

ASSIGNMENTS (PROF. ZIZZA):

STRUMENTI FORMALI PER LA BIOINFORMATICA

INTRODUZIONE E ASSIGNMENT INIZIALI

Seguire il percorso indicato nel testo sottostante e stilare una breve relazione con i contenuti recuperati e elaborati man mano, usando le risorse proposte e recuperabili dalle lezioni (vedi sito del corso).

Link necessari

- NIH: <https://www.ncbi.nlm.nih.gov/>
- UniProt: <https://www.uniprot.org>
- BLAST: (raggiungibile dal sito precedente)
- Needle: https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle
- (dal sito <https://www.ebi.ac.uk/> potete imparare tante cose...)
- Galaxy: <https://usegalaxy.eu/>

- seguire il tutorial introduttivo

<https://usegalaxy.eu/training-material/topics/introduction/tutorials/galaxy-intro-101/tutorial.html#galaxy-basics-for-genomics>

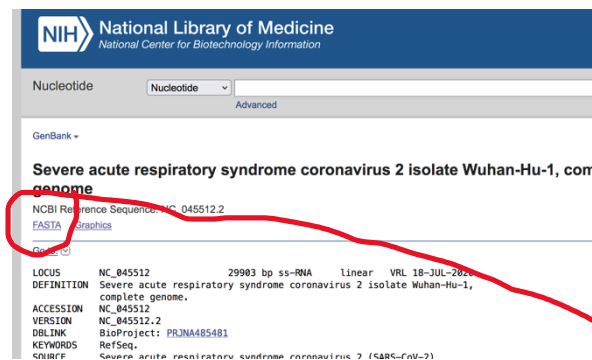
- e quello sull'analisi di qualità dei file FASTA


<https://usegalaxy.eu/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#quality-control>

1. RICERCA DI SEQUENZE IN BANCHE DATI

Proviamo a cercare sul sito NIH la sequenza del SARS Cov2.

- Accedere a <https://www.ncbi.nlm.nih.gov/>
- Nella barra di ricerca, indicare Sars Covid 19 e viene fornita il numero identificativo NC_045512.2 in RefSeq (il database delle sequenze di NCBI). Cliccando sul nome, si apre la pagina di descrizione della sequenza



- Si ritrovano tutti i concetti visti durante la lezione. Cliccando su FASTA, si ottiene la sequenza nucleotidica. Cliccando poi su GenBank si ritorna sul file della sequenza aminoacidica.
- Dal menù laterale si può direttamente usare Blast per effettuare ricerche e confronti. Eseguiamolo e dopo alcuni secondi viene mostrato l'elenco delle sequenza più simili alla nostra, selezionando "Core nucleotide BLAST database", per accelerare la ricerca.
- Aprire il menù "Search Summary" per rileggere le nozioni imparate su BLAST
- Cliccando su Alignments si ha il dettaglio dei vari confronti
- Cliccando su MSA viewer e sul bottone  si ha anche la visualizzazione della sequenza e si possono zoomare le zone di similitudine

E' possibile analizzare anche insiemi di reads suddivisi in più file che derivano da vari "run" di sequenziamento della sequenza in oggetto. Ad esempio, è ottenibile al link <https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR11528307>. un archivio di reads del SarsCov2. Alcune domande sono poste da un assignment di Compeau e le riporto qui come esercizio.

- How many nucleotide bases are found in this file? What is the GC-content of the reads? (That is, the percentage of reads that are either G or C.)

Answer (1 point each): 186.5 million; 38.2%.

- We know that the original SARS-CoV genome from the 2003 outbreak has approximately 30,000 nucleotides. If the SARS-CoV-2 genome has length 30,000 nucleotides, then what is the coverage of this read dataset?

Answer (2 points): 186.5 million/30,000 ~ 6217X

- Clicking on the “Reads” tab will allow us to see ten read-pairs at a time. Each read is in **FASTA format**, meaning that the read has a header beginning with the “>” symbol identifying the read, followed by the read itself. What is the read-pair at spot #15213? Why do you think that one read is shorter than the other?

Answer (3 points; 1 point for first part, 2 points for second part): Reads are shown below. There are lots of reasons why the read could be shorter. The most likely correct one is that the machine at that position started losing signal, but students may have some creative ideas too.

>gnl|SRA|SRR11528307.15213.1 MN01288:4:000H32WJK:1:11101:22090:17095
(Biological)


CAACAAGGCCAAACTGTCAC TAAGAAATCTGCTGCTGAGGCTTCTAAGAAGCCTCG
GCAA
AAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCC
AGAA
CAAACCCAAGGAAATTTTGGGGACCAAGGAA

>gnl|SRA|SRR11528307.15213.2 MN01288:4:000H32WJK:1:11101:22090:17095
(Biological)

TCAATATGCTTATTCAGCAAAATGACTTGAT

- Ogni read può essere visualizzata Cliccando sul menù “Reads”. da cui è possibile anche evidenziare il quality score.

2. ALLINEAMENTO A COPPIE



Pairwise Sequence Alignment
Identify regions of similarity between two biological sequences.

[Needle](#) | [Stretcher](#) | [GGSEARCH2SE0](#) | [Water](#) | [Matcher](#) | [LALIGN](#) | [SSEARCH2SE0](#) | [GeneWise](#) | [Less](#)

STOP: Let's apply this to the same protein (say, hemoglobin subunit alpha) in a few different species. What do you think we will see?

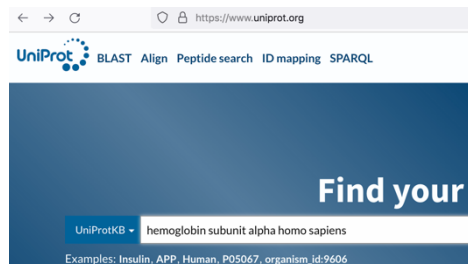
- *Homo sapiens* vs. *Gorilla gorilla gorilla*
- *Homo sapiens* vs. *Bos Taurus* (cow)
- *Homo sapiens* vs. *Danio rerio* (zebrafish)

<https://www.uniprot.org/uniprot/P69905>
<https://www.uniprot.org/uniprot/P01923>
<https://www.uniprot.org/uniprot/P01966>
<https://www.uniprot.org/uniprot/Q90487>

EMBOSS "Needle" server:
https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle

77

- Seguendo il link, si ottengono le sequenze proteiche dell'emoglobina dei 4 organismi. E' possibile accedere al sito UniProt e indicare nella barra del menù la sequenza richiesta



- Selezionare quella richiesta

P69905 · HBA_HUMAN

Protein ¹	Hemoglobin subunit alpha	Amino acids	142 (go to sequence)
Gene ¹	HBA1; HBA2	Protein existence ¹	Evidence at protein level
Status ¹	UniProtKB reviewed (Swiss-Prot)	Annotation score ¹	0.5
Organism ¹	Homo sapiens (Human)		

Entry Variant viewer **820** Feature viewer Genomic coordinates Publications External links

Tools Download Add Community curation (3) Add a publication Entry feedback

Function¹

- Selezionando "Variant Viewer" vengono visualizzate le varianti e le patologie associate

(ult. agg. 1.12.2024)

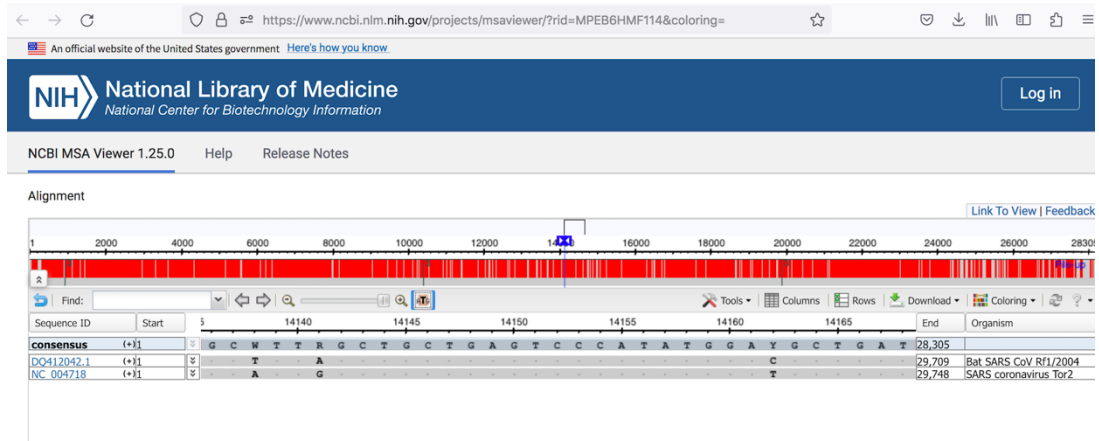
- E' possibile copiare la sequenza, oppure farne il download, scegliendo il formato (testo, FASTA...). Questo si ripete per le varie sequenze.
- Poiché l'obiettivo è fare un confronto a coppie, lanciamo Needle dal sito EMBOSS https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle copiando e incollando le due sequenze. Lasciamo i parametri indicati.
- Analizziamo i risultati dell'allineamento e confrontiamo i valori ottenuti.

Altro esercizio

Altro esempio, può essere quello di estrarre le sequenze della proteina Spike su diverse sequenze di virus della famiglia Coronavirus. Ad esempio, come fatto durante la lezione, selezioniamo da NIH le sequenze di

- Sars Covid 19 (NC_045512.2)
- Sars Coronavirus Tor2 (NC_004718)
- Bat SARS Coronavirus Rf1 (DQ412042.1)
- Bat Coronavirus (NC014470.1)

Eseguire a coppie la chiamata a BLAST, selezionando allineamento a coppie. Analizzare i risultati. Cliccando su MSA Viewer si possono evidenziare anche le sequenze nucleotidiche e i mismatch.



Lo stesso si può fare sulle sequenze proteiche della proteina Spike di ogni virus. Per ottenerle, tornare al sito NIH relativo alla sequenza virale. Accedere a GenBank, dove verrà visualizzata la sequenza. Ad esempio, per Sars Coronavirus Tor 2 il codice è YP_009825051.1 ed è ottenibile dal file

(ult. agg. 1.12.2024)

CDS

```
/db_xref= GeneID:1409000
21492..25259
/gene="S"
/locus_tag="sars2"
/gene_synonym="E2"
/codon_start=1
/product="Spike glycoprotein"
/protein_id="YP_009825051.1"
/db_xref="GeneID:1489668"
/translation="METELLELTITSGSGLRCTTFDDVQAPNYTQHTSSMRGVYYPD
EIFRSDTLTLTQDLFLPFYSNVTGFHTINHTFGNPVVPKDGIFYAATEKSNVVRGWV
FGSTMNKSQSVIIINNSTNVVIRACNFELCDNPFFAVSKPMGTQHTHTIFDNAFNCT
FEYISDAFSLDVSEKSGNFKHLREFVFNKDGFLVYKGYQPIDVVRDLPSGFNTLKP
IFKLPLGINITNFRAILTAFAQDIWGTSAAYFVGYLKPTTFMLKYDENGITIDAV
DCSQNPALAEKCSVKSFEIDKGIYQTSNFRVVPVSGDVVRFPNITNLCPPGEVFNATKF
PSVYAWERKKISNCVADYSVLNSTFFSTFKCYGVSATKLNLCFSNVYADSFVVKGD
DVRQIAPGQTGVIADYNYKLDDFMGCVLAWNTRNIDATSTGNYNYKRYLRHGKLRP
FERDISNVFPSPDGKCTPPALNCYWPLNDYGFYTTTGIGYQPYRVVLSFELLNAPA
TVCGPKLSTDLIKNQCVNFNGLTGTGLTPSSKRFQPFQFGRDVSDFDTSVRDPK
TSEILDISPACAFGGVSVITPGTNASSEVAVLYQDVNCTDVSTAIHADQLTPAWRIYST
GNNVFQTAGCLIGAHEVDTSYECDIPIGAGICASYHTVSLRSTSQKSIVAYTMSLG
ADSSIAYSNNTIAIPTNFSISITTEVMPVMAKTSVDCNMYICGDSFTECANLLQYGS
FCTQLNRALSGIAAEQDRNTREVFQVQKMYKPTLKYFGGFNFSQILPDPLKPTKRS
FIEDLLFNKVTLDAGFMKQYGECLGDINARDLICAQFNGLTVLPPLTDDMIAAYT
AALVSGTATAGWTFGAGAAQIPFAMQMAYRFNGIGVTQNVLYENQKQIANQFNKAIS
QIQESLTTTSTALGKLQDVVNQAALNTLVKQLSSNFGAISSVLNDILSRDKVEAE
VQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYH
LMSFPQAAPHGVVFLHVTYVPSQERNFTTAPAICHEGKAYFPREGVFVFNQTSWFTQ
RNFFSPQIITDNTFVSGNCDVIGIINNNTYDPLQPELDSFKEELDKYFKNHTSPDV
DLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYVWLGFAI
GLIAIMVTILLCCMTSCCCLKGACSCGSCCKFDEDDSEPVLGKVLHYT"
25268..26092
"SPYKE"
```

gene

Si possono confrontare le varie proteine Spyke come fatto per le sequenze virali, analizzando le similarità.

3. ALLINEAMENTO MULTIPLO

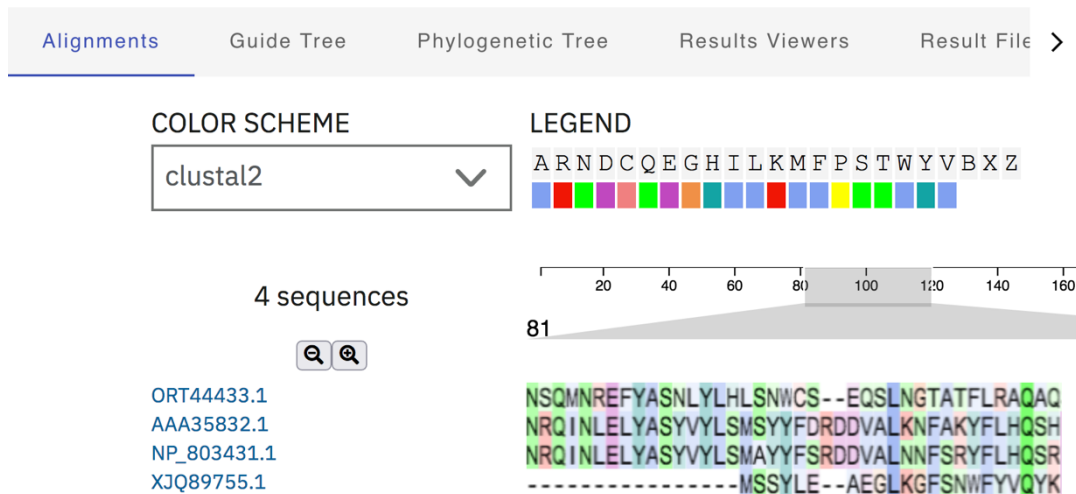
Possiamo considerare

- le 4 sequenze virali della famiglia Coronavirus dell'assignment precedente, oppure
- le 4 proteine Spike copiate e incollate in un file, oppure
- il file sulle sequenze aminoacidiche della ferritina caricate sul sito e anche in questa cartella
- le 7 sequenze di emoglobina indicate nella lezione del 20 novembre 2024, copiate nel file "globine.txt" in questa cartella.

Eseguire ClustalOmega dal sito <https://www.ebi.ac.uk/jdispatcher/msa/clustalo>

Il risultato mostrerà gli allineamenti sia in forma testuale, sia in forma grafica (Alignment) dove zoomando sarà possibile anche visualizzare la sequenza proteica (l'immagine è per le sequenze della ferritina).

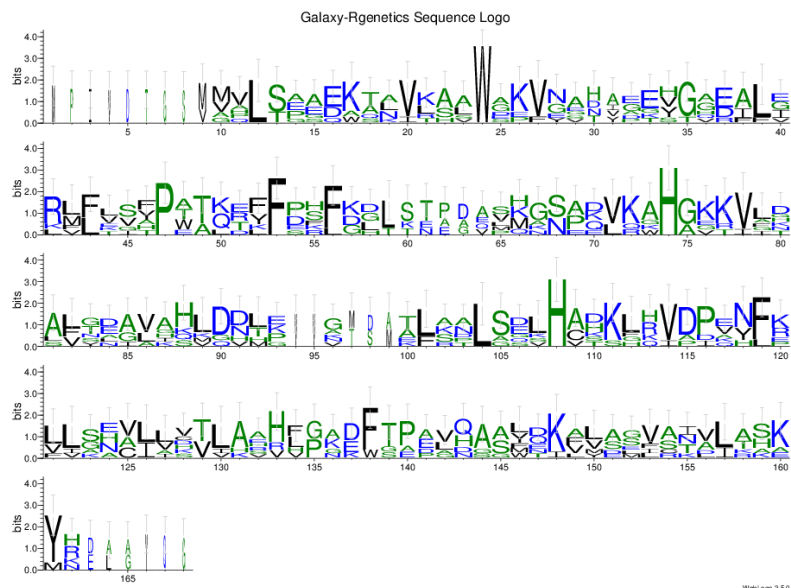
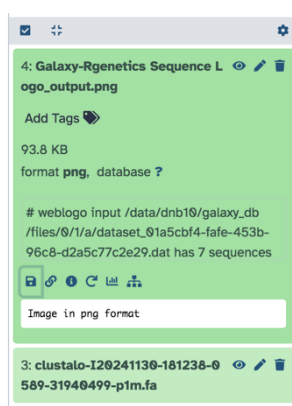
(ult. agg. 1.12.2024)



E' possibile visualizzare l'albero guidato e l'albero filogenetico. E' possibile anche invocare il tool MSView per visualizzare l'allineamento.

Nel menù "Result Files" è possibile dare il download dell'allineamento in formato FASTA e così poter utilizzare il tool **SequenceLogo** di Galaxy.

Ad esempio, il file "globine.txt" è stato copiato per eseguire ClustalOmega dal sito EBI. L'output è stato salvato nel formato FASTA (vedi sopra). Cercando "SequenceLogo" nei tool di Galaxy e caricando il file .fa, l'output viene visualizzato in formato png.



Predisponendo i vari file FASTA delle sequenze da allineare in una cartella, è possibile eseguire il tool da Galaxy e successivamente, attraverso il tool **Select Sequence**, ottenere la sequenza consenso.

4. SEQUENCE ASSEMBLY

Eseguire il tutorial su Galaxy

- <https://usegalaxy.eu/training-material/topics/assembly/tutorials/debruijn-graph-assembly/tutorial.html#de-bruijn-graph-assembly>
- Uso di Quast <https://usegalaxy.eu/training-material/topics/assembly/tutorials/assembly-quality-control/tutorial.html#genome-assembly-quality-control>
- Eseguire gli assignment di Compeau, caricati nella cartella (a,b).

5. MAPPING

Eseguire i tutorial su Galaxy

- <https://usegalaxy.eu/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html#mapping>
- <https://usegalaxy.eu/training-material/topics/sequence-analysis/tutorials/sars-with-galaxy-on-anvil/tutorial.html#sars-cov-2-viral-sample-alignment-and-variant-visualization>