UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Data Analysis

2223-1-F0901D043

Obiettivi

Il candidato sarà in grado di: comprendere aspetti basilari del disegno dello studio, implementare autonomamente analisi statistiche di base, leggere con spirito critico la letteratura scientifica che presenti analisi statistiche descrittive e inferenziali.

Il candidato sarà in grado di: acquisire le conoscenze ed i concetti di base relativi alle metodologie e alle tecniche computazionali per la raccolta, la gestione e l'analisi di dati in biologia molecolare, come i dati di sequenze generati dalle tecnologie Next Generation Sequencing (NGS), e la padronanza dei principali strumenti computazionali necessari per estrarre informazioni di interesse per la ricerca biomedica dalle principali banche dati di sequenze.

Contenuti sintetici

I due moduli di cui si compone il corso si propongono di contribuire alla formazione di un biotecnologo medico che sia in grado di:

- comprendere i principi del disegno sperimentale in medicina e biologia
- conoscere le principali tecniche di analisi statistica dei dati
- utilizzare un software per l'elaborazione dei dati
- compiere l'interpretazione critica dei risultati presentati nella letteratura scientifica.
- essere introdotto alla bioinformatica: motivazioni, problemi e metodologie.
- conoscere le tecnologie NGS
- conoscere le principali basi di dati; accesso, interrogazione, inserimento dati
- conoscere le principali tecniche di analisi dei dati: ricostruzione e annotazione di genomi; confronto di sequenze: algoritmi di allineamento globale, locale e multiplo; ricostruzione di filogenie; analisi del trascrittoma.

Programma esteso

Il modulo di Biostatistica si articola in due parti: la prima relativa alla statistica descrittiva, la seconda alla statistica inferenziale, la terza relativa all'interpretazione di articoli scientifici. Le parti prima e seconda presentano le seguenti caratteristiche:

- includono aspetti metodologici di disegno dello studio e programmazione dell'esperimento
- vengono erogate nella forma di riflessione su particolari esempi applicativi
- prevedono l'uso del pacchetto applicativo per l'analisi dei dati STATA

Parte prima - Generalità sulla statistica descrittiva, Principali rappresentazioni tabellari e grafiche di dati variabili qualitative e quantitative, Indicatori di ordine di grandezze e dispersione di un fenomeno, Distribuzione Gaussiana, Elementi di calcolo delle probabilità.

Parte seconda - Generalità sulla statistica inferenziale, Verifica di ipotesi nulle relative alla media di variabili continue, Test T in disegno semplice ed appaiato, Verifica di ipotesi nulle relative alla associazione per variabili categoriali, Test chi quadrato, Verifica di ipotesi nulle relative alla proporzione di variabili dicotomiche: Test McNemar, Cenni all'analisi della varianza, Studio della potenza del test e calcolo della dimensione del campione.

Il modulo di Bioinformatica si articola in 8 parti:

- La gestione dei dati nelle scienze della vita
- L'informatica essenziale: Algoritmi e programmi, Alfabeti, parole, grafi, Basi di dati
- La tecnologia NGS: Piattaforme NGS di seconda generazione, Piattaforme NGS di terza generazione, formato dei dati genomici, Ricostruzione e annotazione di genomi
- Basi di dati di sequenze molecolari: Basi di dati Genomiche (EMBL – GenBank), Basi di dati di sequenze proteiche (SwissProt, PDB), I sistemi di interrogazione delle Basi di Dati
- Analisi di sequenze in biologia molecolare: Algoritmi di String matching esatto, Allineamento di sequenze, Motivazioni, Matrici a punti, Matrici di sostituzione PAM, BLOSUM, Allineamento globale: Algoritmo di Needleman-Wunsch, Allineamento locale: Algoritmo di Smith-Waterman, Algoritmi euristici: BLAST, Fasta, BWA, Allineamento multiplo; CLUSTALW
- Ricerca di motivi funzionali in sequenze: Alberi di suffissi, Algoritmi di pattern discovery
- Analisi del trascrittoma: Annotazione di geni e trascritti alternativi, Analisi di dati RNA-seq
- Evoluzione molecolare: ricostruzione di alberi filogenetici: Algoritmi di Clustering, k-means, Neighbor joining, UPGMA, Metodi di massima parsimonia, Metodi di massima verosimiglianza

Prerequisiti

Il candidato deve possedere una conoscenza di base dell'uso del personal computer, dell'informatica e di biologia molecolare.

Modalità didattica

Lezioni tradizionali, Quiz on-line, video clip.

Materiale didattico

https://www.pearson.it/opera/pearson/0-7623-fondamenti_di_statistica

- M. Helmer Citterich, F. Ferrè, G. Pavesi, C. Romualdi, G. Pesole, Fondamenti di bioinformatica (Zanichelli editore)
- Dispense fornite dai docenti
- Si raccomanda agli studenti di iscriversi alle 3 pagine elearning (quella del corso, quelle dei moduli)

Periodo di erogazione dell'insegnamento

Primo semestre.

Modalità di verifica del profitto e valutazione

Prova scritta (Biostatistica) e Prova orale (Bioinformatica). Il voto finale verrà calcolato come la media dei voti dei due moduli.

Le comunicazioni relative ad aspetti organizzativi degli appelli verranno date attraverso il forum sulla pagina "Analisi dei Dati".

Orario di ricevimento

Da definire con lo studente via email.

Sustainable Development Goals