Sample Scenario (17, translated)

Evidence for evolution by comparing proteins in different species

Summary

2 Procedure

2.1 Find the sequences of the CFTR protein in several species in the UniProtKB database
UniProtKB is a database that contains all protein sequences listed from public data (~230 million proteins).
N.B.: to study evolution, protein sequences are often used: they are more relevant from an evolutionary point of view. Here we are working with the CFTR protein, which causes cystic fibrosis when it is faulty, but other proteins can be used. (cf. selected proteins)

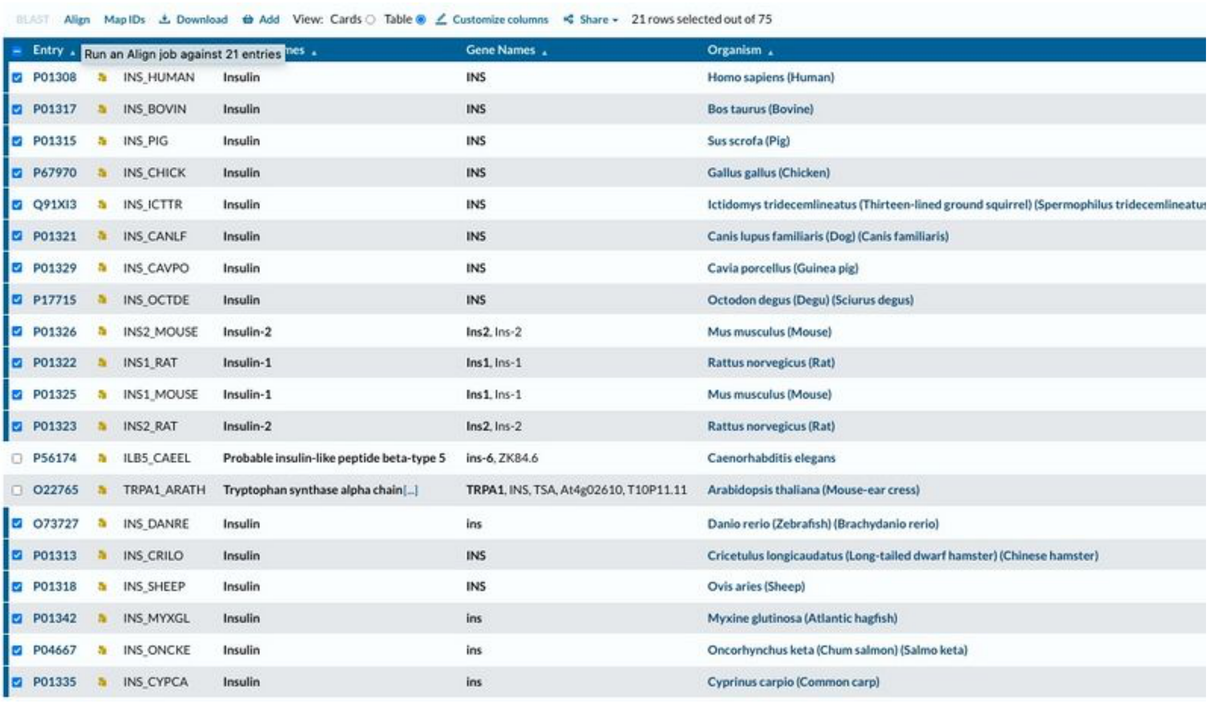
Open UniProtKB ; in Query, type gene: CFTR.

You have to search for proteins by their gene name (1)

UniProt will search for all entries in which the term CFTR *solution* the list is huge.
Select entries "reviewed" solution (*Insulin*) solution CFTR *(as of 23.09.22).*

(1) *N.B.: The name of the protein (cf. list of selected proteins) often differs from the name of the gene. It is best to search for a protein by its gene name, as gene names are 'standardized' by expert committees. UniProtKB is a protein database: the information is focused on protein sequences but the list of gene names is exhaustive. The same gene can sometimes have several names. Example: the CFTR gene has a synonym: abcc7*

2.2 Selecting the same protein in several organisms



BLAST Align Map IDs Download Add View: Cards ○ Table ● Customize columns Share 21 rows selected out of 75

Run an Align job against 21 entries

−	Entry			nes	Gene Names	Organism
☑	P01308		INS_HUMAN	Insulin	INS	Homo sapiens (Human)
☑	P01317		INS_BOVIN	Insulin	INS	Bos taurus (Bovine)
☑	P01315		INS_PIG	Insulin	INS	Sus scrofa (Pig)
☑	P67970		INS_CHICK	Insulin	INS	Gallus gallus (Chicken)
☑	Q91XI3		INS_ICTTR	Insulin	INS	Ictidomys tridecemlineatus (Thirteen-lined ground squirrel) (Spermophilus tridecemlineatu
☑	P01321		INS_CANLF	Insulin	INS	Canis lupus familiaris (Dog) (Canis familiaris)
☑	P01329		INS_CAVPO	Insulin	INS	Cavia porcellus (Guinea pig)
☑	P17715		INS_OCTDE	Insulin	INS	Octodon degus (Degu) (Sciurus degus)
☑	P01326		INS2_MOUSE	Insulin-2	Ins2, Ins-2	Mus musculus (Mouse)
☑	P01322		INS1_RAT	Insulin-1	Ins1, Ins-1	Rattus norvegicus (Rat)
☑	P01325		INS1_MOUSE	Insulin-1	Ins1, Ins-1	Mus musculus (Mouse)
☑	P01323		INS2_RAT	Insulin-2	Ins2, Ins-2	Rattus norvegicus (Rat)
□	P56174		ILB5_CAEEL	Probable insulin-like peptide beta-type 5	ins-6, ZK84.6	Caenorhabditis elegans
□	O22765		TRPA1_ARATH	Tryptophan synthase alpha chain[...]	TRPA1, INS, TSA, At4g02610, T10P11.11	Arabidopsis thaliana (Mouse-ear cress)
☑	O73727		INS_DANRE	Insulin	ins	Danio rerio (Zebrafish) (Brachydanio rerio)
☑	P01313		INS_CRILO	Insulin	INS	Cricetulus longicaudatus (Long-tailed dwarf hamster) (Chinese hamster)
☑	P01318		INS_SHEEP	Insulin	INS	Ovis aries (Sheep)
☑	P01342		INS_MYXGL	Insulin	ins	Myxine glutinosa (Atlantic hagfish)
☑	P04667		INS_ONCKE	Insulin	ins	Oncorhynchus keta (Chum salmon) (Salmo keta)
☑	P01335		INS_CYPCA	Insulin	ins	Cyprinus carpio (Common carp)

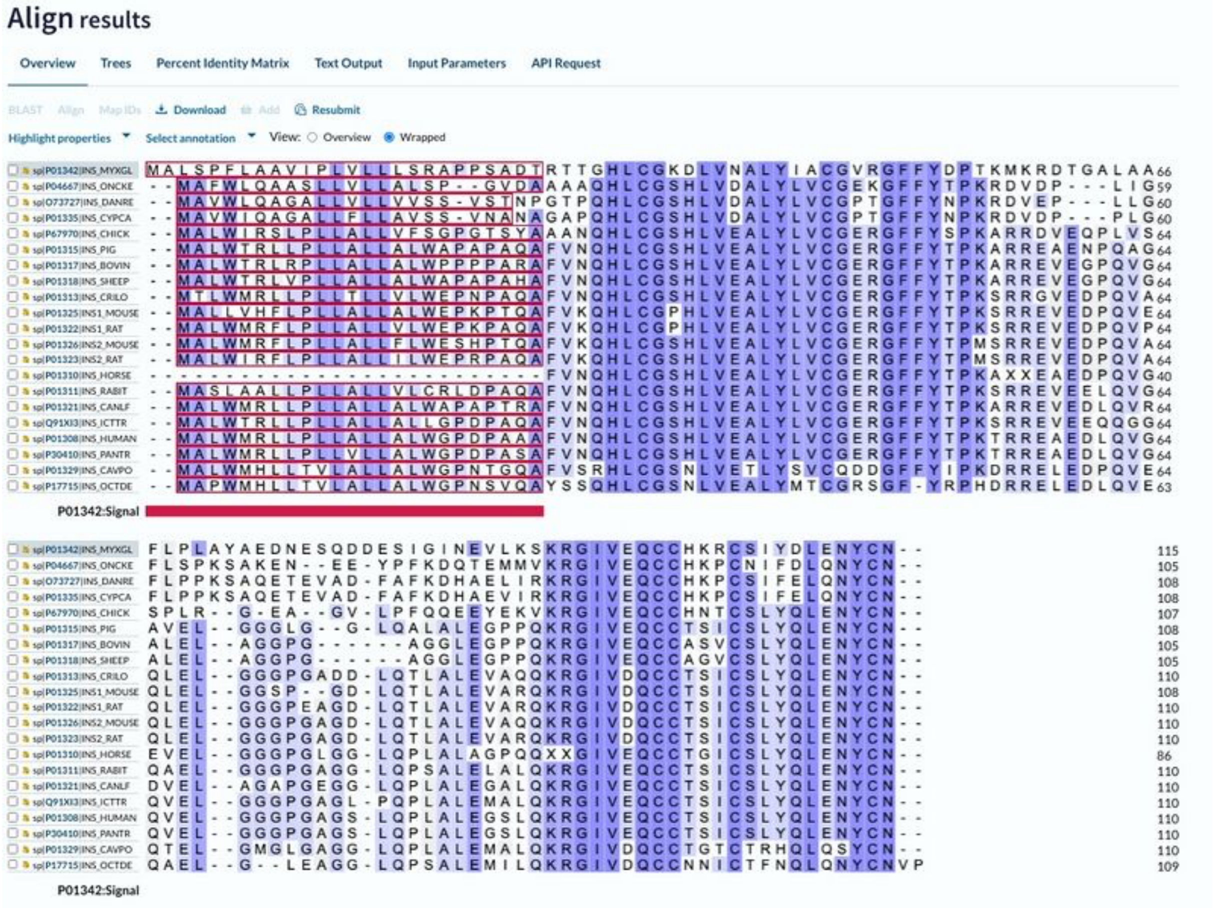
Sample results obtained by searching for 'INS' as a gene name.

Select (tick) from the list the proteins of the species that interests you.
A minimum of 5 protein sequences is required to construct a sequence alignment that makes sense. (cf.list of selected proteins).
It may be wise to impose a few species on each group to facilitate comparisons (e.g., human, chimpanzee, mouse, rat, bovine, horse, Danio rero (fish), fruit fly) and let them choose a few others at will.
[…] The selected sequences are marked with a blue rectangle at the beginning of the line and their number appears in the blue band at the top: e.g. "21 rows selected out of 97".
[…]

2.3 Building an alignment

- Click the "Align" button above the strip at the top of the list. A window displays the sequences that will be aligned in FASTA format.
- Click the button at the bottom right Align XX sequences. In principle do not touch anything After a certain time (several seconds, even minutes) a "Tool results" window appears displaying the alignments carried out during the session.

- • Click "Completed" to display the alignment in a new page. Example for CFTR



View: ○ Overview ● Wrapped

```
☐ sp|P01342|INS_MYXGL   MALSPFLAAVIPLVLLLSRAPPSADTRTTGHLCGKDLVNALYIACGVRGFFYDPTKMKRDTGALAA 66
☐ sp|P04667|INS_ONCKE   --MAFWLQAASLLVLLALSP--GVDAAAAQHLCGSHLVDALYLVCGEKGFFYTPKRDVDP---LIG 59
☐ sp|O73727|INS_DANRE   --MAVWLQAGALLVLLVVSS-VSTNPGTPQHLCGSHLVDALYLVCGPTGFFYNPKRDVEP---LLG 60
☐ sp|P01335|INS_CYPCA   --MAVWIQAGALLFLLAVSS-VNANAGAPQHLCGSHLVDALYLVCGPTGFFYNPKRDVDP---PLG 60
☐ sp|P67970|INS_CHICK   --MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQPLVS 64
☐ sp|P01315|INS_PIG     --MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAENPQAG 64
☐ sp|P01317|INS_BOVIN   --MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEGPQVG 64
☐ sp|P01318|INS_SHEEP   --MALWTRLVPLLALLALWAPAPAHAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEGPQVG 64
☐ sp|P01313|INS_CRILO   --MTLWMRLLPLLTLLVLWEPNPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRRGVEDPQVA 64
☐ sp|P01325|INS1_MOUSE  --MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVEDPQVE 64
☐ sp|P01322|INS1_RAT    --MALWMRFLPLLALLVLWEPKPAQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVEDPQVP 64
☐ sp|P01326|INS2_MOUSE  --MALWMRFLPLLALLFLWESHPTQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVEDPQVA 64
☐ sp|P01323|INS2_RAT    --MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVEDPQVA 64
☐ sp|P01310|INS_HORSE   --------------------------FVNQHLCGSHLVEALYLVCGERGFFYTPKAXXEAEDPQVG 40
☐ sp|P01311|INS_RABIT   --MASLAALLPLLALLVLCRLDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRREVEELQVG 64
☐ sp|P01321|INS_CANLF   --MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEDLQVR 64
☐ sp|Q91XI3|INS_ICTTR   --MALWTRLLPLLALLALLGPDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRREVEEQQGG 64
☐ sp|P01308|INS_HUMAN   --MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVG 64
☐ sp|P30410|INS_PANTR   --MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVG 64
☐ sp|P01329|INS_CAVPO   --MALWMHLLTVLALLALWGPNTGQAFVSRHLCGSNLVETLYSVCQDDGFFYIPKDRRELEDPQVE 64
☐ sp|P17715|INS_OCTDE   --MAPWMHLLTVLALLALWGPNSVQAYSSQHLCGSNLVEALYMTCGRSGF-YRPHDRRELEDLQVE 63
              P01342:Signal

☐ sp|P01342|INS_MYXGL   FLPLAYAEDNESQDDESIGINEVLKSKRGIVEQCCHKRCSIYDLENYCN--  115
☐ sp|P04667|INS_ONCKE   FLSPKSAKEN--EE-YPFKDQTEMMVKRGIVEQCCHKPCNIFDLQNYCN--  105
☐ sp|O73727|INS_DANRE   FLPPKSAQETEVAD-FAFKDHAELIRKRGIVEQCCHKPCSIFELQNYCN--  108
☐ sp|P01335|INS_CYPCA   FLPPKSAQETEVAD-FAFKDHAEVIRKRGIVEQCCHKPCSIFELQNYCN--  108
☐ sp|P67970|INS_CHICK   SPLR--G-EA--GV-LPFQQEEYEKVKRGIVEQCCHNTCSLYQLENYCN--  107
☐ sp|P01315|INS_PIG     AVEL--GGGLG--G-LQALALEGPPQKRGIVEQCCTSICSLYQLENYCN--  108
☐ sp|P01317|INS_BOVIN   ALEL--AGGPG------AGGLEGPPQKRGIVEQCCASVCSLYQLENYCN--  105
☐ sp|P01318|INS_SHEEP   ALEL--AGGPG------AGGLEGPPQKRGIVEQCCAGVCSLYQLENYCN--  105
☐ sp|P01313|INS_CRILO   QLEL--GGGPGADD-LQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN--  110
☐ sp|P01325|INS1_MOUSE  QLEL--GGSP--GD-LQTLALEVARQKRGIVDQCCTSICSLYQLENYCN--  108
☐ sp|P01322|INS1_RAT    QLEL--GGGPEAGD-LQTLALEVARQKRGIVDQCCTSICSLYQLENYCN--  110
☐ sp|P01326|INS2_MOUSE  QLEL--GGGPGAGD-LQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN--  110
☐ sp|P01323|INS2_RAT    QLEL--GGGPGAGD-LQTLALEVARQKRGIVDQCCTSICSLYQLENYCN--  110
☐ sp|P01310|INS_HORSE   EVEL--GGGPGLGG-LQPLALAGPQQXXGIVEQCCTGICSLYQLENYCN--  86
☐ sp|P01311|INS_RABIT   QAEL--GGGPGAGG-LQPSALELALQKRGIVEQCCTSICSLYQLENYCN--  110
☐ sp|P01321|INS_CANLF   DVEL--AGAPGEGG-LQPLALEGALQKRGIVEQCCTSICSLYQLENYCN--  110
☐ sp|Q91XI3|INS_ICTTR   QVEL--GGGPGAGL-PQPLALEMALQKRGIVEQCCTSICSLYQLENYCN--  110
☐ sp|P01308|INS_HUMAN   QVEL--GGGPGAGS-LQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN--  110
☐ sp|P30410|INS_PANTR   QVEL--GGGPGAGS-LQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN--  110
☐ sp|P01329|INS_CAVPO   QTEL--GMGLGAGG-LQPLALEMALQKRGIVDQCCTGTCTRHQLQSYCN--  110
☐ sp|P17715|INS_OCTDE   QAEL--G--LEAGG-LQPSALEMILQKRGIVDQCCNNICTFNQLQNYCNVP  109
              P01342:Signal
```

Alignment of a few CFTR protein sequences from various species - similarity enabled.

Aligned sequences are displayed in table, one sequence (species) per line, 66 per line. (In Overview mode = default)
On a line below, the protein signal is shown in red.
NB: "-" (dash) means that the alignment program introduced a space (insertion/deletion or "gap") to be able to align sequences of different lengths.
To find the name of the species, click on the accession number in blue -> a new window opens with the entry UniProtKB corresponding to this protein and the full name of the species.

Nota Bene

The « Similarity » highlight above the table on the left is enabled by default: similar regions are highlighted (purple) - allowing discussion of those that are more evolutionarily conserved.

Highlight “Physical properties” then “Hydrophobicity”: highlights regions likely to be transmembrane

Additional information and possible questions

The physico-chemical properties of the different amino acids are described *here*

A correspondence table of 3-letter and 1-letter amino acid and codon codes can be accessed *here*.

A phylogenetic tree?

At the top of the page, instead of the "overview" visualization, you can click "Trees" to display trees with different visualizations...

Note that it is a guided tree, used by the program to construct the alignment and is based solely on observed differences between sequences. It is not a proper phylogenetic tree that is much more complex to establish! See for example scenario. Phylogeny, biodiversity and pizza...

See also scenario 19 *Find the date of evolutionary divergence of two species*

3 *Possible student questions:*

- Find the organisms closest to (human, mouse, cow, ...) for each of the proteins studied by the class.
  Try to find another explanation than the common origin and the independent evolution since the separation.
- Why use "protein" sequences rather than nucleotides in this evolutionary context?
- Are there sequences that are similar in a large number of organisms?
- Observing that the sequences of insulin, EPO, a Histone, CFTR, etc. are very similar in many species: what are the possible explanations for this fundamental similarity?
- Insulin varies in many regions, while Histone (H4) is almost identical between very many organisms. perform its functions in DNA winding and regulation.
- Do similar sequences between a large number of organisms indicate that there are no mutations in those places?
- Ask the students to find the time when the species studied separated and compare with the tree obtained. There is no site offering this, but a search on the internet often makes it possible to find an estimate of the last common ancestor between two species.

Discussion of these alignments can illustrate the fundamental concepts of common origin and divergence common to various proteins.

Scenario established with scientific advice from Dr. Marie-Claude Blatter of SIB Swiss Institute of Bioinformatics