Esercitazione: Allineamento di una Sequenza Sconosciuta

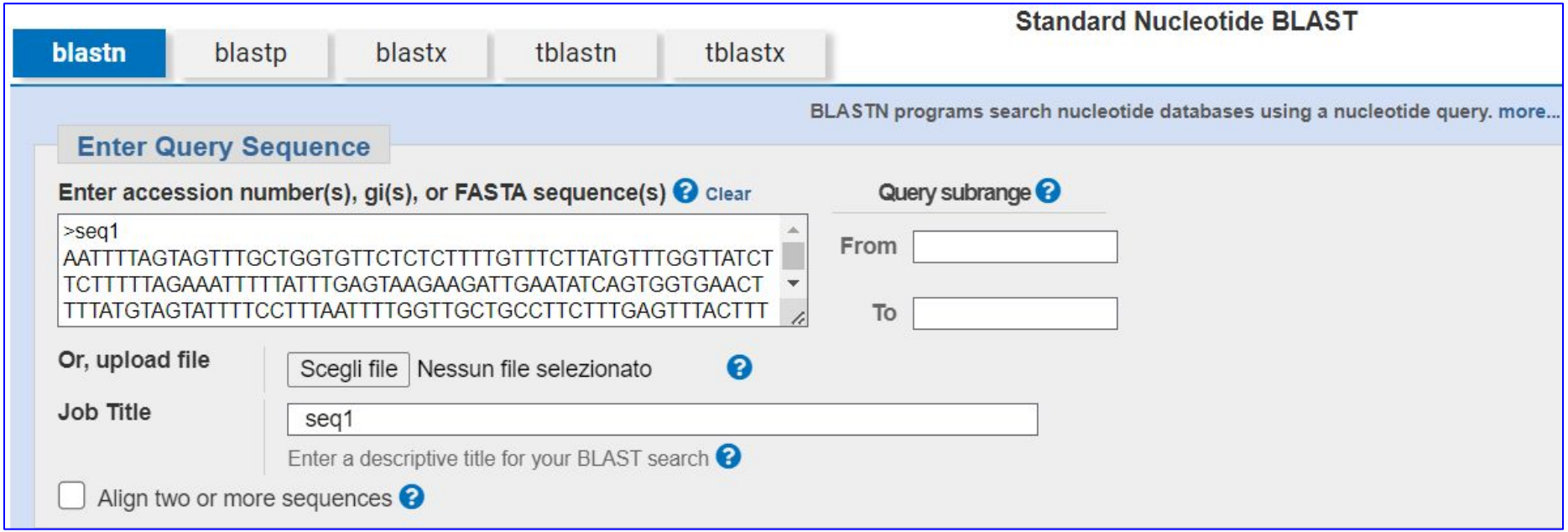
Supponiamo di aver isolato la seguente sequenza di DNA da un campione ambientale e di voler scoprire a quale organismo appartiene o se ha delle similitudini con altre sequenze note:

>seq1
AATTTTAGTAGTTTGCTGGTGTTCTCTCTTTTGTTTCTTATGTTTGGTTATCTTCTTTTTA
GAAATTTTTATTTGAGTAAGAAGATTGAATATCAGTGGTGAACTTTTATGTAGTATTTTCC
TTTAATTTTGGTTGCTGCCTTCTTTGAGTTTACTTTATTATTATGGTTTGATGAATCTTGAT
AGTAATTTGTCAAAGTTACTGGTCATCAGTGGTACTTTACGAGTTTAGTGATATTCCGGG
TAGAATTTGATTCTAAGTCTGTGGATCAGTTGGAGTTAGGTGAGCCTCGTTTGGAGGTT
TAATCGTTGTGTTGTTCCTTGTGATATTAACCGTTTTTGTATGGATGTTATTCATTCTTGT
TGGGCTTTGCCTAGCTATTAAGTTGGATGAGGGGTATTTTGTCTACTGTTTCTTATAGTT
TTCCTACTGTTGGTGATGGTCAATGTTCAGAGATTTGTGGGGCTATAGTTTTATGCCTAT
TGCTCTGGAAGTGTTATTG

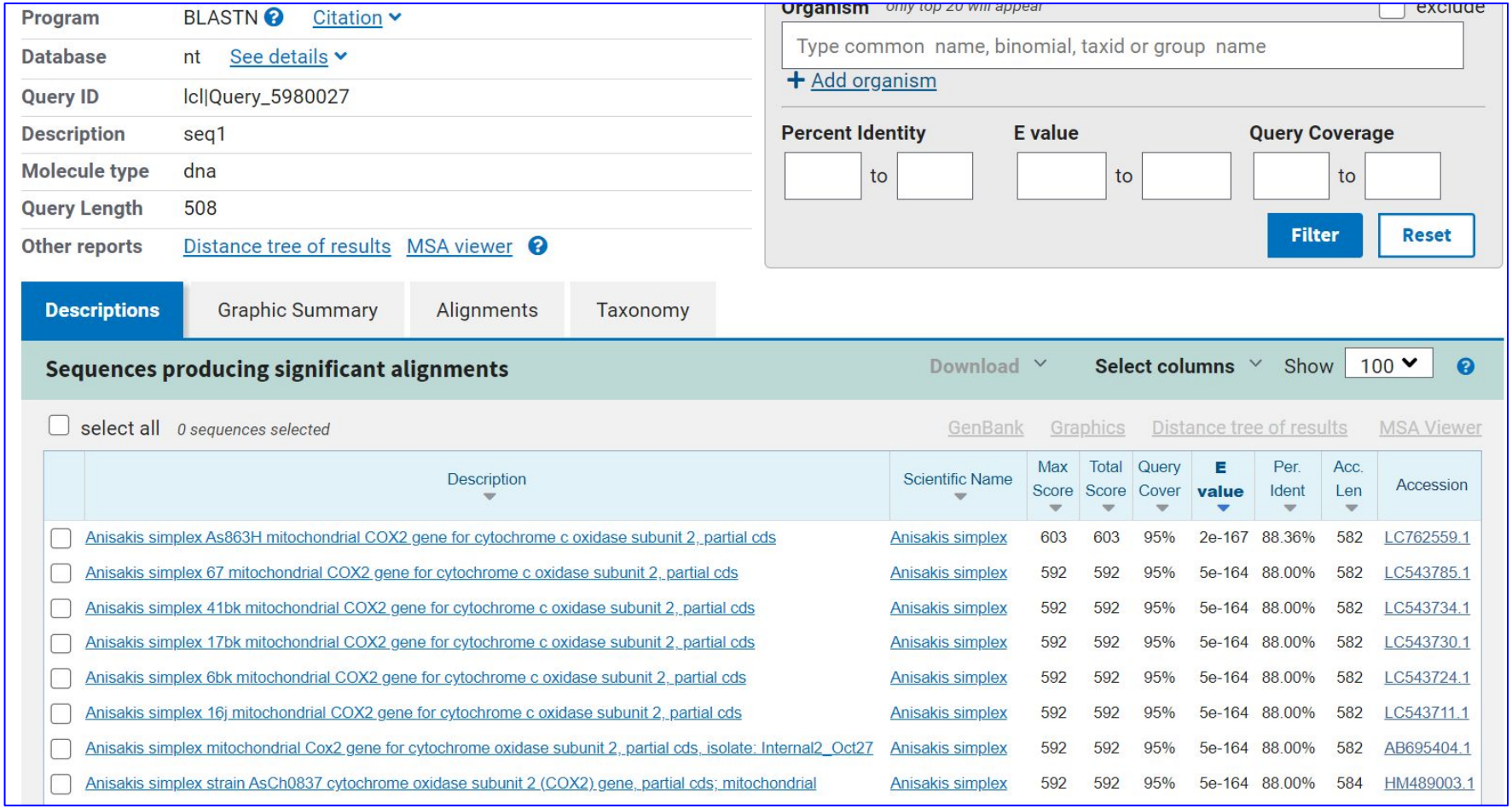
Identificazione mediante il portale blast di NCBI

Accediamo al Blastn di NCBI qui:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome



Screenshot della pagina dei risultati del Blast di NCBI



Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_5980027
Description	seq1
Molecule type	dna
Query Length	508
Other reports	Distance tree of results MSA viewer

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to

E value to

Query Coverage to

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *0 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Anisakis simplex As863H mitochondrial COX2 gene for cytochrome c oxidase subunit 2, partial cds	Anisakis simplex	603	603	95%	2e-167	88.36%	582	LC762559.1
Anisakis simplex 67 mitochondrial COX2 gene for cytochrome c oxidase subunit 2, partial cds	Anisakis simplex	592	592	95%	5e-164	88.00%	582	LC543785.1
Anisakis simplex 41bk mitochondrial COX2 gene for cytochrome c oxidase subunit 2, partial cds	Anisakis simplex	592	592	95%	5e-164	88.00%	582	LC543734.1
Anisakis simplex 17bk mitochondrial COX2 gene for cytochrome c oxidase subunit 2, partial cds	Anisakis simplex	592	592	95%	5e-164	88.00%	582	LC543730.1
Anisakis simplex 6bk mitochondrial COX2 gene for cytochrome c oxidase subunit 2, partial cds	Anisakis simplex	592	592	95%	5e-164	88.00%	582	LC543724.1
Anisakis simplex 16j mitochondrial COX2 gene for cytochrome c oxidase subunit 2, partial cds	Anisakis simplex	592	592	95%	5e-164	88.00%	582	LC543711.1
Anisakis simplex mitochondrial Cox2 gene for cytochrome oxidase subunit 2, partial cds, isolate: Internal2_Oct27	Anisakis simplex	592	592	95%	5e-164	88.00%	582	AB695404.1
Anisakis simplex strain AsCh0837 cytochrome oxidase subunit 2 (COX2) gene, partial cds; mitochondrial	Anisakis simplex	592	592	95%	5e-164	88.00%	584	HM489003.1

A partire dallo screenshot della slide precedente, rispondere alle seguenti domande:

- A) Inferire la sequenza query a quale specie appartiene
- B) Dire di quale gene si tratta
- C) Riportare il Phylum al quale appartiene la specie identificata
- D) verificare che ci sia almeno una release del genoma
- E) Se vi e' il genoma riportare
  - Genome size
  - Number of scaffolds
  - N50
  - GC percent
  - Genome coverage
- F) verificare che ci sia almeno una release del genoma
- G) dire quanti esperimenti di NGS sono presenti nell'SRA e quanti in particolare di trascrittomica

Identificazione della sequenza mediante il software diamond da riga di comando in ambiente linux

Di cosa abbiamo bisogno in input per poter lanciare il comando "diamond blastx"?

Abbiamo bisogno di 2 file:

1) il file della sequenza *query*, seq1.fa,

e

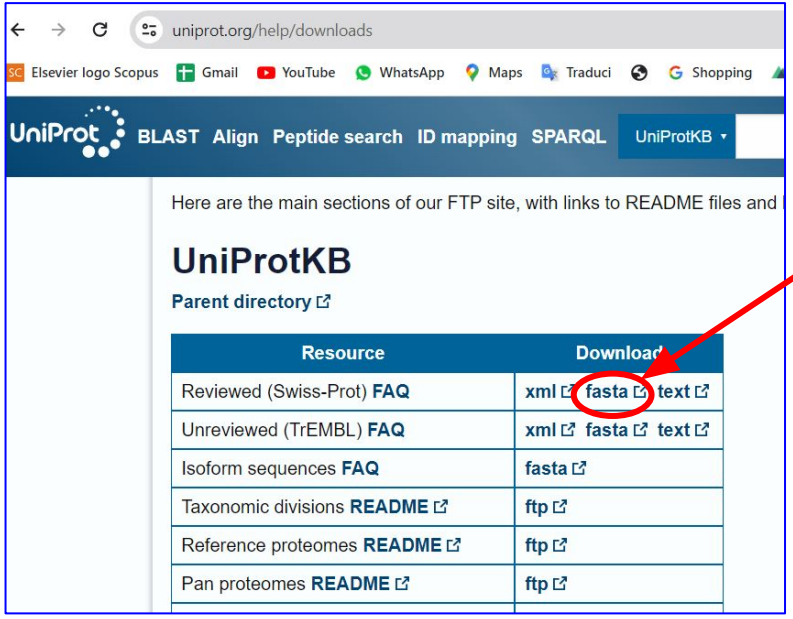
2) il file della banca dati SwissProt in formato fasta da indicizzare

Identificazione della sequenza mediante il software diamond da riga di comando in ambiente linux

- Creazione della sequenza in locale e downloading di SwissProt

Trasferiamo la sequenza “seq1.fa” su una macchina linux e utilizziamo il diamond per allineare con blastx contro la banca dati SwissProt, TR e NR. Sul vostro PC utilizzerete solo la banca dati SwissProt.

- con l’editor “nano” inseriamo la sequenza copiata dalla slide dentro il file seq1.fa
- scarichiamo la banca dati SwissProt ()



copiamo il link con il tasto destro del mouse e scarichiamo il db in locale sulla macchina con il comando wget

wget

https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz

Identificazione della sequenza mediante il software diamond da riga di comando in ambiente linux

- **downloading di diamond**

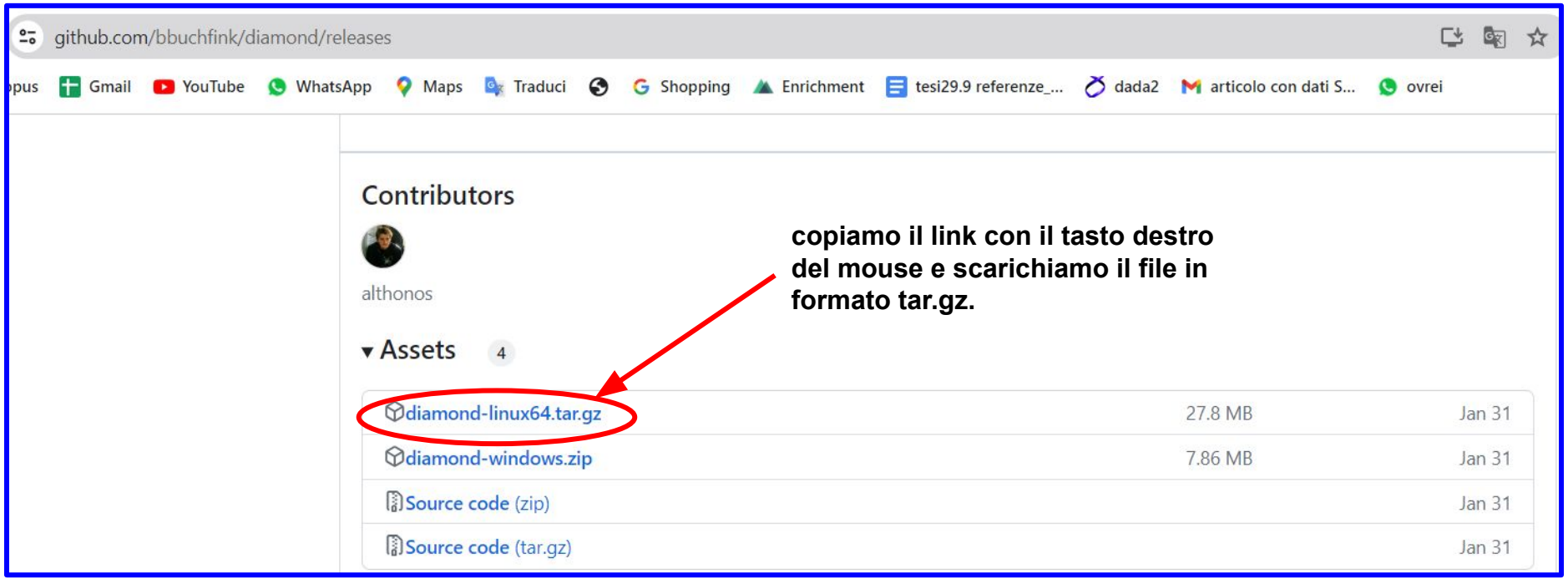
Colleghiamoci al link github delle releases di diamond:

https://github.com/bbuchfink/diamond/releases

GitHub è una piattaforma di hosting per il controllo di versione e la collaborazione tra sviluppatori, che permette loro di lavorare insieme a progetti software da qualsiasi parte del mondo. Fondato nel 2008, GitHub si basa sul sistema di controllo di versione distribuito Git, inventato da Linus Torvalds, il creatore di Linux. GitHub facilita la gestione di progetti software, consentendo agli utenti di caricare codice sorgente, collaborare con altri, tracciare e risolvere problemi, e implementare funzionalità di controllo di versione per tenere traccia e combinare modifiche al codice in un progetto condiviso.

Identificazione della sequenza mediante il software diamond da riga di comando in ambiente linux

- downloading di diamond



Identificazione della sequenza mediante il software diamond da riga di comando in ambiente linux

- estrazione della suite di diamond

Per estrarre un file con estensione `.tar.gz` su un sistema Linux o macOS, puoi usare il comando `tar` dalla linea di comando. Il comando completo per estrarre un file `.tar.gz` è:

tar xzvf diamond-linux64.tar.gz

- `x`: indica a `tar` di estrarre i file.
- `z`: dice a `tar` di decomprimere l'archivio (poiché `.gz` indica che è compresso con gzip).
- `v`: sta per "verbose", facoltativo, fa sì che `tar` mostri i nomi dei file mentre vengono estratti.
- `f`: permette di specificare il nome del file archivio.

Identificazione della sequenza mediante il software diamond da riga di comando in ambiente linux

- estrazione della suite di diamond

DIAMOND SI SCARICA UFFICILAMENTE CON

wget https://github.com/bbuchfink/diamond/archive/refs/heads/master.zip

L'ESSEGUIBILE DI DIAMOND LO TROVATE SUL SITO:

http://deb.scienceontheweb.net/BioInformatica1/lezioni/diamond.o

ED E' SCARICABILE SU LINUX CON IL COMANDO

wget https://deb.scienceontheweb.net/BioInformatica1/lezioni/diamond.o --no-check-certificate

UNA VOLTA SCARICATO RINOMINARE IL FILE CON IL COMANDO:

mv diamond.o diamond

ULTERIORE ALTERNATIVA LINUX MAC/OS

```
brew install diamond
```

SWISSPROT INDEXING

1) **Per prima cosa decomprimiamo il file fasta di SwissProt:**

gunzip uniprot_sprot.fasta.gz

2) una volta fatto il gunzip sul file uniprot_sprot usiamo la funzione makedb di diamond per creare l'indice della banca dati:

diamond makedb --in nomefile.fasta -d nomeDatabase

- `--in nomefile.fasta`: specifica il file (banca dati in formato fasta) di input. Sostituisci `nomefile.fasta` con il nome del tuo file FASTA contenente le sequenze di proteine che vuoi indicizzare.

- `-d nomeDatabase`: specifica il nome dell'indice del database da creare. Sostituisci `nomeDatabase` con il nome desiderato per il tuo database indice. Questo nome verrà utilizzato quando esegui ricerche utilizzando questo indice.

SWISSPROT INDEXING

Applichiamo il comando 2) della slide precedente per creare l'indice della banca dati uniprot_sprot:

diamond makedb --in uniprot_sprot.fasta -d sp

- `--in` ***uniprot_sprot.fasta***: e' la nostra banca dati di sequenze

  di aminoacidi in formato fasta che utilizziamo come input per l'indicizzazione.

- `-d` ***sp***: specifica il nome dell'indice del database (***uniprot_sprot.fasta)*** da creare.

  Questo nome verrà utilizzato quando eseguirete le ricerche su ***uniprot_sprot.fasta,*** utilizzando questo indice che avrà estensione dmnd (***sp.dmnd***).