UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Elementi di Bioinformatica

2627-3-E3101Q116

Obiettivi

Alla fine del corso lo studente avrà compreso le principali problematiche computazionali e tecniche algoritmiche in bioinformatica. Sarà in grado di scrivere programmi Python di modeste dimensioni per risolvere problemi in bioinformatica utilizzando anche dati presenti su basi di dati pubbliche

Conoscenza e capacità di comprensione

Questo insegnamento fornisce conoscenze basilari e capacità di comprensione relativamente a:

- Algoritmi su sequenze biologiche
- Strutture dati per l'indicizzazione di sequenze
- Algoritmi per la ricostruzione di storie evolutive
- Programmazione in Pyhton in campo bioinformatico
- Formati di file usati in bioinformatica

Conoscenza e capacità di comprensione applicate

Alla fine dell'insegnamento gli studenti saranno in grado di:

- Comprendere quali algoritmi e strutture dati utilizzare per affrontare alcuni problemi in bioinformatica
- Scrivere programmi in Python in campo bioinformatico
- Costruire semplici pipeline di analisi di dati bioinformatici

Autonomia di giudizio

Alla fine dell'insegnamento gli studenti saranno in grado di interpretare i risultati di un'analisi bioinformatica.

Abilità comunicative

Alla fine dell'insegnamento gli studenti saranno in grado di comunicare i risultati di un'analisi bioinformatica.

Capacità di apprendere

Alla fine dell'insegnamento gli studenti avranno competenze per:

- apprendere autonomamente nuove librerie e strumenti software in bioinformatica.
- affrontare nuovi problemi computazionali nel campo della genomica.

Contenuti sintetici

Principali problemi e algoritmi in bioinformatica. Pattern Matching. Allineamento di sequenze. Sequenziamento di DNA. Storie evolutive. Gestione di dati (formati standard in Bioinformatica) e banche dati genomiche.

Programma esteso

1. Introduzione ai dati genomici: sequenze primarie di DNA, RNA e proteine (formato standard FASTA).
2. Espressione genica e splicing alternativo (formati standard BED e GTF).
3. Sequenziamento di un genoma e qualità del dato di sequenziamento (formato standard FASTQ).
4. Allineamento globale e locale di due sequenze: algoritmi di Needleman-Wunsch, Smith-Waterman e Gotoh.
5. Allineamento semi-globale di due sequenze.
6. Allineamento con banda di due sequenze.
7. Allineamento multiplo.
8. Pattern matching esatto: algoritmi di Rabin-Karp e di Dömölki (paradigma shift-and).
9. Indicizzazione di testi: suffix tree, suffix array e loro utilizzo per risolvere problemi su stringhe (tra cui pattern matching esatto e ricerca della sottostringa comune più lunga).
10. Assemblaggio di un genoma: grafo di overlap e grafo di de Bruijn.
11. Alberi evolutivi. Modelli a partire da caratteri: algoritmo di Gusfield per la filogenesi perfetta. Modelli a partire da distanze: distanze ultrametriche e distanze additive. Algorithmi UPGMA e Neighbor Joining. Cenni di ricostruzione tramite massima verosimiglianza.
12. Genotipi e Aplotipi. Distinzione fra singolo individuo e pedigree.
13. Introduzione a Python.
14. Pandas (manipolazione di data frame).
15. Biopython.

Prerequisiti

1. Complessità in tempo e spazio.
2. Strutture dati fondamentali: liste, array, alberi di ricerca, dizionari
3. Algoritmi di ordinamento: radix sort, merge sort.
4. Gerarchia delle memorie.
5. Capacità di scrivere un breve programma in qualunque linguaggio di programmazione.

Modalità didattica

Lezioni in aula e attività di laboratorio. Viene utilizzata la piattaforma di e-learning per integrare lo studio individuale tramite arricchimento delle attività in aula e per autovalutazioni in itinere del livello di preparazione ottenuto.

Tutte le attività sono tenute in presenza e vengono registrate, ma non trasmesse in streaming. L'insegnamento è erogato in Italiano. Le attività previste sono:

- 36 ore di lezioni in aula da 2/3 ore svolte in modalità erogativa
- 24 attività di laboratorio da 2 ore svolte in modalità erogativa nella parte iniziale ed in modalità interattiva nella parte successiva

Materiale didattico

Il libro di testo seguito per quasi tutte le lezioni è “Algorithms on Strings, Trees and Sequences”, di Daniel Gusfield, Cambridge Univ. Press. La biblioteca tiene alcune copie del libro di testo, anche come ebook.

Il libro "An Introduction to Bioinformatics Algorithms" di N. Jones, P. Pevzner viene usato esclusivamente come approfondimento per la parte di ricostruzione di filogenesi e di sequenziamento.

Il libro "Theoretical Evolutionary Genetics" di J. Felsenstein viene usato esclusivamente come approfondimento per la parte di ricostruzione di filogenesi. Un libro analogo è Population and Quantitative Genetics by Graham Coop.

Per quanto riguarda Python fare riferimento a Pensare in Python di A. B. Downey (versione inglese Think Python).

Per la parte relativa a Pandas il libro di riferimento è Python Data Science Handbook di VanderPlas

Periodo di erogazione dell'insegnamento

Secondo semestre

Modalità di verifica del profitto e valutazione

La verifica dell'apprendimento consiste di una prova scritta e di una parte progettuale.

La prova scritta è individuale, basata su domande a risposta aperta relative ai contenuti del corso di natura algoritmica. La prova scritta dura un'ora e contiene 4 domande, tra cui lo studente dovrà sceglierne 3 a cui rispondere.
Le competenze che vengono valutate sono: capacità di astrazione, conoscenza degli algoritmi trattati, capacità di analisi della complessità computazionale.
La prova scritta viene valutata in trentesimi in base alla correttezza e completezza delle risposte e alla capacità di individuare gli elementi essenziali di un argomento.

La parte progettuale consiste nella scrittura di un programma in Python e viene svolta individualmente, con successiva presentazione del lavoro fatto. Gli studenti possono scegliere il testo del progetto tra temi alternativi proposti dal docente al termine delle lezioni.

La discussione del progetto avviene dopo avere superato la prova scritta ed entro un anno da tale superamento.

La valutazione finale viene ottenuta tramite media pesata delle votazioni (in trentesimi) ottenute nelle due parti, con peso 50% per la prova scritta e 50% per il progetto. Entrambe le parti devono avere valutazione positiva.

Si ricorda che è necessario essere iscritti alle prove d'esame tramite segreterie online. Non ci saranno eccezioni al riguardo.

Non sono previste prove in itinere.

Orario di ricevimento

Il ricevimento con la prof.ssa Rizzi è in presenza e su appuntamento (scrivere una email a raffaella.rizzi@unimib.it)
Il ricevimento con il prof. Della Vedova è online e su appuntamento: https://www.unimib.it/gianluca-della-vedova

Sustainable Development Goals