Bioinformatics Project 3 - Clustering and Phylogeny

09.12.2024

The goal of the project is to conduct a comprehensive phylogenetic analysis and submit a report with a detailed description of the methods applied, the datasets used, and the results obtained.

1. Clustering:
   (a) Choose an algorithm to work with.
   (b) Create your own dataset:
      i. choose 8 different human proteins,
      ii. run BLAST (online version) on these proteins, and for each protein, download the sequences for seven different organisms. Be careful not to select identical sequences—refer to the tutorial below for guidance,
      iii. remember to include the identifiers of the sequences used in your report, for example, in the form of a table,
   (c) use BLAST (local version) or another comparison method (e.g., MSA) to obtain input data (similarity between sequences). Select the appropriate input format for your chosen clustering method,
   (d) cluster the provided sequences.
   (e) Check the results. Do the clusters correspond to similar proteins identified by BLAST?
2. Phylogenetics:
   (a) use this module to build trees in three different ways:
      i. separate tree for each "group" of proteins
      ii. separate tree for each cluster
      iii. one common tree for all downloaded sequences
   (b) create consensus trees from your two approaches (clusters vs. the groups of proteins downloaded together).
   (c) create different visualizations (from the two consensus trees and the single tree generated from all sequences) to compare the results and draw conclusions.
      i. Color tree branches based on the organism
      ii. Color tree branches based on the protein "group"
   (d) What is your observation? Which approach seems to work best?
   (e) Are the trees similar to each other? Does the evolution look exactly the same in different trees?

3. Report

   (a) The grade will be mainly based on the description of the research tasks and results.

   (b) Prepare a detailed report of your work.

   (c) Include the following points:

      i. Description of your dataset (organisms and proteins that you chose).

      ii. Short description of the methods and algorithms that you used.

      iii. Description of the steps of your analysis.

      iv. Description of your results and conclusions from the project.

      v. Figures showing consensus trees and examples of trees from your groups and clusters (see 2e).

      vi. Tree generated from all the sequences (see 2e).

      vii. Answers to questions 2f and 2g.

4. Grading

   (a) max 10 points

   (b) Dataset preparation 2p

   (c) Clustering 3p

   (d) Constructing trees inside groups and clusters 3p

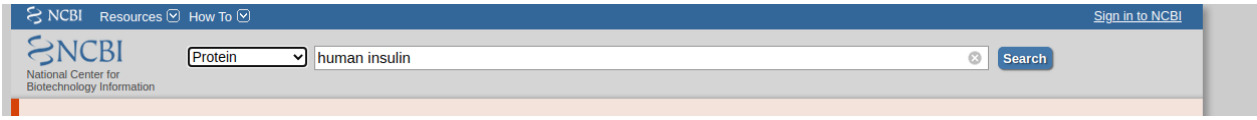
   (e) Consensus trees 2p

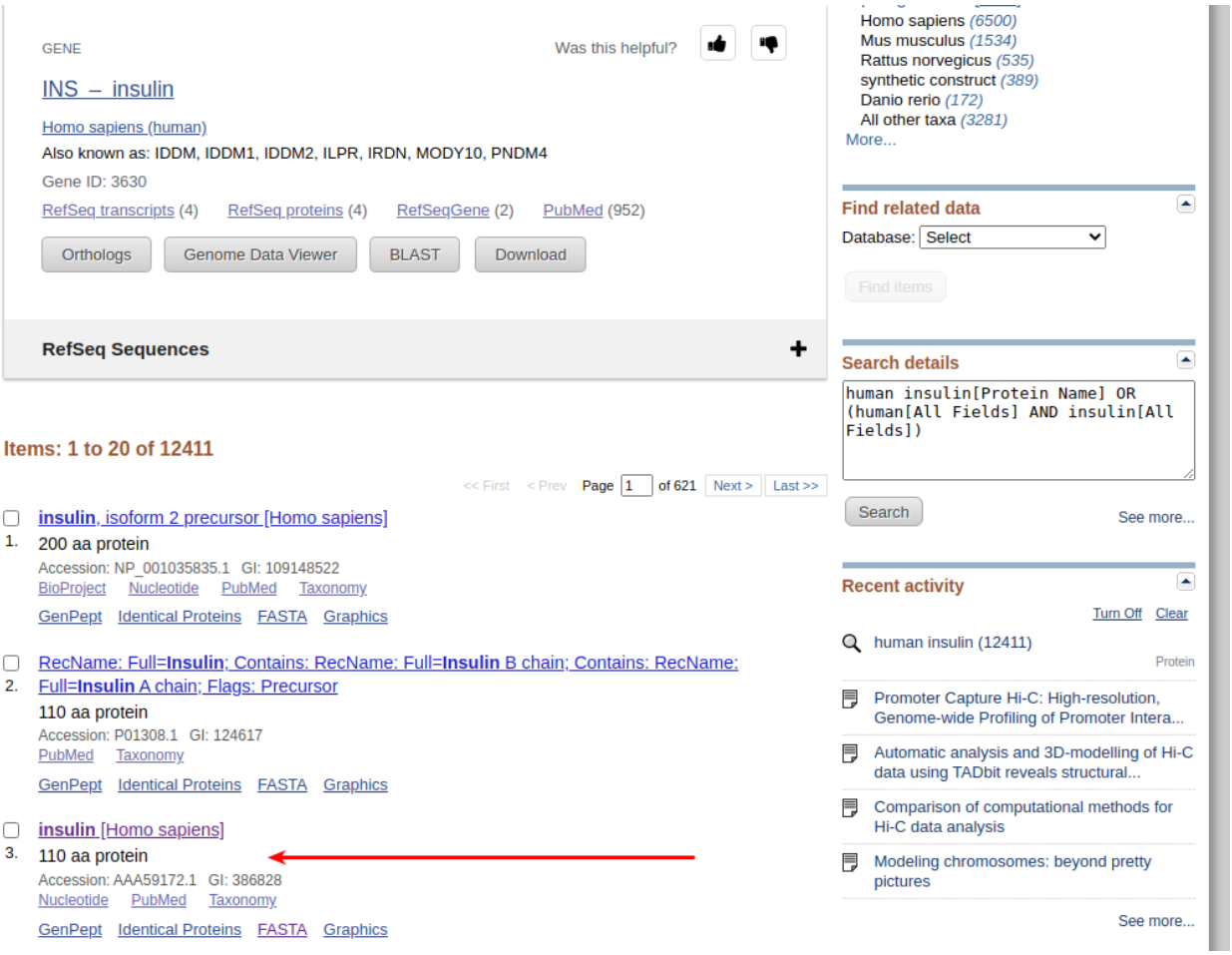
   (f) All of the above will be graded based on both your description in the report and included code.

   (g) five points can be deducted for a careless and inaccurate report

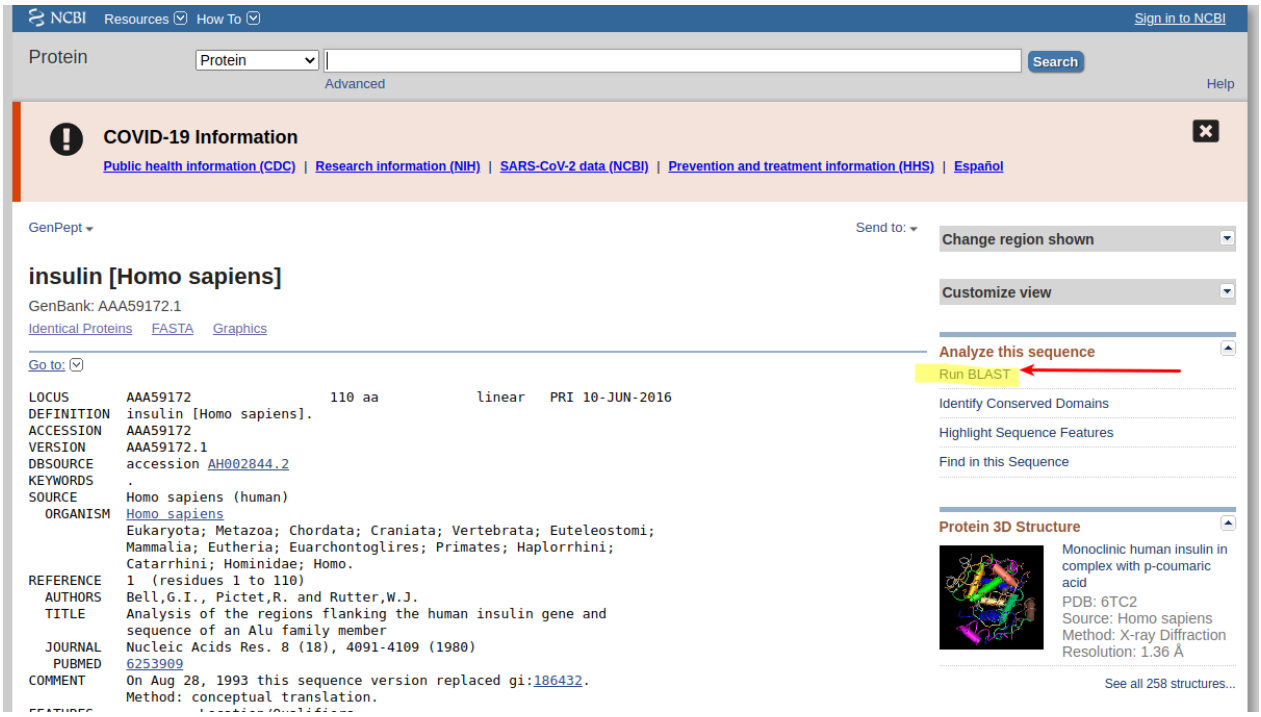
5. Deadline: 20/12/2024 23:59

1. Search for chosen protein on NCBI:

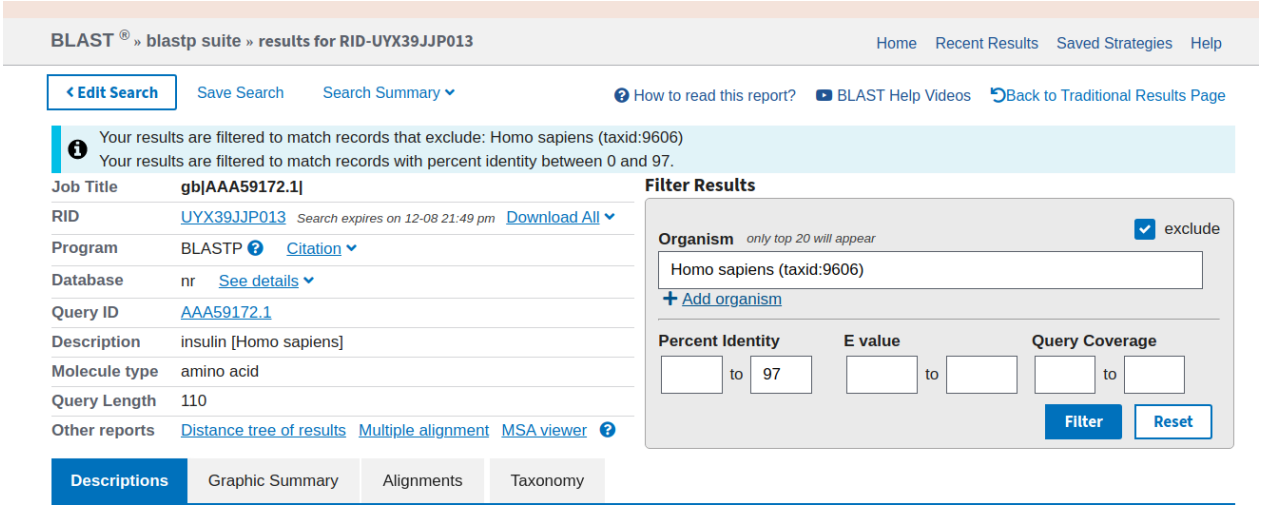




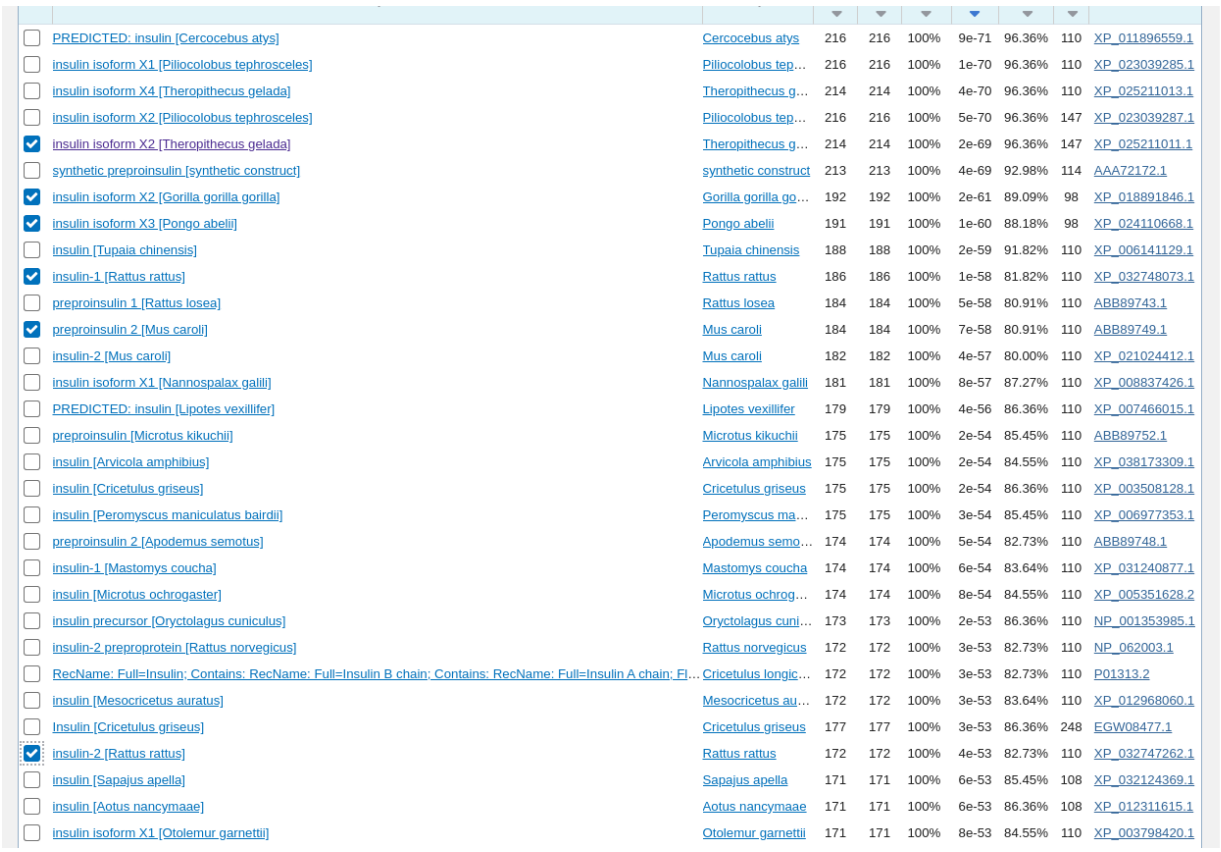
2. Run BLAST on chosen protein.



3. Filter the results of Blast. We don't want to have identical results and our goal is to search for sequences from different organisms than the original.

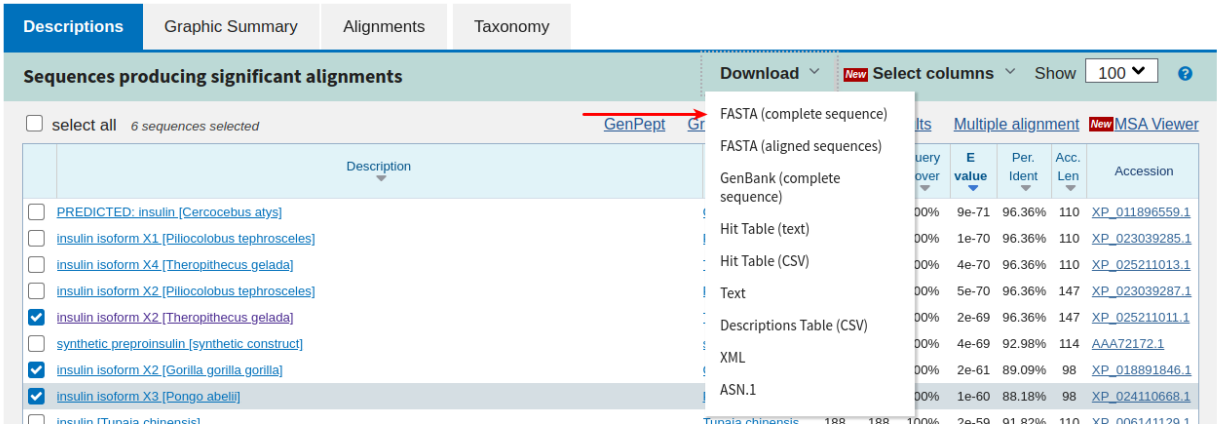


4. Choose sequences from different organisms:



PREDICTED: insulin [Cercocebus atys]	Cercocebus atys	216	216	100%	9e-71	96.36%	110	XP_011896559.1
insulin isoform X1 [Piliocolobus tephrosceles]	Piliocolobus tep…	216	216	100%	1e-70	96.36%	110	XP_023039285.1
insulin isoform X4 [Theropithecus gelada]	Theropithecus g…	214	214	100%	4e-70	96.36%	110	XP_025211013.1
insulin isoform X2 [Piliocolobus tephrosceles]	Piliocolobus tep…	216	216	100%	5e-70	96.36%	147	XP_023039287.1
insulin isoform X2 [Theropithecus gelada]	Theropithecus g…	214	214	100%	2e-69	96.36%	147	XP_025211011.1
synthetic preproinsulin [synthetic construct]	synthetic construct	213	213	100%	4e-69	92.98%	114	AAA72172.1
insulin isoform X2 [Gorilla gorilla gorilla]	Gorilla gorilla go…	192	192	100%	2e-61	89.09%	98	XP_018891846.1
insulin isoform X3 [Pongo abelii]	Pongo abelii	191	191	100%	1e-60	88.18%	98	XP_024110668.1
insulin [Tupaia chinensis]	Tupaia chinensis	188	188	100%	2e-59	91.82%	110	XP_006141129.1
insulin-1 [Rattus rattus]	Rattus rattus	186	186	100%	1e-58	81.82%	110	XP_032748073.1
preproinsulin 1 [Rattus losea]	Rattus losea	184	184	100%	5e-58	80.91%	110	ABB89743.1
preproinsulin 2 [Mus caroli]	Mus caroli	184	184	100%	7e-58	80.91%	110	ABB89749.1
insulin-2 [Mus caroli]	Mus caroli	182	182	100%	4e-57	80.00%	110	XP_021024412.1
insulin isoform X1 [Nannospalax galili]	Nannospalax galili	181	181	100%	8e-57	87.27%	110	XP_008837426.1
PREDICTED: insulin [Lipotes vexillifer]	Lipotes vexillifer	179	179	100%	4e-56	86.36%	110	XP_007466015.1
preproinsulin [Microtus kikuchii]	Microtus kikuchii	175	175	100%	2e-54	85.45%	110	ABB89752.1
insulin [Arvicola amphibius]	Arvicola amphibius	175	175	100%	2e-54	84.55%	110	XP_038173309.1
insulin [Cricetulus griseus]	Cricetulus griseus	175	175	100%	2e-54	86.36%	110	XP_003508128.1
insulin [Peromyscus maniculatus bairdii]	Peromyscus ma…	175	175	100%	3e-54	85.45%	110	XP_006977353.1
preproinsulin 2 [Apodemus semotus]	Apodemus semo…	174	174	100%	5e-54	82.73%	110	ABB89748.1
insulin-1 [Mastomys coucha]	Mastomys coucha	174	174	100%	6e-54	83.64%	110	XP_031240877.1
insulin [Microtus ochrogaster]	Microtus ochrog…	174	174	100%	8e-54	84.55%	110	XP_005351628.2
insulin precursor [Oryctolagus cuniculus]	Oryctolagus cuni…	173	173	100%	2e-53	86.36%	110	NP_001353985.1
insulin-2 preproprotein [Rattus norvegicus]	Rattus norvegicus	172	172	100%	3e-53	82.73%	110	NP_062003.1
RecName: Full=Insulin; Contains: RecName: Full=Insulin B chain; Contains: RecName: Full=Insulin A chain; Fl…	Cricetulus longic…	172	172	100%	3e-53	82.73%	110	P01313.2
insulin [Mesocricetus auratus]	Mesocricetus au…	172	172	100%	3e-53	83.64%	110	XP_012968060.1
Insulin [Cricetulus griseus]	Cricetulus griseus	177	177	100%	3e-53	86.36%	248	EGW08477.1
insulin-2 [Rattus rattus]	Rattus rattus	172	172	100%	4e-53	82.73%	110	XP_032747262.1
insulin [Sapajus apella]	Sapajus apella	171	171	100%	6e-53	85.45%	108	XP_032124369.1
insulin [Aotus nancymaae]	Aotus nancymaae	171	171	100%	6e-53	86.36%	108	XP_012311615.1
insulin isoform X1 [Otolemur garnettii]	Otolemur garnettii	171	171	100%	8e-53	84.55%	110	XP_003798420.1

5. Download the sequences




```
>XP_025211011.1 insulin isoform X2 [Theropithecus gelada]
MAGLLKRLGVSPGAPGQGTWPSAGLRPACLPGHCPSAMALWMRLLPLLALLALWGPDSVPAFVNQHLCGSHLVEALYLVC
GERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
>XP_018891846.1 insulin isoform X2 [Gorilla gorilla gorilla]
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQGSLQPLALEGSLQKRGIV
EQCCTSICSLYQLENYCN
>XP_024110668.1 insulin isoform X3 [Pongo abelii]
MALWMRLLPLLALLALWGPDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQGSLQPLALEGSLQKRGIV
EQCCTSICSLYQLENYCN
>XP_032748073.1 insulin-1 [Rattus rattus]
MALWMRFLPLLALLVVWEPKPAQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVEDPQVPQLELGGGPEAGDLQTL
ALEVARQKRGIVDQCCTSICSLYQLENYCN
>ABB89749.1 preproinsulin 2 [Mus caroli]
MALWMRFLPLVALLFLWESHPTQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVEDPQVAQLELGGGPGAGDLQTL
ALEVAQQKRGIVDQCCTSICSLYQLENYCN
>XP_032747262.1 insulin-2 [Rattus rattus]
MALWIRFLPLLALLVLWEPRPAQAFVKQHLCGSHLVEALYLVCGERGFFYTPVSRREVEDPQVAQLELGGGPGAGDLQTL
ALEVARQKRGIVDQCCTSICSLYQLENYCN
```

6. Remember to also download the reference sequence (Homo sapiens insulin in this case)

7. Repeat the process with different proteins. This time search only for sequences in previously chosen organisms. This way you will get 8 different proteins (or similar proteins) from the same 7 organisms.