UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Modellizzazione dell'espressione genica di singola cellula: dal singolo gene alla loro rete di interazioni

Relatore

Prof. Sandro Azaele

Correlatrice

Dr. Clelia Corridori

Laureando

Nicola Lanzilao

Anno Accademico 2023/2024

Indice

Abstract

L'espressione genica è il processo biologico nel quale le informazioni contenute nel DNA, attraverso i geni, vengono utilizzate per sintetizzare l'mRNA (trascrizione), che viene poi utilizzato per produrre le proteine mediante i ribosomi (traduzione). Da studi ed esperimenti descritti nella letteratura scientifica si conoscono diversi dettagli biologici del processo, come il ruolo dei geni, dell'mRNA e delle proteine, oltre alla produzione di proteine mediante burst. È possibile descrivere le caratteristiche principali dell'espressione genica utilizzando dei processi stocastici appropriati, nei quali la complessità del modello risultante aumenta all'aumentare dei dettagli biologici inclusi. In questa tesi si discuterà di come è possibile descrivere l'espressione genica delle singole cellule a partire da modelli di singolo gene fino a modelli di reti di regolazione genica. Verranno discussi vari modelli e diverse approssimazioni che aiutano a capire come funzionano le reti geniche reali. I risultati analitici verranno corroborati da simulazioni o integrazioni numeriche.

Capitolo 1

Introduzione

L'espressione genica, alla base del funzionamento delle cellule, è il processo attraverso cui l'informazione contenuta in un gene, un segmento di DNA, viene convertita in proteine, macromolecole funzionali che svolgono ruoli vitali nelle cellule e nei tessuti. Questo processo avviene in due fasi cruciali: la trascrizione e la traduzione. Durante la trascrizione, la sequenza nucleotidica di una specifica regione del DNA viene copiata in una molecola di acido ribonucleico (RNA). Successivamente, l'RNA agisce come modello per la sintesi delle proteine durante la traduzione. Secondo il dogma centrale della biologia (mostrato in figura 1.1) questo processo è unidirezionale: l'RNA non può essere utilizzato per produrre altro DNA, così come le proteine non possono modificare RNA e DNA. Tuttavia, in alcuni contesti, come nelle infezioni virali o in processi cellulari specifici, l'unidirezionalità del processo può essere violata, con l'RNA che può essere utilizzato per produrre nuovo DNA tramite la trascrizione inversa, e le proteine che possono influenzare la struttura o la stabilità di RNA e DNA.

Affinché cominci la trascrizione è necessario che l'enzima RNA polimerasi, si leghi al *promotore* situato a monte del gene, dove si trova la sequenza di nucleotidi che indicano l'inizio della sintetizzazione di RNA. Questo processo è governato dalla *cromatina*, una struttura nucleoproteica di circa $10\mu m$ di diametro in cui il DNA è compattato insieme alle proteine. La cromatina governa l'attivazione trascrizionale attraverso un rimodellamento locale della sua struttura. Le proteine che avvolgono il DNA sono di due tipi principali: gli *istoni*, che permettono al DNA di attorcigliarsi attorno a loro grazie alla loro carica positiva (la superficie del DNA è carica negativamente), e le proteine non istoniche, che svolgono ruoli strutturali, enzimatici e regolatori.

La traduzione inizia una volta che l'mRNA raggiunge i ribosomi: l'mRNA viene letto in gruppi di tre nucleotidi alla volta, chiamati *codoni*; ai codoni si legano gli *amminoacidi*, utilizzati per formare le proteine. Le proteine possono essere prodotte in modo costante o in in *burst*, fenomeno per cui una singola molecola di mRNA produce grandi quantità di proteine in tempi brevissimi. Una volta prodotte le proteine lasciano il ribosoma per svolgere le loro funzioni.

Il processo di produzione delle proteine non avviene in modo indipendente per ogni gene. Infatti vi è una vasta gamma di proteine che possono influenzare geni, attivandone o inibendone il processo di trascrizione. Questo complesso sistema di interazioni tra geni tramite le rispettive proteine è noto come rete di regolazione genica (GRN), dove i geni interagiscono e si regolano a vicenda tramite la dipendenza biologica di un gene dalla proteina prodotta da un altro gene. Le GRN coordinano precisamente l'espressione genica in risposta a segnali interni ed esterni, contribuendo alla regolazione del funzionamento cellulare e all'omeostasi dell'organismo.

È possibile misurare sperimentalmente il prodotto delle reti di regolazione genica (GRN) utilizzando dati di trascrittomica (RNA-sequencing), che descrivono il numero di mRNA trascritto da un

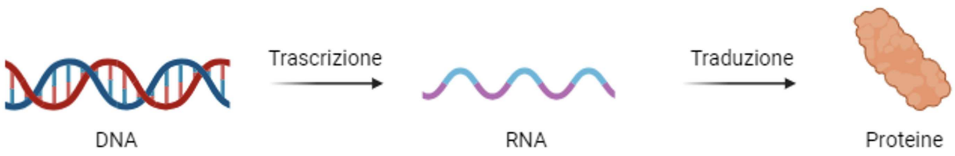


Figura 1.1: Dogma centrale della biologia: il DNA viene trascritto in RNA. L'RNA viene successivamente tradotto in proteine. Immagine creata con BioRender.com.

determinato gene. Fino ad oggi, la maggior parte degli studi di GRN e relativa espressione genica si basava su dati di RNA-sequencing bulk, dove l'informazione sui singoli geni viene ottenuta come quantità media su popolazioni di cellule[1, 2, 15]. Questo tipo di misura non permette di studiare l'eterogeneità presente tra cellule. Tuttavia, grazie all'avanzamento tecnologico, ora siamo in grado di misurare i livelli di mRNA anche a livello di singola cellula, con una tecnica denominata single-cell RNA-sequencing (scRNA-seq) [8, 12]. Tale tecnica permette di misurare l'espressione genica di migliaia di cellule per ogni campione, fornendo quindi una miglior descrizione della variabilità presente nei dati. A partire da dati di scRNA-seq è stato possibile sviluppare modelli che descrivono la dinamica delle cellule[6].

Un'integrazione interdisciplinare tra la biologia e la fisica offre un valido approccio per analizzare la regolazione dell'espressione genica all'interno di una GRN. Attraverso l'applicazione di modelli meccanicistici ispirati alla fisica, è possibile incorporare conoscenze sulle dinamiche molecolari e le interazioni che governano il sistema. Questi modelli mirano a catturare con precisione la dinamica del sistema biologico, facendo leva su formalismi matematici e teorie della fisica. L'adozione di un modello meccanicistico facilita l'interpretazione di dati sperimentali, consente di formulare previsioni teoriche e di inferire informazioni altrimenti non accessibili sperimentalmente. Tuttavia, la sfida principale di questo approccio risiede nell'incorporare dettagli meccanicistici che descrivano fedelmente il fenomeno biologico osservato senza compromettere l'interpretabilità del modello. Un esempio di approccio meccanicistico è il "modello a due stati", utilizzato con successo per dati da sequenziamento di RNA di singola cellula[10].

Il comportamento delle GRN rappresenta un sistema complesso che non può essere compreso analizzando i singoli geni isolatamente. L'obiettivo di questo elaborato è esplorare modelli meccanicistici che incorporano elementi biologici dettagliati per descrivere sia il comportamento di singoli geni che l'interazione tra essi all'interno delle GRN.

Per trattare l'espressione genica si farà riferimento ai processi stocastici: infatti, l'attivazione dei promotori (e conseguentemente la produzione di proteine) non è un processo deterministico, bensì un processo stocastico. Si definisce processo stocastico una collezione di variabili aleatorie che rappresentano l'evoluzione temporale di un sistema soggetto a casualità.

Si presentano 5 modelli che descrivono il funzionamento di un singolo gene (capitolo 2), fino ad arrivare ad un Processo di Markov frammentario-deterministico2.5[5], che tratta l'attivazione del gene come un processo stocastico. Quest'ultimo modello di singolo gene sarà la base per la costruzione della rete di regolazione genica, che descriveremo seguendo il lavoro proposto da Herbach et al. in [5]. Le GRN vengono descritte prima tramite un'approssimazione di campo medio chiamata "approssimazione di Hartree" 3.2[5], omonima dell'approssimazione usata in fisica della materia, dato l'approccio analogo. Successivamente, un modello di GRN con interazioni descritte esplicitamente viene fornito includendo nella descrizione il ruolo regolatore della cromatina nell'espressione genica3.3[5]. Infine, nel capitolo 4 vengono simulati 3 esempi di GRN basandosi sul modello con interazioni esplicite della sezione 3.3.

Capitolo 2

Modelli meccanicistici per singolo gene

Un gene può essere definito come un'unità di informazione che codifica per una specifica funzione biologica, solitamente attraverso la produzione di proteine. La comprensione di come un singolo gene regoli la produzione di proteine è cruciale per decifrare i meccanismi più complessi alla base della vita cellulare. I processi di trascrizione e traduzione, apparentemente lineari, nascondono una complessità regolata da una serie di fattori che possono influenzare l'efficienza e la velocità della produzione proteica. Per modellarli si possono adottare diversi approcci. Un modello deterministico assume che i livelli di mRNA e proteine possano essere descritti con equazioni differenziali, basandosi su tassi di produzione e degradazione costanti. Questo approccio è utile per una visione di alto livello, ma può non catturare l'intera varietà di comportamenti biologici intrinsecamente stocastici. Alternativamente, i modelli stocastici offrono una rappresentazione più fedele di variabilità e casualità dei processi di espressione genica, considerando le fluttuazioni nelle concentrazioni di mRNA e proteine come variabili aleatorie. Questi modelli sono particolarmente utili per esaminare la dinamica dell'espressione genica a livello di singola cellula, dove l'eterogeneità gioca un ruolo fondamentale. In questa sezione si espongono vari approcci, al fine di indagarne le proprietà e i limiti.

Il primo modello, analizzato nella sezione 2.1, prevede un approccio deterministico al problema con livelli sufficientemente elevati di proteine e mRNA. Nella sezione 2.2 si inizia a studiare il problema con un approccio stocastico, focalizzandosi sulla produzione di proteine. L'mRNA viene introdotto con il modello a due stadi, sezione 2.3. Nella sezione 2.4 si include nel sistema mRNA-proteine anche lo stato del gene (attivo o spento), la cui attivazione viene descritta tramite un processo stocastico; la dinamica del sistema viene quindi descritta come un processo di Markov frammentario-deterministico, dove la parte deterministica descrive l'evoluzione temporale dei livelli di mRNA e proteine.

Infine, nella sezione 2.5, si considera il sistema descritto nella sezione precedente come interamente stocastico e si studia la concentrazione delle proteine alla stazionarietà.

Si introduce ora la notazione utilizzata nelle sezioni seguenti:

$$m(t) := \text{concentrazione di mRNA} \qquad p(t) := \text{concentrazione di proteine} \tag{2.1}$$

$$\nu_0 := \text{rate di trascrizione} \qquad \nu_1 := \text{rate di traduzione} \tag{2.2}$$

$$d_0 := \text{rate di morte dell}'\text{mRNA} \qquad d_1 := \text{rate di morte delle proteine} \tag{2.3}$$

2.1 Modello deterministico per l'Espressione Proteica

Si descrive ora un modello deterministico per i processi di trascrizione e traduzione. Tale descrizione è giustificata dal fatto che, se i livelli di mRNA e proteine sono molto elevati, possono essere trattati come variabili continue deterministiche. Questa modello è descritto dal sistema di equazioni differenziali lineari:

$$\begin{aligned} \dot{m} &= \nu_0 - d_0 m \\ \dot{p} &= \nu_1 m - d_1 p \end{aligned} \tag{2.4}$$

Con condizioni iniziali:

$$m(0) = m_0 \quad p(0) = p_0$$

Il risultato della prima equazione viene trovato facilmente per separazione di variabili:

$$m(t) = \left(m_0 - \frac{\nu_0}{d_0}\right) e^{-d_0 t} + \frac{\nu_0}{d_0} \tag{2.5}$$

Sostituendo il risultato della prima equazione nella seconda si ottiene:

$$\dot{p} = \nu_1 \left(\left(m_0 - \frac{\nu_0}{d_0}\right) e^{-d_0 t} + \frac{\nu_0}{d_0} \right) - d_1 p \tag{2.6}$$

La parte omogenea viene risolta per separazione di variabili in maniera analoga alla prima equazione ($p_{om}(t) = De^{-d_1 t}$), la soluzione particolare invece viene trovata calcolando separatamente la parte relativa al termine noto costante e quella relativa al termine noto dipendente dal tempo; ciò è possibile poiché l'equazione è lineare del primo ordine. Per la soluzione particolare relativa alla parte dipendente dal tempo si distinguono due casi: $d_0 \neq d_1$ e $d_0 = d_1$. In entrambi i casi la soluzione si trova mediante integrazione seguendo classici metodi di risoluzione di equazioni lineari del primo ordine.
Si ottiene quindi la prima soluzione particolare:

$$p_{p_{2a}}(t) = e^{-d_1 t} \int_0^t \left(m_0 - \frac{\nu_0}{d_0}\right) e^{(-d_0+d_1)t} dt = \nu_1 \left(m_0 - \frac{\nu_0}{d_0}\right) \left(\frac{e^{-d_0 t} - e^{-d_1 t}}{d_1 - d_0}\right), \tag{2.7}$$

E la seconda:

$$p_{p_{2b}}(t) = e^{-d_1 t} \int_0^t \left(m_0 - \frac{\nu_0}{d_0}\right) dt = \nu_1 \left(m_0 - \frac{\nu_0}{d_0}\right) t e^{-d_1 t}. \tag{2.8}$$

La costante D viene trovata imponendo la condizione iniziale:

$$p_0 = D + \frac{\nu_1 \nu_0}{d_1 d_0} \quad \Rightarrow \quad D = p_0 - \frac{\nu_1 \nu_0}{d_1 d_0}$$

La soluzione finale é quindi:

$$p(t) = \frac{\nu_1 \nu_0}{d_1 d_0} + \left(p_0 - \frac{\nu_1 \nu_0}{d_1 d_0}\right) e^{-d_1 t} + \nu_1 \left(m_0 - \frac{\nu_0}{d_0}\right) F(t) \tag{2.9}$$

dove $F(t) = p_{p_{2a}}(t)$ se $d_0 \neq d_1$ e $F(t) = p_{p_{2b}}(t)$ se $d_0 = d_1$.
Questo modello deterministico non prevede alcun tipo di fluttuazioni nelle concentrazioni di mRNA e proteine, dunque è adatto solo a descrivere l'andamento dei valori medi e non rispecchia il reale sistema biologico mRNA-proteine in oggetto.

2.2 Modello discreto per l'Espressione Proteica

Passiamo ora ad una trattazione stocastica del sistema mRNA-proteine. Per fare ciò introduciamo il modello "a due stati" per la descrizione dell'espressione genica (Fig. 2.1) Le reazioni possibili per le singole molecole di mRNA e proteine in questo sistema sono le seguenti:

$$\begin{aligned} \emptyset &\xrightarrow{\nu_0} \text{mRNA} \\ \text{mRNA} &\xrightarrow{\nu_1} \text{mRNA} + \text{proteina} \\ \text{mRNA} &\xrightarrow{d_0} \emptyset \\ \text{proteina} &\xrightarrow{d_1} \emptyset \end{aligned} \tag{2.10}$$

con ν_0 e ν_1 rate di nascita e d_0 e d_1 rate di morte di mRNA e proteine, rispettivamente. Tali rate dipendono da molti fattori, tra cui la disponibilità di RNA polimerasi, la fase del ciclo cellulare in cui si trova la cellula e l'attività ribosomale. Assumeremo che questi rate non varino nel tempo.

In questa sezione consideriamo una versione semplificata di tale modello che trascura la dinamica dell'mRNA. Per fare ciò consideriamo innanzitutto che un mRNA vive in media un tempo d_0^{-1}, durante il quale produce proteine a un rate ν_1. Durante la sua vita produce quindi mediamente $d_0^{-1}\nu_1$

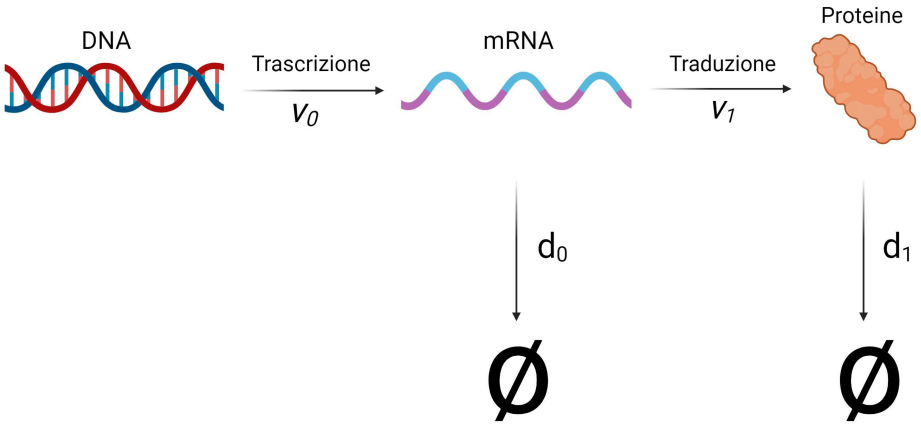


Figura 2.1: Schema del Modello a Due Stadi con v_0, v_1 rate di nascita di mRNA e proteine rispettivamente, d_0, d_1 sono invece i rate di morte, con notazione analoga per i pedici. Immagine creata con BioRender.com.

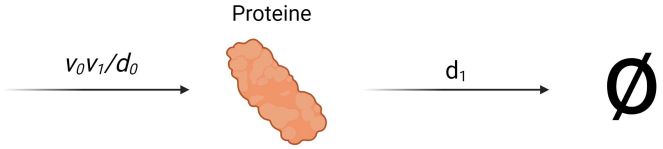


Figura 2.2: Modello di Espressione proteica con $b = \frac{\nu_0\nu_1}{d_0}$ rate di nascita delle proteine e d_1 rate di morte. Immagine creata con BioRender.com.

proteine. Come approssimazione possiamo quindi tralasciare la descrizione della dinamica dell'mRNA e concentrarci su un modello che descriva la sola dinamica di nascita e morte delle proteine (Fig. 2.2). In questo modello il rate di nascita delle proteine è $b_n = \frac{\nu_0\nu_1}{d_0}$ e il rate di morte sarà $d_n = d_1 n$. Si scrive ora la master equation per il sistema, dove $P_n(t)$ è la probabilità che vi siano n proteine al tempo t:

$$\begin{aligned} P_n(t+\delta t) &= P_{n-1}(t)\nu_1\delta t + P_{n+1}(t)d_1(n+1)\delta t + P_n(t)(1-\nu_1\delta t - d_1 n\delta t) \\ \frac{P_n(t+\delta t) - P_n(t)}{\delta t} &= P_{n-1}(t)\nu_1 + P_{n+1}(t)d_1(n+1) - P_n(t)\nu_1 - P_n(t)d_1 n \end{aligned} \tag{2.11}$$

Facendo ora tendere δt a zero otteniamo:

$$\frac{\partial P_n(t)}{\partial t} = \nu_1(P_{n-1}(t) - P_n(t)) + d_1((n+1)P_{n+1}(t) - nP_n(t)) \tag{2.12}$$

Per risolvere l'equazione si introduce la funzione generatrice dei momenti $G(z,t) = \sum_{n\geq 0} z^n P_n(t)$. Sostituendola nella master equation e usando le proprietà della funzione generatrice:

$$\sum_{n\geq 0} z^n P_{n-1}(t) = zG(z,t); \quad \sum_{n\geq 0}(n+1)z^n P_{n+1}(t) = \frac{\partial G(z,t)}{\partial z}; \quad \sum_{n\geq 0} nz^n P_n(t) = z\frac{\partial G(z,t)}{\partial z} \tag{2.13}$$

si riscrive l'equazione:

$$\frac{\partial G(z,t)}{\partial t} + d_1(z-1)\frac{\partial G(z,t)}{\partial z} = \nu_1(z-1)G(z,t) \tag{2.14}$$

L'equazione viene risolta ora con il metodo delle caratteristiche:

$$\begin{aligned} \frac{dz}{dt} &= d_1(z-1), \\ \frac{dG}{dt} &= \nu_1(z-1)G. \end{aligned} \tag{2.15}$$

Dalla prima equazione si trova quindi $z(t) = z_0 e^{d_1 t} + 1 \;\Rightarrow\; z_0 = (z-1)e^{-d_1 t}$. La seconda equazione si risolve sostituendo il risultato ottenuto nella prima e separando le variabili:

$$\begin{aligned} &\frac{dG}{G} = \int_0^t \nu_1 (z_0 e^{d_1 t}) dt \quad \Rightarrow \quad ln(G) = \frac{\nu_1 z_0}{d_1} e^{d_1 t} \\ &G = G_0(z_0) \exp\left(\frac{\nu_1 z_0}{d_1} e^{d_1 t}\right) = G_0(z_0) e^{\frac{\nu_1 (z-1)}{d_1}} \end{aligned} \tag{2.16}$$

Si ricordano ora alcune proprietà della funzione generatrice e si pone come condizione iniziale che al tempo $t = 0$ valga $P_n(0) = \delta^n_{n_0}$:

$$G(1,t) = G_0(0) = 1 \qquad G(z,0) = z^{n_0}$$

Si usa ora la seconda identità e si trova che:

$$G_0(z_0) e^{\frac{\nu_1}{d_1}(z-1)} = z^{n_0} \quad \Rightarrow \quad G_0(z-1) = z^{n_0} e^{-\frac{\nu_1}{d_1}(z-1)} \quad \Rightarrow \tag{2.17}$$

$$\Rightarrow \quad G_0(z) = (z+1)^{n_0} e^{-\frac{\nu_1}{d_1} z} \quad \Rightarrow \tag{2.18}$$

$$\Rightarrow \quad G_0(z_0) = G_0((z-1)e^{-d_1 t}) = ((z-1)e^{-d_1 t} + 1)^{n_0} \exp\left(-\frac{\nu_1}{d_1}(z-1)e^{-d_1 t}\right) \tag{2.19}$$

Si riporta quindi la funzione generatrice:

$$G(z,t) = ((z-1)e^{-d_1 t} + 1)^{n_0} \exp\left(\frac{\nu_1}{d_1}(z-1)(1 - e^{-d_1 t})\right) \tag{2.20}$$

Ricordando ora che $G(z,t) = \sum_{n \geq 0} z^n P_n(t)$, imponendo $n_0 = 0$ e ponendo $\lambda(t) := \frac{\nu_1}{d_1}(1 - e^{-d_1 t})$ si trova che la distribuzione di probabilità che descrive la concentrazione delle proteine è una poissoniana:

$$P_n(t) = e^{-\lambda(t)} \frac{\lambda^n(t)}{n!} \tag{2.21}$$

Tale risultato non è in accordo con i dati empirici, si limita alla descrizione dell'andamento del valore medio della concentrazione di proteine. Il problema di questo modello è che fa un'approssimazione che non rispecchia il sistema reale: non include l'mRNA, la cui presenza è fondamentale per la produzione di proteine.

2.3 Modello a Due Stadi per l'Espressione Genica

Si considerano ora il modello a due stati completo, con mRNA e proteine, descritto nella sezione precedente (Fig. 2.1, Reazioni 2.10).

Si procede quindi scrivendo la master equation per il sistema descritto, con m, n concentrazioni di mRNA e proteine rispettivamente e $P_{m,n}(t)$ la probabilità di avere tali concentrazioni al tempo t. Per allegerire la notazione si scrive $P_{m,n}$ tralasciando la dipendenza dal tempo.

$$\frac{\partial P_{m,n}}{\partial t} = \nu_0(P_{m-1,n} - P_{m,n}) + \nu_1 m(P_{m,n-1} - P_{m,n}) + d_0((m+1)P_{m+1,n} - mP_{m,n}) + d_1((n+1)P_{m,n+1} - nP_{m,n}) \tag{2.22}$$

Si definisce ora la funzione generatrice $G(x,y,t) = \sum_{m \geq 0} \sum_{n \geq 0} x^m y^n P_{m,n}(t)$, utilizzandone le proprietà:

$$\begin{aligned} &\sum_{m,n} x^m y^n P_{m-1,n} = xG \qquad &&\sum_{m,n} m x^m y^n P_{m,n-1} = xy \frac{\partial G}{\partial x} \\ &\sum_{m,n} m x^m y^n P_{m,n} = x \frac{\partial G}{\partial x} \qquad &&\sum_{m,n} (m+1) x^m y^n P_{m+1,n} = \frac{\partial G}{\partial x} \\ &\sum_{m,n} (n+1) x^m y^n P_{m,n+1} = \frac{\partial G}{\partial y} \quad &&\sum_{m,n} n x^m y^n P_{m,n} = x \frac{\partial G}{\partial n} \end{aligned} \tag{2.23}$$

si riscrive la master equation come segue:

$$\frac{\partial G}{\partial t} = \nu_0(x-1)G + \nu_1 x(y-1)\frac{\partial G}{\partial x} + d_0\left(\frac{\partial G}{\partial x} - x\frac{\partial G}{\partial x}\right) + d_1\left(\frac{\partial G}{\partial y} - y\frac{\partial G}{\partial y}\right) \tag{2.24}$$

Si riscrive nuovamente l'equazione mediante il cambio di variabili $u = x-1$, $v = y-1$, $\tau = d_1 t$ e introducendo i seguenti parametri:

$$a := \frac{\nu_0}{d_1} \quad b := \frac{\nu_1}{d_0} \quad \epsilon = \frac{d_1}{d_0} = \frac{\tau_0}{\tau_1} \tag{2.25}$$

con: a efficienza di trascrizione, ovvero il numero medio di burst per ciclo cellulare; b efficienza di traduzione, ovvero il numero medio di proteine prodotte per burst; τ_0 tempo di vita medio dell'mRNA e τ_1 tempo di vita medio delle proteine. Si riporta quindi l'equazione ottenuta:

$$\frac{\partial G}{\partial \tau} = auG + \frac{1}{\epsilon}(bv(u+1) - u)\frac{\partial G}{\partial u} - v\frac{\partial G}{\partial v}$$

Data la complessità analitica dell'equazione ottenuta, per risolverla si considera la sola produzione di proteine o di mRNA. In questo caso consideriamo solo la produzione di mRNA, ponendo $G(u, v = 0, \tau) = g(u, \tau) = \sum_{m\geq 0}(1+u)^m P_m(\tau)$. L'equazione allora diventa:

$$\frac{\partial g}{\partial \tau} + \frac{u}{\epsilon}\frac{\partial g}{\partial u} = aug \tag{2.26}$$

L'equazione viene ora risolta utilizzando il metodo delle caratteristiche:

$$\begin{aligned} \frac{du}{d\tau} &= \frac{u}{\epsilon} \\ \frac{dg}{d\tau} &= aug \end{aligned} \tag{2.27}$$

La prima equazione si risolve facilmente per separazione di variabili e si trova che $u = u_0 e^{\frac{\tau}{\epsilon}} \Rightarrow u_0 = ue^{-\frac{\tau}{\epsilon}}$. Anche la seconda equazione viene risolta in maniera analoga, trovando:

$$g = Cexp(a\epsilon U_0 e^{\frac{\tau}{\epsilon}}) = Cexp(a\epsilon u)$$

con C costante. Si riportano ora alcune proprietà della funzione generatrice, si impone la condizione iniziale $m(0) = k$ e si utilizza $P_k(\tau) = \delta_m^k$:

$$g(0,\tau) = 1 \quad g(u_0,\tau) = (1+u_0)^k$$

Utilizzando ora la seconda proprietà riportata si ricava la costante C:

$$(1+u_0)^k = Ce^{a\epsilon u_0} \quad \Rightarrow \quad C = (1+u_0)^k e^{-a\epsilon u_0} = (1+ue^{-\frac{\tau}{\epsilon}})^k exp(-a\epsilon u e^{-\frac{\tau}{\epsilon}})$$

Si trova quindi la funzione generatrice:

$$g(u,\tau) = (1+ue^{-\frac{\tau}{\epsilon}})^k exp(-a\epsilon u(1-e^{-\frac{\tau}{\epsilon}}))$$

Risolvendo l'equazione alla stazionarietà separando le variabili si trova invece:

$$g(u) = e^{a\epsilon u} \quad \Rightarrow \quad g(x) = e^{\frac{\nu_0}{d_0}(x-1)}$$

Ricordando poi che $g(x) = \sum_{m\geq 0} x^m P_m$ si trova infine la distribuzione di probabilità per l'mRNA alla stazionarietà è una poissoniana:

$$P_m = e^{-\frac{\nu_0}{d_0}}\frac{\left(\frac{\nu_0}{d_0}\right)^m}{m!};$$

Questa approssimazione del sistema, che tiene conto solo del mRNA e non delle proteine, non fornisce quindi risultati in accordo con i dati sperimentali.

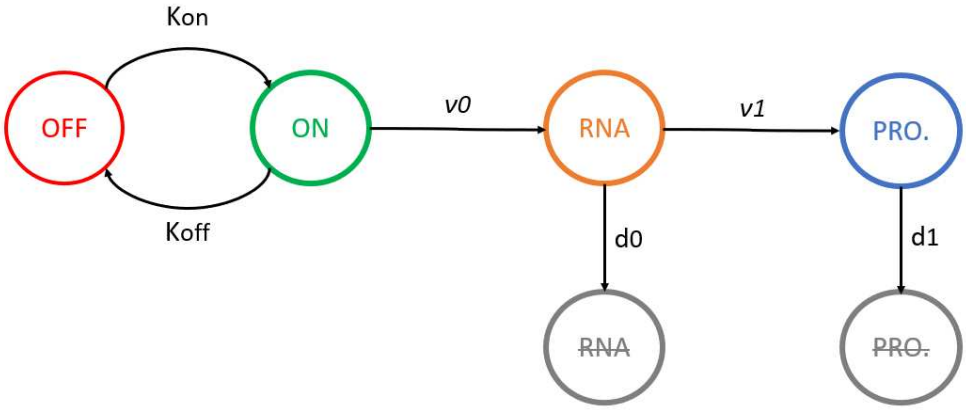


Figura 2.3: Schema del processo di Markov frammentario-deterministico usato per descrivere il comportamento di un singolo gene in questa sezione, con k_{on}, k_{off} rate di accensione e spegnimento del promotore rispettivamente, ν_0 e ν_1 rate di nascita di mRNA e proteine, rispettivamente, d_0 e d_1 rate di morte con notazione analoga ai rate di nascita per i pedici. La realizzazione di questa immagine è ispirata alla fig.1 del paper di Herbach et al. in [5].

2.4 Piecewise-deterministic Markov process model

Si descrive ora la dinamica del gene promotore, che può essere attivo o inattivo, e può passare da uno stato all'altro con rate costanti k_{on} e k_{off} rispettivamente. La trascrizione di mRNA avviene quindi solo durante i periodi di attività del gene, mentre la traduzione avviene mediante le molecole di mRNA. Quando $k_{on} << k_{off}$ e $d_0 << k_{off}$, il periodo di attivazione dei geni può essere visto come istantaneo e la dinamica dell'mRNA è molto più veloce di quella del promotore, implicando che la traduzione avvenga in burst: una molecola di mRNA produce grandi quantità di proteine in tempi brevissimi. Il modello risultante può essere definito dall'insieme delle reazioni chimiche riportate in tabella 2.1. Tale modello stocastico è di difficile trattazione analitica. Tuttavia, nel sistema analizzato

Reazioni	Rate	Interpretazione
$G \to G^*$	k_{on}	Attivazione del gene
$G^* \to G$	k_{off}	Spegnimentno del gene
$G^* \to G^* + M$	ν_0	Trascrizione
$M \to M + P$	ν_1	Traduzione
$M \to \emptyset$	d_0	Morte dell'mRNA
$P \to \emptyset$	d_1	Morte delle proteine

Tabella 2.1: Reazioni chimiche presenti nel modello, dove G,G^* sono rispettivamente gene spento e acceso, M è l'mRNA e P le proteine.

mRNA e proteine sono presenti in quantità sufficienti da permettere di trattarle come variabili continue. L'unico elemento ad essere discreto è il promotore, che risulta quindi essere l'unica variabile stocastica del modello. Matematicamente si introduce nel modello gli stati del promotore acceso ($E = 1$) e spento ($E = 0$), rendendo così il modello un Processo di Markov frammentario-deterministico (PDMP), descritto in figura 2.3. Un PDMP è un processo stocastico la cui evoluzione temporale è governata da salti casuali nel tempo, in mezzo ad essi il sistema segue evoluzioni deterministiche continue descritte da equazioni differenziali ordinarie. Come in un processo di Markov classico, il comportamento futuro del sistema dipende solo dallo stato attuale. Nel nostro caso i salti casuali nel tempo sono l'attivazione e la disattivazione del promotore, in mezzo a tali eventi l'andamento della concentrazione di proteine ed mRNA è di tipo deterministico continuo. Si riporta quindi il sistema di equazioni che rappresenta

l'evoluzione temporale del processo descritto:

$$\begin{aligned} &E : 0 \xrightarrow{k_{on}} 1, \;\; 1 \xrightarrow{k_{off}} 0 \\ &\dot{m} = \nu_0 E - d_0 m \\ &\dot{p} = \nu_1 m - d_1 p \end{aligned} \tag{2.28}$$

Si effettua ora un cambio di variabili in modo che le nuove equazioni siano adimensionali:

$$\begin{aligned} \hat{m} &= \frac{d_0}{\nu_0} m \\ \hat{p} &= \frac{d_0 d_1}{\nu_0 \nu_1} p \end{aligned} \tag{2.29}$$

Per alleggerire la notazione verrà tralasciato il cappuccio nell'analisi. Si riportano quindi le nuove equazioni:

$$\begin{aligned} \dot{m} &= d_0(E - m) \\ \dot{p} &= d_1(m - p) \end{aligned} \tag{2.30}$$

Sapendo che $E \in \{0, 1\}$ si risolve ora il sistema trattando la variabile stocastica E come se fosse costante. Le equazioni vengono risolte in maniera analoga a quella della sezione2.1, ricordando inoltre che si hanno le condizioni inizali $m(t_0) = m_0$ e $p(t_0) = p_0$. Si riportano quindi i risultati ottenuti:

$$\begin{aligned} m(t) &= (m_0 - E)e^{-d_0(t-t_0)} + E \\ p(t) &= (p_0 - E)e^{-d_1(t-t_0)} + E + \frac{d_1}{d_1 - d_0}(m_0 - E)(e^{-d_1(t-t_0)} - e^{-d_0(t-t_0)}) \end{aligned} \tag{2.31}$$

Sappiamo inoltre che la dinamica dell'mRNA è molto più rapida di quella delle proteine ($d_1 << d_0$), si trova quindi:

$$p(t) \approx (p_0 - E)e^{-d_1(t-t_0)} + E$$

che approssima la soluzione del sistema ridotto all'equazione $\dot{p} = d_1(m - p)$, ovvero il sistema che si ottiene considerando l'mRNA come stazionario. Per riottenere le soluzioni del sistema originale basta applicare alle soluzione trovate del sistema adimensionale il cambio di variabili inverso.

2.5 Distribuzione di mRNA o proteine alla stazionarietà

È possibile studiare il comportamento alla stazionarietà di mRNA e proteine per il modello descritto dalle reazioni 2.1. Qui ci concentriamo sulle proteine, anche se il risultato ottenuto può essere trovato in modo analogo per l'mRNA. Si riprende il sistema della sezione2.4 sapendo che nel limite in cui la dinamica dell'mRNA è molto più rapida delle proteine ($d_1 << d_0$) il rumore dell'mRNA è completamente mediato dalle proteine. In questo limite è possibile fare un'approssimazione di scala, considerando quindi l'mRNA come stazionario.

Si riportano quindi le equazioni di Chapman-Kolmogorov che descrivono l'evoluzione temporale del sistema gene-proteine, $P_0(x, t) = P_0$ e $P_1(x, t) = P_1$ sono le probabilità che il promotore sia spento o acceso rispettivamente, e $x(t) = x$ è la concentrazione di proteine nel tempo.

$$\begin{aligned} \frac{\partial P_0}{\partial t} &= d_1 \frac{\partial (x P_0)}{\partial x} + k_{off} P_1 - k_{on} P_0 \\ \frac{\partial P_1}{\partial t} &= -\frac{\partial [(\nu_1 - d_1 x) P_1]}{\partial x} + k_{on} P_0 - k_{off} P_1 \end{aligned} \tag{2.32}$$

Successivamente la soluzione del sistema di equazioni verrà trovata alla stazionarietà. Si riportano inoltre le no flux boundary condition $J_i(x) = 0$ per $x = 0, \frac{\nu_1}{d_1}$ e $i = 0, 1$, con $J_0(x) = -d_1 x P_0$ e $J_1(x) = (\nu_1 - d_1 x) P_1$, al fine di conservare la probabilità totale nel sistema. Inoltre, dalla condizione di

normalizzazione $\int_0^{\frac{\nu_1}{d_1}}(P_0+P_1)dx=1$, si ricava che $\int_0^{\frac{\nu_1}{d_1}}P_0dx=\frac{k_{off}}{k_{on}+k_{off}}$ e $\int_0^{\frac{\nu_1}{d_1}}P_1dx=\frac{k_{on}}{k_{on}+k_{off}}$; infatti, considerando la prima equazione del sistema:

$$d_1\frac{d(xP_0)}{dx}+k_{off}P_1-k_{on}P_0=0 \quad \Rightarrow \quad -d_1P_0-d_1x\frac{dP_0}{dx}=k_{off}P_1-k_{on}P_0$$

$$-d_1\int_0^{\frac{\nu_1}{d_1}}P_0dx-\int_0^{\frac{\nu_1}{d_1}}d_1x\frac{dP_0}{dx}dx=k_{off}\int_0^{\frac{\nu_1}{d_1}}P_1dx-k_{on}\int_0^{\frac{\nu_1}{d_1}}P_0dx \quad \Rightarrow \quad k_{off}\int_0^{\frac{\nu_1}{d_1}}P_1dx=k_{on}\int_0^{\frac{\nu_1}{d_1}}P_0dx$$

$$\int_0^{\frac{\nu_1}{d_1}}(P_1+P_0)\,dx=1 \quad \Rightarrow \quad \int_0^{\frac{\nu_1}{d_1}}P_0\left(1+\frac{k_{on}}{k_{off}}\right)dx=1 \quad \Rightarrow$$

$$\Rightarrow \quad \int_0^{\frac{\nu_1}{d_1}}P_0dx=\frac{k_{off}}{k_{on}+k_{off}} \quad \Rightarrow \quad \int_0^{\frac{\nu_1}{d_1}}P_1dx=\frac{k_{on}}{k_{on}+k_{off}}$$

Si definiscono ora $\alpha_0(x)=-d_1x$, $\alpha_1(x)=\nu_1-d_1x$. Riscriviamo ora il sistema alla stazionarietà in forma vettoriale:

$$\partial_x[\hat{\alpha}(x)\vec{P}(x)]=\hat{K}\vec{P}(x) \tag{2.33}$$

Dove:

$$\vec{P}(x)=\begin{pmatrix}P_0(x)\\P_1(x)\end{pmatrix} \qquad \hat{\alpha}(x)=\begin{pmatrix}\alpha_0(x)&0\\0&\alpha_1(x)\end{pmatrix} \qquad \hat{K}=\begin{pmatrix}-k_{on}&k_{off}\\k_{on}&-k_{off}\end{pmatrix} \tag{2.34}$$

Si definiscono ora il vettore $\vec{q}(x)=\hat{\alpha}(x)\vec{P}(x)$, la matrice $\hat{M}=\hat{K}\hat{\alpha}^{-1}(x)$ e si riscrive nuovamente il sistema:

$$\partial_x\vec{q}(x)=\hat{M}\vec{q}(x) \tag{2.35}$$

In seguito vengono sottointese le dipendenze da x per alleggerire la notazione. Scritto il sistema in questa forma, per risolvere il problema si cercano gli autovalori e gli autovettori di $\hat{M}$. Notando che $det(\hat{M})=0$ e che $tr(\hat{M})=-\left(\frac{k_{on}}{\alpha_0}+\frac{k_{off}}{\alpha_1}\right)$, si trova che gli autovalori sono $\lambda_1=0$, $\lambda_2=-\left(\frac{k_{on}}{\alpha_0}+\frac{k_{off}}{\alpha_1}\right)$. Calcolando gli autovettori, si trova rispettivamente:

$$\vec{v}_1=\begin{pmatrix}1\\\frac{\alpha_1k_{on}}{\alpha_0k_{off}}\end{pmatrix} \qquad \vec{v}_2=\begin{pmatrix}1\\-1\end{pmatrix} \tag{2.36}$$

Per trovare $\vec{q}$, si definisce ora $\tilde{v}_i=e_i^{\phi}\vec{v}_i$, con $i=1,2$ e $\phi_i=\phi_i(x)$: $\vec{q}$ verrà calcolato come combinazione lineare dei $\tilde{v}_i$. Per trovare le funzioni ϕ_i si sfrutta il fatto che i vettori $\tilde{v}_i$ sono autovettori di $\hat{M}$ relativi ai medesimi autovalori dei $\vec{v}_i$. Si calcola ora ϕ_1:

$$\partial_x\tilde{v}_1=\vec{0} \quad \Rightarrow \quad \phi_1'\vec{v}_1e_1^{\phi}+e_1^{\phi}\vec{v}_1'=0 \tag{2.37}$$

Prendendo la prima equazione del sistema si trova che $\phi_1=c$, con $c\in\mathbb{R}$. Si calcola ora ϕ_2:

$$\partial_x\tilde{v}_2=-\left(\frac{k_{on}}{\alpha_0}+\frac{k_{off}}{\alpha_1}\right)\tilde{v}_2 \tag{2.38}$$

Considerando ora la prima equazione del sistema, si trova:

$$\partial_x\phi_2=-\left(\frac{k_{on}}{\alpha_0}+\frac{k_{off}}{\alpha_1}\right)=\frac{k_{on}}{d_1x}-\frac{k_{off}}{\nu_1-d_1x} \quad \Rightarrow \quad \phi_2=\frac{k_{on}}{d_1}log(d_1x)+\frac{k_{off}}{d_1}log(\nu_1-d_1x) \tag{2.39}$$

Si trovano quindi i vettori $\tilde{v}_i$:

$$\tilde{v}_1=e^c\begin{pmatrix}1\\\frac{\alpha_1k_{on}}{\alpha_0k_{off}}\end{pmatrix} \qquad \tilde{v}_2=(d_1x)^{\frac{k_{on}}{d_1}}(\nu_1-d_1x)^{\frac{k_{off}}{d_1}}\begin{pmatrix}1\\-1\end{pmatrix} \tag{2.40}$$

Sia ora $\vec{q} = a\tilde{v}_1 + b\tilde{v}_2$, per definizione di $\vec{q}$ avremo che $\vec{P} = \hat{\alpha}^{-1}(a\tilde{v}_1 + b\tilde{v}_2)$. Vogliamo ora normalizzare le probabilità, affinchè p_0,p_1 siano normalizzabili si vede chiaramente che bisogna porre $a = 0$: se così non si facesse, gli integrali per la normalizzazione divergerebbero in $x = 0$. Rimaniamo quindi con:

$$\vec{P} = \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} = -\frac{b}{d_1} \begin{pmatrix} (d_1 x)^{\frac{k_{on}}{d_1}-1}(\nu_1 - d_1 x)^{\frac{k_{off}}{d_1}} \\ (d_1 x)^{\frac{k_{on}}{d_1}}(\nu_1 - d_1 x)^{\frac{k_{off}}{d_1}-1} \end{pmatrix} \tag{2.41}$$

Ricaviamo ora la costante b sfruttando la condizione di normalizzazione su p_0:

$$\begin{aligned} \int_0^{\frac{\nu_1}{d_1}} P_0 dx = -\int_0^{\frac{\nu_1}{d_1}} \frac{b}{d_1}(d_1 x)^{\frac{k_{on}}{d_1}-1}(\nu_1 - d_1 x)^{\frac{k_{off}}{d_1}} dx = -\frac{b}{d_1^2}\nu_1^{\frac{k_{on}+k_{off}}{d_1}} \int_0^1 t^{\frac{k_{on}}{d_1}-1}(1-t)^{\frac{k_{off}}{d_1}} dt = \\ = -\frac{b}{d_1^2}\nu_1^{\frac{k_{on}+k_{off}}{d_1}} B\left(\frac{k_o n}{d_1}, \frac{k_{off}}{d_1}+1\right) = \frac{k_{off}}{k_{on}+k_{off}} \end{aligned} \tag{2.42}$$

Con $B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$ funzione beta. Usando ora la proprietà della funzione beta $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ (con $\Gamma(z) = \int_0^{+\infty} t^{z-1}e^{-t}$ funzione gamma di Eulero), si calcola b:

$$\begin{aligned} -\frac{b}{d_1^2}\nu_1^{\frac{k_{on}+k_{off}}{d_1}} B\left(\frac{k_o n}{d_1}, \frac{k_o f f}{d_1}+1\right) &= \frac{k_{off}}{k_{on}+k_{off}} \\ b = -d_1^2 \nu_1^{-\frac{k_{on}+k_{off}}{d_1}} \frac{k_{off}}{k_{on}+k_{off}} &\frac{\Gamma\left(\frac{k_{on}+k_{off}}{d_1}+1\right)}{\Gamma\left(\frac{k_{on}}{d_1}\right)\Gamma\left(\frac{k_{off}}{d_1}+1\right)} \\ b = -d_1^2 \nu_1^{-\frac{k_{on}+k_{off}}{d_1}} &\frac{\Gamma\left(\frac{k_{on}+k_{off}}{d_1}\right)}{\Gamma\left(\frac{k_{on}}{d_1}\right)\Gamma\left(\frac{k_{off}}{d_1}\right)} = -\frac{d_1^2 \nu_1^{-\frac{k_{on}+k_{off}}{d_1}}}{B\left(\frac{k_{on}}{d_1}, \frac{k_{off}}{d_1}\right)} \end{aligned} \tag{2.43}$$

Si trova quindi:

$$\vec{P} = -\frac{d_1^2 \nu_1^{-\frac{k_{on}+k_{off}}{d_1}}}{B\left(\frac{k_{on}}{d_1}, \frac{k_{off}}{d_1}\right)} \begin{pmatrix} (d_1 x)^{\frac{k_{on}}{d_1}-1}(\nu_1 - d_1 x)^{\frac{k_{off}}{d_1}} \\ (d_1 x)^{\frac{k_{on}}{d_1}}(\nu_1 - d_1 x)^{\frac{k_{off}}{d_1}-1} \end{pmatrix} \tag{2.44}$$

Ponendo ora $\frac{\nu_1}{d_1} = 1$ si trova la densità di probabilità totale $P(x) = P_0(x) + P_1(x)$:

$$P(x) = \frac{x^{\frac{k_{on}}{d_1}-1}(1-x)^{\frac{k_{off}}{d_1}-1}}{B\left(\frac{k_{on}}{d_1}, \frac{k_{off}}{d_1}\right)}$$

che è una distribuzione beta. Si riportano alcuni esempi di distribuzioni beta in fig.2.4.

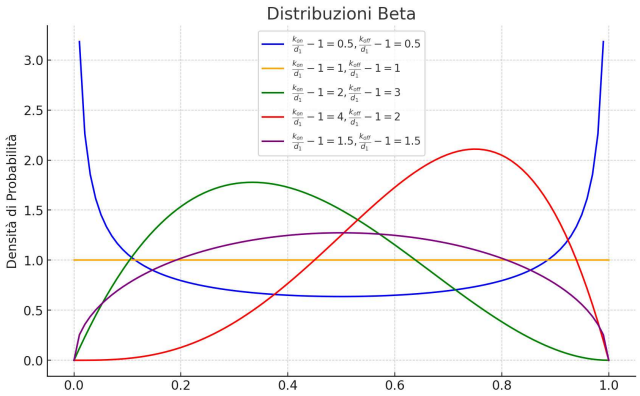


Figura 2.4: Esempi di distribuzioni beta.

Capitolo 3

Modelli per reti di regolazione genica

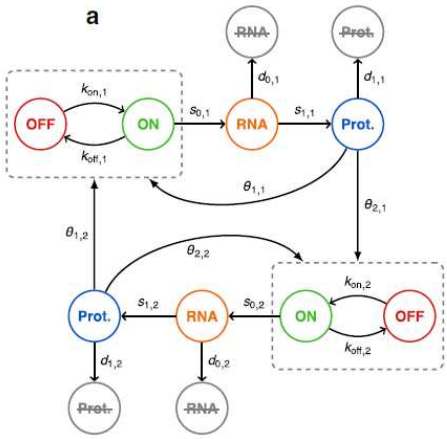


Figura 3.1: Rappresentazione della GRN ottenuta dalla generalizzazione del modello PDMP di singolo gene della sezione 2.4 modellizzando le interazioni tra i geni. L'immagine è presa dalla fig.2a del paper di Herbach et al. in [5], dove $\theta_{i,j}$ rappresenta l'azione del gene j sul gene i, $k_{on,i}$ e $k_{off,i}$ sono i rate di accensione e spegnimento del promotore i rispettivamente, $\nu_{0,i}$ e $\nu_{1,i}$ sono i rate di nascita di mRNA e proteine rispettivamente del gene i, $d_{0,i}$ e $d_{1,i}$ sono i rate di morte del gene i con notazione analoga ai rate di nascita per i pedici. Nella figura, i geni sono all'interno di riquadri grigi, rappresentati dai due possibili stati in cui possono trovarsi.

È noto che le proteine prodotte da un gene possono influenzare l'attività di un altro gene attivandolo o inibendolo. Per avere quindi un modello di descrizione di regolazione genica in una GRN è necessario modellare l'attività di più geni e le loro interazioni, che possono essere attivatorie o inibitorie.

Presentiamo quindi un esempio di modello di GRN a partire dal modello PDMP di singolo gene descritto nella sezione 2.4. Vogliamo includere la dipendenza dell'attività di un gene da parte delle proteine prodotte da altri geni, per farlo includiamo la dipendenza dei rate k_{on} e k_{off} del gene dal livello delle proteine presenti nel sistema. Possiamo quindi descrivere il modello come segue:

$$\begin{aligned} E : 0 &\xrightarrow{k_{on}(p_1,\dots,p_n)} 1, \;\; 1 \xrightarrow{k_{off}(p_1,\dots,p_n)} 0 \\ \dot{m}_i &= \nu_{0,i} E_i - d_{0,i} m_i \\ \dot{p}_i &= \nu_{1,i} m_i - d_{1,i} p_i \end{aligned} \tag{3.1}$$

Viene riportato un esempio di tali interazioni per un sistema composto da 2 geni in figura 3.1. La dipendenza dei rate k_{on}, k_{off} dalle proteine così scritta non è definita, un esempio di funzione che la descrive si vedrà in seguito.

Per trattare i diversi modi di descrivere le GRN nella sezione 3.1 si parlerà di basi di teoria dei grafi. Nella sezione 3.2 si presenterà un metodo per trattare le reti geniche e modellare l'attività genica al fine di trovare la concentrazione di proteine alla stazionarietà anche nel caso di network, detto approssimazione di Hartree. Infine, nella sezione 3.3, si presenta un modello meccanicistico dove la dipendenza dalle proteine dei rate $k_{on,i}$, $k_{off,i}$, viene modellata tramite l'inclusione dello stato della cromatina nel modello.

3.1 Elementi di teoria dei grafi

Una rete genica può essere descritta matematicamente utilizzando un grafo. Un grafo $G = (V, E)$ è composto da un insieme di nodi, o vertici, $V = n_1, ..., n_N$ e un insieme di link, o archi, $E = l_1, ..., l_K$. I nodi sono elementi tra loro distinti, gli archi, o connessioni, sono coppie di elementi di V e descrivono le relazioni tra i nodi. Due nodi legati tramite un arco si dicono vicini. Il numero di vertici N è detto *ordine* del grafo, mentre il numero di archi L è detto *dimensione* di G. Nel nostro caso i nodi saranno i geni mentre gli archi saranno le relazioni di attivazione o inibizione tra i geni.

Il rapporto ρ tra il numero di archi L e il massimo numero possibile di archi nella rete, L_{MAX}, è detto densità del grafo. Se un grafo di ordine N ha zero archi, ed è quindi composto da N nodi isolati, la sua densità è nulla e si dice che il grafo è *vuoto*. Il numero massimo di archi in un grafo di ordine N scala come N^2. Un grafo con un numero di archi vicino quello massimo, $L \sim N^2$, si dice *denso*; se invece $L << N^2$ il grafo si dice *sparso*.

I grafi possono essere classificati in base alla natura degli archi, definiamo quindi:

- I grafi diretti (fig.3.2(b)), dove gli archi sono coppie ordinate distinte (i, j) di elementi distinti di V con una direzione. L'arco potrà essere uscente da i ed entrante in j o viceversa. Il numero massimo di archi in un grafo diretto di ordine N è $L_{MAX} = N(N-1)$;

- I grafi indiretti (fig.3.2(a)), dove gli archi sono coppie non ordinate distinte (i, j) di elementi distinti di V senza una direzione. Il numero massimo di archi in un grafo indiretto di ordine N è $L_{MAX} = \frac{N(N-1)}{2}$.

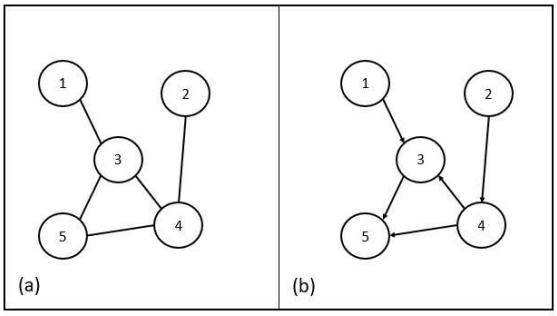


Figura 3.2: (a) Esempio di grafo indiretto. (b) Esempio di grafo diretto.

Gli archi inoltre possono avere o meno un peso. In base a ciò definiamo quindi:

- I grafi pesati (fig.3.3(b)), dove ad ogni arco è associato un valore numerico che definisce l'intensità dell'interazione rappresentata.

- I grafi non pesati (fig.3.3(a)), dove agli archi non sono associati pesi.

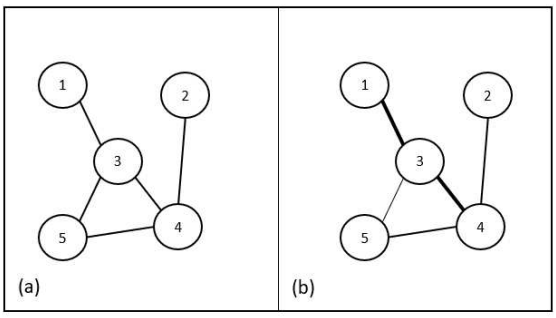


Figura 3.3: (a) Esempio di grafo non pesato. (b) Esempio di grafo pesato.

Ad ogni nodo i di un grafo è associato un *grado* k_i, definito come il numero di archi connessi al nodo i. Se il grafo è diretto si può distinguere tra il grado entrante, il numero di archi entranti in i, e il grado uscente, il numero di archi uscenti da i. Il grado medio di un grafo di ordine N è $< k >= \frac{\sum_{i=1}^{N} k_i}{N}$.

Un modo per rappresentare un grafo è la *matrice delle adiacenze* o *delle interazioni* $\theta = (\theta_{i,j})$, una matrice quadrata NxN dove gli elementi θ_{ij} sono diversi da zero solo se i nodi i, j sono connessi da un arco, in questo caso i nodi si dicono adiacenti. Per i grafi indiretti la matrice è simmetrica. Se il grafo è non pesato gli elementi di matrice potranno assumere solo valori binari, $\theta_{i,j} = 0$ se i e j non sono connessi e $\theta_{i,j} = 1$ altrimenti. Diversamente, se il grafo è pesato i valori numerici rappresenteranno l'intensità dell'interazione. Nel caso delle reti geniche gli elementi di matrice potranno avere valore positivo se tra i due geni vi è una relazione di attivazione, valore negativo se vi è una relazione di inibizione, nullo se l'espressione genica di un gene non influenza l'attività dell'altro gene.

Se in un grafo esiste almeno un cammino tra ogni coppia di nodi, il grafo si dice *connesso*.

Se in un grafo sono presenti self-loop o archi multipli, coppie di nodi connesse tra loro da più di un arco, il grafo viene chiamato $multigrafo$.

3.1.1 Tipologie di rete

Si riportano in questa sezione alcuni esempi rappresentativi di tipologie reti. Non si considerano self-loop o archi multipli, inoltre dove non specificato si considerano solo link indiretti non pesati.

Rete random

Un grafo random, o grafo di Erdös-Rényi, è una rete con gli archi distribuiti in maniera randomica seguendo il modello di Erdös-Rényi. Per generare un grafo random di ordine N con probabilità p che due nodi siano collegati, partiamo da un grafo vuoto di ordine N. Si collega ora ogni coppia di nodi con probabilità p, il grafo così ottenuto a una distribuzione di probabilità binomiale per il grado k dei nodi:

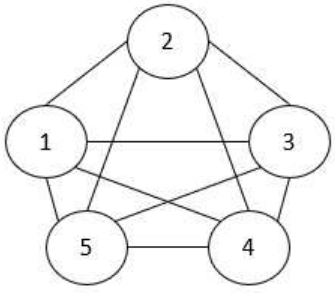
$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{3.2}$$

Il grado medio è quindi $< k >= p(N-1)$ e la varianza è $\sigma_k^2 = p(1-p)(N-1)$. La distribuzione del grado è omogenea, quindi possiamo approssimare $k_i \approx < k >$. Se il grafo random è sparso, il grado di ogni nodo è molto minore di N, $P(k)$ può essere quindi approssimata a una poissoniana:

$$P(k) = \frac{(\lambda)^k e^{-\lambda}}{k!} \tag{3.3}$$

con grado medio $< k >= \lambda = pN$ e varianza $\sigma_k^2 = \lambda$.
La matrice delle adiacenze assumerà dei valori randomici con probabilità p che l'elemento di matrice θ_{ij} sia diverso da zero.

Rete fully connected



$$\theta = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \tag{3.4}$$

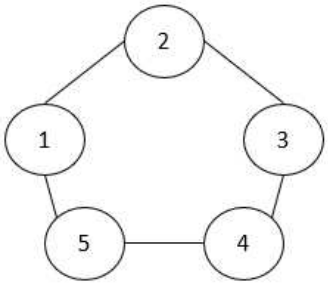
Figura 3.4: Esempio di rete $fully\ connected$ a 5 nodi.

Un grafo *fully connected*, o *completo*, è un grafo dove tutti i nodi sono connessi tra loro, ciascuna coppia è connessa da un singolo arco, la densità del grafo sarà quindi $\rho = 1$. Il grafo ha $\frac{N(N-1)}{2}$ archi. In questo caso tutti gli elementi di matrice θ_{ij} della matrice delle adiacenze saranno diversi da zero.

Si riporta un esempio di rete fully connected a 5 nodi in figura 3.4 con la relativa matrice delle interazioni 3.4.

Rete ad anello e a cascata

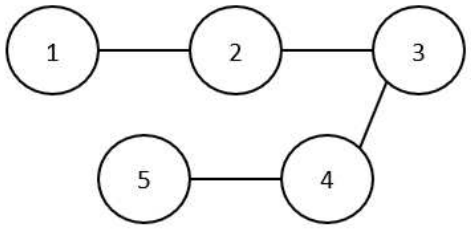
Una rete ad anello è rappresentabile da un grafo tale per cui un nodo i è connesso esattamente ad altri due nodi, $i-1$ e $i+1$. In questo caso gli unici elementi della matrice di adiacenza non nulli saranno gli elementi $\theta_{i+1,i}$, $\theta_{i,i+1}$ $\forall i$, e gli elementi $\theta_{N,1}$, $\theta_{1,N}$ (assumendo che il grafo sia di ordine N). Si riporta un esempio di rete ad anello a 5 nodi in figura 3.5 con la relativa matrice delle interazioni 3.5.



$$\theta = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{3.5}$$

Figura 3.5: Esempio di rete ad anello a 5 nodi.

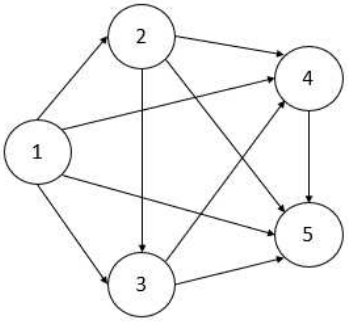
Nel caso in cui $\theta_{N,1}$, $\theta_{1,N} = 0$ la rete è detta a cascata. Si riporta un esempio di rete a cascata a 5 nodi in figura 3.6 con la relativa matrice delle interazioni 3.6.



$$\theta = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{3.6}$$

Figura 3.6: Esempio di rete a cascata a 5 nodi.

Rete gerarchica



$$\theta = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \tag{3.7}$$

Figura 3.7: Esempio di rete gerarchica a 5 nodi.

Per questa tipologia di rete è necessario considerare link diretti dove in $\theta_{i,j}$ i viene influenzato da j. Tale rete si dice gerarchica poiché un nodo influisce solo sull'attività dei nodi con indice di riga superiore al suo: il primo nodo influirà quindi sull'attività di tutti gli altri, il secondo sull'attività di tutti gli altri meno il primo e così via. Si riporta un esempio di rete gerarchica a 5 nodi in figura 3.7 con la relativa matrice delle interazioni 3.7.

3.2 Approssimazione di Hartree per il Modello PDMP

Si presenta ora un metodo per studiare il problema della descrizione delle interazioni in una rete genica. Consideriamo una rete composta da n geni, la sua evoluzione temporale è governata da un sistema di equazioni differenziali alle derivate parziale 2^n-dimensionale. Si considera inoltre di essere nel limite in cui ($d_1 << d_0$), dove è possibile fare un'approssimazione di scala e considerare l'mRNA come stazionario. Riprendendo quindi per ogni gene il modello della sezione 2.5, il sistema è:

$$\begin{aligned} &E_i : 0 \xrightarrow{k_{on,i}} 1, 1 \xrightarrow{k_{off,i}} 0 \\ &\dot{p}_i = d_{1,i}(E_i - p_i) \end{aligned} \tag{3.8}$$

Il sistema non è analiticamente trattabile, per risolverlo si utilizza il approssimazione "di Hartree", così chiamato poichè l'idea è analoga a quella dell'omonimo metodo in fisica della materia[7]: si assume che gli n geni agiscano indipendentemente l'uno dall'altro, ma siano soggetti a un "campo proteomico" medio generato dagli altri geni; il problema, inizialmente 2^n-dimensionale, viene ridotto così a n problemi 2-dimensionali indipendenti.

Prima di scrivere la master equation associata al sistema si riporta la notazione utilizzata nella trattazione che segue: siano $\mathcal{E} = \{0,1\}^n$ e $\Omega = (0,1)^n$. Al tempo t le configurazioni di promotore e proteine saranno rispettivamente $E_t = (e_1, ..., e_n) = e \in \mathcal{E}$, $P_t = (y_1, ..., y_n) = y \in \Omega$. La funzione $u(t,y) = (u_e(t,y))_{e\in\mathcal{E}} \in \mathbb{R}^{2^n} \simeq (\mathbb{R}^2)^{\bigotimes n}$ rappresenta la funzione di densità di probabilità di (E_t, P_t). Si riporta ora la master equation associata al sistema:

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{n} \frac{\partial F_i u}{\partial y_i} = \sum_{i=1}^{n} K_i u \tag{3.9}$$

Con le matrici $F_i(y_i), K_i(y_i) \in \mathcal{M}_{2^n}(\mathbb{R}) \simeq \mathcal{M}_2(\mathbb{R})^{\bigotimes n}$ definite nel seguente modo:

$$\hat{F}_i(y_i) = \mathbb{I}_2 \bigotimes ... \bigotimes \hat{F}^{(i)}(y_i) \bigotimes ... \bigotimes \mathbb{I}_2 \qquad \hat{K}_i(y_i) = \mathbb{I}_2 \bigotimes ... \bigotimes \hat{K}^{(i)}(y_i) \bigotimes ... \bigotimes \mathbb{I}_2 \tag{3.10}$$

$$\hat{F}^{(i)}(y_i) = \begin{pmatrix} -d_{1,i}y_i & 0 \\ 0 & d_{1,i}(1-y_i) \end{pmatrix} \qquad \hat{K}^{(i)}(y_i) = \begin{pmatrix} -k_{on,i}(y) & k_{off,i}(y) \\ k_{off,i}(y) & k_{on,i}(y) \end{pmatrix} \tag{3.11}$$

La somma a sinistra nell'equazione 3.9 è quindi un termine di trasporto deterministico, mentre la somma a destra è il termine stocastico che descrive le transizioni tra le configurazioni possibili dei promotori.

Si riportano inoltre le condizioni al contorno del sistema

$$\forall i\{1, ..., n\}, \;\; \hat{F}_i u = 0 \text{ in } \partial\Omega \tag{3.12}$$

e le condizioni legate alla probabilità

$$u \geq 0 \quad \wedge \quad \forall t \in \mathbb{R}_+, \; \sum_{e\in\mathcal{E}} \int_\Omega u_e(t,y)dy = 1 \tag{3.13}$$

L'approssimazione di Hartree consiste ora nel fissare i valori di y_j per $i \neq j$ con i fissata, come anticipato il problema 2^n-dimensionale diventa così n probelmi 2-dimensionali distinti. La soluzione del problema verrà approssimata dal prodotto tensore delle n soluzioni dei problemi 2-dimensionali ottenute.

Il problema ridotto per il gene i è quindi:

$$\frac{\partial u^i}{\partial t} + \frac{\partial \hat{F}^{(i)} u^i}{\partial y_i} = \hat{K}^{(i)} u^i \tag{3.14}$$

con $u^i(t,y) = (u_0^i(t,y), u_1^i(t,y))^T \in \mathbb{R}_+^2$. Si nota che il problema ridotto può essere risolto analogamente all'ultimo modello per singolo gene trattato nella sezione 2.5: una volta trovate le u^i, si può approssimare la pdf finale alla stazionarietà come:

$$u(t,y) \approx \bigotimes_{i=1}^{n} u^i(t,y) \tag{3.15}$$

3.3 Modello PDMP per GRN con interazioni esplicite

Vogliamo ora descrivere un sistema di più geni senza considerarli indipendenti, come nella sezione precedente, ma considerandone le interazioni. Facendo riferimento al modello a due stati per una GRN, descritto dalle equazioni 3.1, si presenta ora un modello di rete genica dove le interazioni vengono calcolate esplicitamente includendo un nuovo dettaglio biologico, la descrizione dello stato della cromatina. Alla base di questo modello vi è l'idea che la velocità di trascrizione dipenda anche da come è distribuita spazialmente la cromatina, il cui stato influenza l'accessibilità dei promotori per le proteine. Per la trattazione di questo modello si fa riferimento all'articolo[5].

Recenti studi [14, 9, 3] hanno mostrato come un aumento della trascrizione provochi un aumento della frequenza dei burst $\left(\frac{1}{k_{on}}\right)$ piuttosto che dell'ampiezza dei burst $\left(\frac{\nu_0}{k_{off}}\right)$, per questo motivo il modello considerato modifica solo i rate $k_{on,i}$ di ogni gene i, tenendo costanti i $k_{off,i}$. Si assume quindi che la frequenza di attivazione di un gene dipende dalle proteine, mentre il tempo di attivazione è casuale e dipende da k_{off}. Nella discussione che segue si tralascia l'indice i per semplicità.

Consideriamo quindi un insieme di stati della cromatina, a ognuno di essi viene associato uno di due possibili rate di attivazione del promotore k_1 e k_0, con $k_0 << k_1$: si chiameranno stato permissivo e non permissivo rispettivamente. La dinamica di transizione della cromatina tra questi stati è molto più veloce di quella di attività del promotore, si può quindi fare una separazione di scale in cui l'attivazione del promotore dipende da uno stato quasi stazionario della cromatina. Ciò è reso possibile dal fatto che le proteine modificano lo stato in cui si trova la cromatina mediante reazioni "hit-and-run", ovvero di legano e si staccano dalla cromatina molto velocemente, permettendo quindi di considerarla in uno stato quasi stazionario. Al gene i viene associato un $k_{on,i}$ che sarà combinazione lineare di k_0 e k_1, che saranno pesati dalle probabilità p_0 e p_1 (funzioni di $y_1, ..., y_n$) rispettivamente, la cui stima è scopo della trattazione che segue.
In assenza di proteine la cromatina si può trovare in due configurazioni (dette "basali"), quella permissiva e quella non permissiva, verranno indicate rispettivamente con $C_\emptyset^*$ e $C_\emptyset$. In assenza di proteine la dinamica è descritta dalle due reazioni

$$C_\emptyset \xrightarrow{\alpha} C_\emptyset^* \qquad C_\emptyset^* \xrightarrow{\beta} C_\emptyset \tag{3.16}$$

con α, β rate di transizione. Considerando ora la presenza di proteine, sia $I \subset G = 1, ..., n$, la configurazione C_I indica che la cromatina è permissiva nello stato I, C_I^* invece indica che la cromatina è non permissiva nello stato I. Se una proteina P_j ($j \in G \backslash I$) interagisce con la cromatina, può alterare il suo stato e la cromatina può successivamente tornare allo stato precedente secondo le seguenti reazioni:

$$\begin{aligned} C_I^* + P_j &\xrightarrow{a_j} C_{I\cup j}^* + P_j, \quad C_{I\cup j}^* \xrightarrow{b_j} C_I^*, \\ C_I + P_j &\xrightarrow{c_j} C_{I\cup j} + P_j, \quad C_{I\cup j} \xrightarrow{d_j} C_I. \end{aligned} \tag{3.17}$$

Il sistema evolverà con $[C_I], [C_I^*] \in \{0,1\}$ e $\sum_I [C_I] + [C_I^*] = 1$. Si è così costruito un processo di Markov a salti con 2^{n+1} stati possibili, con 2^n stati permissivi e 2^n non permissivi.
Sia ora $S = \{0,1\}^{n+1}$, ogni stato è rappresentato dal vettore $s = (s_0, s_1, ..., s_n)$, con $s_0 = 1$ se la

cromatina è in configurazione permissiva e 0 altrimenti, $s_j = 1$ ($j \in G$) se la cromatina è stata alterata dalla proteina P_j e 0 altrimenti. Se i rate delle reazioni son tutti positivi la master equation che descrive il sistema ha un'unica soluzione stazionaria. La probabilità P_s che il sistema si trovi nello stato s è:

$$P_s = \begin{cases} Z^{-1}\alpha \prod_{j=1}^{n}(\lambda_j[P_j]s_j + 1 - s_j), \text{ se } s_0 = 0 \\ Z^{-1}\beta \prod_{j=1}^{n}(\mu_j[P_j]s_j + 1 - s_j), \text{ se } s_0 = 1 \end{cases}$$

con $\lambda_j = \frac{a_j}{b_j}$, $\mu_j = \frac{c_j}{d_j}$ e Z costante di normalizzazione.
Si stimano ora le probabilità p_0, p_1:

$$\begin{aligned} p_0 &= \sum_{s_1,\dots,s_n} \pi_{(0,s_1,\dots,s_n)} = Z^{-1}\beta \sum_{s_1,\dots,s_n} \prod_{j=1}^{n}(\mu_j[P_j]s_j + 1 - s_j) = Z^{-1}\beta \prod_{j=1}^{n}(\mu_j[P_j] + 1) \\ p_1 &= \sum_{s_1,\dots,s_n} \pi_{(1,s_1,\dots,s_n)} = Z^{-1}\alpha \sum_{s_1,\dots,s_n} \prod_{j=1}^{n}(\lambda_j[P_j]s_j + 1 - s_j) = Z^{-1}\alpha \prod_{j=1}^{n}(\lambda_j[P_j] + 1) \end{aligned} \tag{3.18}$$

Dal fatto che $p_0 + p_1 = 1$, si trova che $Z = \alpha \prod_{j=1}^{n}(\lambda_j[P_j] + 1) + \beta \prod_{j=1}^{n}(\mu_j[P_j] + 1)$. Si trova infine

$$k_{on,i} = \frac{k_1\alpha \prod_{j=1}^{n}(\lambda_j[P_j] + 1) + k_0\beta \prod_{j=1}^{n}(\mu_j[P_j] + 1)}{\alpha \prod_{j=1}^{n}(\lambda_j[P_j] + 1) + \beta \prod_{j=1}^{n}(\mu_j[P_j] + 1)} \tag{3.19}$$

Da questa formula si vede che k_{on} dipenderà dalla proteina P_j solo se $\lambda_j \neq \mu_j$, ovvero quando le reazioni che avvengono alla presenza di P_j hanno velocità sbilanciate, che tendono a favorire o configurazioni permissive ($\lambda_j > \mu_j$) o non permissive ($\lambda_j < \mu_j$).

Finora si è considerato che le P_j interagiscono come monomeri, vogliamo ora includere interazioni sotto forma di altri complessi. Se consideriamo che interagiscono con la cromatina dopo aver formato dimeri o altri complessi, e se queste reazioni di formazione dei complessi sono più veloci della dinamica della cromatina, possiamo tenerne conto sostituendo $[P_j]$ nell'equazione 3.19 con una funzione di $[P_j]$ che rappresenta la concentrazione quasi stazionaria del complesso. Si sceglie quindi di sostituire $[P_j]$ con $[P_j]^{m_j}$, con $m_j > 0$. Si nota che $m_j = 2, 3$ rappresenta in buona approssimazione le interazioni tra le P_j considerandole come dimeri o trimeri rispettivamente, ma in generale m_j non è necessariamente un numero intero.

Fino a questo momento il modello non considerava l'auto-interazione dei geni, per considerarla si pone $\lambda_i \neq \mu_i$ nell'equazione 3.19. Questo però porta a dei problemi di identificabilità: in tali condizioni, non è infatti possibile distinguere tra auto-attivazione, auto-inibizione e livelli basali. Per risolvere questo problema, consideriamo solo i casi di auto-attivazione ponendo $c_i = d_i = 0$ e tenendo solo gli stati rilevanti della cromatina ($C_I^* \forall I$ e C_I tale che $i \notin I$). Il sistema così descritto ha ancora un'unica distribuzione di probabilità alla stazionarietà, che porta alla formula 3.19 per k_{on} nel caso in cui $\mu_i = 0$. Si considera inoltre il fatto che l'auto-attivazione è rilevante solo quando il livello basale è sufficientemente piccolo fisicamente (perché sia possibile un comportamento bistabile), si prende quindi il limite $\alpha << 1$ tenendo fissato $\alpha\lambda_i$. La formula diventa quindi:

$$k_{on} = \frac{k_1\alpha\lambda_i[P_i]^{m_i} \prod_{j\neq i}^{n}(\lambda_j[P_j]^{m_j} + 1) + k_0\beta \prod_{j\neq i}^{n}(\mu_j[P_j]^{m_j} + 1)}{\alpha\lambda_i[P_i]^{m_i} \prod_{j\neq i}^{n}(\lambda_j[P_j]^{m_j} + 1) + \beta \prod_{j\neq i}^{n}(\mu_j[P_j]^{m_j} + 1)} \tag{3.20}$$

con $m_i > 0$ se il gene i può auto-attivarsi e $m_i = 0$ altrimenti
I parametri necessari per l'equazione 3.20 non sono ancora chiaramente interpretabili, per ottenere una forma ridotta si introduce la seguente parametrizzazione: $s_j = \mu_j^{-1/m_j}$, $\theta_j = \log(\lambda_j/\mu_j) \forall j \neq i$ e $s_i = (\beta/\alpha)^{1/m_i}$, $\theta_i = \log(\lambda_i)$. Questi nuovi parametri possono essere interpretati nel modo che segue: s_j può essere visto come soglia per l'influenza della proteina j sul gene i e θ_j caratterizza il tipo di influenza che ha la proteina j a seconda del suo segno (se $\theta_j = 0$ allora k_on non dipende dalla proteina j), s_i e θ_i rappresentano invece il comportamento basale e l'intensità dell'auto-attivazione. Introducendo infine la notazione $y_i = [P_i]$ e introducendo l'indice i per il gene che si sta considerando, la funzione $k_{on,i}$ diventa:

$$k_{on,i} = \frac{k_{0,i} + k_{1,i}\Phi_i(y)(y_i/s_{i,i})^{m_{i,i}}}{1 + \Phi_i(y)(y_i/s_{i,i})^{m_{i,i}}} \tag{3.21}$$

con

$$\Phi_i(y) = \exp(\theta_{i,i}) \prod_{j \neq i} \frac{1 + \exp(\theta_{i,j})(y_j/s_{i,j})^{m_{i,j}}}{1 + (y_j/s_{i,j})^{m_{i,j}}} \tag{3.22}$$

Abbiamo quindi trovato un'espressione dei rate $k_{on,i}$ dove dipendono dalle interazioni tra i geni e quindi dalle proteine.

Capitolo 4

Simulazioni

In questo capitolo si propongono delle simulazioni numeriche del modello per la regolazione genica che considera le interazioni tra i geni (descritto nella sezione 3.3). Tali simulazioni del modello PDMP sfruttano due algoritmi, l'algoritmo di Eulero per simulare la parte deterministica del modello (mRNA e proteine), e l'algoritmo di Gillespie per simulare quella stocastica (i geni) (sezione4.1). Mediante tali simulazioni si vogliono studiare diversi aspetti dell'espressione genica, quale la dinamica temporale dei promotori e dei livelli di mRNA e proteine, o l'influenza di un diverso numero di connessioni e con diverse intensità sull'evoluzione temporale del sistema. Verrà mostrato inoltre come per lo studio delle GRN sia necessario osservarne ogni aspetto, sia l'espressione genica dei singoli geni ma anche quella complessiva del sistema.

4.1 Algoritmo di Gillespie

L'algortimo di Gillespie (o metodo di Monte Carlo cinetico) è una procedura computazionale che serve a simulare traiettorie temporali di processi stocastici a tempo continuo per cui sono noti i rate di transizione [13]. Questo algoritmo è adatto a simulare reazioni chimiche in sistemi biologici, in particolare quando il numero di molecole coinvolte è limitato. In tali scenari, le fluttuazioni casuali hanno un impatto significativo sulla dinamica del sistema, rendendo inadeguati approcci deterministici convenzionali, basati su equazioni differenziali. L'algoritmo di Gillespie supera queste limitazioni, fornendo una rappresentazione accurata della variabilità intrinseca dei sistemi biologici stocastici a scala molecolare.

Consideriamo un sistema dove possono avvenire N reazioni distinte, di cui sono noti i rate di transizione. Lo scopo è calcolare la traiettoria temporale $\bar{n}(t)$ del sistema. Definiamo $P(\bar{n}, t|\bar{n}_0, t_0)$ come la probabilità che il sistema sia nello stato $\bar{n}$ al tempo t con condizioni iniziali $(\bar{n}_0, t_0)$. Ciò che è necessario sapere è (1) quando avverrà la prossima reazione e (2) quale reazione j sarà. Si introduce quindi $P(\tau, j|\bar{n}, t)$, ovvero la densità di probabilità congiunta che descrive la distribuzione del tempo τ fino alla prossima reazione e l'identità della reazione j. La probabilità che la reazione j avvenga nell'intervallo infinitesimale $[t+\tau, t+\tau+\Delta\tau)$ è data da $P(\tau, j|\bar{n}, t)\Delta\tau$, che si compone della probabilità che non si verifichi alcuna reazione nell'intervallo $[t, t+\tau)$ e della probabilità che la reazione j si verifichi nel susseguente intervallo infinitesimale. Si definisce poi $w_j(\bar{n})$ come la propensità della reazione j per il sistema nello stato $\bar{n}$. Tale quantità si calcola come prodotto tra il rate di una reazione per la concentrazione delle molecole reagenti coinvolte nella reazione. Infine, l'intervallo di tempo τ viene diviso in k parti tali che $\Delta t = \tau/k$. Si calcola ora $P(\tau, j|\bar{n}, t)$:

$$\begin{aligned} \lim_{k\to\infty} P(\tau, j|\bar{n}, t)\Delta\tau &= [1 - \sum_j w_j(\bar{n})\Delta t]^k w_j(\bar{n})\Delta\tau; \\ \lim_{k\to\infty} P(\tau, j|\bar{n}, t) &= [1 - \sum_j w_j(\bar{n})\frac{\tau}{k}]^k w_j(\bar{n}) = \exp(-\sum_j w_j(\bar{n})\tau) w_j(\bar{n}) \end{aligned} \tag{4.1}$$

Conseguentemente, la probabilità che una qualunque reazione avvenga nell'intervallo $[t+\tau, t+\tau+\Delta\tau)$ è:

$$P(\tau|\bar{n},t) = \sum_j P(\tau, j|\bar{n},t) = \sum_j w_j(\bar{n})\exp(-\sum_j w_j(\bar{n})\tau) \tag{4.2}$$

τ segue quindi una distribuzione esponenziale di media $1/w_0(\bar{n})$, con $w_0(\bar{n}) = \sum_j w_j(\bar{n})$. La probabilità che avvenga la reazione j è quindi:

$$\frac{P(\tau, j|\bar{n},t)}{P(\tau|\bar{n},t)} = \frac{w_j(\bar{n})}{w_0(\bar{n})} \tag{4.3}$$

Si ha quindi il seguente algoritmo:

1. Imporre le condizioni iniziali $\bar{n} = \bar{n}_0$, $t = t_0$;
2. calcolare $w_j(\bar{n})\forall j$ e $w_0(\bar{n})$;
3. calcolare τ da una distribuzione esponenziale con media $1/w_0(\bar{n})$: $\tau = \frac{1}{w_0(\bar{n})}\log(\frac{1}{r})$, con r numero casuale estratto da una distribuzione uniforme, $r \in (0,1)$;
4. scegliere quale reazione avviene con probabilità uniforme $w_j(\bar{n})/w_0(\bar{n})$;
5. aggiornare lo stato del sistema coerentemente con quanto ottenuto. Tornare al punto 2.

4.2 Simulazioni per GRN con modello PDMP

Facendo riferimento al modello per le interazioni esplicite nella GRN della sezione 3.3 si presentano le GRN simulateb in seguito.

Tramite queste simulazioni abbiamo studiato tre casi particolari di GRN: inizialmente si è simulato un sistema composto da 2 geni, denominato *toggle switch* [4] (in 4.3). La seconda simulazione considera una GRN composta da 4 geni (in 4.4), con l'obiettivo di replicare comportamenti di regolazione genica osservati nelle cellule pluripotenti durante le prime fasi del differenziamento [1]. Infine è stata eseguita la simulazione di una GRN composta da 6 geni, denominata GRN *gerarchica* (in 4.5). Con quest'ultimo modello si vuole indagare la regolazione genica in una rete in cui i geni hanno diversi gradi di interazione.

Ogni GRN è descritta dalla matrice delle interazioni θ, dove l'elemento $\theta_{i,j}$ rappresenta l'azione del gene j sul gene i. Si riportano i parametri comuni a tutte le simulazioni proposte in tabella 4.1 (quando diversi i parametri vengono riportati nella sezione apposita).

Parametri	Valori	Unità di misura
k_0	0.34	h^{-1}
k_1	2.15	h^{-1}
k_{off}	10	h^{-1}
$s_{i,j}$	0.01 $(i \neq j)$	proteine
$s_{i,i}$	0.095	proteine

Tabella 4.1: Parametri utilizzati nelle simulazioni. Si fa riferimento al lavoro di Herbach et al. in [5].

4.3 Simulazione di 2 geni - *toggle switch*

Nel modello *toggle switch* i geni presentano una tendenza reciproca a inibirsi, per questo viene implementata una matrice d'interazione tale che $\theta_{i,i} = 4$ e $\theta_{i,j} = -8$. I pesi delle interazioni sono presi dal paper di Herbach et al. in [5].

Per questa simulazione si utilizzano i parametri riportati nelle tabelle 4.1 e 4.2.

La figura 4.1 riporta i grafici relativi al periodo di attività genica, ai livelli di mRNA e dei livelli di proteine. Dall'analisi dei grafici relativi al periodo di attività del promotore emerge un cambio di stato asincrono tra i due geni, in linea con le previsioni basate sulla configurazione della simulazione. Questa

Parametri	Valori	Unità di misura
ν_0	1000	$\text{mRNA} * \text{h}^{-1}$
ν_1	10	$\text{protein} * \text{mRNA}^{-1} * \text{h}^{-1}$
d_0	0.7	h^{-1}
d_1	0.4	h^{-1}
$m_{i,j}$	2	-

Tabella 4.2: Parametri utilizzati nella simulazione del sistema a 4 geni.

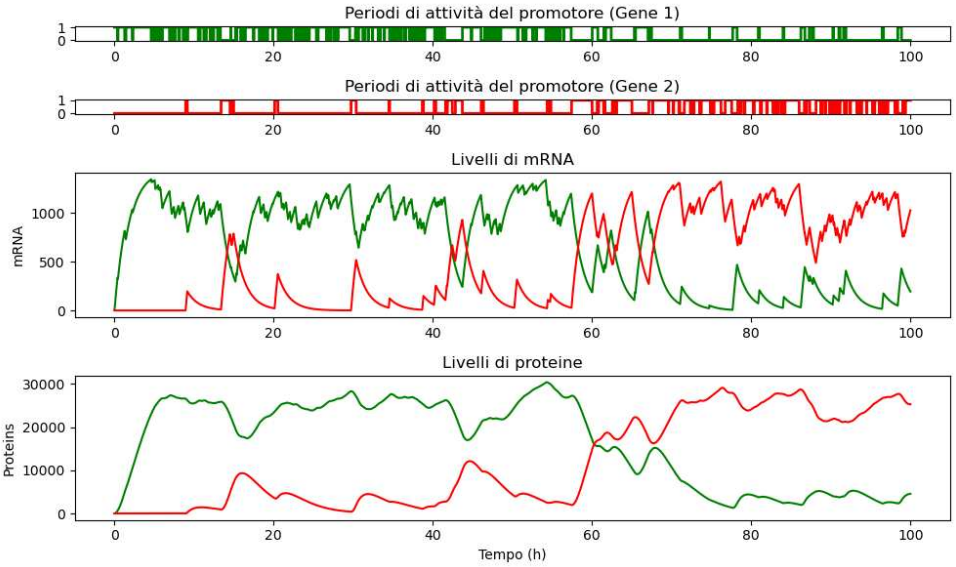


Figura 4.1: Grafici dell' attività dei geni e dei livelli di mRNA proteine nel tempo per il modello *toggle switch* .

dinamica si riflette nell'alternanza dei livelli di mRNA: quando sono alti per un gene, sono bassi per l'altro e viceversa, ciò riflette la loro tendenza a inibirsi reciprocamente.

Anche l'analisi delle distribuzioni di probabilità congiunte, in figura 4.2, evidenzia l'anticorrelazione tra il livello di mRNA (o di proteine) del primo gene e il livello di mRNA (o di proteine) del secondo: un incremento del livello di mRNA (o proteine) per un gene è accompagnato da una diminuzione nel corrispettivo gene, sottolineando nuovamente la relazione di proporzionalità inversa tra le espressioni geniche dei due componenti della GRN. Sono inoltre riportati i coefficienti di Pearson, che per mRNA e proteine valgono rispettivamente $\rho_{mRNA} = -0.961$ e $\rho_{proteine} = -0.983$, confermando l'anticorrelazione data la loro prossimità a -1.

4.4 Simulazione di 4 geni

In questa sezione si simula il comportamento genico osservato nelle prime fasi del differenziamento delle cellule embrionali pluripotenti [11]. In particolare prendiamo ad esempio il differenziamento di cellule pluripotenti murine. Da letteratura [1] è noto che queste cellule vengono guidate nel processo di differenziamento dall'attività di quattro gruppi di geni:

- I geni *naive* iniziano con alti livelli di mRNA che tendono a decrescere nel tempo;
- i geni *formative early* si caratterizzano per partire da livelli bassi o nulli di mRNA, che presentano una crescita elevata immediatamente successiva all'inizio della simulazione;
- i geni *formative late* presentano una dinamica simile ai geni *formative early*, differenziandosi per il momento d'inizio dell'incremento dei livelli di mRNA, che si verifica in una fase temporale successiva rispetto all'inizio della simulazione;
- i geni *other* tendono a mantenere un livello di mRNA stabile nel tempo .

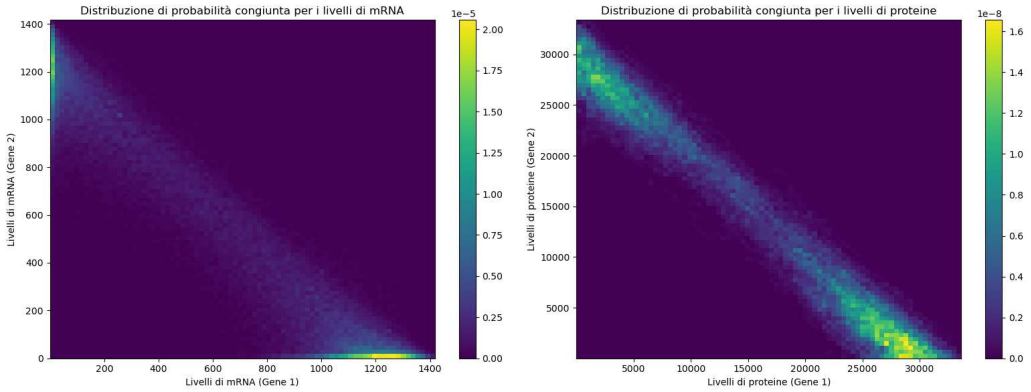


Figura 4.2: Distribuzione di probabilità congiunta dei livelli di mRNA (a sinistra) e dei livelli di proteine (a destra) per il modello *toggle switch*. I coefficiente di Pearson per mRNA e proteine sono rispettivamente $\rho_{mRNA} = -0.961$ e $\rho_{proteine} = -0.983$.

La GRN che abbiamo selezionato come rappresentativa di tale sistema biologico è descritta dalla seguente matrice delle interazioni:

$$\theta = \begin{pmatrix} 4 & -8 & -8 & -2 \\ -8 & 4 & 8 & 4 \\ -10 & 8 & 4 & -2 \\ 2 & -2 & 2 & 1 \end{pmatrix} \tag{4.4}$$

Da tale matrice si nota come tutte le tipologie di geni analizzate tendano ad auto-attivarsi. I geni *naive* tendono a inibire i geni *formative* e ad attivare gli *other*. I geni *formative* a loro volta tendono a spegnere i *naive*, tendono ad attivarsi tra loro e si differenziano tra *formative early* e *late* nell'interazione con gli *other*: gli *early* tendono ad inibire gli *other*, mentre i *late* tendono ad attivarli. Infine, gli *other* tendono ad attivare i *formative early* e ad inibire i *naive* e i *formative late*.

I parametri utilizzati in questa simulazione sono riportati nelle tabelle 4.1 e 4.3.

Parametri	Valori	Unità di misura
ν_0	1000	$\text{mRNA} * \text{h}^{-1}$
ν_1	5	$\text{protein} * \text{mRNA}^{-1} * \text{h}^{-1}$
d_0	0.5	h^{-1}
d_1	0.3	h^{-1}
$m_{i,j}$	2	-

Tabella 4.3: Parametri utilizzati nella simulazione del sistema a 4 geni.

Utilizzando tale matrice l'mRNA segue l'andamento previsto teoricamente, come si può vedere dal grafico 4.3.

I risultati ottenuti nella simulazione sono inoltre confrontabili con i dati sperimentali di RNA-seq bulk di cellule murine pluripotenti in differenziamento nel tempo. I dati sono relativi all'espressione genica di 4 geni, ognuno appartente a uno dei 4 gruppi simulati: il gene *Stat3* è il *naive*, il gene *Zic3* è il *formative early*, il gene *Lef1* è il *formative late* e il gene *Pou5f1* è l'*other*. Tali dati tratti dal lavoro di Carbognin et al. in [1] Questo confronto testimonia che il modello PDMP per GRN è una buona descrizione di un sistema biologico reale con una sua dinamica e delle sue interazioni specifiche.

4.5 Simulazione di una GRN *gerarchica* a 6 geni

In questa sezione l'obiettivo è simulare un modello di GRN composto da 6 geni, il cui comportamento si divide in 2 categorie:

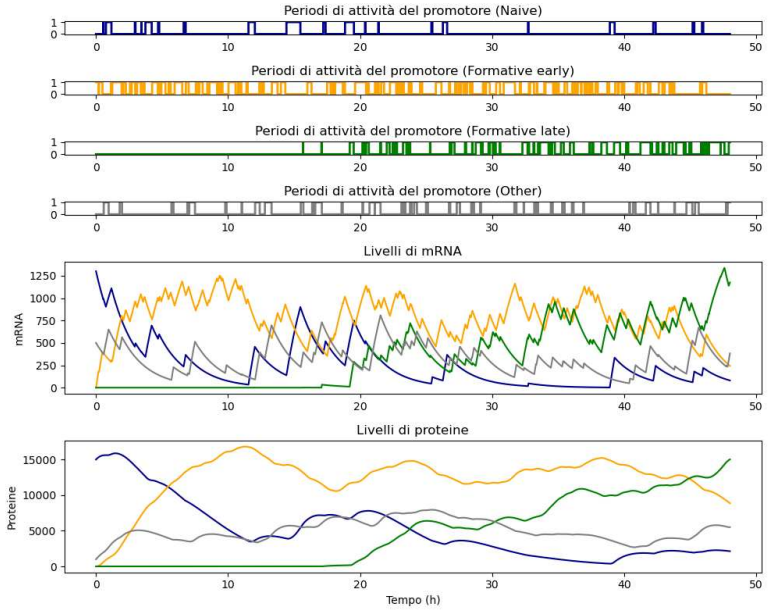


Figura 4.3: Grafici del periodo di attività dei geni, dei livelli di mRNA e dei livelli di proteine per il sistema a 4 geni. Il gene *naive* è di colore blu, il gene *formative early* è di colore arancione, il gene *formative late* è di colore verde e il gene *other* è di colore grigio.

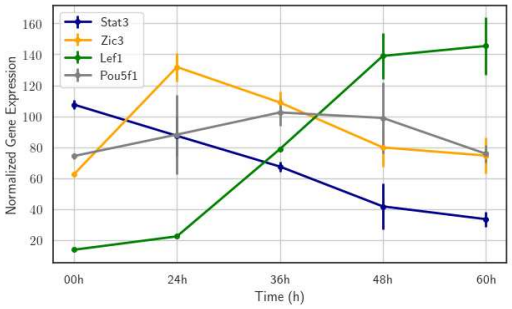


Figura 4.4: Grafico dell'espressione genica di 4 geni appartenenti ai quattro gruppi, *naive* (Stat3), *formative early* (Zic3), *formative late* (Lef1), *other* (Pou5f1). I dati sono tratti dal lavoro di Carbognin et al. in [1], l'errore sul valore medio è calcolato su 2 campioni per ogni gene ad ogni tempo.

- I geni *housekeeping*, che sono sempre attivi e presentano alti livelli di espressione genica. Essi codificano proteine fondamentali per le funzioni cellulari di base, che pertanto devono essere sempre presenti. Un esempio di tali proteine è l'*actina*.
- I geni *regolatori*, che si attivano raramente e tendono ad avere un basso livello di espressione genica. Le proteine che producono svolgono funzioni di regolazione, ad esempio agiscono come fattori di trascrizione.

La matrice d'interazione che rappresenta questa GRN è la seguente:

$$\theta = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ -1 & -2 & 3 & 0 & 0 & 0 \\ -1 & -2 & -3 & 4 & 0 & 0 \\ -1 & -2 & -3 & -4 & 5 & 0 \\ 1 & 2 & -3 & 4 & 5 & 6 \end{pmatrix} \tag{4.5}$$

I parametri utilizzati nella simulazione sono riportati nelle tabelle 4.1 e 4.4.

Dal grafico 4.5 si può notare come i geni 1, 2, 3 si attivino solo in alcuni momenti e con livelli di mRNA e proteine bassi se comparati agli altri tre geni per come li abbiamo definiti nella matrice

Parametri	Valori	Unità di misura
ν_0	100	$\text{mRNA} * \text{h}^{-1}$
ν_1	5	$\text{protein} * \text{mRNA}^{-1} * \text{h}^{-1}$
d_0	0.5	h^{-1}
d_1	0.5	h^{-1}
$m_{i,j}$	6	-

Tabella 4.4: Parametri utilizzati nella simulazione della GRN gerarchica a 6 geni.

delle interazioni θ, essi infatti rappresentano geni con funzioni di regolazione, mentre i geni 4, 5, 6 appartengono alla categoria dei geni *housekeeping*, infatti si attivano molto più frequentemente dei geni *regolatori* mantenendo un livello di espressione genica più elevato nel tempo.

In aggiunta, si presentano le distribuzioni dei livelli di mRNA (fig.4.6) e proteine (fig.4.7) per ciascun gene, oltre le distribuzioni che rappresentano i livelli di mRNA e proteine complessivi del sistema (fig.4.8). Per ogni distribuzione è stato fatto un fit di una distribuzione beta, prevista teoricamente nella sezione 2.5: si nota che i coefficienti r^2, riportati nelle didascalie dei grafici, sono prossimi all'unità, indicando la validità della previsione teorica della distribuzione beta per modellare l'espressione genica del sistema. Dall'analisi degli istogrammi emerge chiaramente che l'espressione genica aumenta all'aumentare dell'indice dei geni, in linea con le tipologie di geni scelte per questo modello. Inoltre, dal confronto delle distribuzioni totali e quelle dei singoli geni, si nota come le prime non siano sufficienti a caratterizzare il sistema. Se tutti i geni si comportassero allo stesso modo le distribuzioni totali sarebbero sufficienti, tuttavia, poiché ogni gene ha un suo ruolo specifico, ciò non è verificato. Allo stesso modo, i singoli geni non possono essere considerati come unità indipendenti, poiché le loro proprietà dipendono dalle interazioni nella GRN.

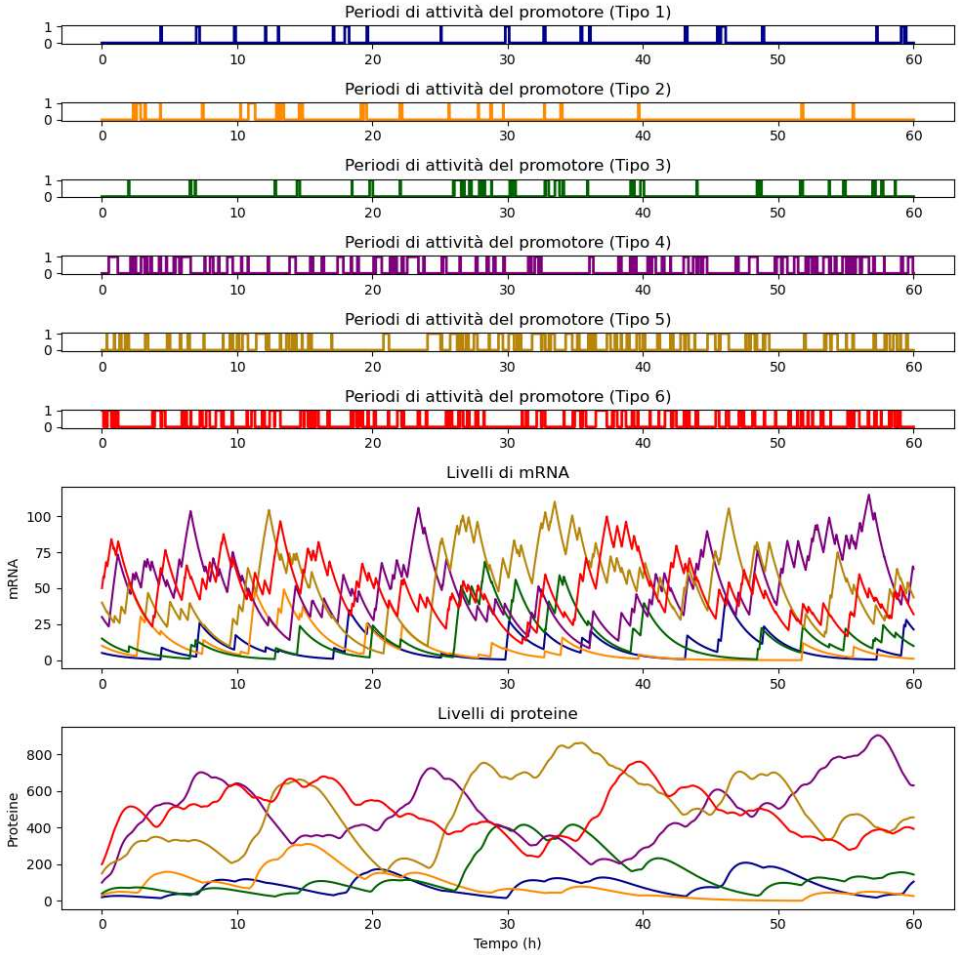


Figura 4.5: Grafico del periodo di attivtà dei geni, dei livelli di mRNA e dei livelli di proteine nella GRN gerarchica. I *regolatori* sono i geni 1, 2, 3 rispettivamente di colore blu, arancione e verde. I geni 4, 5, 6 sono gli *housekeeping*, rispettivamente di colore viola, marrone e rosso.

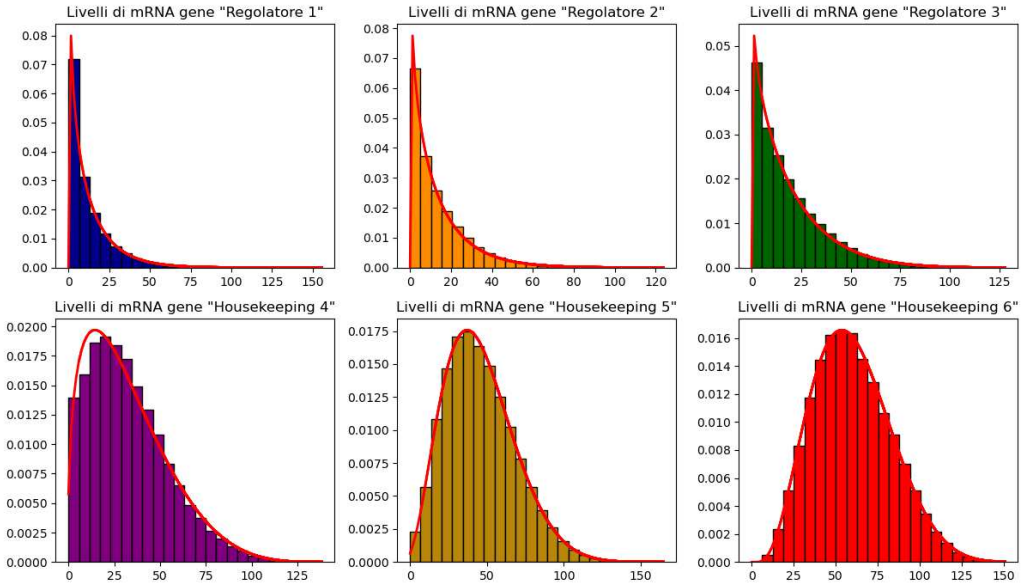


Figura 4.6: Distribuzioni dei livelli di mRNA e relativo fit con distribuzione beta (in rosso) per ogni gene della GRN gerarchica. Valori di r^2 relativi a ogni fit: $r_1^2 = 0.975$, $r_2^2 = 0.996$, $r_3^2 = 0.998$, $r_4^2 = 0.988$, $r_5^2 = 0.999$, $r_6^2 = 0.999$.

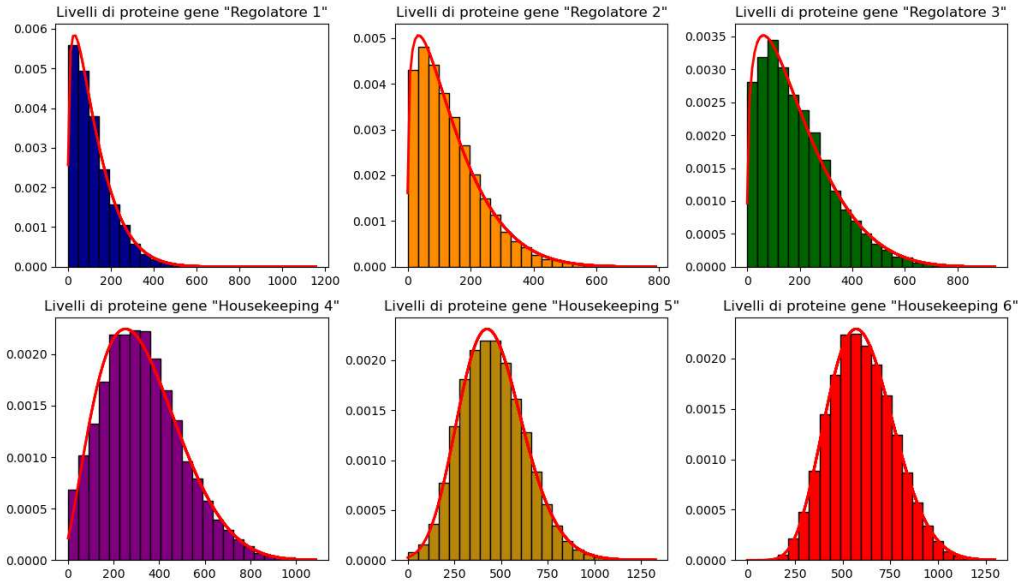


Figura 4.7: Distribuzioni dei livelli di proteine e relativo fit con distribuzione beta (in rosso) per ogni gene della GRN gerarchica. Valori di r^2 relativi a ogni fit: $r_1^2 = 0.999$, $r_2^2 = 0.995$, $r_3^2 = 0.993$, $r_4^2 = 0.986$, $r_5^2 = 0.998$, $r_6^2 = 0.999$.

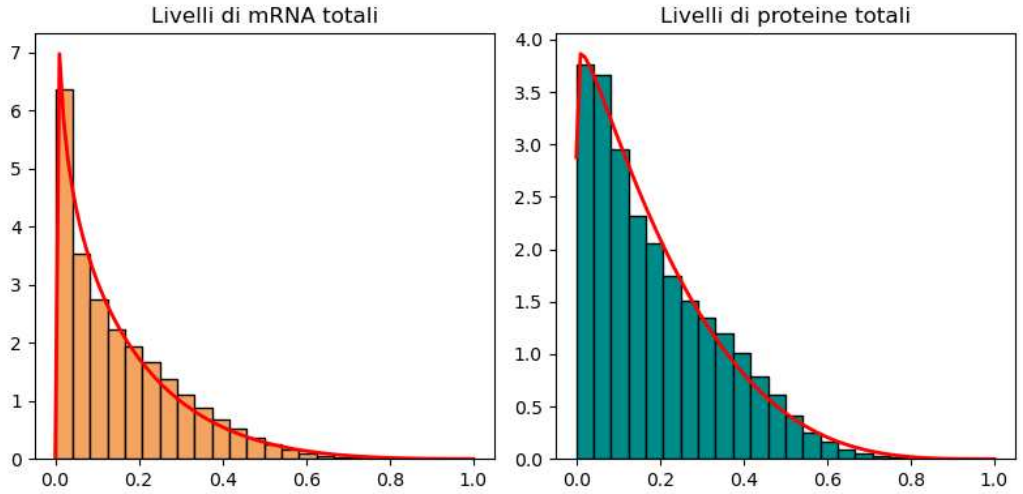


Figura 4.8: Distribuzioni e relativo fit con distribuzione beta (in rosso) dei livelli di mRNA e proteine totali del sistema. Valori di r^2 relativi a ogni fit: $r^2_{mRNA} = 0.986$, $r^2_{proteine} = 0.992$. Per questi grafici i dati sono stati normalizzati.

Capitolo 5

Conclusioni

L'obiettivo di questo elaborato è lo studio della dinamica dell'espressione genica all'interno della cellula attraverso un approccio interdisciplinare che integra modelli fisici per decifrare questo complesso sistema biologico. Per fare ciò, si sono studiati modelli meccanicistici includendo dettagli biologici per descrivere il comportamento prima di singoli geni (capitolo 2), poi di gruppi di geni organizzati in reti di regolazione genica, GRN (capitolo 3).

Particolare attenzione è stata posta nello studio del modello PDMP per GRN 3.3: esso infatti permette un'accurata descrizione meccanicistica del fenomeno di trascrizione e traduzione di ogni gene nella GRN, con una dinamica deterministica per mRNA e proteine, ed una stocastica per i geni. Ciò rende questa descrizione una buona approssimazione alla dinamica biologica che si vuole descrivere e risulta efficiente dal punto di vista computazionale.

Per indagare le capacità del modello di descrivere sistemici biologici e le caratteristiche dell'espressione genica di singola cellula, sono state simulate 3 diverse GRN note. Per studiare l'influenza di interazioni specifiche sul sistema è stata simulata la GRN *toggle switch* 4.3, grazie al quale è stato possibile vedere come l'azione inibitoria di un gene sull'altro, e viceversa, influisce sulla dinamica di entrambi i geni. Successivamente l'implementazione del sistema a 4 geni ha permesso di studiare la capacità dell'approccio di riprodurre andamenti di espressione genica noti sperimentalmente. Infine, si è simulata una GRN *gerarchica* composta da 6 geni. Grazie a tale rete abbiamo osservato i diversi comportamenti dei geni all'interno di tale GRN e il loro impatto sulla frequenza di attivazione dei geni stessi, e sulla loro produzione di proteine ed mRNA. Simulando tale GRN, si è studiato come diversi numeri di connessioni influiscano sul comportamento del singolo gene e sui geni a lui connessi. In quest'ultima simulazione, confrontando le distribuzioni dei livelli di mRNA e proteine dei singoli geni con quelle complessive del sistema, si è vista l'importanza di analizzare il sistema in ogni suo aspetto: i geni non possono essere considerati come singole unità, poiché le loro proprietà dipendono anche dalle loro interazioni nella GRN, tuttavia, le proprietà della GRN non possono essere descritte solo dalle distribuzioni totali del sistema, poiché esse non forniscono alcuna informazione sui diversi contributi di ogni gene.

La cellula è un sistema il cui funzionamento non può essere descritto nella sua totalità mediante la descrizione dei sui singolo elementi come unità isolate. Essa va considerata come un sistema complesso: infatti è solo attraverso l'analisi delle interazioni tra i suoi elementi costitutivi che è possibile comprendere la sua dinamica, che emerge dalle interazioni degli elementi che la compongono.

In conclusione, un approccio interdisciplinare tra biologia e fisica in quest'ambito è fondamentale: modelli fisici diversi permettono l'inclusione di dettagli biologici diversi, cercando di descrivere meccanicisticamente il funzionamento del sistema sfruttando il formalismo matematico e i modelli teorici propri della fisica. Questi modelli possono permettere di rivelare meccanismi biologici precedentemente non osservati o non compresi, mostrando una prospettiva che va oltre i limiti tradizionali di un approccio esclusivamente biologico, coniugando le informazioni sperimentali con i modelli teorici.

Bibliografia

[1] Elena Carbognin, Valentina Carlini, Francesco Panariello, Martina Chieregato, Elena Guerzoni, Davide Benvegnù, Valentina Perrera, Cristina Malucelli, Marcella Cesana, Antonio Grimaldi, et al. Esrrb guides naive pluripotent cells through the formative transcriptional programme. *Nature cell biology*, 25(5):643–657, 2023.

[2] Sara-Jane Dunn, Meng Amy Li, Elena Carbognin, Austin Smith, and Graziano Martello. A common molecular logic determines embryonic stem cell self-renewal and reprogramming. *The EMBO journal*, 38(1):e100003, 2019.

[3] Takashi Fukaya, Bomyi Lim, and Michael Levine. Enhancer control of transcriptional bursting. *Cell*, 166(2):358–368, 2016.

[4] Timothy S Gardner, Charles R Cantor, and James J Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.

[5] Ulysse Herbach, Arnaud Bonnaffoux, Thibault Espinasse, and Olivier Gandrillon. Inferring gene regulatory networks from single-cell data: a mechanistic approach. *BMC systems biology*, 11(1):1–15, 2017.

[6] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

[7] M Moshinsky. How good is the hartree-fock approximation. *American Journal of Physics*, 36(1):52–53, 1968.

[8] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

[9] Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription factors modulate c-fos transcriptional bursts. *Cell reports*, 8(1):75–83, 2014.

[10] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[11] Austin Smith. Formative pluripotency: the executive phase in a developmental continuum. *Development*, 144(3):365–373, 2017.

[12] Fuchou Tang, Catalin Barbacioru, Ellen Nordman, Bin Li, Nanlan Xu, Vladimir I Bashkirov, Kaiqin Lao, and M Azim Surani. Rna-seq analysis to capture the transcriptome landscape of a single cell. *Nature protocols*, 5(3):516–535, 2010.

[13] Raúl Toral and Pere Colet. *Stochastic numerical methods: an introduction for students and scientists*. John Wiley & Sons, 2014.

[14] José Viñuelas, Gaël Kaneko, Antoine Coulon, Elodie Vallin, Valérie Morin, Camila Mejia-Pous, Jean-Jacques Kupiec, Guillaume Beslon, and Olivier Gandrillon. Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts. *BMC biology*, 11(1):1–19, 2013.

[15] Ayako Yachie-Kinoshita, Kento Onishi, Joel Ostblom, Matthew A Langley, Eszter Posfai, Janet Rossant, and Peter W Zandstra. Modeling signaling-dependent pluripotency with boolean logic to predict cell fate transitions. *Molecular systems biology*, 14(1):e7952, 2018.