UNIVERSITÀ DEGLI STUDI DI MILANO

CONCORSO PUBBLICO, PER ESAMI, A N. 1 POSTO DI CATEGORIA D - AREA TECNICA, TECNICO-SCIENTIFICA ED ELABORAZIONE DATI, CON RAPPORTO DI LAVORO SUBORDINATO A TEMPO INDETERMINATO PRESSO L'UNIVERSITÀ DEGLI STUDI DI MILANO - DIPARTIMENTO DI SCIENZE E POLITICHE AMBIENTALI - PROGETTO DIPARTIMENTI DI ECCELLENZA 2023/2027 - CODICE 22379

La Commissione giudicatrice della selezione, nominata con Determina Direttoriale n. 20920 del 19/12/2023, composta da:

Prof.ssa Paola Causin	Presidente
Prof. Federico Zambelli	Componente
Dott. Uliano Guerrini	Componente
Sig.ra Serenella Ricci	Segretaria

comunica le tracce relative alla prima prova scritta:

TRACCIA n. 1

Il candidato descriva le principali disposizioni della policy sulla gestione dei dati della ricerca dell'Università di Milano, discutendone gli ambiti di applicazione, le modalità di trattamento dei dati, e le principali responsabilità degli afferenti all'Università di Milano in merito ai dati della ricerca.

Il candidato consideri il seguente scenario: un ricercatore in ambito socio-economico sta studiando la relazione tra reddito e stato di salute in una popolazione di anziani. Essendo parte di un grande progetto, il ricercatore possiede i dati relativi a 50000 persone, di cui sono stati registrati reddito lordo annuale e un indicatore soggettivo dello stato di salute complessivo.
Quali suggerimenti possono essere dati al ricercatore per analizzare quantitativamente i dati? Il candidato illustri le metodologie che ritiene più opportune, evidenziandone punti di forza e debolezza. Illustri inoltre quali strumenti software possono essere adottati per la analisi dei dati, assumendo che il ricercatore in questione non possieda forti competenze nella programmazione informatica e che necessiti quindi di un significativo supporto in questa fase

TRACCIA n. 2

In luce della policy dell'Università di Milano sulla gestione dei dati della ricerca, il candidato selezioni una categoria di dati biologici con cui ha dimestichezza (ad es. epidemiologici, genomici, trascrittomici, metagenomici, proteomici, cristallografici, etc…) e fornisca un'analisi dell'applicazione dei principi FAIR (Findable, Accessible, Interoperable, Reusable) attraverso esempi concreti o ipotetici.

Il candidato consideri il seguente scenario: un ricercatore in ambito biologico sta studiando campioni di acqua prelevati da zone possibilmente contaminate. Essendo parte di un grande progetto, il ricercatore possiede i dati relativi a 60000 campioni, di cui sono stati registrati pH, durezza e concentrazioni di cinque specie chimiche utili alle analisi.
Quali suggerimenti possono essere dati al ricercatore per analizzare quantitativamente i dati e in particolare stabilire se il campione prelevato è considerato potabile? (si assumano note le soglie massime e/o i range di valori che rendono il campione potabile). Il candidato illustri le metodologie che ritiene più opportune, evidenziandone punti di forza e debolezza. Illustri inoltre quali strumenti

software possono essere adottati per la analisi dei dati, assumendo che il ricercatore in questione non possieda forti competenze nella programmazione informatica e che necessiti quindi di un significativo supporto in questa fase

TRACCIA n. 3

La policy dell'Università di Milano sulla gestione dei dati della ricerca incoraggia la condivisione dei dati scientifici secondo il paradigma FAIR (Findable, Accessible, Interoperable, Reusable). Il candidato illustri come i principi FAIR applicati ai dati della ricerca siano utili a recepire e implementare nella pratica il paradigma dell'Open Science.

Il candidato consideri il seguente scenario: un ricercatore in biologia dispone di immagini ottenute al microscopio confocale di una larga popolazione di cellule endoteliali. In particolare, desidera sviluppare un flusso di lavoro di analisi delle immagini per segmentare accuratamente le cellule e i loro organelli, estrarre i dati morfometrici rilevanti e assegnare gli organelli alle loro cellule.
Quali suggerimenti possono essere dati al ricercatore per sviluppare il flusso di lavoro? Il candidato illustri le metodologie che ritiene più opportune, evidenziandone punti di forza e debolezza. Illustri inoltre quali strumenti software possono essere adottati per l'analisi, assumendo che il ricercatore in questione non possieda forti competenze in questo ambito e che necessiti quindi di un significativo supporto in questa fase.

La Commissione comunica le tracce relative alla seconda prova scritta a contenuto teorico-pratico:

TRACCIA n. 1

In bioinformatica, il formato FASTA è un modo standard di rappresentare sequenze di acidi nucleici (DNA o RNA) mediante semplici stringhe di caratteri. In questo formato, la stringa che rappresenta la sequenza nucleotidica è preceduta da una singola riga di intestazione (header) che inizia con il carattere '>', seguito da un identificatore della sequenza. Nella riga dopo l'intestazione, segue la sequenza di basi nucleotidiche che compone la molecola, rappresentata come una serie di lettere (adenina [A], citosina [C], guanina [G] e timina [T]).
Esempio:
>seq_1
AGCTTAGCTAGCTTACGATCGATCGTACGAT
Un "k-mero" è una sottostringa di lunghezza k presente all'interno di una sequenza di DNA (o RNA). In altre parole, è una sequenza consecutiva di k nucleotidi contenuta nella sequenza più ampia. Ad esempio, nella sequenza "seq_1" sopra riportata, i primi due tri-meri sono costituiti da 'AGC' e 'GCT'. Il valore di k rappresenta la lunghezza di queste sottostringhe.
Il candidato descriva con pseudocodice un algoritmo di analisi di sequenza che implementi quanto segue.
L'algoritmo dovrà analizzare la sequenza S di una molecola di DNA fornita in input in formato FASTA per determinare la frequenza f con cui appare in S ciascun k-mero m di lunghezza k. Il valore di k è determinato dall'utilizzatore tramite un parametro impostabile che può assumere valori interi compresi tra 2 e 12. L'output atteso consiste in un file CSV o TSV su due colonne riportante la frequenza fm per ogni m per cui fm > 0.
Il candidato illustri almeno una parte dell'algoritmo sviluppato utilizzando uno dei seguenti linguaggi di programmazione: C, C++, R o Python.
Infine, il candidato dovrà scrivere una documentazione sintetica per il codice sviluppato, spiegando chiaramente la logica e il flusso dell'algoritmo, oltre a qualsiasi assunzione o decisione presa durante lo sviluppo. La documentazione dovrebbe essere sufficientemente dettagliata da permettere a un altro sviluppatore di

comprendere e potenzialmente estendere o modificare il codice e ad un utilizzatore di capire come eseguire il codice sviluppato.

TRACCIA n. 2

In bioinformatica, il formato FASTA è un modo standard di rappresentare sequenze di acidi nucleici (DNA o RNA) o proteine mediante semplici stringhe di caratteri. In questo formato, la stringa che rappresenta la sequenza nucleotidica è preceduta da una singola riga di intestazione (header) che inizia con il carattere '>', seguito da un identificatore della sequenza. Nella riga dopo l'intestazione, segue la sequenza di basi nucleotidiche che compone la molecola, rappresentata come una serie di lettere (adenina [A], citosina [C], guanina [G] e timina [T]).
La traduzione degli mRNA in proteine da parte del ribosoma funziona attraverso il cosiddetto “codice genetico” che associa a ciascuna tripletta di basi un amminoacido secondo la tabella riportata in calce.
Una Open Reading Frame (ORF) è una sequenza codificante in un mRNA che inizia sempre con una tripletta di start (ATG) e finisce sempre con una tripletta di stop (TAA, TAG, TGA) e contiene un certo numero di altre triplette. La tripletta di start solitamente si trova non esattamente all’inizio della sequenza del mRNA ma dopo un certo numero di basi chiamato regione non tradotta (UTR).
Tra tutte le possibili ORF contenute in un mRNA, quella che solitamente viene tradotta, detta CDS, è quella più lunga, ovvero costituita dal maggior numero di triplette.
Il candidato descriva con pseudocodice un algoritmo di analisi di sequenza che implementi quanto segue.
L'algoritmo dovrà analizzare la sequenza completa in formato FASTA di un mRNA fornito in input e identificare al suo interno tutte le possibili Open Reading Frames (ORF). Tra queste, l’algoritmo dovrà selezionare quella che più plausibilmente rappresenta la Coding Sequence (CDS). Successivamente, l'algoritmo dovrà tradurre la CDS selezionata in una sequenza di amminoacidi. L'output atteso è la sequenza di amminoacidi, in formato FASTA, corrispondente alla CDS identificata.
Il candidato illustri almeno una parte dell’algoritmo sviluppato utilizzando uno dei seguenti linguaggi di programmazione: C, C++, R o Python.
Infine, il candidato dovrà scrivere una documentazione sintetica per il codice sviluppato, spiegando chiaramente la logica e il flusso dell'algoritmo, oltre a qualsiasi assunzione o decisione presa durante lo sviluppo. La documentazione dovrebbe consentire ad un utilizzatore di capire come eseguire il codice sviluppato.
Tripletta Ammino
acido Tripletta Ammino
acido Tripletta Ammino
acido Tripletta Ammino
acido
TTT F TAT S TAT Y TGT C
TTC F TAC S TAC Y TGC C
TTA L TAA S TAA STOP TGA STOP
TTG L TAG S TAG STOP TGG W
CTT L CCT P CAT H CGT R
CTC L CCC P CAC H CGC R
CTA L CCA P CAA Q CGA R
CTG L CCG P CAG Q CGG R
ATT I ACT T AAT N AGT S
ATC I ACC T AAC N AGC S
ATA I ACA T AAA K AGA R
ATG M (START) ACG T AAG K AGG R
GTT V GCT A GAT D GGT G
GTC V GCC A GAC D GGC G
GTA V GCA A GAA E GGA G

GTG V GCG A GAG E GGG G

TRACCIA n. 3

In bioinformatica, il formato FASTA è un modo standard di rappresentare sequenze di acidi nucleici (DNA o RNA) mediante semplici stringhe di caratteri. In questo formato, la stringa che rappresenta la sequenza nucleotidica è preceduta da una singola riga di intestazione (header) che inizia con il carattere '>', seguito da un identificatore della sequenza. Nella riga dopo l'intestazione, segue la sequenza di basi nucleotidiche che compone la molecola, rappresentata come una serie di lettere (adenina [A], citosina [C], guanina [G] e timina [T]).
Esempio:
>seq_1
AGCTTAGCTAGCTTACGATCGATCGTACGAT
Un “k-mero” è una sottostringa di lunghezza k presente all'interno di una sequenza di DNA (o RNA). In altre parole, è una sequenza consecutiva di k nucleotidi contenuta nella sequenza più ampia. Ad esempio, nella sequenza “seq_1” sopra riportata, i primi due tri-meri sono costituiti da ‘AGC’ e ‘GCT’. Il valore di k rappresenta la lunghezza di queste sottostringhe.
Un mismatch tra due k-meri di ugual k è definito come la distanza di Hamming tra le due stringhe, ovvero il numero di posizioni nelle quali i simboli corrispondenti sono diversi. Ad esempio, AAA e AAC hanno una distanza di Hamming di 1.
Il candidato descriva con pseudocodice un algoritmo di analisi di sequenza che implementi quanto segue.
L'algoritmo dovrà prendere in input una sequenza S di una molecola di DNA fornita in formato FASTA, un k-mero K con k compreso tra 3 e 12, e un numero m di mismatch compreso tra 0 e 3. L’algoritmo dovrà riportare in output tutte le posizioni della sequenza S in cui K compare con al più m mismatch.
Il candidato illustri almeno una parte dell’algoritmo sviluppato utilizzando uno dei seguenti linguaggi di programmazione: C, C++, R o Python.
Infine, il candidato dovrà scrivere una documentazione sintetica per il codice sviluppato, spiegando chiaramente la logica e il flusso dell'algoritmo, oltre a qualsiasi assunzione o decisione presa durante lo sviluppo. La documentazione dovrebbe essere sufficientemente dettagliata da permettere a un altro sviluppatore di comprendere e potenzialmente estendere o modificare il codice e ad un utilizzatore di capire come eseguire il codice sviluppato.

Milano, 25 gennaio 2024

La Commissione

Prof.ssa Paola Causin Presidente

Prof. Federico Zambelli Componente

Dott. Uliano Guerrini Componente

Sig.ra Serenella Ricci Segretaria