

Proposta di corso trasversale:

Data analysis and interpretation in molecular medicine: from databases to artificial intelligence

Module 3: PROTEOMICS – From sequence to structure

Proponenti:

- Prof. Enza Maria Valente (coordinatore, corso di dottorato in Translational Medicine)
- Prof. Riccardo Bellazzi (corso di dottorato in Bioingegneria, Bioinformatica e Tecnologie per la Salute)

Obiettivi formativi:

Questo corso trasversale si articolerà in tre moduli indipendenti (Genomics, Transcriptomics, Proteomics), che saranno svolti a rotazione su tre anni. In questo modo il corso sarà fruibile ogni anno a tutti i dottorandi dei cicli attivi, senza ripetizioni.

Alla fine del terzo modulo ("Proteomics"), il dottorando sarà in grado di:

- 1) conoscere ed utilizzare i database di proteine disponibili in rete;
- 2) conoscere gli strumenti informatici di analisi della struttura primaria di una proteina (il formato FASTA, predizione di struttura secondaria, analisi di possibili siti di glicosilazione, allineamento multiplo di sequenze, ecc.)
- 3) conoscere e saper interpretare il formato standard di rappresentazione di una proteina nello spazio;
- 4) conoscere gli strumenti informatici per la predizione della struttura terziaria di una proteina (dai metodi automatici più comuni alla predizione accurata di strutture tridimensionali);
- 5) panoramica sulle tecniche di "drug discovery" *in silico* (software più utilizzati, database di molecole, ecc.)
- 6) avere una panoramica dell'uso e delle limitazioni delle moderne tecniche data-driven basate su reti neurali artificiali in contesti di drug-discovery, quali predittori e generatori.

Specifiche del corso:

Numero di ore/lezioni: 20

Periodo di svolgimento: giugno 2021

Modalità di verifica dell'apprendimento: sessione pratica finale di esercitazione

Potenziali dottorati interessati:

Macroarea di Scienze della Vita: Translational Medicine; Genetica, Biologia Molecolare e Cellulare

Macro-area di Scienze e Tecnologie: Tecnologie per la Salute, Bioingegneria e Bioinformatica; Ingegneria Elettronica, Informatica ed Elettrica, Scienze Chimiche e Farmaceutiche

Programma preliminare:

- Session 1 (8 h): Introduction to proteins and bioinformatics databases and tools

I dottorandi riceveranno lezioni frontali per conoscere i più importanti database di sequenze proteiche, confrontarli ed imparare ad utilizzarne le funzioni più importanti. In particolare, sarà discusso il database di sequenza *Uniprot*, i software di analisi della struttura primaria di una proteina (PSIPRED, XTALPRED, Clustal, ExPasy ProtParam, NetNGlyc, TMHMM)

- Session 2 (4 h): Data analytics and 3D modeling.

I dottorandi riceveranno lezioni frontali che introdurranno diversi approcci metodologici per l'analisi e la generazione di modelli proteici a partire dalla sequenza aminoacidica. Verranno discussi i principali software di predizione della struttura tridimensionale di una proteina (SwissProt, Modeller) e il dottorando imparerà a conoscere ed utilizzare software di visualizzazione di proteine nello spazio (il formato PDB, PyMol, UCSF Chimera). Il dottorando, infine, imparerà a distinguere un buon modello da una predizione poco accurata.

- Session 3 (4 h): Statistical learning and deep artificial neural networks: selected applications in drug discovery

L'*apprendimento statistico* è una branca dell'intelligenza artificiale che consiste nell'uso di modelli matematici per scoprire e sfruttare regolarità presenti in insiemi di dati osservati. L'interesse verso approcci *data-driven* è cresciuto notevolmente grazie alla disponibilità di potenza di calcolo, di dati strutturati, ed al successo di *deep neural networks* (DNN) in problemi notoriamente difficili. In questa lezione i dottorandi esamineranno i principi delle DNNs, mediante architetture di largo interesse ed applicabilità. Saranno passate in rassegna alcune applicazioni descritte nella recente letteratura che hanno applicato architetture DNN a passaggi della *drug discovery pipeline*, e sarà verificato come gli autori le abbiano adattate a problemi quali predizione di tossicità, affinità, generazione di ligandi, ed altre.

- Session 4 (4 h): Hands-on practice

In questa mezza giornata finale del corso, che si svolgerà in aula multimediale, i dottorandi avranno la possibilità di mettere in pratica quanto appreso utilizzando pacchetti software per il machine learning (R, Python) e risolvendo alcuni quesiti che richiedono l'utilizzo dei database genomici.

Possibili speakers

Marco Lolicato, Pavia (PDB, PyMol, Chimera, 3D modeling)

Chiara Platania, Catania (docking and drug discovery)

Toni Giorgino, Milano, CNR (molecular dynamics, statistical learning)