Data Science Lab in Biosciences

UniShare

Davide Cozzi
@dlcgold

Indice

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

Big Data in Biotechnology & Biosciences

Capitolo 3

Making Sense of Biological Data

La **biologia** è una *scienza della vita* che include:

- zoologia
- citologia
- ecologia
- botanica
- microbiologia
- fisiologia
- genetica
- biochimica
- ...

Tale scienza si occupa di capire le strutture, le funzioni, le origini, le interazioni e la tassonomia (ovvero della classificazione) delle creature viventi, non solo dell'uomo.
Si hanno vari livelli d'informazione con tanti approcci per lo studio delle informazioni stesse. È come se avessi un enorme *grafo multilayer* di informazioni, molto *complesso*. In questa scienza non si hanno regole universali, si ha sempre l'eccezione. La validità dei metodi analitici dipende fortemente da caratteristiche, conosciute o meno, dei dati stessi. Alcuni algoritmi hanno sunti diversi a seconda della tipologia del dato biologico. I dati sono tendenzialmente poco conosciuti. La *biologia* prevede quindi studi davvero molto

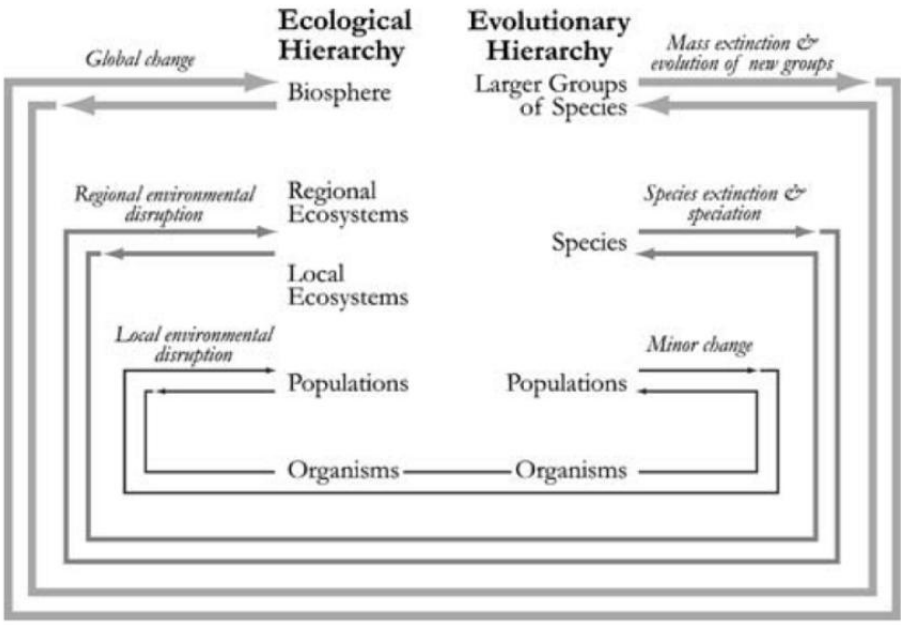

Figura 3.1: Vecchio schema grafico delle gerarchie in biologia e delle loro interazioni.

complicati, più di quanto sembri. Inoltre alcuni fenomeni sono difficilissimi da misurare e hanno una grande incertezza.
Si ha una *struttura gerarchica evoluzionaria della biologia* ma solo perché l'uomo ha strutturato le conoscenze sulla tassonomia. In realtà la vita è meno schematica e più "fluida" di quel che si vuole credere. Si hanno quindi:

- raggruppamento di specie simili
- specie
- popolazione
- organismi

Un'altra gerarchia si basa sui fenomeni che interconnettono i vari livelli appena elencati. Si hanno quindi i cosiddetti **ecosistemi**, avendo una *gerarchia ecologica*. Per l'uomo si ha una *reale*, ovvero dove vive un organismo, praticamente globale, vivendo praticamente ovunque. Si hanno quindi interazioni tra queste varie gerarchie nella biologia, che descrive sia le caratteristiche evolutive che ambientali, come visibile in figura 3.1.
La biologia inoltre evolve molto in fretta anche nel definire i suoi formalismi e i suoi limiti. Le pubblicazioni "invecchiano" molto in fretta.

In biologia si hanno delle *regole*, influenzate comunque dalla percezione umana che crea dei *bias descrittivi* che si ripercuotono sui dati stessi. Anche in questo le nuove tecnologie di sequenziamento permettono di avere dati migliori, in quanto permettono di osservare il fenomeno in modo corretto.
Anche definire l'oggetto della descrizione biologica influisce. Ad esempio nell'uomo ci sono vari livelli per definire ciò che è vivo. Una volta vorrei studiare cellule, altre volte tessuti etc...
L'uomo ha:

- $\sim 10^{14}$ cellule
- ~ 23000 geni
- $\sim 10^{14}$ milioni di batteri con ~ 9 milioni di geni

L'insieme di tutti i batteri, ovvero il *microbioma*, non è qualcosa di trascurabile anche perché senza moriremmo (ci permettono, ad esempio, di poter mangiare cibi in qualsiasi posto del mondo, cosa non così tipica nel mondo umano). In aggiunta abbiamo *microspore/funghi* che devono essere considerati. Definire qualcosa come "vivo" o meno è davvero un problema, che è legato al bisogno umano di categorizzare le cose. Un esempio classico è la classificazione dei *virus*, che ancora non sono stati categorizzati in modo univoco nel mondo scientifico. Interessante è notare che le linee cellulari usate in laboratorio sono per lo più derivanti dallo stesso campione umano, di una donna chiamata Arietta, in quanto vengono ogni volta coltivate, scambiate tra i laboratori e studiate.
Vari video interessanti sul tema sono linkati nelle slide della lezione 1.
Dentro una cellula inoltre abbiamo moltissime fonti di dati e conseguentemente tante informazioni che vanno studiate in modo selezionato e preciso, con le relative tecniche (non perfette e con bias).
Un altro problema aggiuntivo è che in biologia si ha un'idea di "privacy" in quanto non si sa quanto di un certo essere vivente può essere "scambiato" con un altro essere vivente (basti pensare alla donazione degli organi nell'uomo).
Un altro discorso interessante è che ogni cellula contiene una copia del DNA che è quasi uguale ad ogni altra ma appunto solo "quasi", questo a causa di mutazioni, adattamento all'ambiente etc...
Alla base della codifica del DNA, come ben si sa, si hanno le quattro basi azotate:

- Adenina
- Citosima

- Guanina
- Timina

Un **gene** è quindi una certe sequenza di basi azotate identificato da una posizione di inizio e di fine su uno *strand* di DNA. I geni possono avere lunghezze diverse. A partire dai geni si producono *proteine* (ovviamente anch'esse di grandezza diversa) che, dopo una serie di altri eventi biologici/biochimici, porteranno ad un certo fenotipo.
Studiando il genoma si apre anche tutto il discorso etico/biologico relativo alla modifica dei geni stessi, ovvero ottenendo gli *OMG* (che comunque hanno varie classificazioni che influenzano i vari discorsi sia etici che tecnici).
Anche il concetto di **specie** è un aspetto importante. Si parla della cosiddetta **speciazione**, ovvero ogni processo naturale che porta alla creazione di diverse specie. Tecnicamente si definisce **specie** un gruppo di organismi che si può riprodurre tra loro e produrre una prole fertile. Per evitare combinazioni tra specie la natura ha messo sia barriere fisiche che genetiche, tali strategie sono però non sempre funzionanti (basti pensare al mulo, unione di cavallo e asino, che però è sterile). Questa è la definizione universale ma comunque è ormai superata a causa di organismi ermafroditi, riproduzione asessuata etc...
Uno degli eventi dietro la speciazione, oltre a fattori ambientali e/o comportamentali, si ha la cosiddetta *selezione naturale* di Darwin. Lo studio dell'evoluzione è comunque molto complesso in quanto servono dati relativi a tantissimi timestamp. Si usano quindi tecniche Bayesiane con conoscenze a priori. Con la genomica possiamo approfondire il tema della *micro-speciazione*. Questo fa comunque capire quanto lo studio della biologia sia dinamico e non stato, basti pensare che la tassonomia con la quale si nominano i vari esseri viventi, cambia radicalmente nel giro di pochi anni.