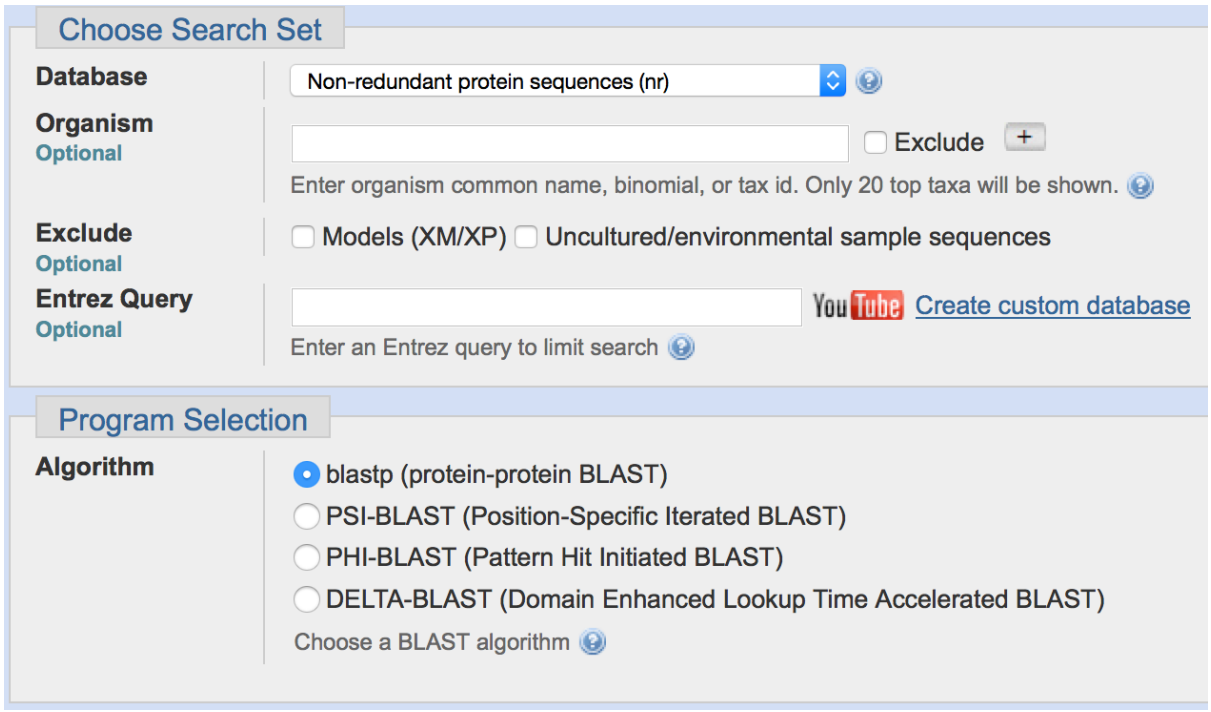
```
Query  1    MEIALVPLENGGAMTVRGGDEARAGCGQATGGELQCPPTAGLSDGPKEPAPKGRGAQRDA  60
            MEIALVPLENGGAMTVRGGDEARAGCGQATGGELQCPPTAGLSDGPKEPAPKGRGAQRDA
Sbjct  1    MEIALVPLENGGAMTVRGGDEARAGCGQATGGELQCPPTAGLSDGPKEPAPKGRGAQRDA  60

Query  61   DSGVRPLPPLPDPGVRPLPPLPEELPRPRRPPPEDEEEEGDPGLGTVEDQALGTASLHHQ  120
            DSGVRPLPPLPDPGVRPLPPLPEELPRPRRPPPEDEEEEGDPGLGTVEDQALGTASLHHQ
Sbjct  61   DSGVRPLPPLPDPGVRPLPPLPEELPRPRRPPPEDEEEEGDPGLGTVEDQALGTASLHHQ  120

Query  121  RVHINISGLRFETQLGTLAQFPNTLLGDPAKRLRYFDPLRNEYFFDRNRPSFDGILYYYQ  180
            RVHINISGLRFETQLGTLAQFPNTLLGDPAKRLRYFDPLRNEYFFDRNRPSFDGILYYYQ
Sbjct  121  RVHINISGLRFETQLGTLAQFPNTLLGDPAKRLRYFDPLRNEYFFDRNRPSFDGILYYYQ  180

Query  181  SGGRLRRPVNVSLDVFADEIRFYQLGDEAMERFREDEGFIKEEEKPLPRNEFQRQVWLIF  240
            SGGRLRRPVNVSLDVFADEIRFYQLGDEAMERFREDEGFIKEEEKPLPRNEFQRQVWLIF
Sbjct  181  SGGRLRRPVNVSLDVFADEIRFYQLGDEAMERFREDEGFIKEEEKPLPRNEFQRQVWLIF  240

Query  241  EYPESSGSARAIAIVSVLVILISIITFCLETLPEFRDERELLRHPPAPHQPPAPAPGANG  300
            EYPESSGSARAIAIVSVLVILISIITFCLETLPEFRDERELLRHPPAPHQPPAPAPGANG
Sbjct  241  EYPESSGSARAIAIVSVLVILISIITFCLETLPEFRDERELLRHPPAPHQPPAPAPGANG  300

Query  301  SGVMAPPSGPTVAPLLPRTLADPFFIVETTCVIWFTFELLVRFFACPSKAGFSRNIMNII  360
            SGVMAPPSGPTVAPLLPRTLADPFFIVETTCVIWFTFELLVRFFACPSKAGFSRNIMNII
Sbjct  301  SGVMAPPSGPTVAPLLPRTLADPFFIVETTCVIWFTFELLVRFFACPSKAGFSRNIMNII  360

Query  361  DVVAIFPYFITLGTELAEQQPGGGGGGQNGQQAMSLAILRVIRLVRVFRIFKLSRHSKGL  420
            DVVAIFPYFITLGTELAEQQPGGGGGGQNGQQAMSLAILRVIRLVRVFRIFKLSRHSKGL
Sbjct  361  DVVAIFPYFITLGTELAEQQPGGGGGGQNGQQAMSLAILRVIRLVRVFRIFKLSRHSKGL  420

Query  421  QILGKTLQASMRELGLLIFFLFIGVILFSSAVYFAEADNQGTHFSSIPDAFWWAVVTMTT  480
            QILGKTLQASMRELGLLIFFLFIGVILFSSAVYFAEADNQGTHFSSIPDAFWWAVVTMTT
Sbjct  421  QILGKTLQASMRELGLLIFFLFIGVILFSSAVYFAEADNQGTHFSSIPDAFWWAVVTMTT  480

Query  481  VGYGDMRPITVGGKIVGSLCAIAGVLTIALPVPVIVSNFNYFYHRETDHEEPAVLKEEQG  540
            VGYGDMRPITVGGKIVGSLCAIAGVLTIALPVPVIVSNFNYFYHRETDHEEPAVLKEEQG
Sbjct  481  VGYGDMRPITVGGKIVGSLCAIAGVLTIALPVPVIVSNFNYFYHRETDHEEPAVLKEEQG  540

Query  541  TQSQGPGLDRGVQRKVSGSRGSFCKAGGTLENADSARRGSCPLEKCNVKAKSNVDLRRSL  600
            TQSQGPGLDRGVQRKVSGSRGSFCKAGGTLENADSARRGSCPLEKCNVKAKSNVDLRRSL
Sbjct  541  TQSQGPGLDRGVQRKVSGSRGSFCKAGGTLENADSARRGSCPLEKCNVKAKSNVDLRRSL  600

Query  601  YALCLDTSRETDL  613
            YALCLDTSRETDL
Sbjct  601  YALCLDTSRETDL  613
```

The BLAST output can be viewed in three different ways: a graphic summary, a list of descriptions and a list of alignments. Each of these views presents you with different information.
The graphic summary provides the user with a quick overview of the results. A colour-coded bar shows the alignment score (colour) while the length of the bar represents the extend of the alignment. In this case we can clearly see that our search returned a number of very good hits, all are coloured in red (= very high score) and align over the entire queried region. Additionally, the BLAST search also simultaneously identifies a number of families and domains contained within the queried region. These can be found at the top of the graphic summary. In this case, the query seems to contain an ion-transporting domain as well as BTB and ion_trans 2 superfamily regions.
In the description, a more in depth overview of the results is found. In essence, it is a list of all the results that have been returned by the BLAST search, sorted by E-value. In this overview you can find the description of the hits, their max score, total score, query cover, E-value , % identity and accession number.

- The max score is the highest alignment bit-score for a matching query and database segment.
- The total score is the sum of the alignment bit-scores for all the matching query and database segments (this only differs from the max score if multiple, separate matching segments are found between the query and a database sequence).
- The E-value indicates the numbers of alignments you'd expect to find by chance with the same or higher bit-score.
- % identity tells you how many exact matches have been found between the query and the database hit.

Finally, the alignment overview shows a detailed overview of a particular database hit. Most scores that are present in the description can also be found here in addition to extra information on the alignment and the alignment itself.
For this exercise, we find a lot of hits for potassium voltage-gated channels. Based on E-value and query cover, it's impossible to discriminate between hits; however, the score and identity for the top hit are still slightly better: potassium voltage-gated channel subfamily A member 5 [Homo sapiens]; in fact, the top hit is even identical to the submitted query!

Go back to the BLASTp page. A number of options are available to fine-tune your search.



For each of the following search options, run a new BLASTp search. Find out what influence the chosen option has on the search and why? Do the results make sense? (Hint: You can check the `"Show results in a new window"` box next to the BLAST button to easily compare different searches!)

- Set the database to UniProtKB/SwissProt (keep it for the next two searches) Why would you use SwissProt over other databases?

> The major differences between these two searches is the removal of all the predicted results in the results list. The reason for this is that the SwissProt database only contains strictly curated entries, while the previous BLAST looked into all non-redundant protein sequences available, including the poorly annotated or predicted ones. The usefulness of applying this kind of restriction to the search space usually depends on the research question, but it never hurts to first look for highly annotated entries that might already be available.

- Search for related *human* sequences. What do you see and how can you explain this result?

> This search space restriction removes all non-human proteins from the results. Interestingly, the results still return a number of highly significant results for other human voltage-gated potassium channel proteins as well. This gives us a strong indication that these proteins have been strongly conserved during evolution and are very similar to each other.

- Search for all related *non-human* sequences. Take a closer look at the hits. What could explain the results?

Removing all human entries results in a list that is strongly dominated by voltage-gated potassium channel subfamily A. Top hits all come from the subfamily member 5. Upon closer inspection, we see that these hits come from related mammals such as for example the ferret, rabbit or rat. You can find this taxonomy data by clicking on the accession number on the right of the description entry. Scroll down in the resulting page to the Organism entry to find the more information about the organism from which the hit originates.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accessior
RecName: Full=Potassium voltage-gated channel subfamily A member 5; AltName: Full=Voltage-gated potassium chann...	Mustela putorius f...	1039	1039	100%	0.0	89.23%	601	P79197.1
RecName: Full=Potassium voltage-gated channel subfamily A member 5; AltName: Full=Voltage-gated potassium chann...	Oryctolagus cunic...	991	991	100%	0.0	86.53%	598	P50638.1
RecName: Full=Potassium voltage-gated channel subfamily A member 5; AltName: Full=RCK7; AltName: Full=RK4; AltN...	Rattus norvegicus	987	987	100%	0.0	86.02%	602	P19024.1
RecName: Full=Potassium voltage-gated channel subfamily A member 5; AltName: Full=Voltage-gated potassium chann...	Mus musculus	982	982	100%	0.0	86.18%	602	Q61762.2
RecName: Full=Potassium voltage-gated channel subfamily A member 3; AltName: Full=MK3; AltName: Full=Voltage-gat...	Mus musculus	653	653	69%	0.0	76.60%	528	P16390.3
RecName: Full=Potassium voltage-gated channel subfamily A member 2; AltName: Full=KC22; AltName: Full=Voltage-ga...	Oryctolagus cunic...	641	641	67%	0.0	77.11%	499	Q09081.2
RecName: Full=Potassium voltage-gated channel subfamily A member 3; AltName: Full=RCK3; AltName: Full=RGK5; Alt...	Rattus norvegicus	641	641	69%	0.0	76.36%	525	P15384.2
RecName: Full=Potassium voltage-gated channel subfamily A member 2; AltName: Full=MK2; AltName: Full=Voltage-gat...	Mus musculus	640	640	67%	0.0	77.11%	499	P63141.1
RecName: Full=Potassium voltage-gated channel subfamily A member 1; AltName: Full=RBKI; AltName: Full=RCK1; Alt...	Rattus norvegicus	637	637	71%	0.0	72.48%	495	P10499.1

B) A case study

Imagine that you are trying to grow an *E. coli* culture. After a few attempts with poor results, you decide to examine the peptide content of the culture. You find an unidentified peptide that was not present in previous, growing *E. coli* cultures.

Can you explain why the cultures grew so poorly? What database did you use during your search? Can you include/exclude some organisms? Make sure to look at the alignments and understand what all the different scores and values mean.

```
>Cell lysate peptide
AASVAVDIAYEGDWKTDGFMIGVGYKF
```

Since we already know that the peptide was not present in previous *E. coli* cultures, it probably originated from another source. As such, we can exclude *E. coli* from the list of organisms to look at. As usual, entries in the SwissProt part of the UniProt KB are more biologically/experimentally supported, so we will look here first. BLAST the peptide against the UniProt KB while excluding *E. coli* results in a list of rather bad hits. However, the top hit is still relatively good. It stems from a ds DNA virus, the *Escherichia* phage lamba. This explains why the culture wasn't growing well: it was contaminated with a phage.

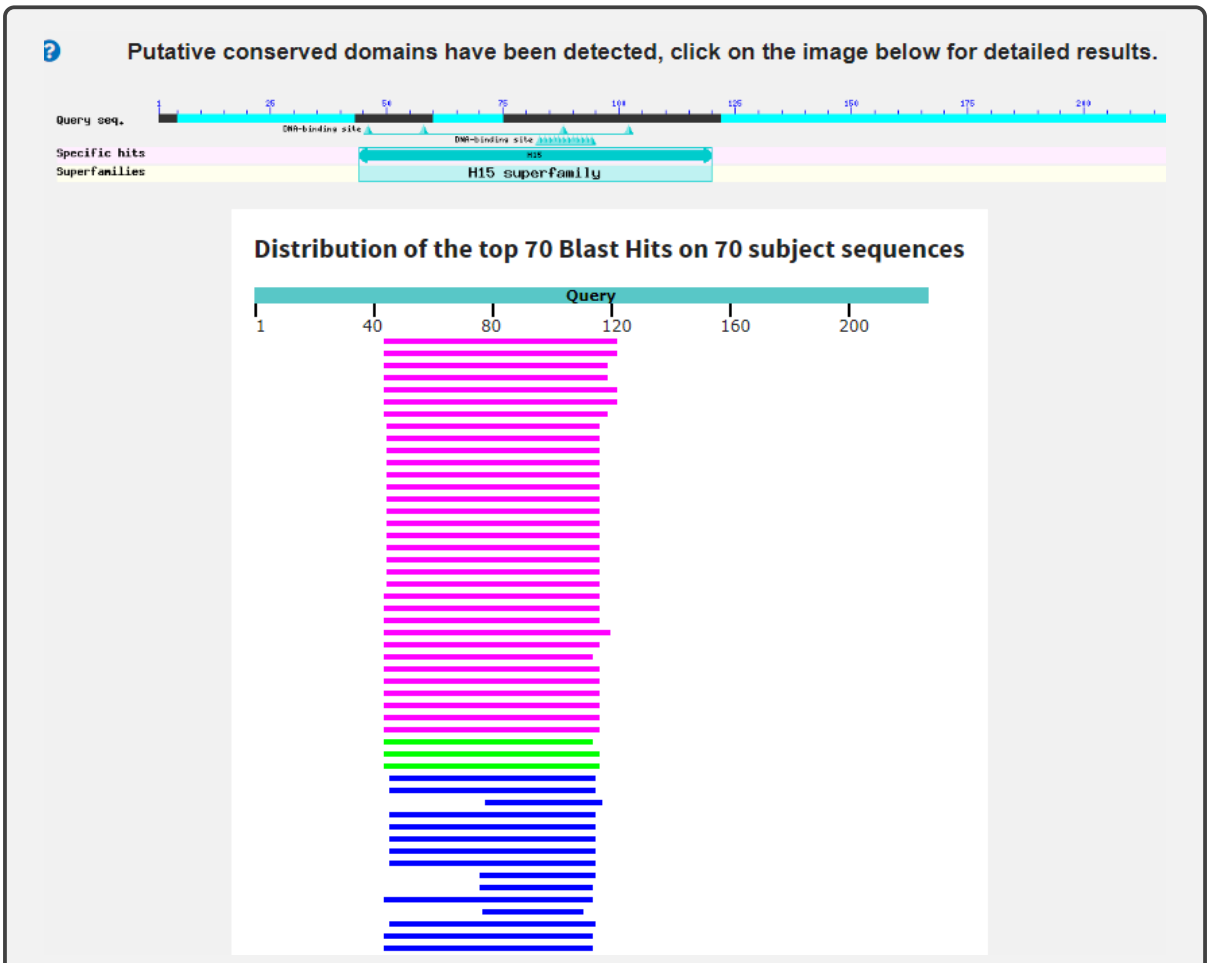
C) Algorithm parameters

In exercise A and B, we compared different searches based on the search parameters. However, the BLAST algorithm also depends on a number of algorithm parameters. These can be changed in a drop-down menu under the BLAST button.

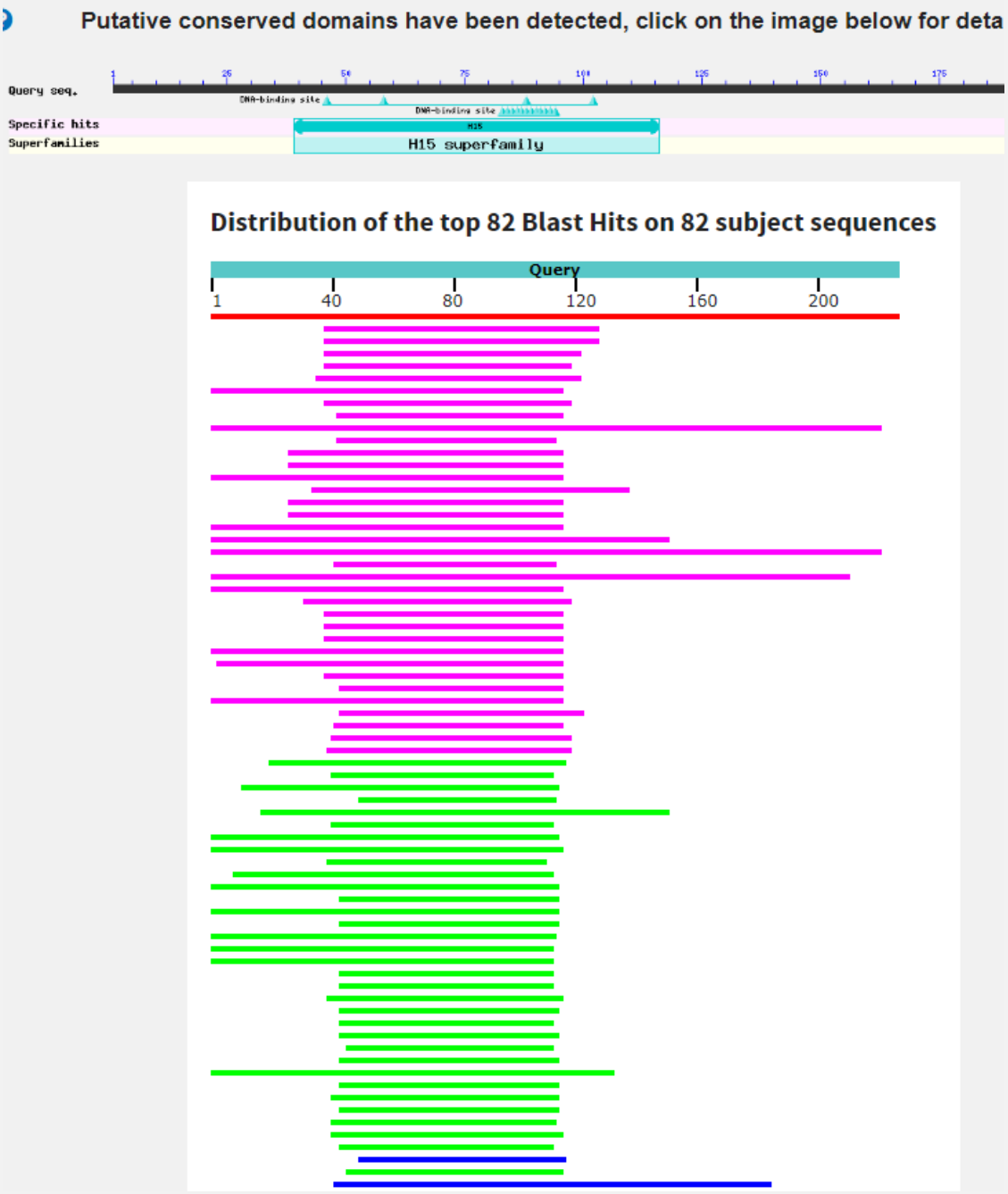
Try BLASTing the following sequence against the UniProt database. Repeat the BLAST without the `"Low complexity regions"` filter on.

```
>gi|121900|sp|P08287.2|H11L_CHICK RecName: Full=Histone H1.11L
MSETAPAPAAEAAPAAAPAPAKAAAKKPKKAAGGAKARKPAGPSVTELITKAVSASKERKGLSLAALKKA
LAAGGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFRLSKKPGEVKEKAPKKKASAAKPKKPAAKK
PAAAAKKPKKAVAVKKSPKKAKKPAASATKKSAKSPKKVTKAVKPKKAVAAKSPAKAKAVKPKAAKPKAA
KPKAAKAKKAAAKKK
```

How does this influence the search? Can you explain the difference between both results?



Graphic overview of results with filtering of low complexity regions.



Graphic overview of results without filtering out low complexity regions.

Low complexity masking removes segments of the sequence from the BLAST search that have low complexity (in essence, very repetitive content or content with a lot of the same amino acids e.g. PPCDPPPPPKDKKKKDDGPP). In our case, a disproportionately high amount of A and K is present at both borders of the sequence. Although the list of top results doesn't really change a lot in this example with or without filtering, the effect of the mask can clearly be seen in the graphic overview (only the middle of the sequence is retained). Some of the hits in the bottom figure (without mask) show extensive regions of alignment with database hits; however, regions of low complexity are usually found in multiple unrelated proteins and thus aren't very biologically informative. Therefore, E-values aren't representative in the unfiltered search.

Look at alignment between the top hit of the filtered BLAST search and the query sequence. Create your own pairwise alignment (use the tools you saw during the alignment session!) between the hit sequence and the query sequence. How does this alignment compare to the alignment created in the BLAST results?

Range 1: 1 to 225 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
370 bits(949)	6e-129	Compositional matrix adjust.	225/225(100%)	225/225(100%)	0/225(0%)

```
Query  1    MSETAPAPAAEAAPAAAPAPAKAAAKKPKKAAGGAKARKPAGPSVTELITKAVSASKERK  60
            MSETAPAPAAEAAPAAAPAPAKAAAKKPKKAAGGAKARKPAGPSVTELITKAVSASKERK
Sbjct  1    MSETAPAPAAEAAPAAAPAPAKAAAKKPKKAAGGAKARKPAGPSVTELITKAVSASKERK  60

Query  61   GLSLAALKKALAAGGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFRLSKKPGEVK  120
            GLSLAALKKALAAGGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFRLSKKPGEVK
Sbjct  61   GLSLAALKKALAAGGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFRLSKKPGEVK  120

Query  121  EKAPKKKASAAKPKKPAAKKPAAAAKKPKKAVAVKKSPKKAKKPAASATKKSAKSPKKVT  180
            EKAPKKKASAAKPKKPAAKKPAAAAKKPKKAVAVKKSPKKAKKPAASATKKSAKSPKKVT
Sbjct  121  EKAPKKKASAAKPKKPAAKKPAAAAKKPKKAVAVKKSPKKAKKPAASATKKSAKSPKKVT  180

Query  181  KAVKPKKAVAAKSPAKAKAVKPKAAKPKAAKPKAAKAKKAAAKKK  225
            KAVKPKKAVAAKSPAKAKAVKPKAAKPKAAKPKAAKAKKAAAKKK
Sbjct  181  KAVKPKKAVAAKSPAKAKAVKPKAAKPKAAKPKAAKAKKAAAKKK  225
```

Alignment without low complexity mask.

Range 1: 40 to 117 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
112 bits(281)	8e-29	Compositional matrix adjust.	78/78(100%)	78/78(100%)	0/78(0%)

```
Query  44   SVTELITKAVSASKERkglslaalkkalaaggYDVEKNNSRIKLGLKSLVSKGTLVQTKG  103
            SVTELITKAVSASKERKGLSLAALKKALAAGGYDVEKNNSRIKLGLKSLVSKGTLVQTKG
Sbjct  40   SVTELITKAVSASKERKGLSLAALKKALAAGGYDVEKNNSRIKLGLKSLVSKGTLVQTKG  99

Query  104  TGASGSFRLSKKPGEVKE  121
            TGASGSFRLSKKPGEVKE
Sbjct  100  TGASGSFRLSKKPGEVKE  117
```

Alignment with low complexity mask.

Although the alignment itself doesn't change a lot, the region that is aligned is now much shorter. This mainly affects the associated bits-score and E-value of the alignment, which are much worse with the low complexity filter on, but more accurately reflect the lack of biological significance of the results.
The pairwise alignment should result in these same alignment as the one without a complexity filter.

D) An unknown DNA sequence

Now go to the BLASTn page. Perform a MEGABLAST with the following sequence:

```
>Unknown DNA sequence
ATTCGTTCCAACACTTTTGTTGCGGAACTGAAAGGCAAACA
GCCAGGCGAAGTTGAAGTGCCGGTTATTGGCGGTCACTCTG
GTGTTACCATTCTGCCGCTGCTGTCACAGGTTCCTGGCGTTA
GTTTTACCGAGCAGGAAGTGGCTGATCTGACCAAACGTATC
CAGAACGCAGGTACTGAAGTGGTTGAAGCGAAAGCCGGTG
GTGGGTCTGCAACCCTGTCTATGGGCCAGGCAGCTGCACG
TTTTGGTCTGTCTCTGGTTCGTGCACTGCAGGGCG
```

Can you tell from which organism this sequence originates and which gene does it encode? If not, try adjusting the `Max target sequences` parameter. What is the full sequence?

We find a lot of top hits with perfect identity and equally great E-values (close to zero). These hits results from *Escherichia coli* genomes and correspond to partial coding DNA sequences (cds) for the malate dehydrogenase gene in different *E. coli* strains. Based on this sequence alone, it's not possible to further discriminate between strains.

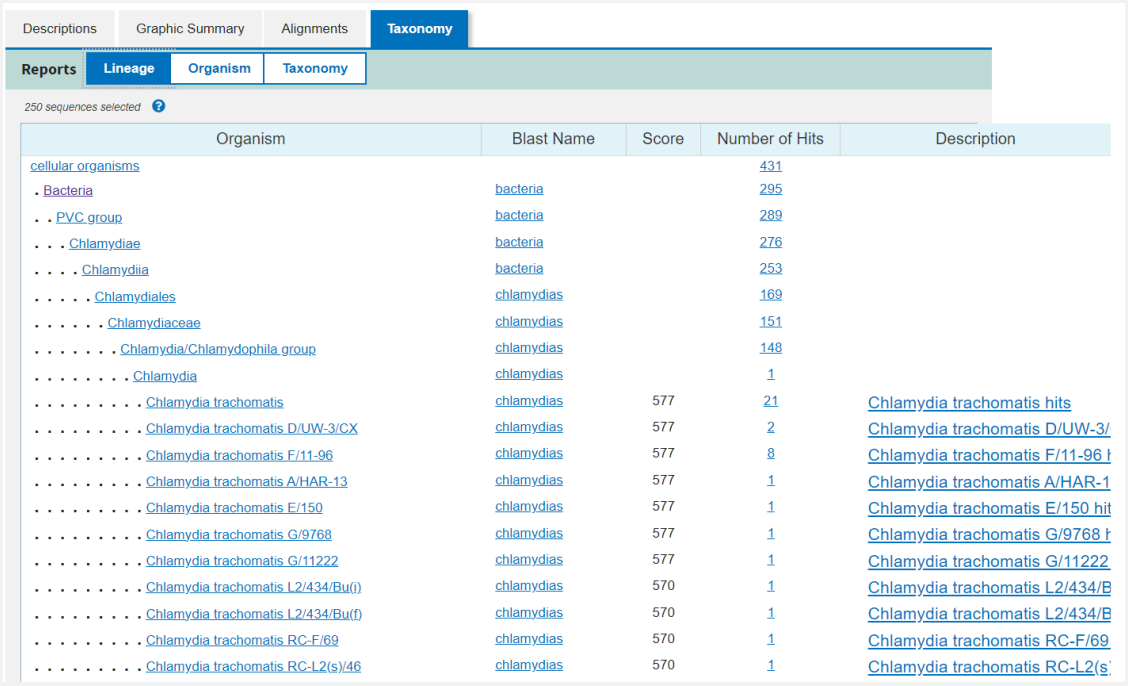
E) Another case study

You have sequenced a part of the coding DNA sequence of this *Chlamydia trachomatis* gene. Can you find out with which protein it is associated? Use a translated BLAST (BLASTx) and set the `Max target sequences` parameter to 250. Take a look at the Lineage report (select the Lineage under the Taxonomy tab). Do you see anything unexpected?

```
>Chlamydia trachomatis gene
ATGTTGAAAATTGATTTAACAGGAAAAATTGCTTTCATAGCCG
GCATAGGCGATGATAACGGGTATGGCTGGGGCATTGCCAAAA
TGTTAGCAGAAGCAGGCGCAACCATACTTGTGGGGACCTGGG
TTCCTATCTATAAAATTTTCTCTCAATCTTTGGAGTTAGGAAAA
TTCAATGCATCTCGTGAACTCTCCAATGGAGAATTGCTAACTTT
CGCTAAAATCTATCCCATGGATGCCAGTTTCGACACCCCAGAA
GATATTCCTCAGGAAATTTTGGAAAATAAACGTTACAAAGATCT
TTCTGGGTACACTGTATCCGAAGTTGTAGAACAGGTGAAAAAAC
ATTTTGGACACATTGATATTCTTGTTCACTCTTTAGCAAACAGTC
CGGAAATTGCTAAACCATTACTTGATACCTCTCGTAAAGGCTAT
CTTGCCGCCTTAAGTACATCCAGCTACTCCTTTATCAGCCTTCT
```

```
CTCTCATTTTGGCCCAATTATGAATGCAGGAGCTAGCACCATCT
CTCTAACTTATCTTGCTTCCATGCGTGCTGTTCCAGGGTATGGCG
GAGGAATGAACGCAGCAAAAGCTGCTTTAGAAAGTGATACAAAA
GTACTGGCTTGGGAAGCCGGCCGACGTTGGGGAGTCCGAGTGAA
TACTATCTCGGCAGGGCCATTAGCTAGCCGTGCAGGAAAAGCTA
TTGGATTTATTGAGAGAATGGTGGATTACTACCAAGACTGGGCTC
CACTACCTTCTCCAATGGAAGCTGAGCAAGTAGGCGCAGCAGCA
GCCTTCTTAGTCTCTCCCCTAGCTAGCGCAATTACGGGAGAAACT
CTCTATGTGGATCACGGAGCCAATGTGATGGGCATAGGTCCAGAA
ATGTTTCCTAAGGATTAA
```

Our top hit indicates that this gene encodes for the enoyl-ACP reductase protein. The lineage report shows that a lot of hits map to *Chlamydia trachomatis* strains and even other *Chlamydia species*, but the most surprising aspect of this taxonomy report is that we can also find that this gene has hits in a number of plant species! One of the possible explanations for this phenomenon could be horizontal gene transfer.

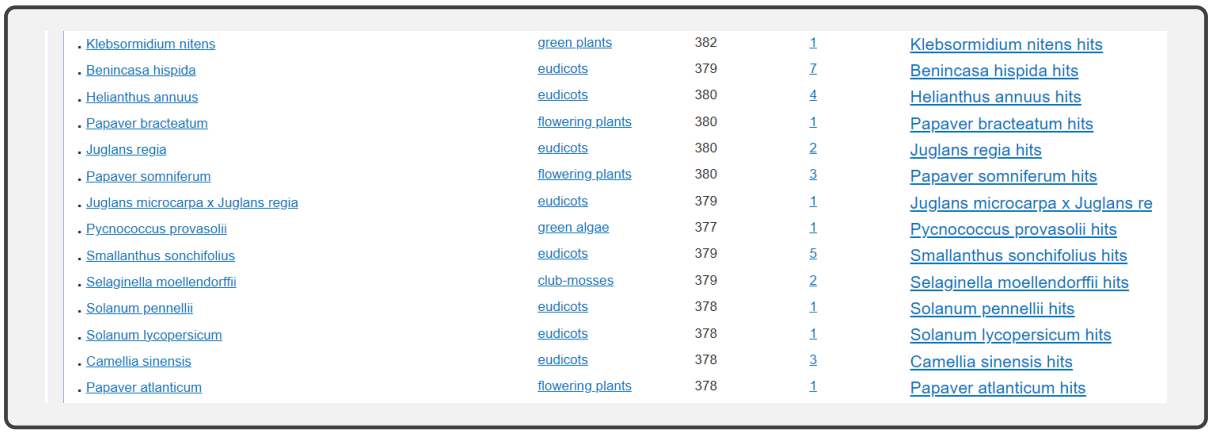


Descriptions | Graphic Summary | Alignments | **Taxonomy**

Reports | **Lineage** | **Organism** | **Taxonomy**

250 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
cellular organisms			431	
. Bacteria	bacteria		295	
. . PVC group	bacteria		289	
. . . Chlamydiae	bacteria		276	
. . . . Chlamydiia	bacteria		253	
. . . . . Chlamydiales	chlamydias		169	
. . . . . Chlamydiaceae	chlamydias		151	
. . . . . Chlamydia/Chlamydophila group	chlamydias		148	
. . . . . Chlamydia	chlamydias		1	
. . . . . Chlamydia trachomatis	chlamydias	577	21	Chlamydia trachomatis hits
. . . . . Chlamydia trachomatis D/UW-3/CX	chlamydias	577	2	Chlamydia trachomatis D/UW-3/
. . . . . Chlamydia trachomatis F/11-96	chlamydias	577	8	Chlamydia trachomatis F/11-96 h
. . . . . Chlamydia trachomatis A/HAR-13	chlamydias	577	1	Chlamydia trachomatis A/HAR-1
. . . . . Chlamydia trachomatis E/150	chlamydias	577	1	Chlamydia trachomatis E/150 hit
. . . . . Chlamydia trachomatis G/9768	chlamydias	577	1	Chlamydia trachomatis G/9768 h
. . . . . Chlamydia trachomatis G/11222	chlamydias	577	1	Chlamydia trachomatis G/11222
. . . . . Chlamydia trachomatis L2/434/Bu(j)	chlamydias	570	1	Chlamydia trachomatis L2/434/B
. . . . . Chlamydia trachomatis L2/434/Bu(f)	chlamydias	570	1	Chlamydia trachomatis L2/434/B
. . . . . Chlamydia trachomatis RC-F/69	chlamydias	570	1	Chlamydia trachomatis RC-F/69
. . . . . Chlamydia trachomatis RC-L2(s)/46	chlamydias	570	1	Chlamydia trachomatis RC-L2(s)



Klebsormidium nitens	green plants	382	1	Klebsormidium nitens hits
Benincasa hispida	eudicots	379	7	Benincasa hispida hits
Helianthus annuus	eudicots	380	4	Helianthus annuus hits
Papaver bracteatum	flowering plants	380	1	Papaver bracteatum hits
Juglans regia	eudicots	380	2	Juglans regia hits
Papaver somniferum	flowering plants	380	3	Papaver somniferum hits
Juglans microcarpa x Juglans regia	eudicots	379	1	Juglans microcarpa x Juglans re
Pycnococcus provasolii	green algae	377	1	Pycnococcus provasolii hits
Smallanthus sonchifolius	eudicots	379	5	Smallanthus sonchifolius hits
Selaginella moellendorffii	club-mosses	379	2	Selaginella moellendorffii hits
Solanum pennellii	eudicots	378	1	Solanum pennellii hits
Solanum lycopersicum	eudicots	378	1	Solanum lycopersicum hits
Camellia sinensis	eudicots	378	3	Camellia sinensis hits
Papaver atlanticum	flowering plants	378	1	Papaver atlanticum hits

Home exercise: Unknown species

Chicken or the egg?

Fragments of a fossil, that were not enough to identify the bird-like species, were found and analysed with mass spectrometry. You were given the sequence and asked to analyse this. What are the outcomes and how would you report this? What search are you using and are you specifying any database?

```
>Unkown Sequence
GLPGESGAVGPAGPIGSR
GVQPPGPQGPR
```

To confirm some of their hypotheses, they have sent you some analysed samples from curated sources.
> B. canadensis
GLPGESGAVGPAGPPGSR
> S. camelus
GPPGESGAVGPAGPIGSR
> G. gallus
GLPGESGAVGPAGPIGSR

Based on this information, would you change your interpretation and why?

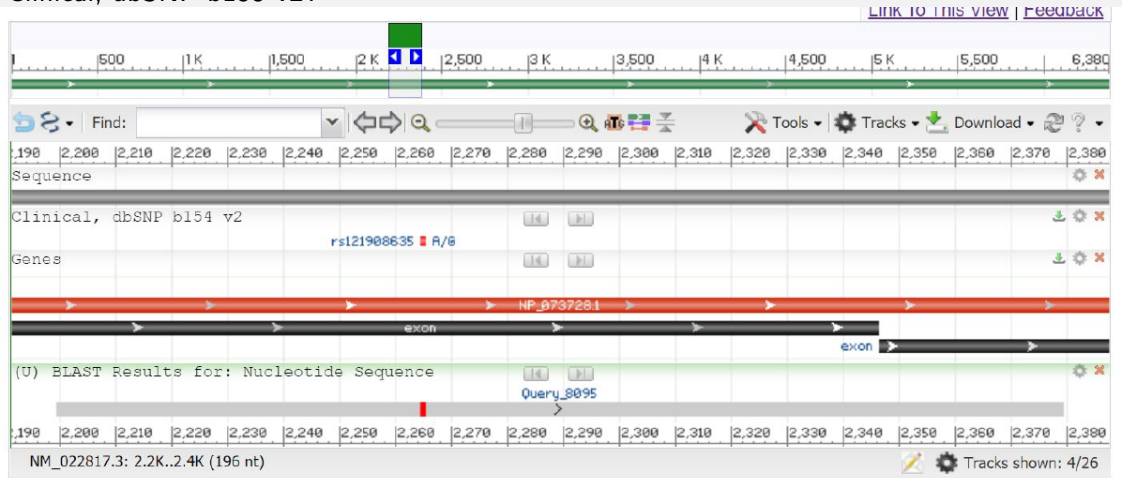
Based on the first sequence you get a hit:
`Full=Collagen alpha-2(I) chain; AltName: Full=Alpha-2 type I collagen [Tyrannosaurus rex] with accession number: P0C2W4.1.`
Suggesting that this might be an actual fossil of a T-rex. However, if we align this with some known species, and also in the BLAST we see a high number of alignments that have a 100%. This is also the case when an alignment is made with the other three species. Collagen is highly similar. And making a conclusion on a species is difficult and it can not be said that this is indeed a T. rex.

Sleep deprived
A research group identified a gene from patients with disturbed sleeping patterns:

```
gggtgaacag ccgcacggga gtaggtacgc acctgacctc gctggcactg
ccgggcaagg cagagggtgt ggcgtcgctc accagccagt gcagctacag
cagcaccatc gtccatgtgg gagacaagaa gccgcagccg gagttagaga
tggtggaaga tgctgcgagt gggccagaat
```

Perform a BLAST and identify which single nucleotide polymorphisms this patient carries. Does this cause a difference on the protein sequence that the patient expresses? See if the gene you identified can really cause sleep disorder.

A Blastn search of this gene results in a good hit that is relevant for this question: *Homo sapiens period circadian regulator 2 (PER2), mRNA)*. When clicking on 'graphics' for this entry (go to Alignments and click on 'Graphics' of the first result). You will see a typical genome browser with the query sequence and the snp indicated with a red mark.
A fast method to check the clinical significance of this snp is to load a track that shows all dbnsp entries of clinical significance (via tracks shown (bottom right) -> Variation -> Check 'Clinical, dbSNP b155 v2'.



If you look up this dSNP entry (*rs121908635*), you will see that this SNP results in *Familial advanced sleep phase syndrome 1*.

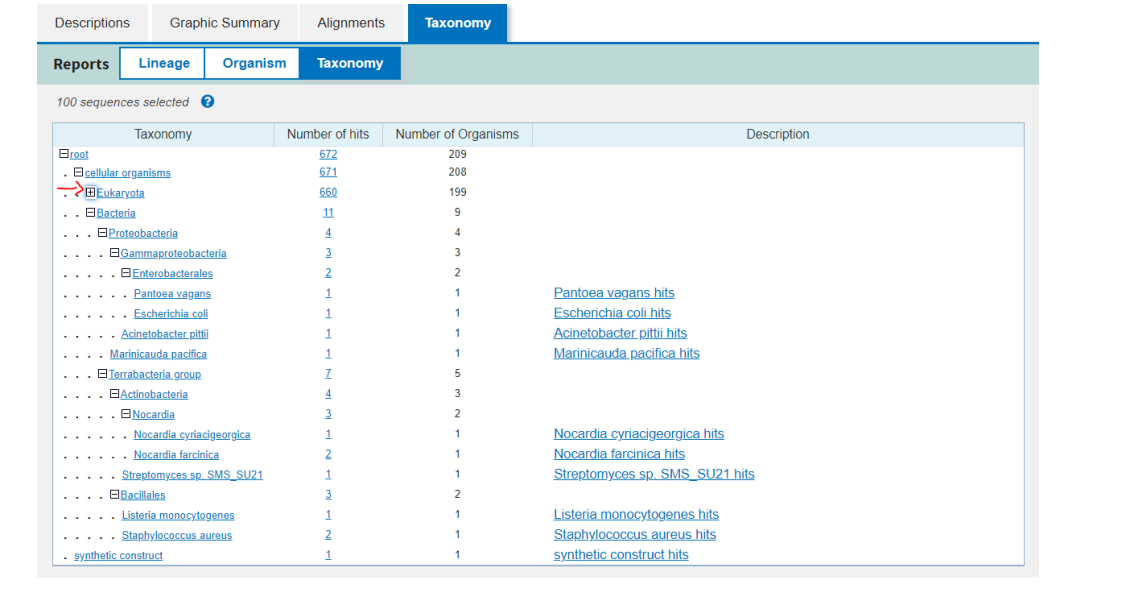
Ubiquitin

Ubiquitin is a regulatory protein that is ubiquitously expressed in eukaryotes. Ubiquitination (or ubiquitylation) refers to the post-translational modification of a protein by the covalent attachment (via an isopeptide bond) of one or more ubiquitin monomers. The most prominent function of ubiquitin is labeling proteins for proteasomal degradation. Besides this function, ubiquitination also controls the stability, function, and intracellular localization of a wide variety of proteins.
Do you think it is a protein found only in vertebrates? Use BLAST to find out how conserved ubiquitin is. As a start point use this human ubiquitin sequence:

```
>human ubiquitin
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
```

Using BLAST find out whether or not ubiquitin is expressed only by eukaryotes.

Take a look at the taxonomy view of your blastp search. Ubiquitin is almost exclusively found in eukaryotes. Some prokaryotes harbour sequences that show some homology to human ubiquitin, but very rarely.



Taxonomy	Number of hits	Number of Organisms	Description
⊟root	672	209	
. ⊟cellular organisms	671	208	
. . ⊞Eukaryota	660	199	
. . ⊟Bacteria	11	9	
. . . ⊟Proteobacteria	4	4	
. . . . ⊟Gammaproteobacteria	3	3	
. . . . . ⊟Enterobacterales	2	2	
. . . . . . Pantoea vagans	1	1	Pantoea vagans hits
. . . . . . Escherichia coli	1	1	Escherichia coli hits
. . . . . Acinetobacter pittii	1	1	Acinetobacter pittii hits
. . . . Marinicauda pacifica	1	1	Marinicauda pacifica hits
. . . ⊟Terrabacteria group	7	5	
. . . . ⊟Actinobacteria	4	3	
. . . . . ⊟Nocardia	3	2	
. . . . . . Nocardia cyriacigeorgica	1	1	Nocardia cyriacigeorgica hits
. . . . . . Nocardia farcinica	2	1	Nocardia farcinica hits
. . . . . Streptomyces sp. SMS_SU21	1	1	Streptomyces sp. SMS_SU21 hits
. . . . ⊟Bacillales	3	2	
. . . . . Listeria monocytogenes	1	1	Listeria monocytogenes hits
. . . . . Staphylococcus aureus	2	1	Staphylococcus aureus hits
. synthetic construct	1	1	synthetic construct hits

Section 5: Sequence motifs

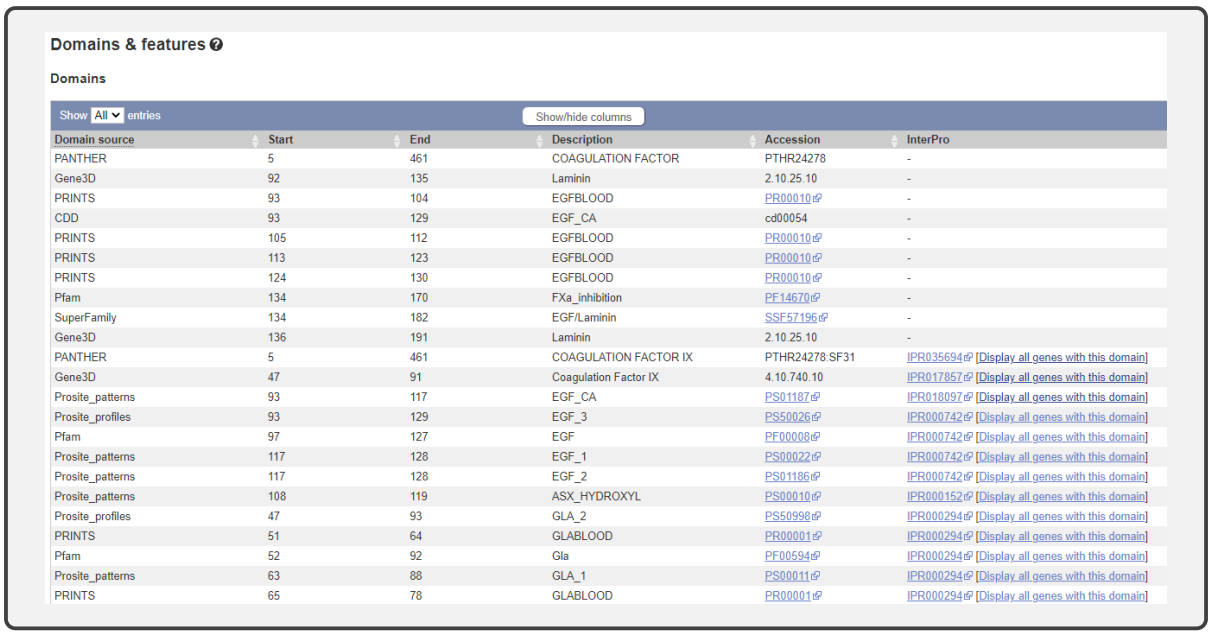
Exercise 5.1: Basic motif exercises

Multiple databases can be used to identify motifs and domains in biological sequences. During the following exercises, we will look into using some of these databases to find out more about these motifs and domains.

A) Identifying the genomic location of domains & motifs.

Ensembl can be found at `http://www.ensembl.org`.

Look for the human F9-201 transcript in the Ensembl database and click the right result. On the left side, select `"Domains & features"` to show all the annotated domains within this transcript. Choose one of the domains and display all the genes encoding for proteins with the same domain (hint: IPR018097 works good in this case). How many genes can you find with this domain on chromosome 22?



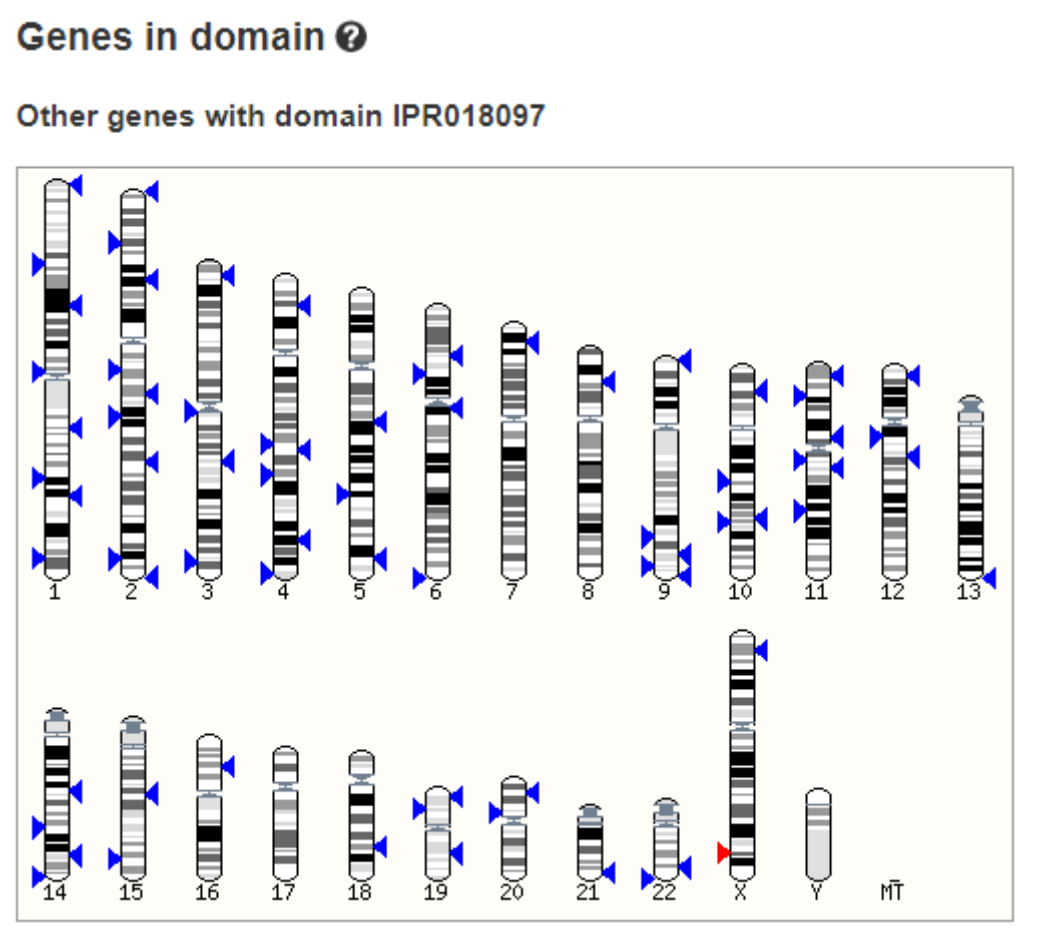
Domains & features

Domains

Show All entries — Show/hide columns

Domain source	Start	End	Description	Accession	InterPro
PANTHER	5	461	COAGULATION FACTOR	PTHR24278	-
Gene3D	92	135	Laminin	2.10.25.10	-
PRINTS	93	104	EGFBLOOD	PR00010	-
CDD	93	129	EGF_CA	cd00054	-
PRINTS	105	112	EGFBLOOD	PR00010	-
PRINTS	113	123	EGFBLOOD	PR00010	-
PRINTS	124	130	EGFBLOOD	PR00010	-
Pfam	134	170	FXa_inhibition	PF14670	-
SuperFamily	134	182	EGF/Laminin	SSF57196	-
Gene3D	136	191	Laminin	2.10.25.10	-
PANTHER	5	461	COAGULATION FACTOR IX	PTHR24278:SF31	IPR035694 [Display all genes with this domain]
Gene3D	47	91	Coagulation Factor IX	4.10.740.10	IPR017857 [Display all genes with this domain]
Prosite_patterns	93	117	EGF_CA	PS01187	IPR018097 [Display all genes with this domain]
Prosite_profiles	93	129	EGF_3	PS50026	IPR000742 [Display all genes with this domain]
Pfam	97	127	EGF	PF00008	IPR000742 [Display all genes with this domain]
Prosite_patterns	117	128	EGF_1	PS00022	IPR000742 [Display all genes with this domain]
Prosite_patterns	117	128	EGF_2	PS01186	IPR000742 [Display all genes with this domain]
Prosite_patterns	108	119	ASX_HYDROXYL	PS00010	IPR000152 [Display all genes with this domain]
Prosite_profiles	47	93	GLA_2	PS50998	IPR000294 [Display all genes with this domain]
PRINTS	51	64	GLABLOOD	PR00001	IPR000294 [Display all genes with this domain]
Pfam	52	92	Gla	PF00594	IPR000294 [Display all genes with this domain]
Prosite_patterns	63	88	GLA_1	PS00011	IPR000294 [Display all genes with this domain]
PRINTS	65	78	GLABLOOD	PR00001	IPR000294 [Display all genes with this domain]

The Ensembl domains & features page for F9-201 shows a number of entries for multiple domain types together with their location within the transcript. By clicking on the display all genes with this domain, you can find all the genomic locations within the human genome where this domain is encoded. For example, the EGF-CA domain can be found dispersed throughout the genome, as seen below:



The location of the F9-201 transcript, from which we started, is indicated in red while the other sites are marked in blue. Scrolling down on this page also returns a list of the genes that encode for this domain together with its genomic location.
The genomic location can be used to find all genes with the same EGF-CA domain on chromosome 22:

ENSG00000159307	Chromosome 22:43197280-43343372	SCUBE1	signal peptide, CUB domain and EGF like domain containing 1
ENSG00000077942	Chromosome 22:45502238-45601135	FBLN1	fibulin 1 [Source:HGNC Symbol;Acc:HGNC:3600]
ENSG00000184164	Chromosome 22:49918167-49927540	CRELD2	cysteine rich with EGF like domains 2 [Source:HGNC Symbol;A

B) Prosite: A curated and annotated protein functional profiles and patterns database.

`http://prosite.expasy.org/scanprosite/` is a useful place to start that allows multiple domain related questions to be answered. We will explore the default example, the human enterokinase protein. Enter its UniProt accession number (P98073) into the white box and start the Prosite scan.

After searching Prosite for patterns and profiles within the sequence of human Enterokinase, we find 13 hits within its sequence: 8 profiles and 5 patterns.

Give an overview of all the different *profiles* and *patterns* found in the enterokinase protein.

The following distinct profiles were found:

- SEA domain profile
- LDL-receptor class A (LDLRA) domain profile (2 times)
- CUB domain profile (2 times)
- MAM domain profile
- SRCR domain profile
- Serine proteases, trypsin domain profile

The following distinct patterns were found:

- LDL-receptor class A (LDLRA) domain signature
- MAM domain signature
- Serine proteases, trypsin family, histidine active site
- Serine proteases, trypsin family, serine active site

Click on one of the profiles (for example, the TRYPSIN_DOM) to find an extensive description of the profile. You can do the same for the patterns. Find the following information:

- Which proteins are known to contain a SEA domain?

  A list with some proteins known to contain a SEA domain can be obtained:
  - Vertebrate agrin
  - Mammalian enterokinase
  - Animal perlecan
  - Some vertebrate epithelial mucins
  - Mammalian cell surface antigen 114/A10

- What is the pattern used to represent the LDLRA_1 pattern?

We can find this information by clicking on the accession number of the pattern in the upper left corner of its entry box.

PS01209 **LDLRA_1** *LDL-receptor class A (LDLRA) domain signature* :

197 - 221:	[confidence level: (0)]	`CIkadlf.CDgevNCpdgsDEDnkm..C`
655 - 677:	[confidence level: (0)]	`CVplvnl.CDghlHCedg.SDEad...C`

In the technical section, we can find the following information on the pattern used to represented LDLRA_1:

Consensus pattern:

```
C-[VILMA]-x(5,6)-C-[DNH]-x(3)-[DENQHT]-C-x(3,4)-[STADEW]-
[DEH]-[DE]-x(1,5)-C
```

The 4 `C`s are involved in disulfide bonds

- How many proteins in the UniProt knowledge base are known to belong to the SRCR_2 profile?

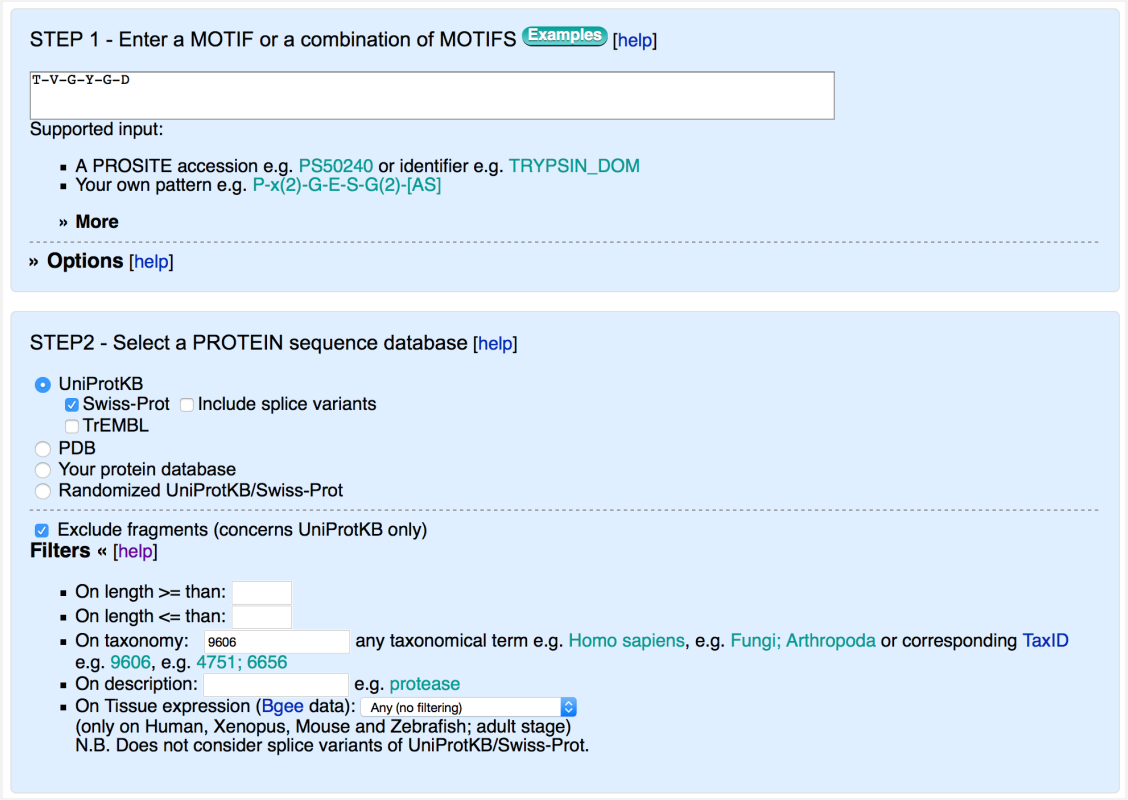
Similarly to the previous exercise, click on the accession number for the SRCR_2 profile. Again, we can find the information we are looking for in the technical section:

- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 104
- detected by PS50287: 98 (true positives)
- undetected by PS50287: 6 (6 false negatives and 0 `'partial'`)

Another way of scanning the contents of Prosite is by inputting your own sequence pattern. This is particularly useful for finding motifs that you have defined yourself.

Suppose you want to find all proteins containing the following, very simple, motif: `T-V-G-Y-G-D`. Make sure to not include isoforms and to exclude fragments. We will further refine our search by only looking for motifs in the human proteome (taxonomy number 9606). Can you find a link between the resulting hits?

Your input window should look similar to the one below. Take note that multiple extra search and filter options are available, though we only use a few of those.



This search results in a list of 23 hits in 23 sequences. Almost all of those hits are voltage-gated potassium channels. The TVGYGD sequence is a part of the strictly conserved selectivity filter of the channel that is required to maintain its selectivity for potassium ions. We also find a calcium-activated potassium channel and the septin-14 protein, which seems to be completely unrelated to potassium channels.

You can download all the matched sequences from UniProt by clicking the "`Matched UniProtKB entries`" link at the bottom of the page.

C) Interpro: A collection of other databases.

While Prosite specializes itself in gathering information on functional profiles and patterns, InterPro integrates several sequence signature databases such as Prosite, Pfam, Blocks, ProDom and PRINTS. You can find InterPro at `http://www.ebi.ac.uk/interpro/`. Searching InterPro is really simple! Just enter a protein sequence into the search box and hit search.

Using InterPro, analyze the following unknown squirrel protein:

```
>squirrel_seq
MALPARLVPLCCLALLALPAQSCGPGRGPVGRRRYVRKQLVPLLYKQ
FVPSVPERTLGASGPAEGRVARGSERFRDLVPNYNPDIIFKDEENSG
ADRLMTERCKERVNALAIAVMNMWPGVRLRVTEGWDEDGHHAQDSLH
YEGRALDITTSDRDRNKYGLLARLAVEAGFDWVYYESRNHVHVSVKA
GTVGGGCFRETEAAQLWGDARGLRELHRAWVLAADAAGRVVPTPVLL
FLDRDLQRRASFVAVETERPPRKLLLTPWHLVFAARGPAPAPGDFAP
VFARRLRAGDSVLAPGGDALRPARVARVAREEAVGVFAPLTAHGTLL
VNDVLASCYAVLESHQWAHRAFAPLRLLHALGALLPGGAVQPTGMHW
YSRFLYRLAEELLG
```

Look for clues about the protein's function within the domains and features of its sequence and find out what protein family it belongs to.

Along with a number of unintegrated signatures, we find that the protein is part of the hedgehog protein family and contains a number of annotated domains. The description of the family tells us:

> Hedgehog proteins are a family of secreted signal molecules required for embryonic cell differentiation. They are synthesised as inactive precursors with an N-terminal signalling domain linked to a C-terminal autoprocessing domain. The three-dimensional structure of the autolytic domain of the hedgehog protein of shows similarity with the beta-strand core of intein splicing domains. It has hence been termed the hint (Hedgehog/Intein) domain [PMID: 9489693].
>
> The activated hedgehog protein is involved in segment polarity and cell to cell communication, and plays important roles in both early embryogenesis and metamorphosis [*PMID: 8166882*, *PMID: 1280560*, *PMID: 1394430*, *PMID: 1340474*].
>
> The protein is expressed in embryo stripes [PMID: 8166882], and in several groups of cells belonging to the fore-gut, hind-gut and various other unidentified tissues. Maximum expression is seen in embryos 6-12 hours after fertilization, and in pupae 1-24 hours after puparium formation.

The protein is made up of an N-terminal signaling domain and a C-terminal Hint domain. You can click on their entries to find more information about each of them.

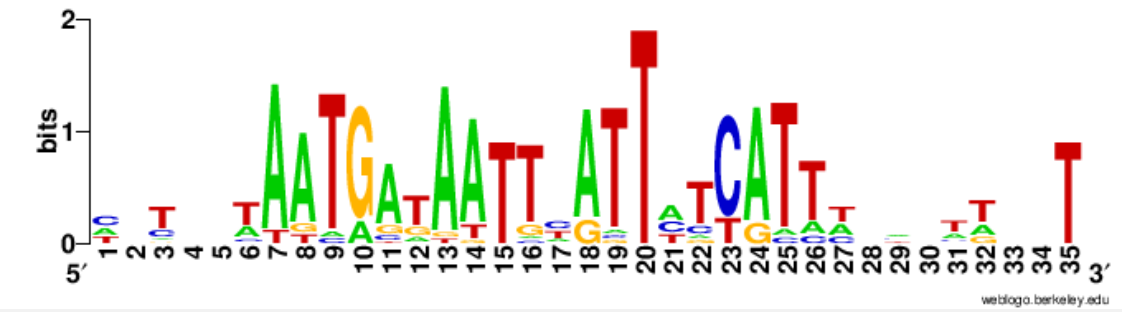
Exercise 5.2: Motif Visualization and Searching

A) Visualisation

The Weblogo tool is an easy way to visualize conserved motives in a set of sequences. The tool calculates the frequency of each base in each position and makes a visualization of this. Note that the sequences have to be aligned for this. If the sequences are NOT aligned, you should perform a multiple sequence alignment first, or use a *de novo* motif discovery tool.

- Navigate to `http://weblogo.berkeley.edu/` and select `"create"`.
- Download the file `"motifseq.fasta"` and open with a text editor (e.g. Notepad or 'Kladblok'), select all sequences and copy them to the input box on the top of the screen.
- Under Advanced logo options: Sequence type, select DNA/RNA
- Press create logo in the bottom.

Using the sequences from `"motifseq.fasta"` will result in an error message because not all sequences have an equal length. You will need to perform a multiple sequence alignment first. This can be done with a tool of your choice (find one in the exercises from previous week if you don't remember). Inputting the motif (e.g. in the ClustalW format) should give you a plot that looks similar to this:

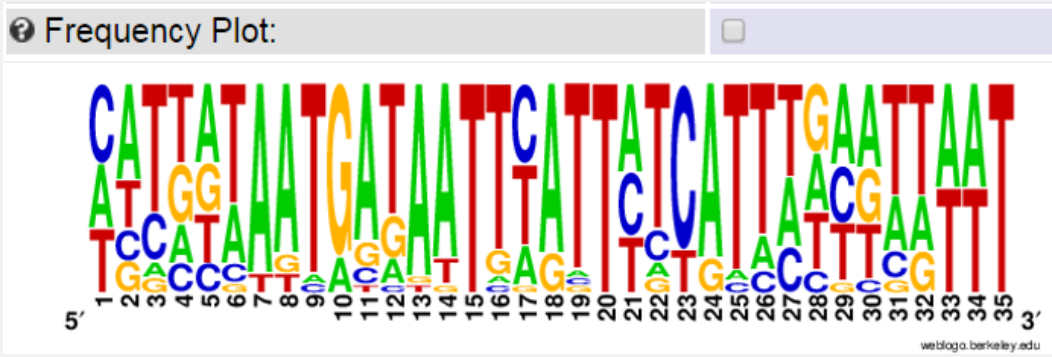


- Note that the y-axis is expressed in a bit score, what does this mean?

  The bit score stands for the (^{2}log of the) maximum entropy for the given sequence position. The maximum entropy you can have is exclusively 1 base on a certain position (perfectly conserved), in which case the bit score here is 2. If you have 50% of one base and 50% of another, the bit score is 1, etc. More you do not have to know about this (the definition of entropy is quite complex). In short, it is a measure for conservation at each site.

- You can change the y-axis by highlighting the option `"frequency plot"` What does this change to your plot? What is the benefit of bit scores versus frequency and vica versa?

  Using bit scores provides additional information about sequence conservation (information content at each site), which is lost in the frequency plots where the stack height is always the same (1).

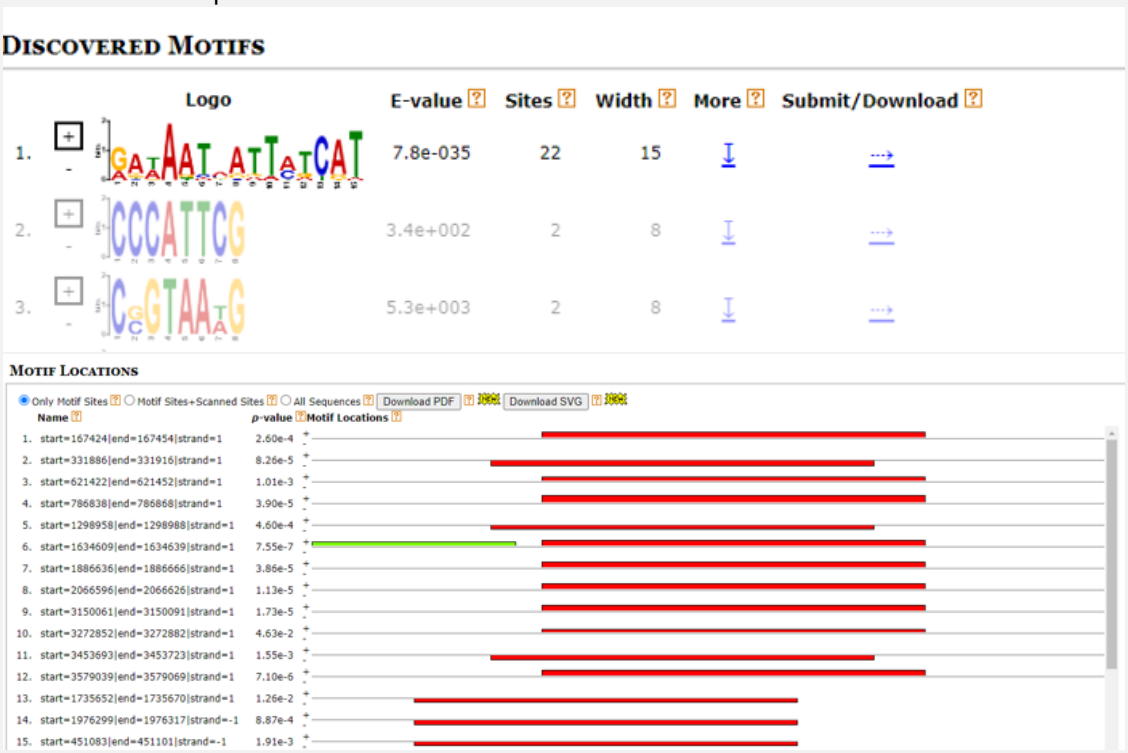


B) Searching Suppose now you want the find motives in the file `"motifseq2.fasta"`. These sequences were sent to you by a colleague who does not always respect the best practices. Therefore, it will not directly be accepted by the MEME tool, so you will have to do a little bit of very common debugging.

- Navigate to `http://meme-suite.org/tools/meme`
- Use MEME, a *de novo*, motif discovery tool, to try to find a motif in these sequences. You can leave the general settings on their defaults.

What motives can you find? Is there 1 motif that clearly jumps out? Look closely at both types of output measures to justify your conclusion. Explore the output of the tool!

Your colleague used spaces instead of underscores in the description lines of the FASTA sequences. This results in duplicate ID's. Performing a 'find and replace' in your text editor should solve this problem.



The first motif has a very low E value and clearly jumps out. This motif occurs in more than 80% of our sequences (22 out of 26). You have to look at a combination of significance and how often a motif occurs.
The most interesting feature is the motif distribution plot. Using this plot, you can check for every sequence where exactly the found motif occurs.

Homework

Human heat shock proteins are indispensable for the cell to cope with stress, and play roles in various processes such as response to toxic compounds and UV, starvation and inflammation.
We are interested in how these proteins are regulated, and therefore downloaded their promoter regions (we already did this for you: HSP_promoters.fasta). Those sequences precede the actual gene, and typically have important gene regulatory functions (e.g. attract proteins needed for gene transcription, give signals on proper timing of transcription etc.). We wonder whether there is a common motif present in these promoter sequences, and if so, what the biological function of this

motif would be.

Interpretation hint: have a look at common promoter motifs on this website: gpminer

Since promoter regions do not necessarily align well, tools relying on MSA such as Weblogo are not useful in this case. However, we can use de novo motif discovery instead. When we enter the promoter sequences in the MEME suite and use standard parameters, the following output is found:

DISCOVERED MOTIFS

	Logo	E-value	Sites	Width	More	Submit/Download
1.		2.9e-003	9	10		
2.		3.7e+001	7	10		
3.		4.9e+003	6	13		

Stopped because requested number of motifs (3) found.

Only the upper one has a decent E value, and can be considered as a potential motif. If we have a look at the gpminer website, a number of typical promoter motifs are listed, among which the TATA box, which can be recognized in this motif. A quick google search learns us the following: "The TATA motif is the binding site for the TATA-binding protein and other transcription factors in some eukaryotic genes. From here on, gene transcription starts."