

## **Linee guida BITS per l'insegnamento della Bioinformatica a livello accademico**

### **Premessa**

Nel 2021, la Società Italiana di Bioinformatica (BITS) ha avviato una ricognizione sugli insegnamenti di Bioinformatica, col duplice scopo di verificare il grado di uniformità nei percorsi formativi attivati presso le varie sedi universitarie, e di capire come si colloca la formazione in Bioinformatica in Italia rispetto alla situazione europea ed internazionale. Tale ricognizione, su base volontaria, ha coinvolto docenti di 20 atenei di diverse parti d'Italia, permettendo l'analisi di circa 50 insegnamenti di Bioinformatica erogati in vari corsi di laurea, sia afferenti all'area delle scienze della vita e della salute, sia all'area informatico-ingegneristica. I risultati, rielaborati in forma aggregata, sono visibili nel documento "Rielaborazione-risultati-survey-2021.pdf".

Da questa ricognizione, seppure non pienamente rappresentativa, è emersa una notevole disomogeneità degli insegnamenti di Bioinformatica, sia per quanto riguarda il carico didattico attribuito che per i contenuti. In particolare, il numero di CFU attribuiti all'insegnamento di Bioinformatica varia da 2 a 12, mentre la selezione dei contenuti appare condizionata dallo specifico contesto entro cui l'insegnamento è erogato.

Alla luce di quanto emerso, e considerando che l'uso di approcci e tecnologie bioinformatiche diventa sempre più importante in ambito scientifico, BITS propone alcune linee guida per l'insegnamento di Bioinformatica a livello accademico. Tali linee guida comprendono contenuti raccomandati di riferimento per i diversi livelli di laurea e le diverse macroaree scientifiche coinvolte, e raccomandazioni aggiuntive per collocare l'insegnamento della Bioinformatica entro un contesto che ne permetta l'ottimale fruizione da parte degli studenti e delle studentesse.

### **Contenuti raccomandati per l'insegnamento della Bioinformatica**

La Bioinformatica è una materia chiaramente interdisciplinare, la cui collocazione didattica è riscontrabile in corsi di laurea appartenenti sia all'ambito biomedico, sia all'ambito informatico-ingegneristico. Per questa sua peculiarità, la società BITS ritiene opportuno differenziare le competenze che le studentesse e gli studenti dei due diversi ambiti devono conseguire. Di conseguenza, sono stati delineati diversi contenuti raccomandati, riferiti sia ad un insegnamento da collocarsi in una laurea triennale, sia un insegnamento da collocarsi in un percorso di laurea magistrale, nelle due diverse macroaree sopra menzionate. Nel caso di un insegnamento da erogare in una laurea triennale, il carico didattico ritenuto adeguato è pari a 6 CFU; nel caso di un insegnamento di bioinformatica in una laurea magistrale, il carico didattico consigliato è di 8 CFU, e comunque di non meno di 6 CFU. All'interno dei contenuti raccomandati sono state identificate quelle conoscenze e abilità ritenute irrinunciabili, con un'indicazione relativa al monte ore opportuno da dedicare.

I contenuti raccomandati sono riportati in calce al presente documento, e se ne consiglia la visione da parte di tutti/e i/le docenti incaricati/e di un insegnamento di Bioinformatica, prima di stilare le relative schede di insegnamento.

### **Raccomandazioni per il contesto didattico entro cui erogare gli insegnamenti di Bioinformatica**

Al fine di contestualizzare al meglio un insegnamento di Bioinformatica ed agevolare il conseguimento degli obiettivi didattici, BITS ritiene utile accompagnare l'elenco dei contenuti raccomandati con le seguenti considerazioni generali:

1) In primo luogo, l'apprendimento di una materia interdisciplinare come la Bioinformatica trae enorme vantaggio dalle relazioni didattiche che tale insegnamento riesce a stabilire, sia con insegnamenti che

permettono di acquisire conoscenze preliminari rispetto ad essa, sia con insegnamenti dove le competenze in bioinformatica possono essere praticamente applicate, fornendo concretezza all'apprendimento. In particolare, per l'area biomedica, BITS suggerisce che gli studenti e le studentesse acquisiscano, prima dell'insegnamento di Bioinformatica, non solo competenze in Biologia cellulare, molecolare e biochimica, ma anche solide competenze in matematica, statistica ed informatica, in modo da possedere una formazione di base sia teorica (funzioni, algoritmi, test statistici) sia pratica (calcolo, sistemi Unix/Linux, ecc.) sulla quale incardinare le basi dell'insegnamento. Analogamente, per l'area Informatica/Ingegneristica, BITS considera opportuno far precedere il corso di Bioinformatica, oltre che da corsi di Matematica, Statistica e informatica, anche da un corso introduttivo di Biologia cellulare, molecolare e biochimica che chiarisca agli studenti e alle studentesse quali entità si stanno apprestando a trattare con strumenti informatici, quale ruolo esse svolgono a livello biologico e come vengono generati i dati da analizzare;

2) L'istituzione di un solo insegnamento di Bioinformatica nel ciclo triennale o magistrale costringe a comprimere eccessivamente gli argomenti. BITS auspica che i corsi di studio di durata triennale possano includere almeno un insegnamento di Bioinformatica di 6 CFU, dedicato ai fondamenti della bioinformatica, e che i corsi di studio di livello magistrale possano includere un insegnamento di Bioinformatica di almeno 8 CFU, dedicato alle applicazioni contemporanee, quali le "scienze omiche", la bioinformatica strutturale e la biologia dei sistemi;

3) È importante che le studentesse e gli studenti acquisiscano consapevolezza piena del significato delle procedure bioinformatiche per la ricerca, analisi e trattamento dei dati; per questo motivo, si raccomanda di evitare il più possibile un approccio d'uso degli strumenti bioinformatici (in particolare, di quelli liberamente disponibili online) di tipo "black box". BITS suggerisce inoltre di affrontare l'insegnamento della Bioinformatica non limitandosi ad un approccio pratico meramente descrittivo, ma delineando progetti di ricerca, opportunamente dimensionati alle competenze da acquisire, che permettano alle studentesse e agli studenti non solo di applicare degli strumenti, ma anche di valutare in modo critico i risultati, cogliendo le differenze che diversi tipi di strumenti e di approcci possono avere sul dato ottenuto;

4) BITS raccomanda inoltre di iniziare a sensibilizzare le studentesse e gli studenti sull'importanza della gestione corretta, condivisione e riproducibilità dei dati scientifici, presentando i principi di "data FAIRness", "data sharing", "open science" e di etica nella ricerca scientifica, anche a livello precoce nel percorso formativo.

BITS invita caldamente le sue socie ed i suoi soci, nonché le colleghe ed i colleghi docenti incaricati di erogare insegnamenti di Bioinformatica, ad aderire a queste raccomandazioni. La Società effettuerà una revisione periodica di questo documento, sulla base delle evoluzioni della disciplina, del feedback ricevuto dai/dalle docenti, e considerando anche il contesto internazionale.

Il Direttivo BITS

## CONTENUTI RACCOMANDATI PER L'INSEGNAMENTO DELLA BIOINFORMATICA PER CdL TRIENNALE – AREA BIOMEDICA

### Organizzazione del corso (6 CFU – 48 ore). In grassetto i contenuti ritenuti essenziali

- 1) Elementi di Informatica e statistica essenziale (8 ore):
  - a) Elementi di architettura dei calcolatori, hw, sw di base e sw applicativo (anche reti e cloud)
  - b) Algoritmi; potenza di calcolo e efficienza degli algoritmi
  - c) Elementi di probabilità e statistica (Media, Mediana, Probabilità a priori e a posteriori, Teorema di Bayes, Verosimiglianza)
  
- 2) Organizzazione e gestione dei dati (8 ore):
  - a) Data base e DBMS: (Struttura dei DB, DB relazionali, progettazione e inserimento dati, metodologie di accesso e interrogazione)
  - b) Banche dati genomiche e proteomiche**
    - i) GenBank - ENA – DDBJ ((INSDC)): formato entry file, inserimento sequenze, ricerca sequenze**
    - ii) Genome browsers: ENSEMBL, UCSC**
    - iii) UniprotKB (Swiss-Prot, TrEMBL)**
    - iv) Esercitazioni: accesso e ricerca incrociata di sequenze e dati**
  
- 3) L'analisi dei dati (20 ore):
  - a) **Confronto di sequenze (10 ore)**
    - i) Allineamento (locale o globale): definizioni; allineamento ottimo, punteggi di allineamento, Matrici di sostituzione**
    - ii) Algoritmi di allineamento: programmazione dinamica, algoritmi euristici (BLAST, FASTA)**
    - iii) Esercitazioni: ricerche per similarità in banche dati**
  - b) Filogenie e predizioni di patterns (10 ore)**
    - i) Allineamenti multipli**
    - ii) Costruzione di profili e predizione di patterns**
    - iii) Metodi di costruzione di alberi filogenetici**
    - iv) Esercitazioni: ricerche in banche dati di profili e patterns, allineamenti multipli**
  
- 4) La bioinformatica strutturale (12 ore):
  - a) Predizione di strutture secondarie
    - i) Metodi statistico-probabilistici (Chou e Fasman, GOR)
    - ii) Metodi basati su intelligenza artificiale (reti neurali, HMM)
  - b) Predizione di strutture terziarie:
    - i) Approcci template-based
    - ii) Approcci template-free
  - c) Interazioni proteina-ligandi:
    - i) Docking
    - ii) Computer-assisted drug design

## **CONTENUTI RACCOMANDATI PER L'INSEGNAMENTO DELLA BIOINFORMATICA PER CdL MAGISTRALE – AREA BIOMEDICA (SENZA INSEGNAMENTO PRECEDENTE IN CDS TRIENNALE)**

*Conoscenze pregresse fondamentali:*

0.1 Elementi di architettura dei calcolatori, hw, sw di base e sw applicativo (anche reti e cloud)

0.2 Algoritmi; potenza di calcolo e efficienza degli algoritmi

0.3 Elementi di probabilità e statistica (Media, Mediana, Probabilità a priori e a posteriori, Teorema di Bayes, Verosimiglianza)

0.4 Elementi di Machine learning e clustering

**Organizzazione del corso (8 CFU – 64 ore). In grassetto i contenuti ritenuti essenziali**

- 1) Introduzione: motivazioni, crescente importanza dei dati in biologia e medicina, la professione di bioinformatico (2 ore)
- 2) **La generazione dei dati (6 ore)**
  - a) **Piattaforme di sequenziamento degli acidi nucleici: le tre generazioni**
  - b) **Dal dato “fisico” al dato “simbolico”: sw di “base calling”**
  - c) **Copertura, qualità delle read, formati FASTQ e FASTA**
  - d) **Dalle read alla sequenza: algoritmi di assemblaggio**
  - e) **Annotazione del genoma**
- 3) **Organizzazione e gestione dei dati (12 ore):**
  - a) **Data base e DBMS: Struttura dei DB, DB relazionali, progettazione e inserimento dati, metodologie di accesso e interrogazione**
  - b) **Banche dati genomiche e proteomiche**
    - i) **Banche dati genomiche (GenBank - ENA – DDBJ)**
    - ii) **Banche dati proteomiche (UniprotKB, Swiss-Prot, TrEMBL – PDB)**
    - iii) **Genome browsers: ENSEMBL, UCSC**
    - iv) **Esercitazioni: accesso e ricerca incrociata di sequenze e dati**
- 4) **L’analisi dei dati (30 ore)**
  - a) **Confronto di sequenze**
    - i) **Allineamento (locale o globale): definizioni; allineamento ottimo, punteggi di allineamento, Matrici di sostituzione**
    - ii) **Algoritmi esatti di allineamento: programmazione dinamica**
    - iii) **Algoritmi euristici (BLAST, FASTA), ricerche per similarità in banche dati**
  - b) **Filogenie e predizioni di patterns**
    - i) **Allineamenti multipli**
    - ii) **Costruzione di profili e predizione di patterns**
    - iii) **Metodi di costruzione di alberi filogenetici**
    - iv) **Esercitazioni**
  - c) **Analisi del trascrittoma**
    - i) **Annotazione di geni e trascritti alternativi**
    - ii) **Analisi di dati RNA-seq**
    - iii) **La struttura dell’RNA**
    - iv) **Esercitazioni**
  - d) **Proteomica e interazioni proteiche**
    - i) **Identificazione delle proteine**
    - ii) **Banche dati di interazioni proteiche, metodi di predizione di interazioni proteiche**
    - iii) **Esercitazioni**
  - e) **Systems biology: il ruolo dei modelli e cenni a network biology (6 ore)**
- 5) **La bioinformatica strutturale (14 ore):**
  - a) **Predizione di strutture secondarie**

- b) Predizione di strutture terziarie**
- c) Interazioni proteina-ligandi
  - i) Docking
  - ii) Computer-assisted drug design

## CONTENUTI RACCOMANDATI PER L'INSEGNAMENTO DELLA BIOINFORMATICA PER CdL TRIENNALE – AREA ING-INF

Organizzazione del corso (6 CFU – 48 ore). In grassetto i contenuti ritenuti essenziali

- 1) Introduzione alla bioinformatica (2 h)
- 2) **Elementi base di biologia molecolare e genomica (2 h)**
- 3) **Sequenziamento del DNA, NGS, reads (4 h)**
- 4) **Banche dati genomiche e proteomiche: struttura e utilizzo (8 h)**
  - a) **GenBank**
  - b) **UniprotKB (Swiss-Prot, TrEMBL)**
  - c) **Esercitazioni: accesso e ricerca incrociata di sequenze e dati**
- 5) **Programmazione in R/Biopython/Julia (A scelta) (8 h)**
- 6) **Algoritmi di allineamento di sequenze (8 h)**
  - a) **Allineamento (locale, globale, multiplo): definizioni; allineamento ottimo, punteggi di allineamento, Matrici di sostituzione**
  - b) **Algoritmi di allineamento: programmazione dinamica, algoritmi euristici (BLAST, FASTA)**
  - c) **Esercitazioni: ricerche per similarità in banche dati, allineamento a coppie e multiplo**
- 7) **Algoritmi per la predizione di struttura 2D e 3D delle proteine (8 h)**
  - a) **Predizione di strutture secondarie: Metodi statistico-probabilistici (Chou e Fasman, GOR)**
  - b) **Predizione di strutture terziarie: Approcci template-based (homology modeling)**
- 8) **Generazione ed analisi dei principali dati omici e sviluppo di pipelines per la bioinformatica (8 h)**
  - a) **Dati genomici, proteomici, interattomici**
  - b) **Cenni alle tecnologie NGS, Microarray, Mass Spectrometry**
  - c) **Cenni alle metodologie per l'analisi dei dati: Esperimenti caso-controllo, Classificazione, Clustering**
  - d) **Principali packages e workflow per la bioinformatica (es. Bioconductor, Galaxy, Bio-Linux, Bio-Python)**

**CONTENUTI RACCOMANDATI PER L'INSEGNAMENTO DELLA BIOINFORMATICA PER CdL MAGISTRALE – AREA ING-INF (SENZA INSEGNAMENTO PRECEDENTE IN TRIENNALE)**

**Organizzazione del corso (8 CFU – 64 ore). In grassetto i contenuti ritenuti essenziali**

- 1) Introduzione alla bioinformatica (2 h)
- 2) **Concetti fondamentali di biologia molecolare e genomica (4 h)**
- 3) **La generazione dei dati (4 h)**
  - a) **Piattaforme di sequenziamento degli acidi nucleici: le tre generazioni**
  - b) **Dal dato “fisico” al dato “simbolico”: sw di “base calling”**
  - c) **Copertura, qualità delle read, formati FASTQ e FASTA**
  - d) **Dalle read alla sequenza: algoritmi di assemblaggio**
  - e) **Annotazione del genoma**
- 4) **Organizzazione e gestione dei dati (6 h)**
  - a) **Banche dati genomiche e proteomiche: formato entry file, inserimento sequenze, ricerca sequenze**
    - i) GenBank - ENA – DDBJ ((INSDC))
    - ii) UniprotKB (Swiss-Prot, TrEMBL)
- 5) Programmazione in R/Biopython/Julia (A scelta) (8 h)
- 6) Strutture dati per la bioinformatica: Alberi di suffissi e di affissi, grafi, trasformata di Burrows-Wheeler (6 h)
- 7) **L’analisi dei dati (24 h)**
  - a) **Confronto di sequenze**
    - i) **Allineamento (locale o globale): definizioni; allineamento ottimo, punteggi di allineamento, Matrici di sostituzione**
    - ii) **Algoritmi esatti di allineamento: programmazione dinamica**
    - iii) **Algoritmi euristici (BLAST, FASTA), ricerche per similarità in banche dati**
  - b) **Filogenie e predizioni di patterns**
    - i) **Allineamenti multipli**
    - ii) **Costruzione di profili e predizione di patterns**
    - iii) **Metodi di costruzione di alberi filogenetici**
    - iv) **Esercitazioni**
  - c) **Analisi del trascrittoma**
    - i) **Annotazione di geni e trascritti alternativi**
    - ii) **Analisi di dati RNA-seq**
    - iii) **La struttura dell’RNA**
  - d) **Proteomica e interazioni proteiche**
    - i) **Identificazione delle proteine**
    - ii) **Banche dati di interazioni proteiche, metodi di predizione di interazioni proteiche**
    - iii) **Esercitazioni**
  - e) **Systems biology: il ruolo dei modelli (4 ore)**
- 8) **La bioinformatica strutturale (10 ore)**
  - a) **Predizione di strutture secondarie**
  - b) **Predizione di strutture terziarie**
  - c) **Interazioni proteina-ligandi**
    - i) **Docking**
    - ii) **Computer-assisted drug design**
- 9) **Generazione ed analisi dei principali dati omici e Sviluppo di pipelines per la bioinformatica (8 h)**

- a) Dati genomici, proteomici, interattomici
- b) Cenni alle tecnologie NGS, Microarray, Mass Spectrometry
- c) Cenni alle metodologie per l'analisi dei dati: Esperimenti caso-controllo, Classificazione, Clustering
- d) Principali packages e workflow per la bioinformatica (es. Bioconductor, Galaxy, Bio-Linux, Bio-Python)