



HPC-Annotator

Un'applicazione parallela per
l'annotazione di trascrittomi

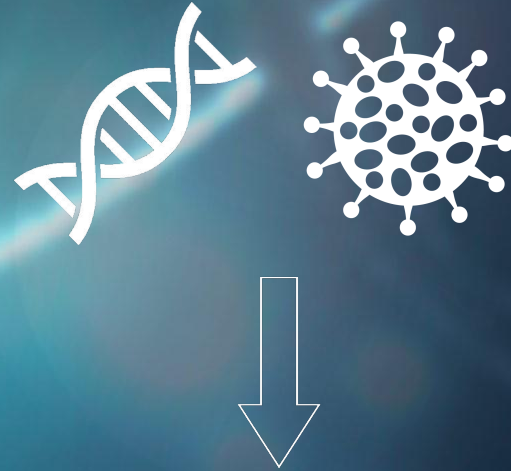


Introduzione

Motivazioni dietro lo
sviluppo di HPC-Annotator

I dati biologici

Le sequenze biologiche sono memorizzate in file in formato Multi-FASTA



Formato Multi-FASTA

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCCTCTTTCTTATCATTTGACATTTAACTCTGGGGCAGGTCCTCGCTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCCGCCCAAGTGGCCCCGGGGCTTGATTTTGTCTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGTGGAGGAGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGCTGCTGCTAAAGCCACTCGCGACCGCAAAAATGCA
GGAGGTGGGGACGCACCTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTCCC
GCGGTTGCAAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Il processo di annotazione

File
Multi-FASTA



Database delle
sequenze note



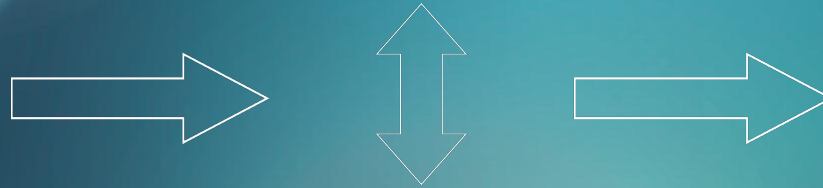
Informazioni delle sequenze simili
alle sequenze del file
Multi-FASTA

Il processo di annotazione



Biologo

Software di
annotazione



Database di
sequenze



Risultati



C'è un problema !

Le operazioni di annotazione richiedono
risorse computazionali ingenti.

I supercomputer



Biologo



Supercomputer



Software di
annotazione
(BLAST)



Database di
sequenze



C'è un altro problema !

I calcoli effettuati durante il processo di annotazione possono richiedere lo stesso molto tempo.

Logica dell'applicazione

Funzionamento ad alto livello
dell'algoritmo

Qual'è l'idea?

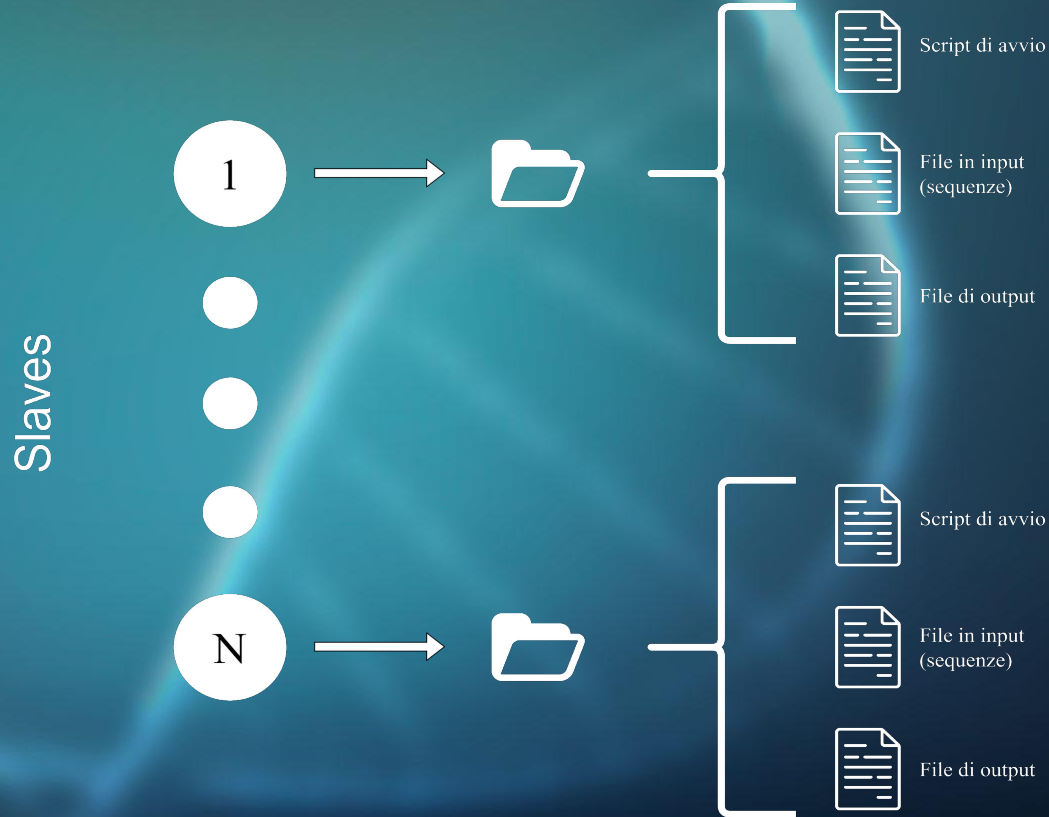




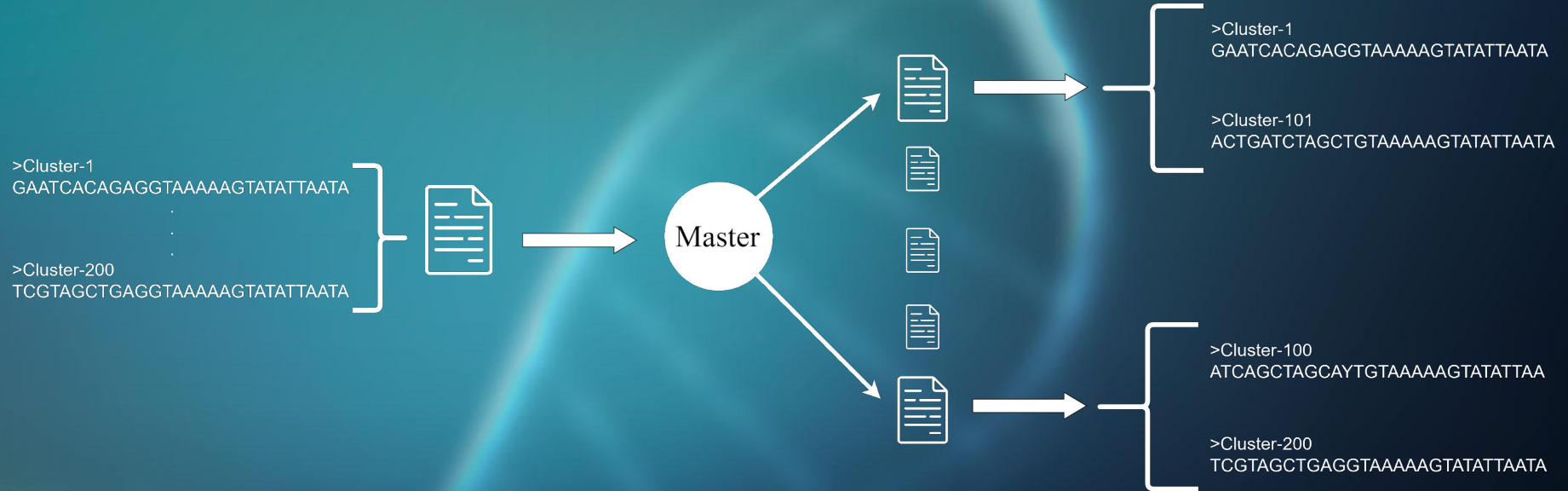
Vantaggi di questo approccio

- Tempi di esecuzione notevolmente ridotti.
- Possibilità di bypassare i limiti temporali imposti dallo scheduler.

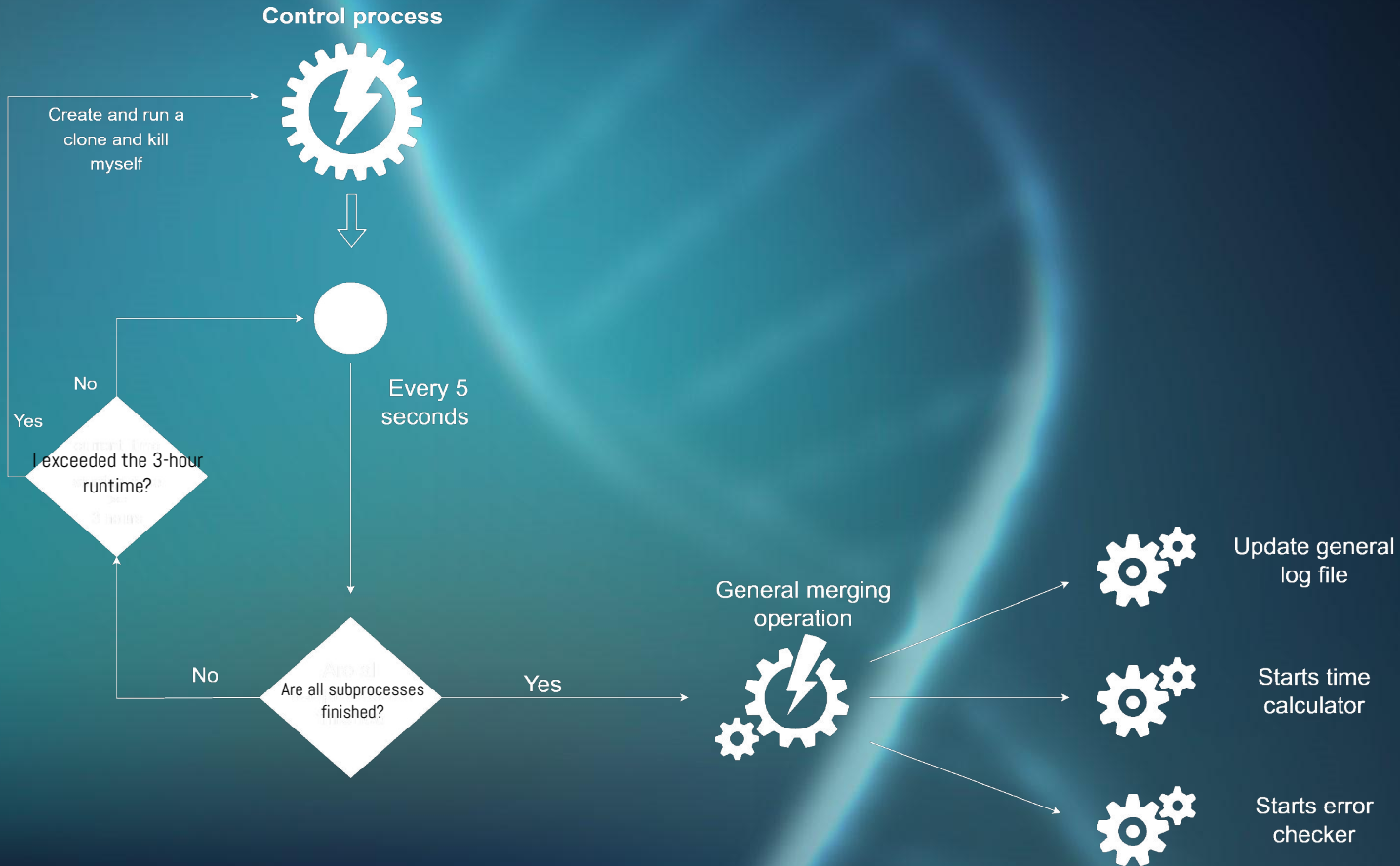
Cosa avviene in dettaglio



Cosa avviene in dettaglio



Il processo di controllo



Analisi delle prestazioni

Benchmark e analisi sui tempi di
esecuzione dell'applicazione

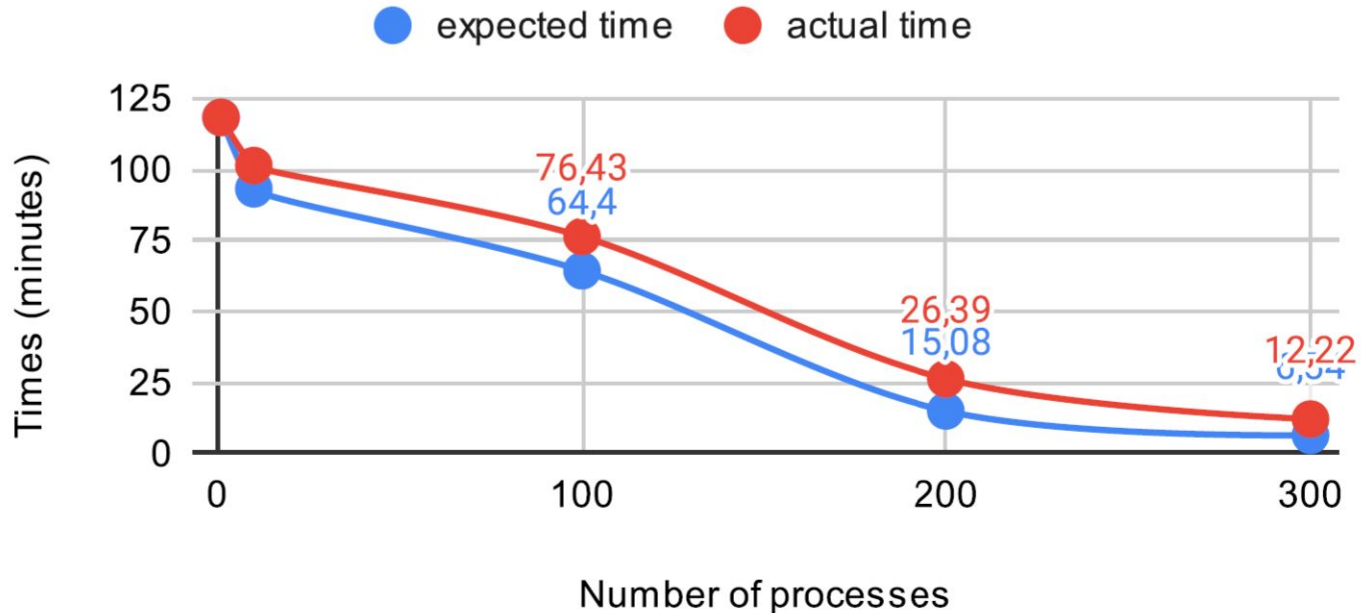
Analisi del trascrittoma di Hyla sarda

| | | | HYLA SARDA | | |
|-----------|-----------|---------|------------|-------------------------|-----------------------|
| Sequences | Processes | Diamond | Database | Expected time (minutes) | Actual time (minutes) |
| 1295741 | 1 | yes | Swiss-Prot | 118,32 | 118,32 |
| 1295741 | 10 | yes | Swiss-Prot | 93,21 | 101,38 |
| 1295741 | 100 | yes | Swiss-Prot | 64,4 | 76,43 |
| 1295741 | 200 | yes | Swiss-Prot | 15,08 | 26,39 |
| 1295741 | 300 | yes | Swiss-Prot | 6,54 | 12,22 |

Grafico dei tempi di esecuzione

Hyla Sarda time analysis

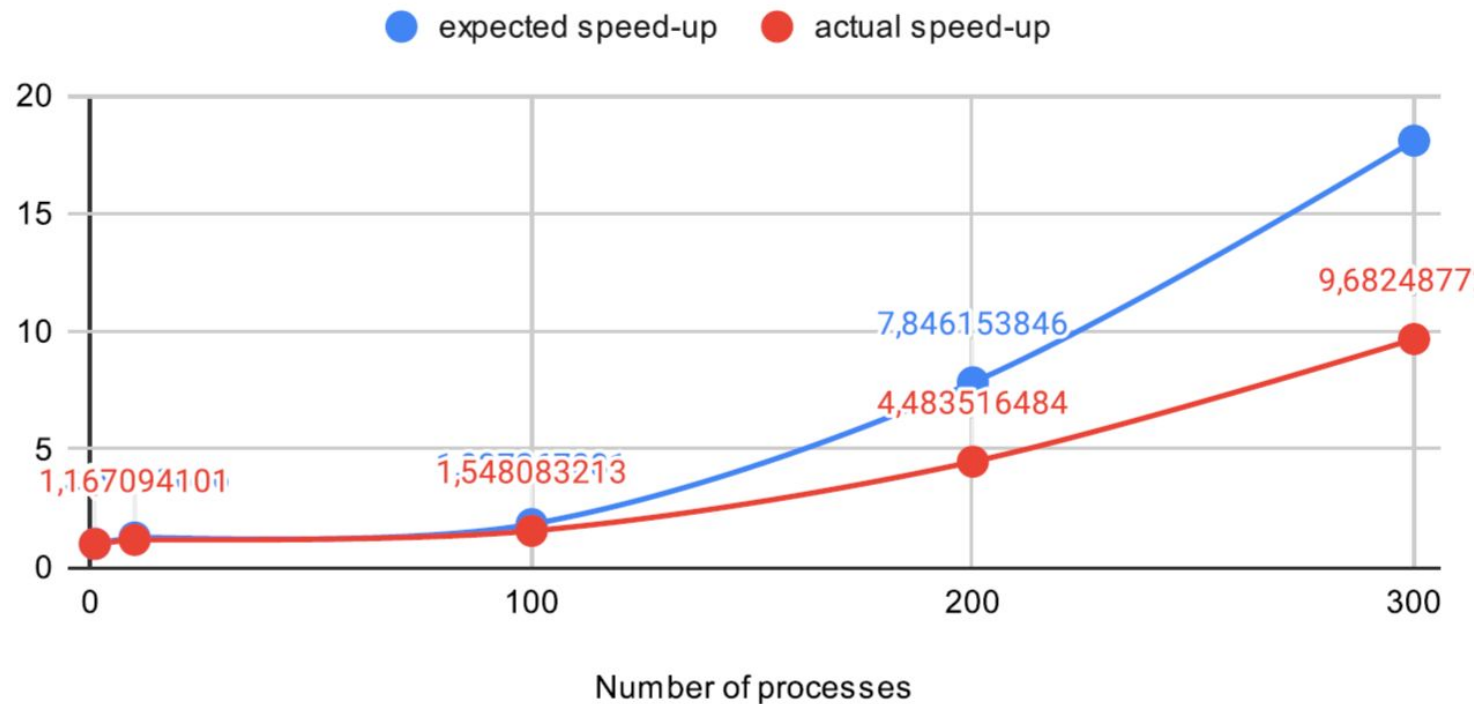
On Swiss-prot database



Speed-Up

Hyla Sarda Speed-Up analysis

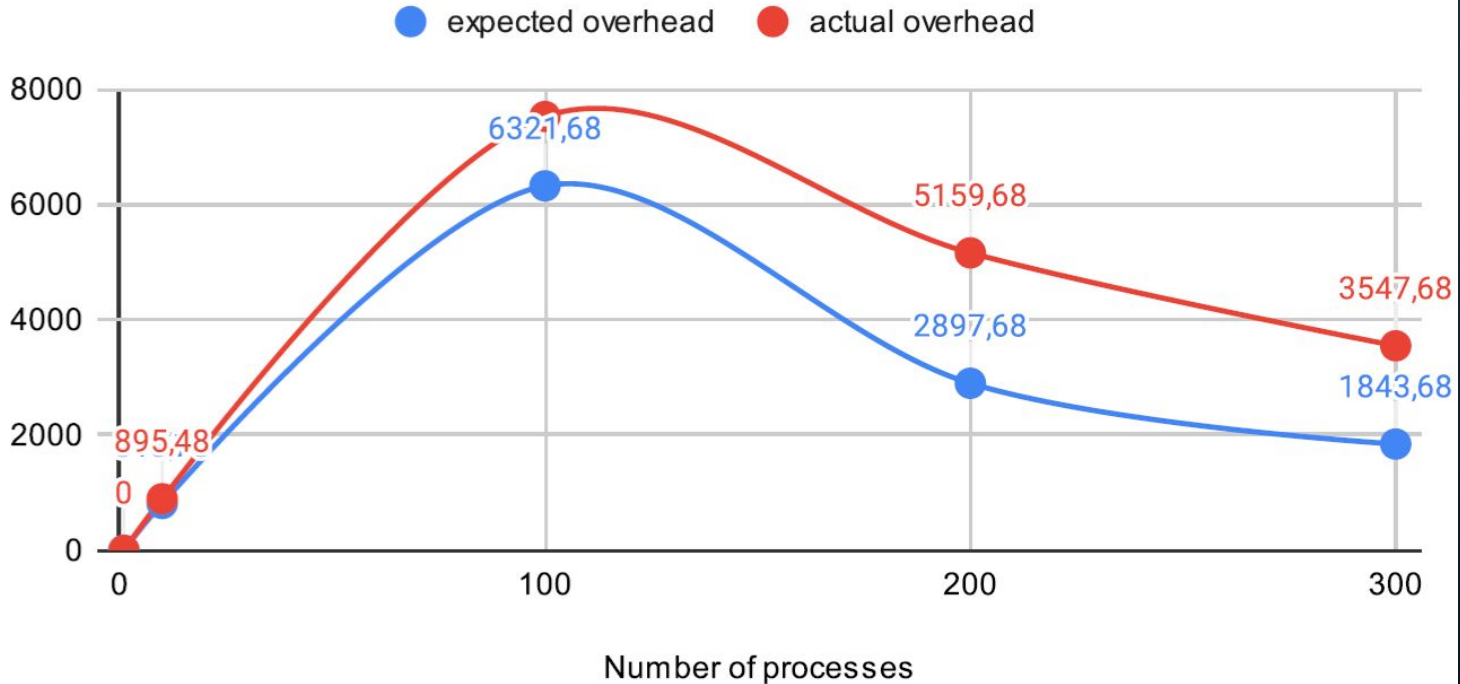
On Swiss-prot database



Overhead

Hyla Sarda Overhead analysis

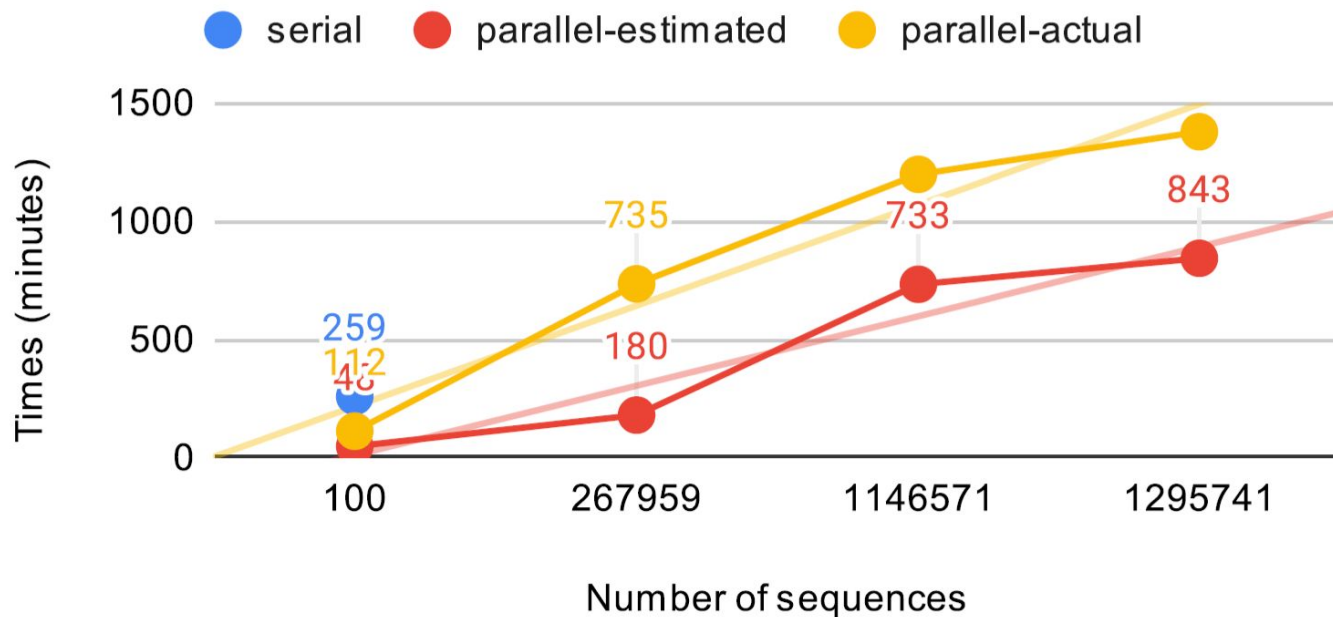
On Swiss-prot database



Benchmark sulla banca dati TrEMBL

Diamond time analysis

On TrEMBL database

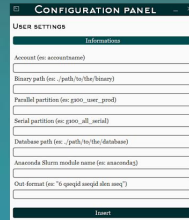


Interfaccia grafica

Tramite la GUI l'utente genera il
codice di HPC-Annotator

Interfacce

Pannello di configurazione



The Configuration Panel is a web-based interface for setting up the HPC-ANNOTATOR. It contains several input fields for user settings, including account, binary path, parallel partitions, serial partitions, database path, and output format. A 'Load' button is at the bottom.

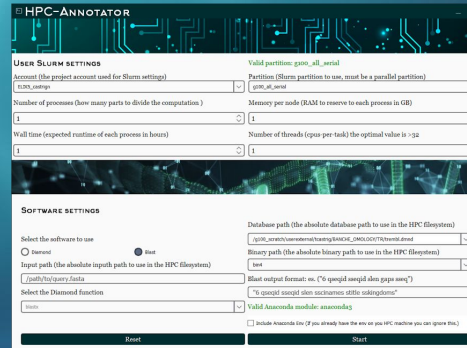
| CONFIGURATION PANEL | |
|--|--|
| USER SETTINGS | |
| Informations | |
| Account (see accountname) | |
| Binary path (see ./path/to/the/binary) | |
| Parallel partitions (see gpus, user, gpus) | |
| Serial partitions (see gpus, all, serial) | |
| Database path (see ./path/to/the/database) | |
| Access to the module name (see accountid) | |
| Out format (see "hpcannotator output file name") | |
| Load | |

JSON

Database json di configurazione

Acquisizione del file json

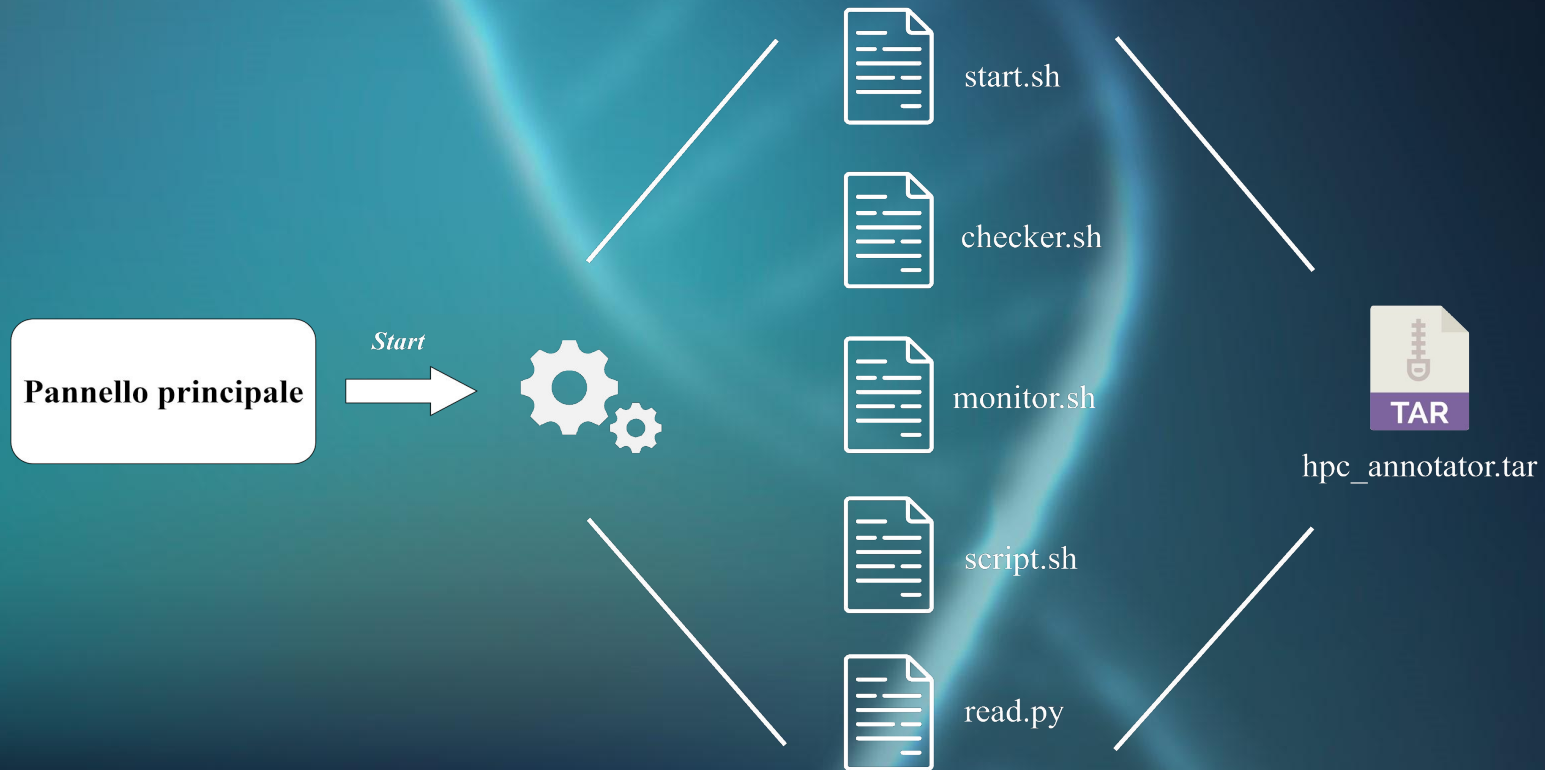
Pannello principale



The HPC-ANNOTATOR main panel is a web-based interface for running the annotation process. It contains two main sections: USER SLURM SETTINGS and SOFTWARE SETTINGS. The USER SLURM SETTINGS section includes input fields for account, partition, number of processes, memory per node, wall time, and number of threads. The SOFTWARE SETTINGS section includes input fields for database path, binary path, input path, output format, and output file name. A 'Start' button is at the bottom.

| HPC-ANNOTATOR | |
|--|--|
| USER SLURM SETTINGS | |
| Account (the project account used for Slurm settings) | Partition (Slurm partition to use, must be a parallel partition) |
| Number of processes (how many parts to divide the computation) | Memory per node (RAM to reserve to each process in GB) |
| Wall time (expected runtime of each process in hours) | Number of threads (cpus per task) the optimal value is >32 |
| Start | |
| SOFTWARE SETTINGS | |

Funzionamento dell'interfaccia



Tool di post-processing

Jupyter notebook per il
post-processing dei dati di
annotazione

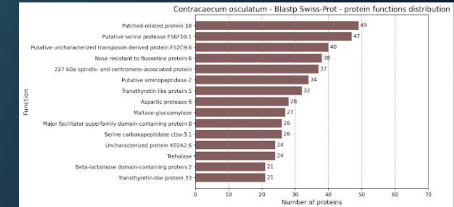
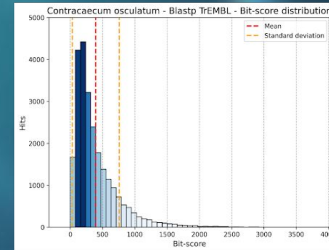
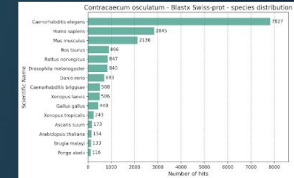
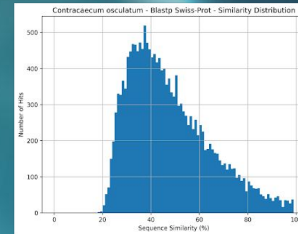
Generazione di report e grafici



Output
annotazione

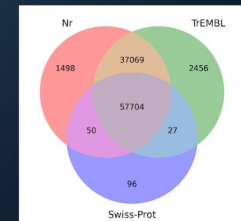


Notebook post-processing

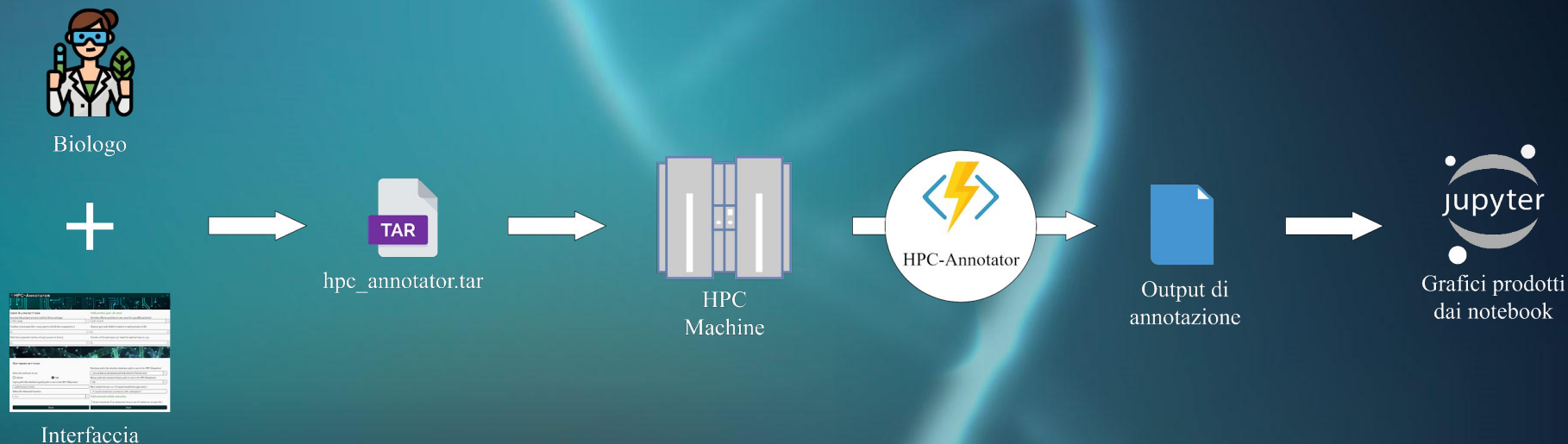


Contracaecum osculatum - Hits Table

| Database | Number of BLASTX results | Number of BLASTP results |
|------------|--------------------------|--------------------------|
| NR | 29694 (80.3 %) | 24366 (65.9 %) |
| TrEMBL | 29904 (80.9 %) | 24600 (66.5 %) |
| Swiss-prot | 20660 (55.9 %) | 17239 (46.6 %) |



Pipeline bioinformatica



Grazie per l'attenzione

L'intero software è disponibile con licenza gratuita ed open-source a questo indirizzo:

<https://github.com/lorenzo-arcioni/HPC-Annotator>

Candidato: Lorenzo Arcioni
Responsabile: Tiziana Castrignanò
Corresponsabile: Paolo Bottoni