

Data and Computational Biology

UniShare

Davide Cozzi
@dlcgold

Indice

1	Introduzione	2
2	Introduzione alla Biologia Computazionale	3
2.1	Accenni di biologia molecolare	3
2.1.1	DNA ed RNA	3
2.1.2	Esoni, Introni e Splicing alternativo	5
3	Esempio del Repressilator	13
3.1	Il Modello Biologico	13
3.2	Il Modello Matematico	15

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/dlccgold/Appunti>.

Capitolo 2

Introduzione alla Biologia Computazionale

Materiale tratto dalla tesi.

2.1 Accenni di biologia molecolare

2.1.1 DNA ed RNA

Prima di iniziare la trattazione più squisitamente computazionale è bene dare un'introduzione, dal punto di vista biologico, di quanto trattato.

Il **DNA**, sigla corrispondente ad **acido desossiribonucleico**, è un acido nucleico contenente le informazioni necessarie al corretto sviluppo di un essere vivente. Dal punto di vista chimico questa particolare macromolecola si presenta nella tipica **struttura a doppia elica**, formata da due lunghe catene di nucleotidi, dette **strand**. Nel dettaglio i singoli nucleotidi sono formati da un **gruppo fosfato**, dal **desossiribosio**, uno **zucchero pentoso**, e da una **base azotata**. Si hanno, inoltre, 4 tipi diversi di basi azotate:

1. **Adenina**, indicata con la lettera *A*
2. **Citosina**, indicata con la lettera *C*
3. **Guanina**, indicata con la lettera *G*
4. **Timina**, indicata con la lettera *T*

Si hanno quindi due **strand**, uno detto **forward strand** (indicato solitamente col simbolo “+”) e uno detto **backward strand** (indicato solitamente col simbolo “−”) che sono direzionati nel verso opposto (in termini tecnici si ha che il forward strand va da 5’ UTR a 3’ UTR, mentre il backward strand da 3’ UTR a 5’ UTR) e sono *appaiati* mediante coppie ben precise di basi azotate. Infatti, secondo il **modello di Watson-Crick**, si ha che:

- l’**Adenina** si appaia con la **Timina** e viceversa
- la **Citosina** si appaia con la **Guanina** e viceversa

Questo accoppiamento permette di poter studiare i due **strand** come uno “complementare” all’altro. Infatti, conoscendo la sequenza di basi azotate di uno **strand**, è possibile ricavare la sequenza dell’altro mediante la tecnica del **Reverse&Complement** dove, preso uno strand, si converte ogni sua base secondo il seguente schema:

- le *A* diventano *T*
- le *T* diventano *A*
- le *C* diventano *G*
- le *G* diventano *C*

Esempio 1. Vediamo, per completezza, un esempio di **Reverse&Complement**.

Prendiamo una sequenza genomica $S = \text{"TAGGCCATATGAC"}$ e definiamo la funzione $RC(x)$ come la funzione che, presa in ingresso una stringa x costruita sull’alfabeto $\Sigma = \{A, C, G, T\}$ (quindi una sequenza genomica), restituisce la **Reverse&Complement** della stessa. Si ha quindi che:

$$RC(S) = \text{"ATCCGGTATACTG"}$$

Per riferirci al **DNA**, contenuto in una data cellula di un essere vivente, usiamo il termine **genoma**, che a sua volta viene organizzato in diversi **cromosomi**. Si definisce **gene** una particolare regione di un **cromosoma** in grado di codificare una proteina.

Ai fini della trattazione del progetto, è necessario introdurre anche l’**RNA**, sigla corrispondente ad **acido ribonucleico** (avendo il **ribosio** come zucchero pentoso), ovvero una molecola, simile al **DNA**, dotata di una singola catena nucleotidica, sempre con 4 tipi di basi azotate (anche se si ha l’**Uracile**, che si indica con la lettera *U*, al posto della **Timina**). Tra i compiti dell’**RNA** si ha quello della codifica e decodifica dei **geni**.

2.1.2 Esoni, Introni e Splicing alternativo

Per ottenere una **proteina** da un **gene** si hanno 3 passaggi:

1. La **trascrizione**, fase dove la sequenza del gene è copiata nel **pre-messenger RNA (pre-mRNA)**. Nel dettaglio viene selezionato uno dei due strand del gene e un enzima, chiamato **RNA Polimerasi**, procede alla trascrizione della sequenza selezionata creando il **pre-mRNA**. In questa fase la *Timina* viene sostituita dall'Uracile. È bene introdurre subito che in questo progetto non si terrà mai conto, a fini di semplificazione, del passaggio tra Timina e Uracile in quanto verrà usata sempre la *Timina*.
2. Lo **splicing**, fase dove vengono rimosse le parti non codificanti dalla molecola di **pre-mRNA**, formando il **messenger RNA (mRNA)**, detto anche **trascritto**. Per poter trattare al meglio questa fase bisogna parlare in primis di **esoni** e **introni**. In prima analisi si potrebbe dire, peccando di precisione, che gli **esoni** sono le sezioni codificanti di un gene mentre gli **introni** sono le porzioni non codificanti. Solo gli esoni formano il trascritto. Si ha, inoltre, che le prime due basi di un introne sono dette 5', nell'uomo solitamente si ha la coppia *GT*, mentre le ultime due, solitamente *AG* nell'uomo, sono dette 3' e sono meglio identificate come **siti di taglio (splice sites)**. Quindi un esone, in realtà, non coincide esattamente con una regione codificante, detta **CDS**, a causa di queste particolari coppie di basi. Si notifica però che, come spesso accade, i termini vengono usati in sovrapposizione.
3. La **traduzione**, fase dove viene effettivamente codificata la proteina a partire da una sezione dell'**m-RNA**. Bisogna quindi nominare particolari sequenze nucleotidiche di cardinalità 3: i **codoni**. Tali triplette sono tradotte in amminoacidi che, concatenati, formano le proteine. Esistono particolari codoni che sono utili al fine di riconoscere l'inizio e la fine della *sintesi proteica*. In particolare si ha un codone d'inizio, detto **start codon**, che solitamente corrisponde alla tripletta *AUG*, mentre, per il codone di fine, detto **stop codon**, solitamente si ha una tripletta tra *UAA*, *UAG* e *UGA*.

In realtà, un gene è in grado di sintetizzare più di una proteina mediante il cosiddetto **splicing alternativo**, che consiste in diverse varianti dell'evento

di splicing al fine di ottenere diversi trascritti. Si descrivono le principali modalità di splicing alternativo:

- L'**exon skipping**, ovvero *salto dell'esone*, dove un esone (o anche più esoni) può essere escluso dal trascritto primario oppure dove un nuovo esone (o più nuovi esoni) può essere incluso nello stesso.
- L'**alternative acceptor site**, ovvero *sito di taglio alternativo 3'*, dove una parte del secondo esone può essere considerata non codificante o, alternativamente, una porzione dell'introne adiacente può essere considerata codificante.
- L'**alternative donor site**, ovvero *sito di taglio alternativo 5'*, dove una parte del primo esone viene considerata non codificante o, alternativamente, una porzione di introne adiacente può essere considerata codificante.
- I **mutually exclusive exons**, ovvero *esoni mutuamente esclusivi*, dove solo uno di due esoni viene conservato nel trascritto.
- L'**intron retention**, ovvero *introne trattenuto*, dove un certo introne viene incluso nel trascritto primario.

Le varie modalità di splicing alternativo non si escludono a vicenda, rendendo lo studio di tale fenomeno assai complesso.

La **biologia** nasce come una disciplina altamente **descrittiva** mentre altre discipline, come, ad esempio, informatica, matematica o fisica, sono discipline **generaliste**. In biologia infatti si parte dai dati e dagli esperimenti per descrivere un fenomeno ed inferire la teoria su di esso. Questo è un discorso più di **filosofia della scienza**.

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* (che rispetto alle *DNAseq* sono più semplici da sequenziare e studiare e servono a vedere cosa sintetizza effettivamente una cellula), o alla cosiddetta **single-cell analysis**, una tecnica più recente, sviluppata negli ultimi 5 anni. I costi di sequenziamento variano a seconda del numero di basi da sequenziare ed è in calo negli anni. 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. Saper modellare fenomeni biologici è essenziale anche per poter capire come eventualmente funzionano tecniche di machine learning dedicate, come le reti neurali. Ovviamente le conoscenze, i tempi (ma anche gli spazi), gli strumenti da usare e sviluppare etc. . . variano al variare del tipo di studio. Ad ogni problema è associato un miglior strumento di modellistica.

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. Si hanno i **mitocondri**, che si occupano di generare *energia* tramite *glicolisi* e sono studiati in ambito filogenetico, in quanto provengono unicamente dalla madre, permettendo la *filogenesi materna*
- **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

Si hanno cellule nell'uomo, come quelle del sangue, dove non si ha un nucleo e non si ha riproduzione. D'altro canto si hanno anche cellule, come quelle dell'occhio, che non cambiano mai nel corso della vita.

In aggiunta si hanno anche i cosiddetti **archaea**.

Tratto da Wikipedia.

Gli archèi o archèobatteri, nome scientifico Archaea (dal termine del greco antico che significa antico) o Archaeobacteria che significa "batteri antichi", sono una suddivisione sistematica della vita cellulare. Possono considerarsi regno o dominio a seconda degli schemi classificativi, ma mostrano strutture biochimiche tali da considerarsi un ramo basilare, presto distaccatosi dalle altre forme dei viventi. Nonostante il nome attribuito a questo taxon, gli archaea non sono i procarioti più antichi mai apparsi sulla Terra, ma sono stati preceduti dagli eubatteri. Essendo costituiti da singole cellule mancanti di nucleo, per forma e dimensioni molto simili ai batteri, sono stati in passato classificati come procarioti o monere assieme ad essi. Originariamente furono ritrovati negli ambienti più estremi, ma successivamente sono stati trovati in tutti gli habitat, compreso l'intestino umano, nel caso del Methanobrevibacter smithii.

Nonostante non sia del tutto sicura la filogenesi del gruppo, gli archei sono quindi (insieme agli eucarioti e agli eubatteri) uno dei tre fondamentali gruppi

degli esseri viventi nella classificazione di Woese. Tesi recenti propongono di considerare Archea ed Eukaryota un unico regno, contrapposto ai Bacteria, in quanto all'origine degli eucarioti vi sarebbe l'endosimbiosi mitocondriale.

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*). Si hanno migliaia di trilioni di cellule nell'uomo.

Uno dei problemi è “allungare” il DNA che normalmente è incredibilmente avvolto su se stesso (e solo in fase di divisione cellulare si riconosce la forma a “X” dei cromosomi, altrimenti è ancora più avvolto su se stesso).

Dal DNA, nel nucleo, si ottiene l'RNA che esce, verso il citoplasma, dove, nei ribosomi, viene usato per sintetizzare le proteine.

Si hanno alcune specie interessanti dal punto di vista genomico e modellistico:

- **Saccharomyces cerevisiae**, ovvero il lievito da birra, con un piccolo genoma, *12 Mb*
- **Caenorhabditis elegans**, un “verme” di cui si è studiato l'intero sviluppo. Gli esemplari femmina hanno poco meno di mille cellule, 959, mentre i maschi poco di più, 1033. Si ha un genoma di *97 Mb*
- **Drosophila melanogaster** un altro modello molto usato, con un genoma di *180 Mb*
- **Homo sapiens**, con un genoma di *3200 Mb*
- **Mus musculus**, ovvero il topo, che ha un genoma molto simile a quello umano e quindi è molto usato negli studi in laboratorio. Si ha un genoma di *3300 Mb*
- **Arabidopsis thaliana**, ovvero la Veccia, che viene usata come modello base per studiare le piante. Si ha un genoma di *125 Mb*
- **Fritillaria assyriaca**, ovvero la Fritillaria, che ha il più lungo genoma conosciuto, di *120000 Mb*. Le piante normalmente hanno un genoma più lungo a causa dell'evoluzione, in quanto conservano molte informazioni che potrebbero essergli utili in futuro, anche in un futuro molto lontano, dovendo sopravvivere anche al fatto che non possono muoversi

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 trasporta 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:

... once ‘information’ has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.

L’unidirezionalità viene parzialmente infranta in caso di mutazioni del DNA ma questo non accade in fase di replicazione. Questa assunzione è una buona ipotesi dal punto di vista pragmatico.

Vediamo anche il pensiero di Sidney Brenner, biologo molto famoso: geni, proteine e cellule sono il *linguaggio macchina* della vita quindi per una corretta simulazione servono questi elementi, altrimenti il programma è una mera imitazione:

... his must not simply be another way of describing the behaviour. For example it is quite easy to write a computer program that will produce a good copy of worms wriggling on a computer screen. But the program, when we examine it, is found to be full of trigonometrical calculations and has nothing in it about neurons or muscles. The program is an imitation; it manipulates the image of a worm rather than the worm object itself. A proper simulation must be couched in the machine language of the object, in genes, proteins and cells.

... The reader may complain that I have said nothing more than ‘carry on with conventional biochemistry and physiology’. I have said precisely that, but I want the new information embedded into biochemistry and physiology in a theoretical framework, where the properties at one level can be produced by computation from the level below.

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. Leibler¹, 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-

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

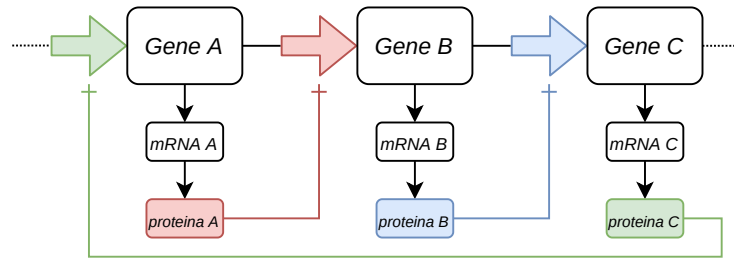


Figura 3.1: Schema di base del Repressilator, con le frecce bidimensionali che rappresentano l'azione di inibizione delle proteine.

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

- *TetR* prodotta dal gene *tetR*
- *λC_I* prodotta dal gene *λC_I*
- *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 sì 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 sì 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 β -D-1-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:

$$\frac{dm_i}{dt} = -m_i + \frac{\alpha}{1 + p_j^n} + \alpha_0$$

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:

$$\frac{dp_i}{dt} = \beta(m_i - p_i)$$

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, sia 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*

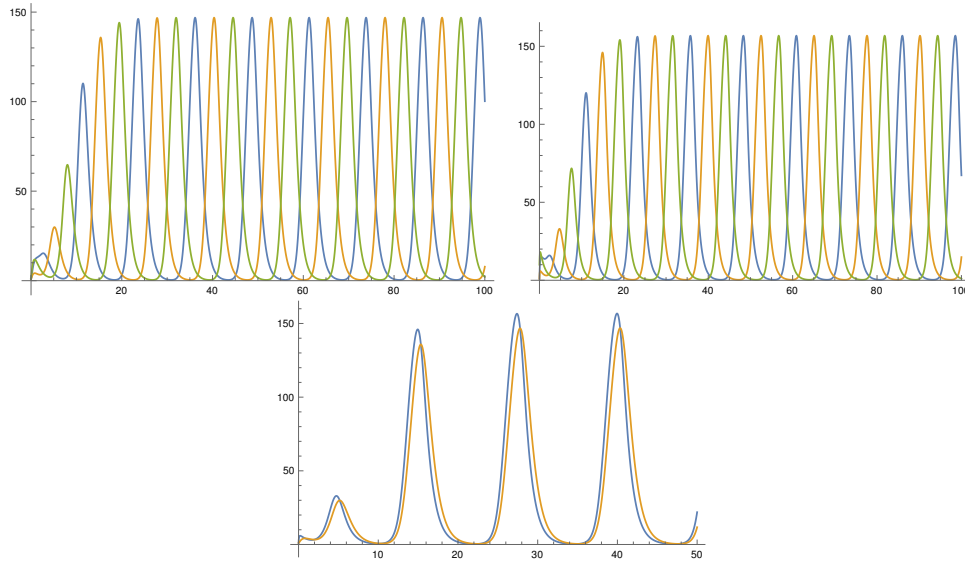


Figura 3.2: Grafici relativi al modello del Repressilator ottenuti tramite Mathematica. In primis, a sinistra la quantità di repressore/proteina rispetto al tempo, a destra quella di mRNA (nel primo caso per le 3 proteine e nel secondo per i 3 mRNA). I grafici cambiano drasticamente quando l'insieme 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.