Data and Computational Biology, Old Version

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

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".

Capitolo 3

Esempio del Repressilator

Introduciamo un esempio che rientra nell'ambito della *synthetic biology*, di M. B. Elowitz e S. Leibler[1], che sarà rivisto sotto diversi aspetti durante il corso. Questo è un esempio di un sistema biologico "ingegnerizzato", uno dei primi esempi di sistema biologico, di **biologia sintetica**.

3.1 Il Modello Biologico

In questo sistema si hanno tre geni, che per praticità chiamiamo *gene A*, *gene B* e *gene C*, ognuno dei quali, dopo essere trascritti e tradotti producono il rispettivo *mRNA* e poi, nel citoplasma, tali *mRNA* vengono usati per sintetizzare le tre rispettive *proteine*.
Quello che succede è che la trascrizione dei 3 geni può partire solo se non c'è proteina attaccata ad una sezione, detta *promotrice del processo di trascrizione*. Tale proteina è detta anche *promotore* o *inibitore*. Diciamo quindi che:

- per il *gene A* non deve esserci la *proteina C* attaccata per avere la trascrizione del gene stesso
- per il *gene B* non deve esserci la *proteina A* attaccata per avere la trascrizione del gene stesso
- per il *gene C* non deve esserci la *proteina B* attaccata per avere la trascrizione del gene stesso

È quindi un processo ciclico. Nel dettaglio del Repressilator le proteine (pro-

[1]M. B. Elowitz, S. Leibler, A synthetic oscillatory network of transcriptional regulators, Nature 403(20), January 2000

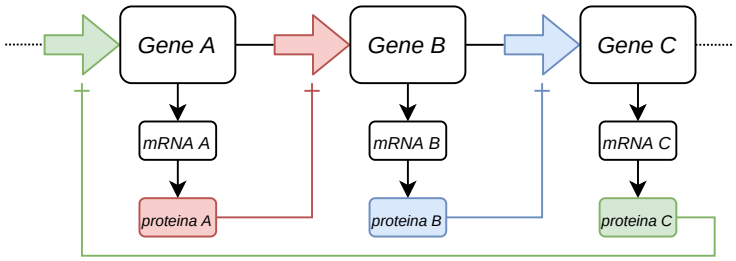

Figura 3.1: Schema di base del Repressilator, con le frecce bidimensionali che rappresentano l'azione di inibizione delle proteine.

dotte dai rispettivi geni che si indicano con la prima lettera minuscola) sono, in ordine (*A, B, C*):

- $TetR$ prodotta dal gene $tetR$
- ΛcI prodotta dal gene λcI
- $LacI$ prodotta dal gene $lacI$

Il punto fondamentale, come visibile in figura 3.1, è capire che se sto producendo una grande quantità di una certa proteina allora sicuramente non avrò produzione di quella di cui tale proteina inibisce la trascrizione del gene e così via. Nel nostro caso se si produce tanta *proteina A* non avremo produzione di *proteina B* e di conseguenza avremo produzione della *proteina C*, ma nel momento in cui questa terza viene prodotta cala la produzione della *proteina A* comportando la produzione della *proteina B* etc.... Ho, in pratica, un sistema oscillatorio, con 3 proteine che si reprimono l'una con l'altra.
La rappresentazione "su carta" di questo comportamento è abbastanza semplice, come vedremo, modellandola tramite un insieme di equazioni differenziali. Il problema è passare dalla teoria alla pratica. Questo sistema "ingegnerizzato", di equazioni differenziali, è in grado di confermare quanto visualizzabile poi tramite esperimenti.
Vediamo quindi come viene effettivamente costruito il sistema sperimentale usando delle colonie di E. Coli, sfruttando la loro biologia. Nei batteri il DNA non è, come detto, racchiuso nel nucleo ma "circola" in una regione, detta *nucleoide*, abbastanza accessibile all'interno del citoplasma. Nei batteri il DNA circola in forme dette **plasmidi** quindi potenzialmente si può sintetizzare un particolare plasmide e inserirlo in un batterio, il quale lo userà per sintetizzare proteine. Prima è stato comunque pensato il modello matematico e poi stato effettivamente costruito l'esperimento (al contrario dell'ordine con cui si stanno ora spiegando quindi).

I due ricercatori hanno costruito due plasmidi (di cui per ora non approfondiamo i dettagli):

- un plasmide che codifica il *Repressilator*, ovvero che contiene i 3 geni che codificano le 3 proteine. Prima di ogni gene si ha attaccata una *zona di induzione*
- un plasmide che codifica un *Reporter*, che codifica una particolare proteina, detta **green fluorescent protein (*Gfp*)**. La *Gfp* è una proteina usata spesso in quanto fa si che un certo sistema diventi fluorescente, di colore verde, una volta che viene illuminato con una certa luce (un laser ad una determinata frequenza). Questo plasmide fa si che, quando $TetR$ è presente in abbondanza la trascrizione del gene *gfp* viene bloccata e quindi diminuisce la quantità di *Gfp*. Quindi, come $TetR$ oscilla per il sistema di *mutua repressione*, si vedrà al microscopio un'oscillazione della fluorescenza della colonia di batteri.

Si ha un ulteriore "trucco". Se si lascia una colonia di E. Coli senza alcun controllo si avrebbe che ogni batterio inizierebbe il ciclo per conto suo, in modo non sincrono, impedendo una corretta visualizzazione della fluorescenza. Questo trucco è quello di inibire la produzione di $LacI$, interferendo con la sua espressione, usando un'ulteriore induttore, detto $IPTG$ ($isopropyl\ \beta\text{-}D\text{-}1\text{-}thiogalactopyranoside$), e ottenendo così la sincronia delle cellule dopo questo impulso iniziale di $IPTG$ (che poi decade velocemente lasciando tutti gli E. Coli nello stesso stato iniziale).

3.2 Il Modello Matematico

Facciamo quindi un passo indietro e vediamo il modello matematico del Repressilator.
Per prevedere il comportamento complessivo del sistema ingegnerizzato, si è quindi scritto un modello matematico che rappresenta la variazione dell'RNA e delle proteine espresse.
Per farlo indichiamo (**questo indice va sistemato**):

- α, proteine/cellula dal promotore non represso
- α_0, proteine/cellula dal promotore represso
- β, rapporto proteina/velocità di decadimento dell'$mRNA$

- n, *coefficiente di cooperatività di Hill* (nel caso del Repressilator si ha $n = 2$)
- m_i, i-esimo *mRNA*
- p_i, i-esima proteina che funge da repressore

L'intero sistema viene modellato con *coppie di equazioni differenziali*. Si hanno quindi:

- un'equazione che ci rappresenta la velocità di variazione dell'i-esimo mRNA:
$$\frac{\mathrm{d}m_i}{\mathrm{d}t} = -m_i + \frac{\alpha}{1 + p_j^n} + \alpha_0$$
Tale velocità dipende dalla quantità che già si ha di mRNA, dalla presenza della proteina che lo reprime (essendo sotto nella frazione al crescere il termine tende a zero, mentre al diminuire tende a 1)
- un'equazione che ci rappresenta la velocità di variazione dell'i-esima proteina che funge da repressore:
$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \beta(m_i - p_i)$$
Tale velocità dipende da quanto mRNA si ha a disposizione meno la quantità di proteina che si ha a disposizione in quel dato momento. Maggiore è la quantità di mRNA e maggiore è la produzione fino a che la proteina stessa non supera un certo livello di quantità, avendo che "satura"

Nelle formule forse indici delle proteine sbagliati.
In ordine si hanno, per i geni:

Indice	1	2	3
i	$lacI$	$tetR$	λcI
j	λcI	$lacI$	$tetR$

Con "velocità di variazione" si intende in pratica un tasso di cambio di concentrazione delle due *specie molecolari*, ovvero un'entità che osserviamo nel modello (in questo caso mRNA o proteina).
Le concentrazioni si esprimono con l'unità di misura K_M e il tempo in τ_{mRNA}, ovvero la velocità di trascrizione. Integrando numericamente le due equazioni differenziali otteniamo un comportamento periodico.

L'esperimento è stato fatto poi osservando come tutto questo diventa osservabile in una colonia di E. Coli, opportunamente trattata, usando delle foto (dove si è osservato anche un drift verso l'alto nel grafico oscillatorio a causa del fatto che la colonia si espande).
La conoscenza di tipo matematico deve però essere trasferita in un esperimento reale che funzioni (e i ricercatori devono essere in grado di manipolare entrambi gli aspetti, si quello della modellazione matematica che quello più biologico e chimico). In questo caso per ottenere oscillazioni stabili servono determinati prerequisiti:

- usare inibitori artificiali piccoli, con la cosiddetta *low leakiness*
- la velocità di decadimento di proteine e mRNA doveva essere simile, per ottenere l'oscillazione. Questo si ottiene attaccando *ssrA* ad ogni repressore
- servono curve di repressione piuttosto "ripide". Per questo si è usato un promotore con multipli *binding sites* (arrivando alla scelta di quelle date proteine), usando repressori cooperativi (questo è rappresentato con il parametro n)
- usare un *Reporter* non stabile, attaccando una variante di *ssrA* a *Gfp*

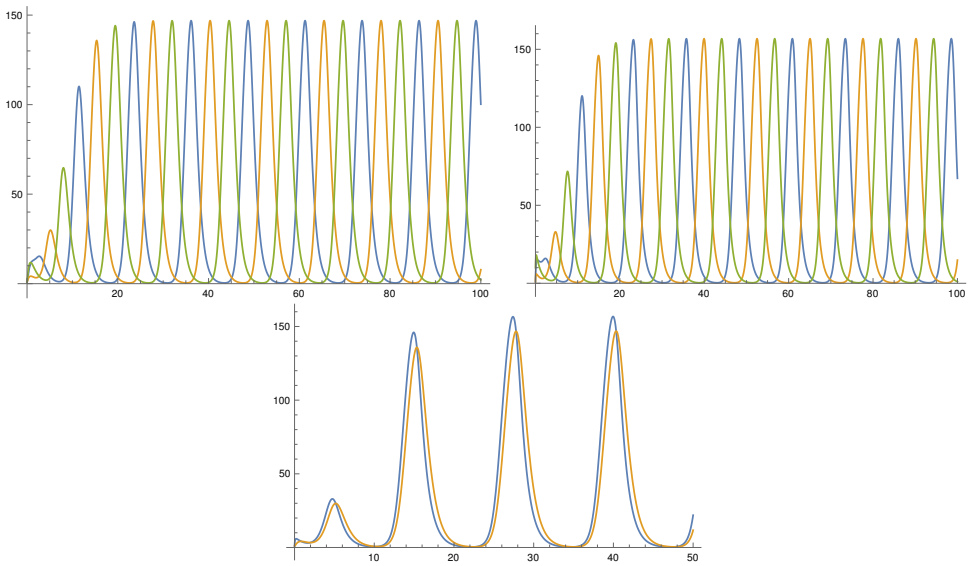

Figura 3.2: Grafici relativi al modello del Repressilator ottenuti tramite Mathematica. In primis, a sinistra la quantità di repressore/proteina rispetto al tempo, a destra quella di mRNA (nel primo caso per le 3 proteine e nel secondo per i 3 mRNA). I grafici cambiano drasticamente quando l'insieme dei valori dei parametri viene modificato. In basso le quantità di mRNA (nel caso di $tetR$) rispetto al repressore/proteina (in questo caso ovviamente $TetR$) associata rispetto al tempo. Si nota che c'è un piccolo delay nel grafico, che rappresenta il tempo di traduzione. Le scale dei tre grafici sono indicative. I parametri sono specificati nel notebook di Mathematica presente nella pagina Moodle.

Capitolo 4

Modellazione di Sistemi Biologici

Cerchiamo ora di capire come classificare i problemi, come analizzarli e comprenderli (anche tramite machine learning) e avere coscienza delle risorse online disponibili per la tematica.
Buona parte della ricerca in biologia computazionale ha come obiettivo quello di ottenere il passaggio dai risultati di laboratorio alle applicazioni cliniche (ed è qualcosa di molto complesso). Per quanto ci sia interesse verso tutte le patologie la più interessante e più studiata (soprattutto in questo corso) è il **cancro**. Un esempio di un sistema particolare dove i tumori si sviluppano è quello delle cosiddette **cripte coloniche (*colonic crypts*)**, avendo che questo sistema è relativamente semplice da studiare dal punto di vista computazionale.
Le *cripte coloniche* si trovano nell'intestino e sono delle sorta di "pozzetti", morfologicamente divisibili in varie aree. Alla base delle cripte ci sono delle **cellule staminali epiteliali**, che sono quelle che poi danno luogo ai tessuti dell'epitelio.
Dal punto di vista matematico tutti gli essere viventi sono di topologia isomorfa a dei tubi.
Tornando al discorso delle cellule staminali si ha che essere si suddividono e, man mano che si suddividono tendono a spingere verso l'alto le cellule che si trovano "al di sopra" di loro. Man mano che tali cellule vengono spinte anch'esse tendono a dividersi spingendo le altre cellule verso il *lumen della cripta*. In questo processo di suddivisione queste cellule si differenziano e le cellule staminali danno luogo ad una progenie che possiamo, dal punto di vista in primis computazionale, rappresentare come un *albero*. Si hanno le cellule di tipo diverso, più o meno differenziate che continuano a salire verso la superficie dell'epitelio e poi tendono a salire su quelli che sono detti i *villi*

intestinali. Questo è un interessante processo che può essere simulato, tra i vari modi, in modo tale che si simuli cosa accade quando le varie differenziazioni non funzionano perché, ad esempio, si ha una cellula che ha acquisito una mutazione, mutazioni che danno luogo ad una crescita non corretta, ad una *displasia*, che è la fase iniziale da cui poi si sviluppano i *tumori del colon*. Si vuole quindi fare queste simulazioni e farle in modo il più fedele possibile. Per capire se una cellula si sta comportando in modo corretto o meno dobbiamo misurarne il comportamento. In primis vogliamo misurare due cose, tra le tante:

1. **gene expression**
2. **gene alterations**, ovvero le varie mutazioni del genome, le cosiddette le *copy number variations* etc...

La tecnologia a disposizione per queste tematiche si è molto evoluta ma tra le tante tecnologie si segnalano:

- *microarrays* per l'espressione genica, usati però molti anni fa essendo una delle prime tecnologie per misurare, in modo indiretto ma parallelo, l'espressione dei geni
- *Next Generation Sequencing (NGS)* per praticamente qualsiasi cosa, anche per l'espressione genica, in modo diretto tramite particolari esperimenti (**nella rec non ho capito il nome di tali esperimenti**). NGS ha avuto molta fama da circa il 2006 in poi, con il monopolio poi di Illumina, anche se di recente si hanno nuove tecnologie che stanno rivoluzionando il settore (che producono read più lunghe)

Parliamo un secondo dei **microarrays**.
Questa è una tecnologia non più utilizzata, essendo di inizio anni duemila, che però è utile per spiegare come si procede a fare un certo tipo di misure, con una tecnologia che è stata poi ripresa da Illumina.
Questo strumento si basa su una griglia a cui sono attaccate delle "sonde" lunghe circa 25 nucleotidi e venivano usati per caratterizzare i geni. Si producono infatti segnali luminosi di diversa intensità e diversa lunghezza d'onda in una griglia, da cui si può ricavare una griglia numerica che dà informazioni in merito alla luce di ogni punto.