Multiple Sequence Alignments and JalView.

Rafael Torres-Perez
- Bioinformatics for Genomics and Proteomics (BioinfoGP) -

1. Introductory Example [*Words.(ods/xlsx)*].

```
acqua
aqua
agua
aigua
auga
apa
```

More for you to practice:

vasser
water
water
vatten
vann

Al-Kwarizmi
algoritmo
algorithm

2. Insulin Example

Uniprot Align: https://www.uniprot.org/align/

2.1. Uniprot IDs:

```
P01308
P01315
P01317
P67970
```

```
>sp|IA25-54-IB90-110
FVNQHLCGSHLVEALYLVCGERGFFYTPKTGIVEQCCTSICSLYQLENYCN
```

2.2. CDS for principal transcripts:

```
>INS-CDS-TRANSCRIPT-202-HUMAN
ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGAC
CCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTAC
CTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGAC
CTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTG
GCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCATCTGC
TCCCTCTACCAGCTGGAGAACTACTGCAACTAG

>INS-CDS-TRANSCRIPT-202-BOVIN
ATGGCCCTGTGGACACGCCTGGCGCCCCTGCTGGCCCTGCTGGCGCTCTGGGCCCCCGCC
CCGGCCCGCGCCTTCGTCAACCAGCATCTGTGTGGCTCCCACCTGGTGGAGGCGCTGTAC
CTGGTGTGCGGAGAGCGCGGCTTCTTCTACACGCCCAAGGCCCGCCGGGAGGTGGAGGGC
CCCCAGGTGGGGGCGCTGGAGCTGGCCGGAGGCCCGGGCGCGGGCGGCCTGGAGGGGCCC
CCGCAGAAGCGTGGCATCGTGGAGCAGTGCTGTGCCAGCGTCTGCTCGCTCTACCAGCTG
GAGAACTACTGTAACTAG

>INS-CDS-TRANSCRIPT-201-PIG
ATGGCCCTGTGGACGCGCCTCCTGCCCCTGCTGGCCCTGCTGGCCCTCTGGGCGCCCGCC
CCGGCCCAGGCCTTCGTGAACCAGCACCTGTGCGGCTCCCACCTGGTGGAGGCGCTGTAC
CTGGTGTGCGGGGAGCGCGGCTTCTTCTACACGCCCAAGGCCCGTCGGGAGGCGGAGAAC
CCTCAGGCAGGTGCCGTGGAGCTGGGCGGAGGCCTGGGCGGCCTGCAGGCCCTGGCGCTG
GAGGGGCCCCCGCAGAAGCGTGGCATCGTGGAGCAGTGCTGCACCAGCATCTGTTCCCTC
TACCAGCTGGAGAACTACTGCAACTAG

>INS-CDS-TRANSCRIPT-201-CHICKEN
ATGGCTCTCTGGATCCGATCACTGCCTCTTCTGGCTCTCCTTGTCTTTTCTGGCCCTGGA
ACCAGCTATGCAGCTGCCAACCAGCACCTCTGTGGCTCCCACTTGGTGGAGGCTCTCTAC
CTGGTGTGTGGAGAGCGTGGCTTCTTCTACTCCCCCAAAGCCCGACGGGATGTCGAGCAG
CCCCTAGTGAGCAGTCCCTTGCGTGGCGAGGCAGGAGTGCTGCCTTTCCAGCAGGAGGAA
TACGAGAAAGTCAAGCGAGGGATTGTTGAGCAATGCTGCCATAACACGTGTTCCCTCTAC
CAACTGGAGAACTACTGCAACTAG
```

3. Global and local alignments.

Ensembl: https://www.ebi.ac.uk/Tools/psa/

```
>P29600|SUBS_BACLE Subtilisin Savinase - Bacillus lentus
AQSVPWGISRVQAPAAHNRGLTGSGVKVAVLDTGISTHPDLNIRGGASFVPGEPSTQDGN
GHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMHVA
NLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGSISYPARYANAMAVGATDQNNNR
ASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPSWSNVQI
RNHLKNTATSLGSTNLYGSGLVNAEAATR

>P41363|ELYA_BACHD Thermostable alkaline protease precursor - Bacillus halodurans
MRQSLKVMVLSTVALLFMANPAAASEEKKEYLIVVEPEEVSAQSVEESYDVDVIHEFEEI
PVIHAELTKKELKKLKKDPNVKAIEKNAEVTISQTVPWGISFINTQQAHNRGIFGNGARV
AVLDTGIASHPDLRIAGGASFISSEPSYHDNNGHGTHVAGTIAALNNSIGVLGVAPSADL
YAVKVLDRNGSGSLASVAQGIEWAINNNMHIINMSLGSTSGSSTLELAVNRANNAGILLV
GAAGNTGRQGVNYPARYSGVMAVAAVDQNGQRASFSTYGPEIEISAPGVNVNSTYTGNRY
VSLSGTSMATPHVAGVAALVKSRYPSYTNNQIRQRINQTATYLGSPSLYGNGLVHAGRAT
Q
```

```
>sp|P29144|TPP2_HUMAN Tripeptidyl-peptidase 2
MATAATEEPFPFHGLLPKKETGAASFLCRYPEYDGRGVLIAVLDTGVDPGAPGMQVTTDG
KPKIVDIIDTTGSGDVNTATEVEPKDGEIVGLSGRVLKIPASWTNPSGKYHIGIKNGYDF
YPKALKERIQKERKEKIWDPVHRVALAEACRKQEEFDVANNGSSQANKLIKEELQSQVEL
LNSFEKKYSDPGPVYDCLVWHDGEVWRACIDSNEDGDLSKSTVLRNYKEAQEYGSFGTAE
MLNYSVNIYDDGNLLSIVTSGGAHGTHVASIAAGHFPEEPERNGVAPGAQILSIKIGDTR
LSTMETGTGLIRAMIEVINHKCDLVNYSYGEATHWPNSGRICEVINEAVWKHNIIYVSSA
GNNGPCLSTVGCPGGTTSSVIGVGAYVSPDMMVAEYSLREKLPANQYTWSSRGPSADGAL
GVSISAPGGAIASVPNWTLRGTQLMNGTSMSSPNACGGIALILSGLKANNIDYTVHSVRR
ALENTAVKADNIEVFAQGHGIIQVDKAYDYLVQNTSFANKLGFTVTVGNNRGIYLRDPVQ
VAAPSDHGVGIEPVFPENTENSEKISLQLHLALTSNSSWVQCPSHLELMNQCRHINIRVD
PRGLREGLHYTEVCGYDIASPNAGPLFRVPITAVIAAKVNESSHYDLAFTDVHFKPGQIR
RHFIEVPEGATWAEVTVCSCSSEVSAKFVLHAVQLVKQRAYRSHEFYKFCSLPEKGTLTE
AFPVLGGKAIEFCIARWWASLSDVNIDYTISFHGIVCTAPQLNIHASEGINRFDVQSSLK
YEDLAPCITLKNWVQTLRPVSAKTKPLGSRDVLPNNRQLYEMVLTYNFHQPKSGEVTPSC
PLLCELLYESEFDSQLWIIFDQNKRQMGSGDAYPHQYSLKLEKGDYTIRLQIRHEQISDL
ERLKDLPFIVSHRLSNTLSLDIHENHSFALLGKKKSSNLTLPPKYNQPFFVTSLPDDKIP
KGAGPGCYLAGSLTLSKTELGKKADVIPVHYYLIPPPTKTKNGSKDKEKDSEKEKDLKEE
FTEALRDLKIQWMTKLDSSDIYNELKETYPNYLPLYVARLHQLDAEKERMKRLNEIVDAA
NAVISHIDQTALAVYIAMKTDPRPDAATIKNDMDKQKSTLVDALCRKGCALADHLLHTQA
QDGAISTDAEGKEEEGESPLDSLAETFWETTKWTDLFDNKVLTFAYKHALVNKMYGRGLK
FATKLVEEKPTKENWKNCIQLMKLLGWTHCASFTENWLPIMYPPDYCVF
```

4. DNA sequences for JalView.

Reference: https://www.nature.com/articles/s41564-020-0688-y/figures/6

OCR tools: Search on Google "*ocr on line*", there's a lot!

Ensembl Tools for MSA:
https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/Multiple+Sequence+Alignment
https://www.ebi.ac.uk/Tools/msa/

GeneBank IDs:

```
AY278741
KF367457
KF569996
KY417151
KY417146
KY417144
KC881005
MN908947
```

KY417142
JX993988
JX993987
DQ648857
KY417147
KY417143
DQ071615
KY417148
GQ153543
GQ153548
KJ473815
KF294457
KJ473816
KY417145
MG772934
MG772933
KJ473811
KJ473814
DQ412042
KJ473812
DQ648856
NC014470

5. Protein sequences for JalView. (File “CoutardFastaProtein.fa”)

Reference: https://www.sciencedirect.com/science/article/pii/S0166354220300528

Genomes_IDs
AY391777
EF065513
KF514433.1
KF530114.1
MG772934
NC_004718.3
NC_006577.2
NC_014470.1
NC_019843.3
NC_045512.2

S Protein IDs
AAR01015.1
ABN10911.1
AGT21367.1
AGT51394.1
AVP78042.1
NP_828851.1
YP_173238.1
YP_003858584.1
YP_009047204.1
YP_009724390.1