Data and Computational Biology

UniShare

Davide Cozzi
@dlcgold

Indice

1 Introduzione **2**

2 Introduzione alla Biologia Computazionale **3**

Capitolo 1

Introduzione

Questi appunti sono presi a lezione. Per quanto sia stata fatta una revisione è altamente probabile (praticamente certo) che possano contenere errori, sia di stampa che di vero e proprio contenuto. Per eventuali proposte di correzione effettuare una pull request. Link: `https://github.com/dlcgold/Appunti`.

Capitolo 2

Introduzione alla Biologia Computazionale

La **biologia** nasce come una disciplina altamente **descrittiva** mentre altre discipline, come, ad esempio, informatica, matematica o fisica, sono discipline **generaliste**.
I biologi propongono **modelli**, come ad esempio i *pathway*, che sono il diretto risultato di osservazioni sperimentali e interpretazione dei risultati.

Definizione 1. *Un* ***pathway (*****percorso*****) biologico*** *è una serie di interazioni tra molecole in una cellula che porta a un determinato prodotto o un cambiamento in una cellula. Tale* percorso *può innescare l'assemblaggio di nuove molecole, come un grasso o una proteina. I* percorsi *possono anche attivare e disattivare i geni o stimolare una cellula a muoversi. I pathway più comuni sono coinvolte nel metabolismo, nella regolazione dell'espressione genica e nella trasmissione dei segnali e svolgono un ruolo chiave negli studi avanzati di genomica.*
Tra le principali categorie si hanno:

- *Metabolic pathway*
- *Genetic pathway*
- *Signal transduction pathway*

Un altro aspetto chiave negli ultimi 25 anni è stato quello della mole di dati prodotti, tramite, ad esempio, **Next Generation Sequencing (*NGS*)**, con la produzione di *DNAseq* e *RNAseq*, o alla cosiddetta **single-cell analysis**. Tutte queste tecnologie *high-throughput* usate in biologia computazionale e in bioinformatica richiedono una forte conoscenza algoritmica, matematica e statistica per la gestione di questa enorme quantità di dati (essendo anche

nell'ambito **big data**) in ambito biomedico. Ovviamente le conoscenze, i tempi (ma anche gli spazi), gli strumenti da usare e sviluppare etc... variano al variare del tipo di studio.
Un altro aspetto non trascurabile è la scala di misura di ciò che viene studiato, ad esempio:

- *organismi*, ad esempio per gli organismi multicellulari si passa da $10\mu m$ a $50/85m$
- *tessuti*, ad esempio per i tessuti umani siamo in un range maggiore di $10^4\mu m^3$
- *cellule*, ad esempio per quelle umane si va da $30\mu m^3$ a $10^6\mu m^3$ con:
  - membrane
  - nuclei
  - ribosomi
  - mitocondri e cloroplasti
  - altri organelli e strutture intracellulari
  - proteine
  - materiale genomico (DNA e RNA e affini strutture: ad esempio istoni)
  - ...

Parlando di tipi di organismi distinguiamo in primis:

- **eucarioti**. In questo caso si hanno cellule più complesse, con numerosi organelli e soprattutto il **nucleo**, dove sono contenute le informazioni
- **procarioti**, come i *batteri*. In questo caso si hanno cellule piccole e semplici. Non hanno un nucleo ma solo una regione, detta **nucleoide**, dove sono contenute le informazioni

In aggiunta si hanno anche i **virus**.
Per ulteriori informazioni sui tipi di organismi guardare online.
Parlando di DNA si ha che ogni cellula umana contiene circa 2 metri di DNA e un organismo umano contiene moltissime cellule rendendo lo studio del DNA davvero complesso (anche dal punto di vista computazionale si hanno file di genomi davvero molto pesanti, di centinaia di MB).

Riprendere da appunti di Bioinformatica il passaggio da DNA a RNA e da RNA a Proteine.
Ad essere interessanti non sono solo le dimensioni di ciò che viene studiato ma anche i vari **tempi**. Vediamo una piccola tabella d'esempio:

Proprietà	**E. coli**	**Uomo**
diffusione di proteine in una cellula	$0.1s$	$\sim 100s$
trascrizione di un gene	$\sim 1m$ $(80\frac{bp}{s})$	$\sim 100s$
generazione di una cellula	da $30m$ a ore	da $20h$ a statico
transizione di stato proteico	da $1\mu s$ a $100\mu s$	da $1\mu s$ a $100\mu s$
rate di mutazione	$\sim \frac{10^{-9}}{\frac{bp}{generazione}}$	$\sim \frac{10^{-8}}{\frac{bp}{anno}}$

Qualche nota:

- i tempi di trascrizione di un gene umano includono i tempi di preprocessamento dell'$mRNA$
- per la generazione di una cellula di E. Coli si hanno 30 minuti in presenza di nutrienti
-

Si studiano quindi i vari **modelli** per la biologia computazionale che possono essere di varie tipologie:

- **continui**, tramite equazioni differenziali ordinarie
- **discreti**
- **stocastici**

Si studiano, in ottica analisi di cancro, anche **grafi mutazionali** e **evoluzioni clonali** (tramite Single-cell analysis).
Un aspetto fondamentale è costituito dall'RNA, che trasposta le informazioni dal DNA (contenuto nel nucleo) al citoplasma della cellula, dove funge da intermediario per il processo di sintesi delle proteine.

Teorema 1 (Dogma principale di Francis Crick)**.** *Si ha quindi il dogma principale della biologia molecolare:*

il flusso d'informazione è unidirezionale

ovvero, in termini più estesi:

una volta che le "informazioni" sono passate nelle proteine, non possono uscirne nuovamente. Il trasferimento di informazioni da acido nucleico ad acido nucleico, o da acido nucleico a proteina, può essere possibile, ma il trasferimento da proteina a proteina, o da proteina ad acido nucleico è impossibile. Per "informazione" si intende qui la precisa determinazione della sequenza, sia delle basi nell'acido nucleico che dei residui amminoacidici nella proteina.

Geni, proteine e cellule sono il *linguaggio macchina* della vita.
Veniamo quindi alla distinzione delle due branche di studio. **Bioinformatica** e **Biologia (del Sistema) Computazionale** sono due aspetti sovrapposti del modo in cui usiamo l'approccio computazionale alla Biologia e alla Medicina, manipolando oggetti simili ma con enfasi diversa e diverse scale spazio-temporali. In entrambe si usano ontologie, formalismi descrittive ma anche, lato più pratico, database. Nel dettaglio:

- la **Bioinformatica** si occupa in primis dell'**analisi di sequenze** ovvero, tra le altre cose, di studio del genoma, RNA folding, folding di proteine e studio dei database necessari a questi studi. Si usano algoritmi di pattern matching e altri metodi di analisi delle stringhe
- la **Biologia (del Sistema) Computazionale** studia, tra le altre cose:
  - modelli e inferenze sulle proprietà dei sistemi, studiando simulazioni e nuove proprietà
  - ricostruzione di reti metaboliche e regolatorie e di modelli di progressione

  Si usano, ad esempio, metodi di machine learning per l'analisi dei dati prodotti e si simulano modelli biologici in modo sia deterministico che stocastico (tramite ad esempio Gillespie e Monte Carlo) e si fa analisi di raggiungibilità

D'altro canto, tecniche come la **Polymerase chain reaction (*PCR*)** ed altre sono appannaggio di biologi e biotecnologi. L'interesse per un biologo computazionale e per un bioinformatico è quello di aiutare altri ricercatori a svolgere le proprie attività. Ad esempio i biologi traggono vantaggio in ottica di acquisire conoscenze di base o anche al ricevere strumenti atti al progettare e pianificare esperimenti. Gli esperimenti biologici sono costosi sia dal punto di vista dei materiali che di persone e tempo.

In biologia computazionale si è quindi interessati a comprendere, anche in termini computazionali, l'interazione complessiva di:

- processi intracellulari (regolatori e metabolici)
- cellule singole
- popolazioni cellulari

Un altro compito dei biologi computazionali è quello di capire cosa succede quando si ha la possibilità di perturbare un sistema e vedere quali sono gli effetti della perturbazione, in particolare vedere cosa succede usando un dato farmaco piuttosto che un altro per intervenire su una certa patologia, parlando, in questo caso, del cosiddetto **momento traslazionale** della **medicina traslazionale**. Con "momento" ci si riferisce al trasferimento di conoscenze delle attività di pura ricerca alle **attività di produzione**, ovvero all'*attività clinica*, con il passaggio alla "vita vera".