SAPIENZA
UNIVERSITÀ DI ROMA

Analisi statistica di una rete metabolica: gli effetti di un *knockout* in *E.coli*

Facolta di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea Magistrale in Fisica

Candidato

Giulio Alessandrini
Matricola 1354389

Relatore

Professor Enzo Marinari

Correlatore

Dottor Matteo Figliuzzi

Anno Accademico 2011-2012

Analisi statistica di una rete metabolica:gli effetti di un *knockout* in *E.coli*
Tesi di Laurea Magistrale. Sapienza – Università di Roma

Questa tesi è stata composta con LaTeX e la classe Sapthesis.

Versione: 13 marzo 2012

Email dell'autore: giulio.a@gmail.com

Analisi statistica di una rete metabolica: gli effetti di un *knockout* in *E.coli*

Giulio Alessandrini

26 aprile 2012

Ai miei amici

Indice

Elenco delle figure

Parte I

Introduzione

Lo scopo di questa tesi è analizzare la rete metabolica del batterio *Escherichia coli* utilizzando un approccio statistico che permetta di avere un quadro complessivo del comportamento della rete e della sua risposta a una perturbazione.

Le reti metaboliche racchiudono l'insieme delle reazioni chimiche che avvengono in una cellula; anche nei casi più semplici si tratta di centinaia di processi che coinvolgono altrettante componenti, i metaboliti cellulari. Capire come è fatto e come funziona il metabolismo non è quindi semplice, ma è un passo fondamentale per spiegare il funzionamento delle cellule, che sono il mattone fondamentale degli organismi complessi.

La ricostruzione di una rete metabolica si basa sull'integrazioni di informazioni provenienti da fonti diverse: il sequenziamento del genoma; lo studio delle proteine, degli RNA e degli enzimi; le misure di concentrazione dei metaboliti all'interno della cellula [16]. I modelli di metabolismo cellulare si basano su questi dati ma allo stesso tempo forniscono un'infrastruttura in cui interpretarli in maniera unitaria. Per verificarne la correttezza e studiarne le proprietà è necessario un sistema per simularne il funzionamento ed è in quest'ottica che nascono i diversi metodi matematici per lo studio dei flussi metabolici (cioè le velocità delle reazioni chimiche) basati sull'utilizzo dei coefficienti stechiometrici. La *flux balance analisys* (FBA) è stata introdotta negli anni '80 [7, 14] e si è sviluppata fino a diventare uno dei metodi principali in questo campo di studi [13, 23]. Tuttavia, sebbene consenta di fare a meno di misure di concentrazioni e coefficienti cinetici, conduce a sistemi di equazioni sottodeterminati, che devono essere vincolati attraverso la misura diretta di alcuni flussi. Per ottenere una soluzione univoca è richiesta inoltre l'imposizione dall'esterno di una funzione obiettivo che viene massimizzata con metodi di programmazione lineare.

Metodi come la FBA sono molto potenti, ma richiedono condizioni e assunzioni che vanno oltre la struttura della rete. Un approccio alternativo si basa sull'adattamento alle reti metaboliche del problema di Von Neumann. Il modello proposto da Von Neumann è particolarmente semplice e permette di trovare il tasso di espansione globale massimo di un sistema economico dato un insieme di beni e processi produttivi senza richiedere nulla più che la struttura del sistema e una serie di assunzioni minimali; inoltre fornisce i valori dell'intensità di ogni processo e del costo di ogni bene [24].
Utilizzando versione ridotta del modello in cui i metaboliti e

le reazioni chimiche prendono il posto dei beni e dei processi di produzione è stato trovato che il tasso di espansione massimo previsto dal il modello sulla rete metabolica di *Escherichia coli* è compatibile con lo stato stazionario [12]. Inoltre in questa situazione è presente un intero *ensemble* di configurazioni che risolvono il modello, a differenza di quanto succede per grafi random [4] o per le soluzioni di FBA. Il modello è stato applicato e risolto tramite esplorazione diretta per trovare gli stati ottimali del metabolismo del globulo rosso [3]. Tramite l'algoritmo numerico `minover` [4, 11] è stato analizzato lo spazio delle soluzioni relativo al metabolismo di *E.coli* [2] e più in generale le proprietà statistiche del modello [8].
Si tratta quindi di un metodo che è già stato applicato con successo nel settore di studio a cui siamo interessati e può essere utilizzato come base di lavoro; allo stesso tempo è relativamente giovane e lascia aperti molti spazi di approfondimento.

In questa tesi ci proponiamo di analizzare le proprietà delle soluzioni del modello di Von Neumann ed il comportamento dell'algoritmo `minover` con un maggior livello di precisione. A tal fine applicheremo tale algoritmo in topologie di reti artificiali particolarmente semplici da permettere una caratterizzazione analitica dello spazio delle soluzioni. Lo scopo principale di questa sezione sarà sviluppare una conoscenza più approfondita del modello, controllare la corrispondenza delle soluzioni numeriche con quelle analitiche e approfondire le capacità dell'algoritmo nell'esplorazione dello spazio delle soluzioni, caratteristica importante se vogliamo eseguire un'analisi statistica sui dati.

La seconda parte del lavoro verrà dedicata allo studio del metabolismo centrale di *E.coli*. Cercheremo di caratterizzare lo spazio delle soluzioni e di confrontare i valori dei flussi predetti con misure sperimentali. Per effettuare questo confronto abbiamo deciso di simulare una perturbazione della rete dovuta all'inibizione di una specifica reazione, la piruvato chinasi, su cui è presente una nutrita letteratura (ad esempio [1, 10, 15, 19, 20]). La ristrutturazione dei flussi in seguito a questo *knockout* è particolarmente importante perché interessa una sezione chiave: il metabolismo del carbonio. La nostra analisi si focalizzerà sullo studio dei rapporti tra i flussi del batterio sano e quello in cui una mutazione genetica inibisce la reazione.
Tenteremo inoltre un raffronto diretto dei valori assoluti previsti dal modello, tenendo presente che i dati sperimentali sono stati rielaborati utilizzano sotto-reti limitate al metabolismo

centrale mentre noi stiamo considerando l'intera rete metabolica del batterio. In questo modo ci proponiamo di capire entro quali limiti e in quali situazioni il nostro modello e le nostre assunzioni minimali riescano ad essere predittive sulle reali distribuzioni dei flussi.

Capitolo 1

Metabolismo cellulare

Un organismo vivente è in grado di utilizzare fonti di energia esterne per manipolare gli elementi al suo interno, producendo ciò che gli serve per sopravvivere, conservarsi e, in ultima istanza, riprodursi. Il complesso di reazioni chimiche che svolgono queste funzioni è chiamato *metabolismo*. Prima di intraprendere una descrizione del metabolismo cellulare riteniamo utile richiamare brevemente alcuni concetti di chimica necessari per trattare l'argomento.

1.1 Le reazioni chimiche nella cellula

Una reazione chimica è un processo che trasforma un gruppo di uno o più reagenti, gli elementi o i composti presenti nello stato iniziale, in un nuovo insieme di sostanze, i prodotti. Questa trasformazione è dovuta alle interazioni tra gli elettroni degli orbitali più esterni degli atomi coinvolti. Le reazioni chimiche sono quindi caratterizzate da scale di energia dell'ordine dell'elettronvolt, non abbastanza perché nella trasformazione si possa modificare la natura degli atomi: gli atomi presenti all'inizio di una reazione sono gli stessi dello stato finale. Questo comporta che né la massa totale né la carica totale possano variare nel corso di una reazione chimica.

Prendiamo come esempio il caso di una reazione che produce le sostanze C e D a partire da A e B. Questa si può scrivere in forma di equazione utilizzando come variabili le concentrazioni delle sostanze espresse in unità di misura opportune. Nel nostro esempio scriveremo

$$\mathrm{a[A]} + \mathrm{b[B]} \longrightarrow \mathrm{c[C]} + \mathrm{d[D]}.$$

In questa formula la freccia indica la direzione della reazione mentre a, b, c e d sono dei *coefficienti stechiometrici* che assicurano il bilanciamento delle masse e delle cariche tra i due lati dell'equazione.

La direzione della freccia dipende invece dal bilancio energetico ed entropico della reazione, è quindi determinata dalle leggi della termodinamica. All'interno di una cellula sana la pressione viene mantenuta costante e le variazioni di temperatura sono modeste. Il potenziale termodinamico che guida la transizione è quindi l'energia libera (G): una reazione procede spontaneamente solo se

$$\Delta G = \Delta H - T\Delta S < 0.$$

I vari addendi indicano la differenza di energia (ΔH) e di entropia (ΔG) e la temperatura (T) del sistema.
Esiste quindi una distinzione tra le reazioni che possono avvenire in entrambi i sensi, dette reversibili, e quelle che procedono lungo una direzione sola. Per quanto appena detto, affinché una reazione sia reversibile ΔG deve poter assumere sia valori positivi sia valori negativi. Nel caso limite in cui $\Delta G = 0$ le due reazioni, diretta e inversa, procedono alla stessa velocità e le concentrazioni non variano nel tempo. Questa condizione si definisce di *equilibrio chimico*.

Il segno di ΔG ci dà informazioni sulla direzione della reazione ma non permette di quantificare la velocità a cui avviene. Questa quantità (chiamata anche flusso della reazione) è influenzata da molti fattori, alcuni esterni: come temperatura o pressione; altri legati direttamente alle specie chimiche coinvolte:

- fattori geometrici, come l'area di contatto tra i reagenti o una particolare orientazione spaziale necessaria per la reazione,
- fattori energetici, alcune reazioni a $\Delta G < 0$ necessitano infatti di un *energia di attivazione* per superare una barriera di potenziale prima di procedere spontaneamente.

Altre sostanze che non prendono parte alla reazione possono agire su questi fattori favorendo o inibendo la reazione, queste sostanze sono chiamate *catalizzatori*. Dato che il tempo medio per superare una barriera di potenziale E è $O(e^{\beta E})$, la presenza di un catalizzatore puù aumentare o diminuire la velocità di reazione di diversi ordini di grandezza.

Nei sistemi biologi il ruolo dei catalizzatori è assunto da particolari proteine chiamate *enzimi*. La loro presenza permette di controllare in modo molto preciso le concentrazioni di tutti i composti presenti all'interno della cellula e di portare velocemente le concentrazioni verso uno stato di equilibro. Uno degli esempi più semplici di reazione catalizzata da un enzima è la cinetica di Michaelis-Menten, che può essere rappresentata schematicamente come

$$[\mathrm{S}] + [\mathrm{E}] \underset{k_-}{\overset{k_+}{\rightleftharpoons}} [\mathrm{SE}] \xrightarrow{k_{cat}} [\mathrm{P}] + [\mathrm{E}]. \tag{1.1}$$

Lo stato iniziale, chiamato substrato, si lega all'enzima tramite un processo reversibile formando lo stato intermedio [SE]: questo complesso viene quindi convertito in un prodotto lasciando libero l'enzima.

Ad ogni processo nello schema precedente corrisponde un coefficiente che ne quantifica la velocità. I parametri k_+ e k_- possono essere calcolati a partire dalla legge di Arrhenius

$$k = Ae^{-E_a/k_BT},$$

che dipende dal rapporto tra l'energia di attivazione E_a e la scala di energia a cui avviene la reazione k_BT. Il terzo parametro, k_{cat} è legato al numero di molecole di substrato che l'enzima riesce a convertire per unità di tempo.[1]

Se facciamo l'ipotesi che la velocità di una reazione sia proporzionale al prodotto delle concentrazioni del reagenti (legge di azione di massa per le reazioni del primo ordine), è possibile costruire il schema di equazioni differenziali

$$\begin{cases} \dfrac{d[\mathrm{S}]}{dt} = -k_+[\mathrm{S}][\mathrm{E}] + k_-[\mathrm{SE}], \\ \dfrac{d[\mathrm{E}]}{dt} = -k_+[\mathrm{S}][\mathrm{E}] + k_-[\mathrm{SE}] + k_{cat}[\mathrm{SE}], \\ \dfrac{d[\mathrm{SE}]}{dt} = k_+[\mathrm{S}][\mathrm{E}] - k_-[\mathrm{SE}] - k_{cat}[\mathrm{SE}], \\ \dfrac{d[\mathrm{P}]}{dt} = k_{cat}[\mathrm{SE}], \end{cases} \tag{1.2}$$

che fornisce la velocità a cui variano le concentrazioni di prodotto e substrato. Sommando la seconda e terza reazione si

[1] Nella maggior parte degli studi di cinetica chimica le variazioni di temperatura sono abbastanza contenute per poter ipotizzare che E_a sia indipendente da T, similmente la dipendenza da T del fattore pre-esponenziale A è trascurabile rispetto all'esponenziale.

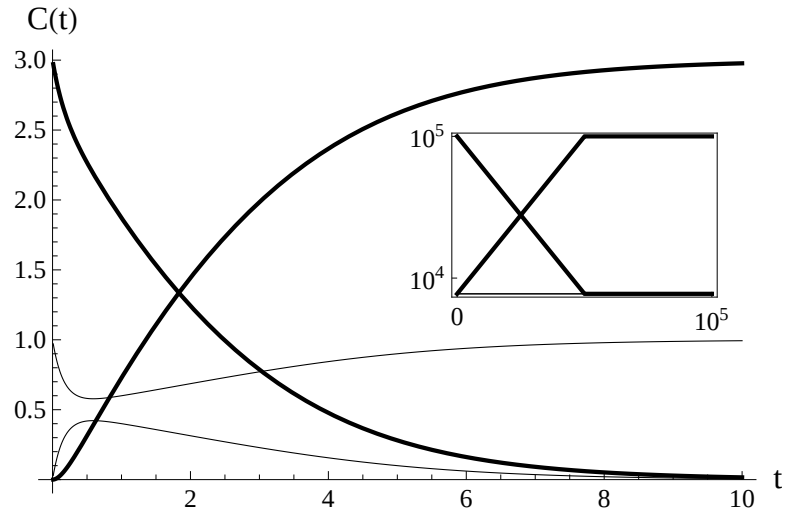


Figura 1.1: Concentrazioni delle specie chimiche in funzione del tempo, soluzioni dal sistema 1.2. Le linee spesse rappresentano il substrato e il prodotto ($[\mathrm{S}]_0 = 3$, $[\mathrm{P}]_0 = 0$); le linee sottili l'enzima ed il complesso enzima-substrato ($[\mathrm{E}]_0 = 1$, $[\mathrm{SE}]_0 = 0$). I coefficienti cinetici sono $k_+ = k_+ = k_{cat} = 1$.

ottiene

$$\frac{d[\mathrm{E}]}{dt} + \frac{d[\mathrm{SE}]}{dt} = 0 \quad \Rightarrow \quad [\mathrm{E}] + [\mathrm{SE}] = [\mathrm{E}]_0,$$

che è la la legge di conservazione della massa dell'enzima.

Il sistema 1.2, però, non è integrabile. Questo fatto ha spinto Michaelis e Menten ad introdurre l'approssimazione che il substrato e il complesso siano in equilibrio chimico,

$$k_+[\mathrm{S}][\mathrm{E}] = k_-[\mathrm{SE}].$$

Ciò avviene quando il substrato raggiunge l'equilibrio molto più rapidamente del tasso di creazione del prodotto ($k_{cat}/k_- \ll 1$). In alternativa si può assumere che la concentrazione del complesso non vari durante il processo,

$$k_+[\mathrm{S}][\mathrm{E}] = k_-[\mathrm{SE}] + k_{cat}[\mathrm{SE}].$$

Questa approssimazione, detta di stato quasi-stazionario, è valida quando la concentrazione dell'enzima è molto minore di quella del substrato.

In entrambi i casi troviamo una formula per la velocità della reazione che dipende direttamente dalle concentrazioni [S]

ed $[\mathrm{E}]_0$.

$$\frac{d[\mathrm{P}]}{dt} = k_{cat}[\mathrm{E}]_0 \frac{[\mathrm{S}]}{K + [\mathrm{S}]}, \quad K = \begin{cases} \frac{k_-}{k_+} & \text{equilibro,} \\ \frac{k_- + k_{cat}}{k_+} & \text{stato quasi-staz.} \end{cases} \tag{1.3}$$

L'approssimazione di stato quasi-stazionario è usata molto spesso poiché tipicamente la concentrazione dell'enzima è molti ordini di grandezza minore di quella del substrato (nel riquadro piccolo della figura 1.1 $[\mathrm{S}]/[\mathrm{E}] \sim 10^5$).

1.2 Il metabolismo

Il metabolismo cellulare è composto da un grandissimo numero di reazioni simili a quelle descritte nel paragrafo precedente. L'insieme di queste reazioni serve a utilizzare la materia che entra nella cellula per sostenerne la vita e permetterne la riproduzione. Questa funzione è svolta attraverso due fasi (figura 1.2):

- il **catabolismo** raggruppa le reazioni che demoliscono i nutrienti assorbiti dalla cellula e trasferiscono parte dell'energia liberata ad un complesso di molecole utilizzate come serbatoio – sono i metaboliti di scambio (*currencies*) come ATP o NADH;

- nell'**anabolismo** queste molecole vengono utilizzate per alimentare le reazioni con cui vengono prodotte le macromolecole necessarie alla cellula.[2]

I due gruppi di reazioni non sono indipendenti. Infatti, oltre ai metaboliti di scambio, molti prodotti finali del catabolismo vengono utilizzati come mattoni per la costruire le macromolecole necessarie.

Questo complesso di reazioni segue una dinamica molto complicata, che dipende dalle concentrazioni di substrati ed enzimi che a loro volta sono connesse alle velocità delle reazioni in un complesso sistema di *feedback*. Da quanto detto è chiaro che avere un quadro globale dei i flussi metabolici è di grande aiuto per comprendere il funzionamento e le proprietà di una cellula. In questo senso l'organizzazione dei flussi

[2] I metaboliti sono composti chimici creati come prodotto intermedio del metabolismo. Solitamente questo termine è riservato a molecole piccole.

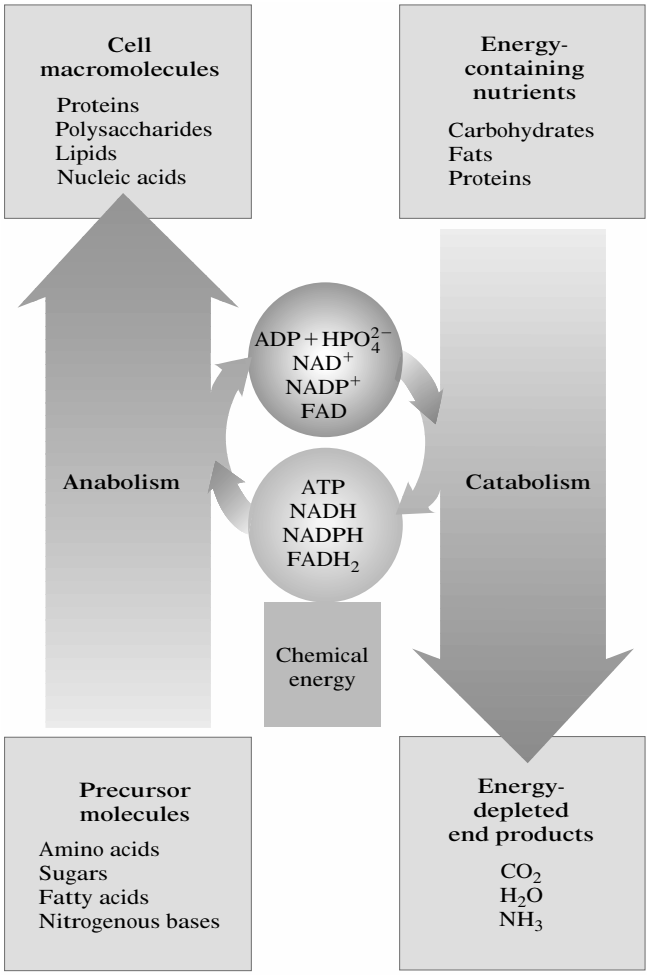


Figura 1.2: Relazione tra anabolismo e catabolismo.

può essere considerata come una rappresentazione del fenotipo cellulare, risultato di molti e differenti processi tra tutte le specie chimiche presenti.

Metabolismo del carbonio

Alcune sezioni del metabolismo hanno un'importanza particolare, sia perché sono passaggi chiave nell'approvigionamento energetico delle cellule, sia perché i flussi delle reazione comprese in queste sezioni possono essere misurati tracciando i gruppi di atomi di carbonio lungo le loro trasformazioni (capitolo 2).

Glicolisi Una delle vie principali del metabolismo del carbonio è la conversione del glucosio in piruvato, la *glicolisi*. Questo processo, che può avvenire solo in presenza di ossigeno, comporta la degradazione del glucosio (composto da sei atomi di carbonio) in due molecole di piruvato secondo la reazione

$$\begin{aligned} \mathrm{C_6H_{12}O_6} + 2\,[\mathrm{NAD}]^+ + 2\,[\mathrm{ADP}] + 2\,[\mathrm{P}]_\mathrm{i} \longrightarrow \\ \longrightarrow 2\,\mathrm{CH_3COCOO^-} + 2\,[\mathrm{NADH}] + 2\,\mathrm{H^+} + 2\,[\mathrm{ATP}] + 2\,\mathrm{H_2O}. \end{aligned}$$

Questa formula non rappresenta una reazione reale bensì la somma di tutte le trasformazioni che avvengono lungo la via.

- In una prima fase preparatoria il glucosio viene convertito in sequenza in glucosio-6-fosfato (G6P), fruttosio-6-fosfato (F6P), fruttosio-1,6-bifosfato (FDP) e gliceraldeide-3-fosfato (G3P) al costo di due molecole di ATP.

- La seconda fase, o fase di rendimento, procede dall'ossidazione della G3P attraverso l'1,3-bisfosfoglicerato (13DPG), il 3-fosfoglicerato (3PG) e il 2-fosfoglicerato (2PG) per terminare nella fosforilazione del fosfenolpiruvato in piruvato (PEP $\longrightarrow$ PYR). Questa serie di reazioni comporta la creazione di quattro molecole di ATP e la riduzione di due molecole di NAD^+ in NADH.

Il guadagno netto della glicolisi è quindi di due molecole di ATP e due molecole di NADH associate alla produzione di due molecole di piruvato che può essere utilizzato per produrre l'amminoacido alanina o, negli organismi eucarioti, portato all'interno di un mitocondrio.

Ciclo di Krebs Nei mitocondri avviene la decarbossilazione ossidativa del piruvato che ne permette la conversione in acetil-CoA. Quest'ultimo è un elemento essenziale del *ciclo di Krebs* (chiamato anche ciclo degli acidi tricarbossilici o ciclo dell'acido citrico), che ha un'importanza fondamentale in tutte le cellule che utilizzano ossigeno nel processo della respirazione cellulare.[3]

Il ciclo di Krebs è l'anello di congiunzione delle vie metaboliche responsabili della degradazione (catabolismo) dei carboidrati, dei grassi e delle proteine in anidride carbonica e acqua con la formazione di energia chimica. Inoltre fornisce molti precursori per la produzione di alcuni amminoacidi e di altre molecole fondamentali per la cellula.

Schematicamente il ciclo si compone delle seguenti fasi:

- Comincia con il trasferimento di un gruppo acetile a due atomi di carbonio dall'acetil-CoA al composto a quattro C ossalacetato (OAA) per la formazione di un composto a sei C (acido citrico).

- L'acido citrico passa attraverso una serie di trasformazioni chimiche ($\longrightarrow$ succinato $\longrightarrow$ fumarato $\longrightarrow$ malato) che comportano la perdita di due atomi di carbonio sotto forma di CO_2.

- La maggior parte dell'energia resa disponibile dai passi ossidativi del ciclo viene trasferita sotto forma di elettroni energetici al gruppo NAD^+ per formare NADH. Per ogni gruppo acetile che entra nel ciclo vengono prodotte tre molecole di NADH.

- Alla fine di ogni giro l'ossalacetato viene rigenerato a partire dal malato ed il ciclo continua.

La somma di tutte queste reazioni risulta essere

$$\begin{aligned}[\mathrm{AcCoA}] + 3\,[\mathrm{NAD}]^+ + \mathrm{Q} + [\mathrm{GDP}] + [\mathrm{P}]_\mathrm{i} + 2\,\mathrm{H_2O} \longrightarrow \\ \longrightarrow \mathrm{CoA{-}SH} + 3\,[\mathrm{NADH}] + 3\,\mathrm{H}^+ + \mathrm{QH_2} + [\mathrm{GTP}] + 2\,\mathrm{CO_2}\end{aligned}$$

[3] La respirazione cellulare è il meccanismo attraverso cui la cellula, in presenza di ossigeno, è in grado di ricavare energia utilizzabile nei processi vitali attraverso la rottura dei legami chimici negli elementi che assorbe.

e se consideriamo tutte anche le reazioni necessarie per l'ossidazione del piruvato e quelle della glicolisi (con l'esclusione di quelle nella catena respiratoria) possiamo scrivere la seguente reazione di ossidazione del glucosio:

$$[\mathrm{G6\,P}] + 10\,[\mathrm{NAD}]^{+} + 2\,[\mathrm{Q}] + 2\,[\mathrm{ADP}] + 2\,[\mathrm{GDP}] + 4\,[\mathrm{P}]_{\mathrm{i}} + 2\,\mathrm{H_2O} \longrightarrow$$
$$\longrightarrow 10\,[\mathrm{NADH}] + 10\,\mathrm{H}^{+} + 2\,[\mathrm{QH_2}] + 2\,[\mathrm{ATP}] + 2\,[\mathrm{GTP}] + 6\,\mathrm{CO_2}.$$

Si stima che la produzione di ATP a partire da una singola molecola di glucosio ammonti a circa 30 unità in quanto la nicotinammide adenina dinucleotide (NADH) è convertita in ulteriore ATP attraverso una catena di trasporto degli elettroni.

Via dei penstoso fosfati Una via parallela alla glicolisi che consente di generare zuccheri pentosi (a cinque atomi di carbonio) e molecole ossidanti NADPH è la via dei pentoso fosfati. Anche questo *pathway* si compone di due fasi.

- Una prima fase ossidativa in cui il G6P viene convertito in 6-fosfogluconato (6PG) e quindi in ribulosio-5-fosfato (RU5P-D); da quest'ultimo zucchero si può ottenere facilmente il ribosio-5-fosfato (R5P) o lo xilulosio-5-fosfato (XU5P-D).

- Una seconda fase non ossidativa in cui i pentosi sono riconvertiti in esosi e rientrano nella glicolisi sotto forma di F6P e G3P.

1.3 Stati stazionari del metabolismo

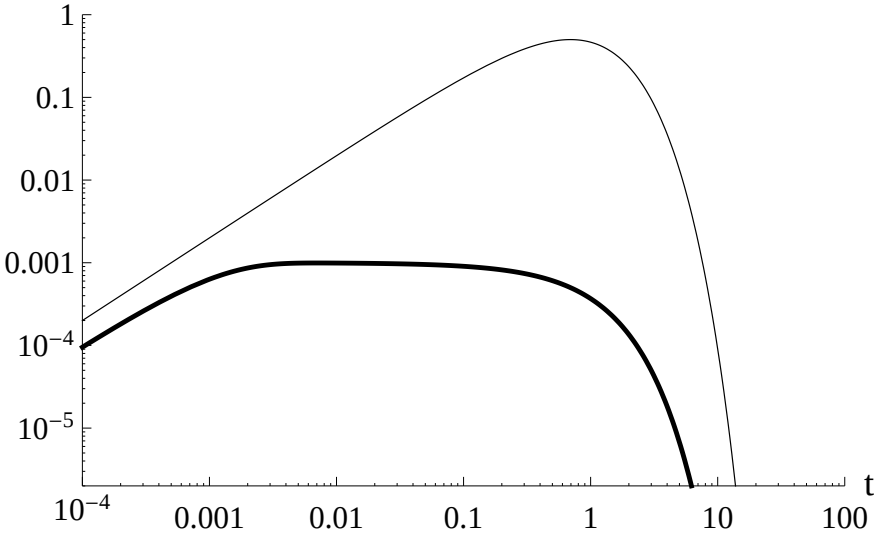
Nel paragrafo 1.1 abbiamo visto come la velocità delle reazioni possa essere ottenuta da sistemi di equazioni differenziali. Questi sistemi non possono essere integrati esplicitamente e quindi si rende necessario fare delle approssimazioni. L'assunzione di stato quasi-stazionario che abbiamo introdotto è la più utilizzata poiché tipicamente le concentrazioni dei substrati sono mantenute alte dalle altre reazioni del metabolismo. L'altra condizione necessaria è la conoscenza di $[E]_0$, [S] e le costanti cinetiche. Purtroppo questi dati sono di difficile accesso sperimentale e se vogliamo farne a meno sono necessarie ulteriori approssimazioni.

Un'ulteriore passo in avanti in questa direzione consiste nell'estendere la condizione di stazionarietà dal composto substrato-enzima alle concentrazioni di tutti i metaboliti della cellula. Rinunciare alla dimensione temporale riduce molto le possibilità di analisi (non si possono studiare transienti, comportamenti ciclici o rapidità di adattamento a perturbazioni esterne) ma ha il pregio di rendere trattabili le equazioni. Siamo infatti in grado di passare da sistemi di migliaia di equazioni differenziali fortemente accoppiate a delle corrispondenti espressioni algebriche. Inoltre, visto che le reazioni chimiche sono lineari nelle concentrazioni, queste espressioni possono essere risolte con le tecniche della algebra lineare.

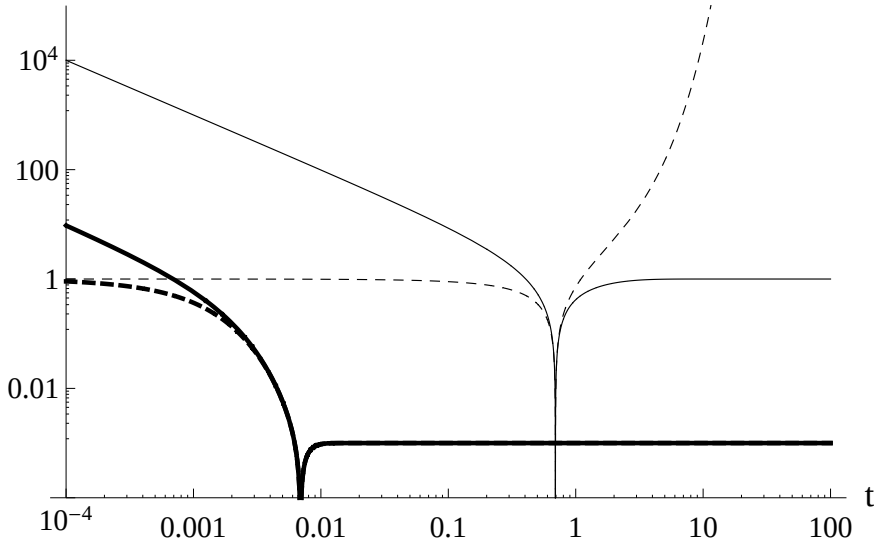
Siccome stiamo imponendo una condizione molto forte, conviene verificare subito in quale regime vale.
Immaginiamo una situazione molto semplificata in cui siano presenti due reazioni in serie, ad esempio un a catena di decadimenti

$$\mathrm{A} \xrightarrow{k_1} \mathrm{B} \xrightarrow{k_2} \mathrm{C}.$$

Possiamo immaginare la prima reazione come un nutriente importante che entra nella cellula, ad esempio il glucosio, e la seconda come la produzione di biomassa cellulare che viene sottratta al metabolismo. Le soluzione si trova scrivendo un sistema analogo a 1.2, che però questa volta è risolvibile analiticamente. Imponendo $d[\mathrm{B}]/dt = 0$ possiamo trovare delle soluzioni approssimate e confrontarle con quelle analitiche. In questo modo è possibile stabilire che l'approssimazione vale quando $[\mathrm{A}] \gg [\mathrm{B}]$ o, equivalentemente, $k_2 \gg k_1$. Se la seconda reazione procede molto velocemente, infatti, dopo un transiente iniziale B non fa in tempo ad accumularsi e la sua concentrazione tende a rimanere costante (figura 1.3(a), nel-



(a) Concentrazione dell'elemento intermedio.



(b) Rapporti tra le derivate delle concentrazioni. Le linee tratteggiate sono $\left|\frac{d[\mathrm{B}]/dt}{d[\mathrm{A}]/dt}\right|$, le linee solide $\left|\frac{d[\mathrm{B}]/dt}{d[\mathrm{C}]/dt}\right|$.

Figura 1.3: Concentrazioni dell'elemento intermedio e variazioni relative. Linee normali: $k_1 = 2$, $k_2 = 1$. Linee spesse: $k_1 = 1$, $k_2 = 10^3$.

l'intervallo $10^{-3} < t < 1$).
È importante notare come in questa approssimazione non sia richiesto rigorosamente $d[\mathrm{B}]/dt = 0$ ma piuttosto che la variazione di [B] sia piccola rispetto alla variazione di [A] e di [C]. Nella figura 1.3(b) possiamo vedere come, se le condizioni sui coefficienti sono soddisfatte, dopo il picco di [B] si verifichi proprio questa situazione.

Da quando detto si capisce che possiamo assumere la condizione stazionarietà quando in una coltura c'è abbondanza di nutrimento e la produzione di biomassa avviene a tasso costante.

All'inizio del paragrafo abbiamo scritto che allo stato stazionario è possibile ottenere i valori dei flussi da tecniche di algebra lineare. Presentiamo un esempio semplice per capire meglio la portata di questa affermazione.
Una rete metabolica è composta da quattro reazioni che interessano tre metaboliti:

$$\begin{aligned} 2\,\mathrm{M}_1 &\longrightarrow \mathrm{M}_2 \\ \mathrm{M}_2 &\longrightarrow \mathrm{M}_1 + \mathrm{M}_3 \\ \mathrm{M}_2^{\mathrm{ec}} &\longrightarrow \mathrm{M}_2 \\ \mathrm{M}_3 &\longrightarrow \mathrm{M}_3^{\mathrm{ec}} \end{aligned}$$

La terza e la quarta reazione portano un metabolita rispettivamente dentro (reazione di *uptake*) e fuori (reazione di scarto) dall'ambiente in cui avvengono le reazioni.[4]
Nell'approssimazione di stato stazionario le concentrazioni non devono cambiare, quindi se chiamiamo c_μ il tasso di produzione del metabolita M_μ possiamo esprimere la condizione di stazionarietà (o di stabilità) come

$$c_\mu = 0, \quad \forall \mu. \tag{1.4}$$

Utilizziamo ora le condizioni di linearità per esprimere il tasso di produzione:

$$c_\mu = \sum_i (b_{i\mu} - a_{i\mu}) s_i;$$

$b_{i\mu}$ è il coefficiente stechiometrico per il metabolita M_μ qualora sia prodotto dalla reazione R_i, altrimenti è 0; $a_{i\mu}$ è il coefficiente stechiometrico nel caso M_μ sia substrato della reazione R_i, altrimenti è 0; s_i indica la velocità della reazione i-esima. Definiamo $\xi_{i\mu} = b_{i\mu} - a_{i\mu}$ e raccogliamo questi coefficienti in una

[4] L'apice `ec` sta per "extra-cellulare"

matrice in cui le righe rappresentano i metaboliti e le colonne le reazioni

$$\Xi = \begin{pmatrix} -2 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$

possiamo allora riscrivere la condizione 1.4 come

$$\Xi \cdot \vec{s} = \vec{c_\mu} = \vec{0},$$

che risolta da

$$s_2 = 2s_1, \quad s_3 = s_1, \quad s_4 = 2s_1.$$

Siamo ora in grado di apprezzare la potenza di questo metodo: se in una rete sono presenti N reazioni e P metaboliti, attraverso le condizioni di stabilità possiamo ridurre i gradi di libertà da N a $(N - P)$.

I gradi di libertà che non riusciamo ad eliminare analiticamente vanno fissati attraverso gli esperimenti. Diventa quindi rilevante disporre di metodi robusti e affidabili, sia numerici sia sperimentali, che permettano di ricavare questi valori in diverse condizioni di coltura o diverse configurazioni della rete metabolica.

Nel prossimo capitolo presenteremo un importante metodo sperimentale per ricavare le velocità delle reazioni mentre nel capitolo terzo ci concentreremo sulle diverse tecniche numeriche, spiegando in dettaglio quella usata in questo lavoro.

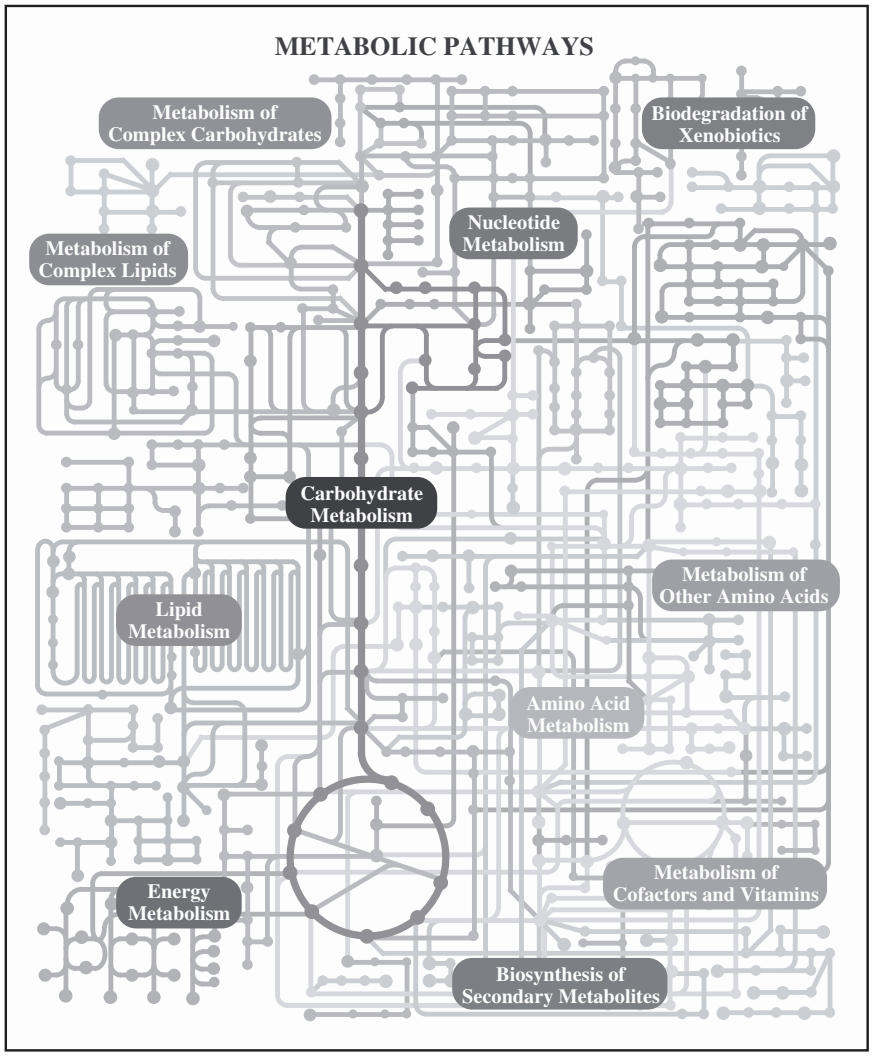


Figura 1.4: Una rappresentazione del metabolismo cellulare dal KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAY database – `http://www.genome.jp/kegg/kegg2.html`.

Capitolo 2

Metodi sperimentali

Nel capitolo precedente abbiamo visto che possiamo imporre una serie di condizioni legate al bilancio della massa, in modo da costringere i valori dei flussi in base alla topologia di una specifica rete. Questo metodo tuttavia non è sufficiente ad eliminare tutti i gradi di libertà e rende necessario porre ulteriori vincoli sperimentali ai valori dei flussi.

Una volta fatte le debite approssimazioni possiamo cercare di misurare le concentrazioni dei substrati e degli enzimi per utilizzarli in equazioni come la 1.3, o misurare substrato e prodotto. Purtroppo solo in pochi casi questi valori sono accessibili direttamente: principalmente per composti che vengono assorbiti dalla cellula o metaboliti che sono espulsi. Per tutti gli altri flussi bisogna appoggiarsi a misurazioni indirette.

2.1 ^{13}C-labelling

Una delle tecniche più promettenti per effettuare queste misure è la marcatura con atomi di carbonio 13. Questa tecnica si basa sulla possibiltà di seguire il carbonio lungo le vie metaboliche attive fino a differenti prodotti sintetizzati dalla cellula e stimarne la distribuzione utilizzando la spettroscopia (NMR o di massa) per osservare la posizione degli atomi di ^{13}C o la distribuzione degli isotopomeri.[1]

È possibile utilizzare direttamente alcuni substrati così marcati e monitorarli tra i metaboliti intracellulari, tra i prodotti extracellulari o nelle proteine della cellula. Tuttavia questo sistema ha degli svantaggi, dal momento che richiede numerosi passaggi di purificazione, ed è molto costoso, poiché necessita

[1]Gli isotopomeri, o isotopi isomeri, sono composti che contengono lo stesso numero di isotopi ma in posizioni differenti.

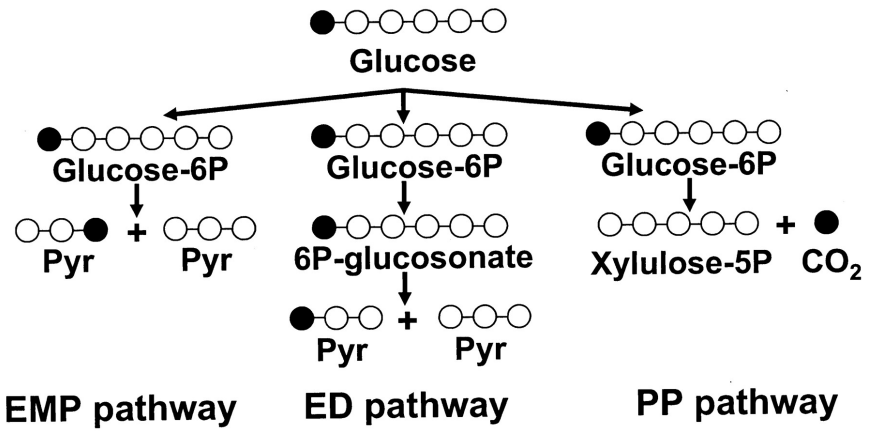


Figura 2.1: Come le misurazioni del motivo di arricchimento del ^{13}C possono essere usate per identificare le vie metaboliche attive. EMP, Embden-Meyerhof-Parnas; ED, Entner-Doudoroff; PP, pentofostati.

di un grosso numero di substrati marcati.
In alternativa si può usare la *bioynthetically directed fractional ^{13}C labelling* [21]. In questo approccio la cellula viene nutrita con una mistura di carbonio marcato e non marcato e la distribuzione finale del ^{13}C negli amminoacidi viene analizzata attraverso uno spettro di correlazione NMR. Questa strategia permette di tracciare i frammenti di carbonio contigui che provengono da una singola molecola attraverso la rete di reazioni cellulari, fornendo un'immagine dettagliata dell'utilizzo delle molecole precursori. Siccome la posizione dei frammenti di carbonio rimasti intatti in un determinato metabolita dipende molto spesso da indicazioni su quale quale via sia stata impiegata per la sua sintesi, è possibile ottenere informazioni sia riguardo all'effettiva topologia della rete, sia riguardo ai rapporti tra i flussi in diversi punti chiave del metabolismo centrale del carbonio. Oltretutto gli esperimenti possono essere effettuati ad un costo minore, in quanto è necessario marcare solo una frazione del carbonio utilizzato.

2.2 Descrizione del metodo

Tecniche di misura

I tre metodi più utilizzati per quantificare l'arricchimento isotopico dei metaboliti intracellulari sono la risonanza magne-

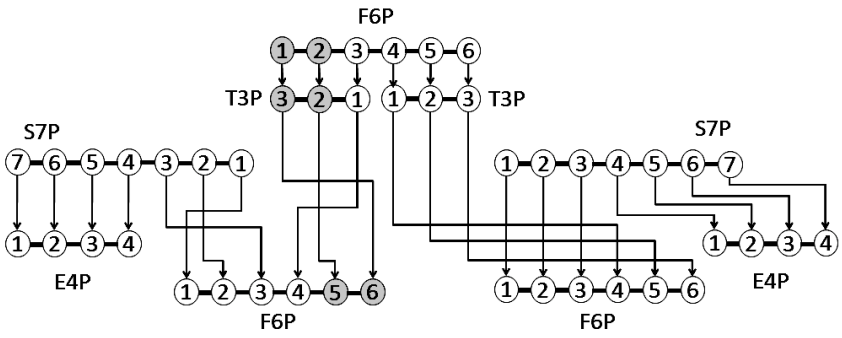


Figura 2.2: Tracciamento degli atomi di ^{13}C, i cerchi più scuri, attraverso la via dei pentofosfati — fonte `http://en.wikipedia.org/wiki/Isotopic_labeling`.

tica nucleare 2D e gascromatografia o cromatografia liquida seguita dalla spettrometria di massa.

2D-NMR Le misure di spettroscopia di correlazione (COSY), o risonanza magnetica nucleare bidimensionale (2D-NMR), sono basate sulla separazione spettrale dei segnali dei protoni legati al carbonio 13 (1H,13C-NMR) e a quella degli atomi di ^{13}C legati ad altri atomi di ^{13}C (13C,13C-NMR). La struttura fine rappresenta le interazioni degli atomi marchiati con i loro vicini anch'essi marchiati.
In dipendenza dal numero degli accoppiamenti scalari ^{13}C-^{13}C possono essere rilevate le seguenti strutture:

- singoletto (s), nessun accoppiamento scalare;
- doppietto (d), un solo accoppiamento scalare ^{13}C-^{13}C;
- tripletto (t), due accoppiamenti identici;
- doppio doppietto (dd), due accoppiamenti ^{13}C-^{13}C differenti.

L'analisi via NMR è caratterizzata da una bassa sensitività, ne consegue che i marcatori nei metaboliti intracellulari vanno inferiti dall'accumulo di componenti di biomassa come gli amminoacidi coinvolti nella sintesi delle proteine, conoscendo i *pathway* di biosintesi dai precursori agli amminoacidi è possibile dedurre la distribuzione dei marcatori nei metaboliti primari.
Tra tutta la biomassa coinvolta nella sintesi proteica, gli amminoacidi contengono la maggior informazione sulla marcatura

dei metaboliti primari poiché essi sono ottenuti da una gran quantità di precursori diversi originati a partire dalla glicolisi, la via dei pentoso fosfati ed il ciclo di Krebs. Siccome questi precursori sono localizzati in compartimenti della cellula prestabiliti, la misura dell'arricchimento isotopico è locale.
I tempi in cui i costituenti della biomassa vengono completamente sostituiti sono molto lenti (dell'ordine delle ore). Sono quindi necessari lunghi tempi di assorbimento del ^{13}C per ottenere una distribuzione costante dei marcatori (stato stazionario isotopico) e siccome la sostituzione del ^{12}C con il ^{13}C avviene principalmente attraverso la formazione di nuova biomassa e distruzione della vecchia, tutto il substrato di coltura deve essere marchiato. Questi tempi lunghi sono uno svantaggio dell'analisi NMR, rendono il metodo inadatto a osservare transienti metabolici e richiedono grandi quantità di ^{13}C, con gli alti costi conseguenti ($\sim$170\$/g).

GC-MS Un'altra tecnica spesso usate per tracciare il ^{13}C e la gascromatografia-spettrometria di massa (GC-MS). Questo tipo di analisi si divide in due fasi: inizialmente il campione viene scaldato e separato nelle sue componenti attraverso la gascromatografia, successivamente queste componenti sono identificate grazie alla loro massa (MS).
La GC-MS è utilizzata infatti per distinguere tra isotopomeri con differenti masse atomiche, che sono quindi legati al numero di atomi di ^{13}C presenti nella molecola, ma non alla loro posizione.
Analogamente alla NMR la GC-MS determina la marcatura dei metaboliti primari indirettamente, attraverso le misure sugli amminoacidi da cui vengono sintetizzati, dunque anche questa tecnica necessita di lunghi tempi per raggiungere lo stato stazionario isotopico. Un vantaggio non trascurabile è, invece, la possibilità di frammentare gli amminoacidi per ottenere ulteriori informazioni sulla posizione dei marcatori.

LC-MS Una tecnica relativamente recente per determinare direttamente la distribuzione del ^{13}C è la cromatografia liquida abbinata alla spettrometria di massa (LC-MS). A differenza delle tecniche precedenti, questo metodo permette di stimare direttamente la distribuzione isotopomera dei metaboliti primari, evitando errori relativi alle assunzioni sulle vie biosintetiche percorse dagli amminoacidi. I tempi di sostituzione dei metaboliti sono dell'ordine dei secondi, rendendo quindi ne-

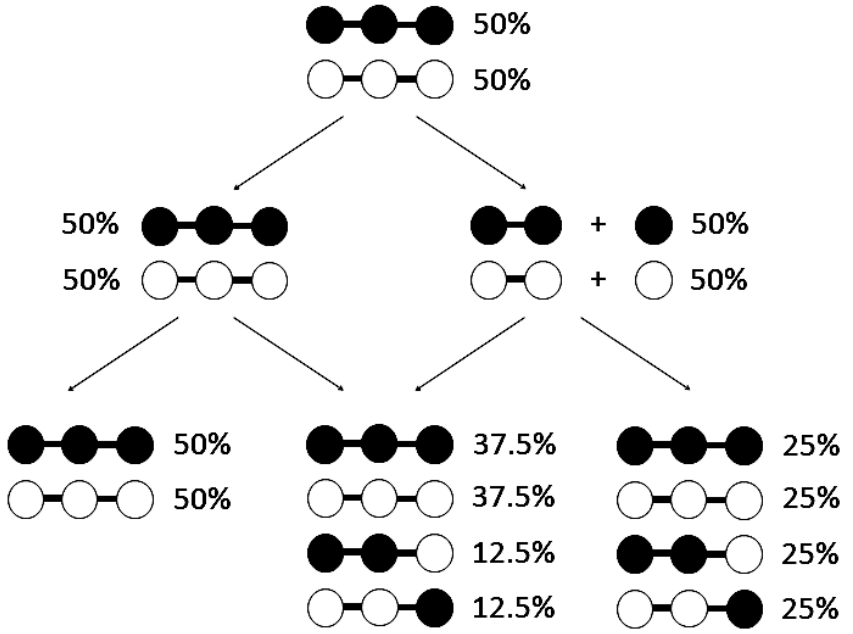


Figura 2.3: Esempio di come una distribuzione di substrato marcato e non marcato evolve attraverso le reazioni — fonte `http://en.wikipedia.org/wiki/Isotopic_labeling`.

cessario un substrato di ^{13}C molto ridotto (rispetto alla NMR e alla GC-MS). Questi tempi brevi richiedono però misurazioni rapide per determinare correttamente la distribuzione del ^{13}C nei metaboliti.

Analisi dei flussi

Una volta misurate le distribuzioni dei marcatori all'interno della cellula è necessario derivare da queste i flussi metabolici. Per ottenere questo risultato si utilizzano principalmente due approcci: uno basato sulla *local flux analysis* e l'altro sul *whole isotopomer modeling*.
Il primo metodo consiste nel determinare i flussi intracellulari attorno ad un nodo costituito da un singolo metabolita interpretando le distribuzioni di ^{13}C dei metaboliti circostanti, solitamente a mezzo di equazioni algebriche.
Il secondo approccio punta a stimare tutti i flussi contemporaneamente attraverso un modello di rete metabolica predefinito. In questo caso i valori sono ottenuti tramite un approccio inferenziale e la soluzione non è unica; inoltre, visto che i flus-

si sono ottenuti da una minimizzazione degli errori sull'intera rete, potrebbero essere subottimali relativamente ad ogni nodo considerato separatamente.

Local node flux analysis Un metodo interessante basato su questo approccio è l'*analisi dei rapporti dei flussi metabolici* (METAFoR) [17, 22]. Tramite questo metodo le misure di distribuzione del ^{13}C sono utilizzate per quantificare il partizionamento dei flussi in vari nodi convergenti del metabolismo con metodi probabilistici.
Facciamo un esempio pratico. In un campione con il 10% di ^{13}C e il 90% di glucosio non marcato la probabilità di trovare due atomi adiacenti entrambi marcati è del 1.2% se i due nuclei provengono da fonti di glucosio differenti, mentre sale al 10% qualora siano parte di un frammento derivato dal medesimo precursore. Quindi l'accoppiamento ^{13}C-^{13}C degli atomi terminali in un frammento con due C è dominato dal singoletto nel caso di una presenza casuale del ^{13}C, poiché gli atomi circostanti sono per la maggior parte ^{12}C, e da un doppietto nel caso opposto.
I singoli rapporti così individuati sono relativamente indipendenti gli uni dagli altri e non richiedono ulteriori informazioni fisiologiche; inoltre possono essere utilizzati in abbinamento a tecniche numeriche come la FBA per introdurre ulteriori condizioni sui flussi.[2]
Il metodo è stato in seguito adattato all'utilizzo di dati provenienti dalla GC-MS [9] e ha permesso di identificare 14 flussi appartenenti al metabolismo centrale di *E.coli*. In questo caso il metodo è particolarmente semplice e ne daremo una descrizione più approfondita.
Consideriamo la situazione in cui due reazioni convergano su di un medesimo metabolita

$$\begin{aligned} R_1: &\quad \mathrm{A} \longrightarrow \mathrm{C}, \\ R_2: &\quad \mathrm{B} \longrightarrow \mathrm{C}, \\ R_3: &\quad \mathrm{C} \longrightarrow \ldots \end{aligned}$$

La conservazione della massa vale sia per il metabolita nel suo complesso per i diversi isotopomeri (le cui masse vengono raccolte nel vettore delle distribuzioni di massa — MDV) e

[2]Una spiegazione del metodo FBA si trova al paragrafo 3.1

fornisce un vincolo al rapporto tra i flussi in ingresso:

$$s_1 + s_2 = s_3,$$
$$s_1 \mathrm{MDV}_A + s_2 \mathrm{MDV}_B = s_3 \mathrm{MDV}_C,$$

da cui si ricava immediatamente

$$\frac{s_1}{s_2} = \frac{\mathrm{MDV}_C - \mathrm{MDV}_B}{\mathrm{MDV}_A - \mathrm{MDV}_C}.$$

Un svantaggio di questo approccio nasce dal fatto che non tutti i flussi relativi ai metaboliti che si stanno studiando sono direttamente misurabili, in questo caso vanno derivati a partire dai metaboliti precedenti o successivi assumendo che la posizione del marcatore resti immutata. Un altro svantaggio, inerente alla struttura stessa del metodo, è la limitazione ai soli flussi convergenti, ossia in cui due reazioni terminano sul medesimo metabolita.

Whole isotopomer modeling Questo approccio si appoggia fondamentalmente sulla formulazione proposta da Schmidt e altri [18] che permette di comparare i dati sperimentali da spettri NMR con simulazioni per lo stato stazionario delle distribuzioni di isotopomeri. Questo procedimento è reso necessario dal fatto che queste misure conducono molto spesso a sistemi sovradeterminati.

Il vantaggio di questo metodo è la possibilità di analizzare contemporaneamente tutti i flussi presenti nella rete studiata; inoltre fissando nel modello la velocità delle reazioni extracellulari (*uptake*, produzione di biomassa, ...) i valori dei flussi determinati sono assoluti.
Considerare la rete metabolica nel suo insieme comporta anche degli svantaggi: a causa dell'altro grado di connettività di queste reti un errore nelle misure localizzato causato da un'incorretta ricostruzione della rete (reazioni sbagliate o mancanti) può propagarsi facilmente ed alterare le stime in tutta la rete.

Capitolo 3

Metodi numerici

I continui progressi nella ricostruzione di reti metaboliche pressoché complete hanno aperto la porta ad una nuova generazione di analisi, rendendo possibile la determinazione dei flussi metabolici attraverso esperimenti *in silico*. In questo capitolo descriveremo due tra i metodi numerici utilizzabili a questo scopo. Come abbiamo spiegato precedentemente, i risultati hanno significato esclusivamente nella fase stazionaria del metabolismo.

Richiamiamo la notazione che abbiamo utilizzato nell'introdurre gli stati stazionari in quanto da questo punto in avanti verrà utilizzata estensivamente.
Indicheremo con N il numero di reazioni presenti nella rete e con P il numero dei metaboliti. Utilizzeremo sempre indici latini per riferirci alle reazioni mentre le lettere greche saranno riservate ai metaboliti. Con c_μ indicheremo il tasso di crescita del del metabolita M_μ, il cui coefficiente stechiometrico relativo alla reazione R_i, che procede con velocità s_i, sarà $\xi_{i\mu}$. Questi coefficienti possono essere scritti come una matrice Ξ di dimensione $P \times N$.

3.1 FBA

Il metodo attualmente più utilizzato per lo studio *in silico* dei flussi metabolici è la *flux balance analysis* [13]: tramite tecniche di programmazione lineare si cerca uno stato stazionario in cui i flussi siano compatibili con i requisiti imposti dai coefficienti stechiometrici e dalle misure sperimentali.

Abbiamo visto che le condizione di stazionarietà per P metaboliti vincolano P gradi di libertà su N totali. Per fissare gli $(N - P)$ gradi di libertà rimanenti bisogna trovare il modo di

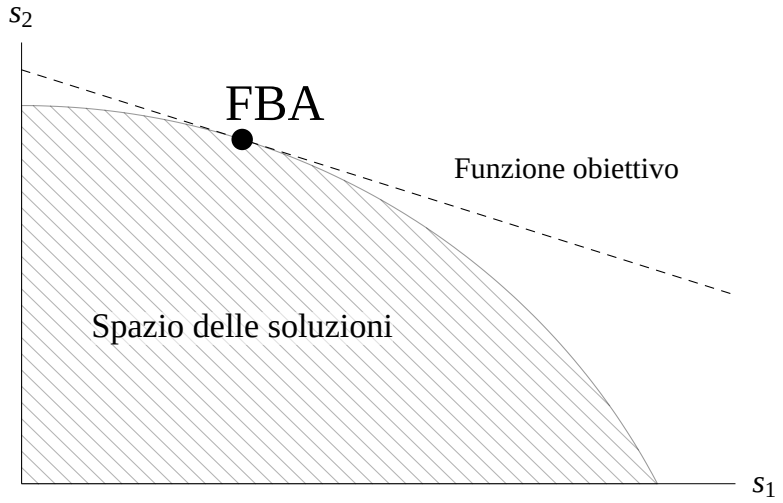


Figura 3.1: Rappresentazione dello spazio delle soluzione per un rete con due reazioni, la funzione obiettivo da massimizzare e la soluzione trovata dalla FBA.

inserire nel modello teorico le informazioni sui flussi tratte dagli esperimenti. Nell'FBA queste informazioni sono codificate in dei limiti superiori ed inferiori ai valori permessi dei flussi,

$$\alpha_i < s_i < \beta_i.$$

Questo sistema permette anche di specificare l'irreversibilità di una reazione ponendo uno dei due limiti uguale a zero.

Dal punto di vista topologico, prima di porre i vincoli i flussi soluzione del sistema sono rappresentati da un vettore $\vec{s}$ elemento di uno spazio vettoriale N-dimensionale. Aggiungere le condizioni di conservazione della massa conduce ad un sottospazio a $(N - P)$ dimensioni.
A questo livello conta esclusivamente la direzione del vettore: se $\vec{s}$ è una soluzione del problema, allora lo è anche $\lambda\vec{s}$. La situazione cambia con l'introduzione dei limiti 3.1, che riducono lo spazio permesso ad un politopo convesso.[1]

Benché ogni punto di questo oggetto geometrico sia soluzione del sistema, tipicamente si desidera trovare una soluzione univoca che sia di rilevanza biologica, cioè in cui alcuni metaboliti vengano prodotti nella corretta proporzione. L'insieme di metaboliti e le proporzioni desiderate sono rappresentati da una *funzione obiettivo*, che equivale ad una combinazione di una o più reazioni o una reazione ulteriore.

[1]Infatti, se $\vec{s}_1$ $\vec{s}_2$ sono soluzioni, lo è anche una loro combinazione convessa $[\alpha\vec{s}_1 + (1 - \alpha)\vec{s}_2]$, per $0 \leq \alpha \leq 1$.

Solitamente viene scelta la produzione di biomassa cellulare, ma in modelli di rete più limitati è possibile imporre degli obiettivi particolari (ad esempio una certa produzione di ATP e NADH nel metabolismo del carbonio).
La funzione obiettivo comporta la selezione di una direzione particolare nello spazio dei flussi (a N o N+1 dimensioni a seconda di come si è definita la funzione) e la ricerca della soluzione corrisponde a trovare il massimo lungo questa direzione compatibilmente ai vincoli. Se la funzione obiettivo è lineare si tratterà di un punto sulla frontiera del politopo (figura 3.1). Siccome il politopo è convesso ogni massimo locale sarà anche un massimo globale garantendo così l'unicità della soluzione, nel caso questa esista.
I motivi per cui la soluzione può non esistere sono l'inconsistenza dei vincoli o il fatto che il politopo non sia limitato nella direzione del gradiente della funzione obiettivo. Entrambi questi problemi si risolvono con una scelta attenta degli intervalli 3.1.

Se un punto di vista geometrico può essere utile nella visualizzazione mentale, a livello numerico la FBA corrisponde ad un problema di ottimizzazione lineare per il sistema

$$\max_{\vec{s}} f(\vec{s}), \quad \Xi \cdot \vec{s} = 0, \quad \alpha_i < s_i < \beta_i; \tag{3.1}$$

la cui soluzione può essere trovata molto rapidamente dai moderni algoritmi di *linear programming*.

L'aspetto più delicato di questo metodo riguarda la scelta della funzione da massimizzare. Qualora si tratti della produzione di biomassa, ad esempio, tutte le componenti necessarie a creare la biomassa saranno comprese tra i reagenti e i metaboliti che alla fine del processo rimarranno nella rete compariranno come prodotti. I coefficienti stechiometrici di questa reazione fittizia vanno però ricavati dai dati sperimentali ovvero sono essi stessi oggetto di studio.

Un approccio di questo tipo potrebbe essere considerato arbitrario, ciò nonostante molte predizioni di FBA si sono rivelate consistenti ai dati sperimentali (ad esempio [5]).

MOMA

Dato un certo genotipo, per sua natura la FBA trova lo stato metabolico ottimale che massimizza un certo obiettivo, per esempio la produzione di biomassa. Se questa assunzione si

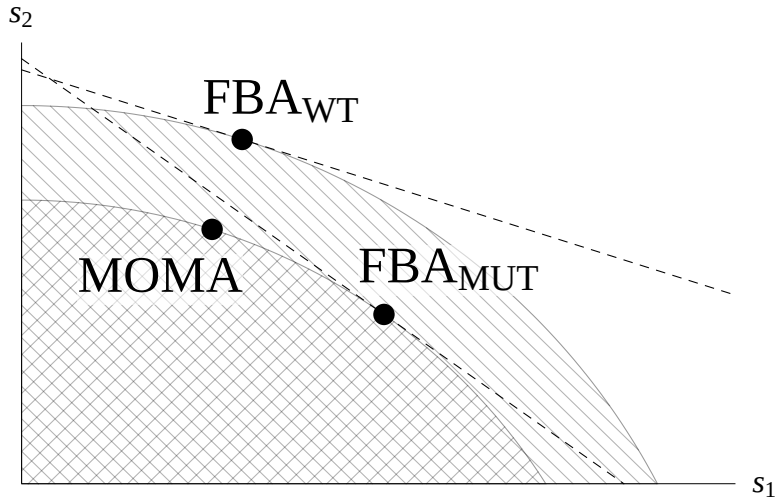


Figura 3.2: Dopo il *knockout* lo spazio delle soluzioni e la funzione obiettivo sono modificati. La MOMA trova una soluzione subottimale ma che minimizza la distanza quadratica dalla soluzione per il caso non perturbato.

rivela essere fondata nel caso di una cellula soggetta a milioni di anni di pressione selettiva, non è altrettanto chiaro se lo sia nel caso di mutanti creati in laboratorio nei quali certe reazioni chimiche sono fortemente soppresse attraverso la manipolazione genetica.

Un approccio alternativo alla FBA che tiene conto di questa possibilità è stato proposto in [19] col nome di *minimization of metabolic adjustment* (MOMA). In questo caso non si scandaglia lo spazio delle soluzioni alla ricerca della soluzione ottimale, bensì si cerca la soluzione al sistema perturbato più vicina alla soluzione del sistema non perturbato. Se chiamiamo s_{wt} la soluzione del sistema non perturbato (*wild type*), le soluzioni nella MOMA sono i flussi $\vec{s}$ che rendono minima

$$D(\vec{s}, \vec{s}_{wt}) = \|\vec{s} - \vec{s}_{wt}\|_2,$$

soggetta agli sessi vincoli visti nel paragrafo precedente.
Il risultato in questo caso non dipende esplicitamente dalla funzione obiettivo, ma solo indirettamente attraverso $\vec{s}_{wt}$. Gli autori sono in grado di dimostrare che il loro metodo predice correttamente gli effetti letali dell'inattivazione di specifici geni in *E.coli*, contrariamente alla FBA.

3.2 Modello di Von Neumann

Il terzo modello che presentiamo è quello effettivamente utilizzato nella nostra tesi, quindi ci prenderemo lo spazio per una descrizione più dettagliata. Il modello è stato elaborato da John Von Neumann nel 1937 per analizzare il profilo di produzione e i prezzi all'equilibrio di un certo numero di beni in un sistema economico, dato un insieme di processi produttivi disponibili [24].

Le assunzioni di base sono molto semplici: nell'economia sono presenti P beni ed N processi utilizzati per produrre questi beni l'uno dall'altro. Ogni bene M_μ è prodotto in quantità $b_{i\mu}$ da uno o più processi R_i a partire da una quantità $a_{i\mu}$ degli altri beni. Il tutto avviene secondo una struttura

$$R_i \colon \sum_{\mu=1}^{P} a_{i\mu} M_\mu \to \sum_{\mu=1}^{P} b_{i\mu} M_\mu \quad i = 1, \ldots, N,$$

con la condizione $a_{i\mu} + b_{i\mu} > 0$.

I problemi a cui Von Neumann cercava una risposta sono:

1. quali processi vengono effettivamente usati;
2. quali sono le velocità relative alle quali vengono prodotti i vari beni;
3. quali sono i prezzi dei beni;
4. qual è il tasso di interesse globale.

Se chiamiamo la velocità con cui procede ogni processo $s_i \geq 0$, possiamo scrivere la produzione della rete al tempo t relativa al bene M_μ come

$$O_\mu(t) = \sum_i b_{i\mu} s_i(t)$$

e il consumo come

$$I_\mu(t) = \sum_i b_{i\mu} s_i(t).$$

Affinché il sistema sia autocatalitco, cioè ciascun bene possa essere ottenuto a partire dagli altri beni presenti nell'insieme considerato, va richiesto

$$c_\mu(t) = O_\mu(t) - I_\mu(t+1) \geq 0$$

ad ogni tempo e per ogni μ. In questo modo l'input richiesto al tempo $t+1$ è interamente coperto dai beni prodotti al tempo t.
Seguendo l'approccio di Von Neumann, cerchiamo gli stati in cui l'economia cresce senza cambiare struttura ($s_1 : s_2 : \cdots : s_N$) e i vari s_i possono cambiare solo se moltiplicati per una costante ϱ, il coefficiente di espansione per l'intera economia. L'unica evoluzione possibile è allora $s_i(t+1) = \varrho s_i(t)$, che implica $s_i(t) = \varrho^t s_i$ e $c_\mu(t) = \varrho^t c_\mu$.
Possiamo quindi riscrivere la condizione di stabilità del sistema come

$$c_\mu = \sum_i (b_{i\mu} - \varrho a_{i\mu}) s_i \geq 0, \quad \forall \mu. \tag{3.2}$$

Per rispondere alle ultime due domande dell'elenco precedente, Von Neumann introduce un vettore dei prezzi $\vec{p}$ soggetto alle condizioni:

$$\sum_\mu (b_{i\mu} - \beta a_{i\mu}) p_\mu \leq 0, \quad \forall i.$$

che implicano che non sia possibile trarre profitto da alcun processo. Questo non significa non guadagnare soldi dalla vendita delle merci ma solo che questo guadagno non può essere maggiore dell'interesse che si otterrebbe lasciando il denaro in una banca.[2]

In aggiunta a questa condizione e alla 3.2, si richiede che qualunque bene prodotto in eccesso abbia prezzo nullo e qualunque processo che produce beni di valore più basso del costi attualizzati $\sum_\mu \beta a_{i\mu}$ venga abbandonato.

Per riassumere, considerando $\vec{s}$ un vettore colonna, $\vec{p}$ un vettore riga e raggruppando in matrici i coefficienti $a_{i\mu}$ e $b_{i\mu}$, il modello di Von Neumann può essere riassunto con

$$B\vec{s} \geq \varrho A\vec{s} \tag{3.3}$$
$$\vec{p}B \leq \beta \vec{p}A \tag{3.4}$$
$$\vec{p}(B - \varrho A)\vec{s} = 0 \tag{3.5}$$
$$\vec{p}(B - \beta A)\vec{s} = 0 \tag{3.6}$$
$$\vec{s} \geq 0 \quad \vec{p} \geq 0 \tag{3.7}$$
$$A + B > 0 \tag{3.8}$$

assieme alle condizioni di normalizzazione

$$\|\vec{s}\|_1 = S \quad \|\vec{p}\|_1 = P$$

[2]Un'assunzione valida in un regime di concorrenza perfetta.

necessarie vista l'invarianza di scala del problema.

Tramite una generalizzazione del teorema del punto fisso di Brouwer, Von Neuman riesce a dimostrare che una soluzione a questo problema esiste e che il coefficiente di crescita dell'economia ϱ è uguale al fattore di interesse β.

3.3 Il modello di VN su reti metaboliche

Il lettore avrà notato che nella descrizione del modello di Von Neumann (VN) abbiamo utilizzato una notazione analoga a quella usata nel paragrafo 3.1 per riferirci ai sistemi biologici. È nostra intenzione infatti rimarcare un parallelismo tra questi due sistemi, in modo da poter applicare il modello alle reti metaboliche.

Analogie e differenze

I flussi Le reazioni chimiche all'interno del cellula possono essere facilmente associate a dei processi di produzione: entrambi connettono uno stato di input ad uno di output tramite una trasformazione lineare. In un caso $a_{i\mu}$ e $b_{i\mu}$ rappresentano le quantità di beni necessarie e risultanti a un processo produttivo in un'unità tempo; nell'altro i coefficienti stechiometrici di substrato e prodotti.

Nel modello di Von Neumann si ha $s_i > 0$, quindi i processi non sono invertibili ma è possibile ritornare alla situazione iniziale combinandone più di uno. Alcune reazioni chimiche sono invece reversibili e per descriverle matematicamente dobbiamo scegliere una convenzione per indicare la reazione inversa.
Se scegliamo di utilizzare un singolo flusso che può assumere valori sia positivi sia negativi, lo spazio delle soluzioni non è convesso a $\varrho \neq 1$, in quanto l'elemento di matrice associato alla reazione reversibile

$$\xi_{\mu i} = \begin{cases} b_{\mu i} - \varrho a_{\mu i}, & \text{se } s_i > 0 \text{ e} \\ a_{\mu i} - \varrho b_{\mu i}, & \text{se } s_i < 0, \end{cases}$$

dipende da ϱ in maniera asimmetrica nei due casi (una dimostrazione rigorosa si trova in appendice a [8]).[3]

[3]Ricodiamo che per insieme convesso si intende un insieme che qualora contenga s_1 e s_2, conterrà anche una qualunque combinazione $[\alpha s_1 + (1 - \alpha)s_2]$, dove $0 < \alpha < 1$.

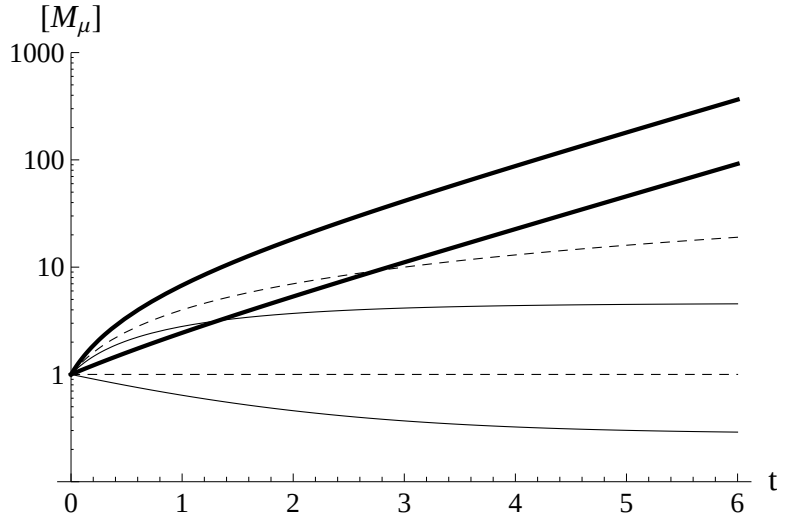


Figura 3.3: Diverse evoluzioni temporali per $[\mathrm{M}_\mu]$. Le due linee spesse in alto hanno $\varrho > 1$ e differenti c_μ; la linee tratteggiate $\varrho = 1$ e $c_\mu > 0$ (la più in alto, crescita lineare) o $c_\mu = 0$; le linee sottili $\varrho < 1$ e differenti c_μ.

Se, al contrario, vogliamo mantenere la convessità, bisogna spezzare in due le reazioni reversibili in modo da utilizzare esclusivamente flussi positivi (come nella versione originale del problema).

Come spesso accade nella modellizzazione fisica è utile associare la soluzione del problema al minimo assoluto di una funzione energia, la convessità è allora una proprietà desiderabile in quanto garantisce che ogni minimo relativo sia anche un minimo assoluto e che questi formino un insieme connesso.
Adotteremo quindi la seconda soluzione, malgrado essa stessa non sia esente da controindicazioni nel caso di configurazioni particolari che emergeranno durante lo studio del modello tramite un algoritmo numerico.

Le concentrazioni e il tasso di espansione Nella descrizione della FBA, i c_μ rappresentavano la varazione singoli metaboliti nell'unità di tempo, cioè $c_\mu = d[M_\mu]/dt$. In VN non è esattamente così in quando, se definiamo i tassi di crescita

$$c_\mu = \sum_i (b_{i\mu} - \varrho a_{i\mu}) s_i,$$

questo parallelismo vale solamente a $\varrho = 1$. La derivata effettiva della concentrazione del metabolita M_μ è

$$\frac{d[\mathrm{M}_\mu]}{dt} = \varrho^t \left[c_\mu + (\varrho - 1) \sum_i a_{i\mu} s_i \right] = \varrho^t \, \tilde{c}_\mu(\varrho) \log \varrho,$$

$$\tilde{c}_\mu(\varrho) = \left[c_\mu + (\varrho - 1) \sum_i a_{i\mu} s_i \right] / \log \varrho.$$

Possiamo quindi integrare questa formula per trovare come variano le concentrazioni nel tempo,

$$\begin{aligned} [\mathrm{M}_\mu] &= \int dt \; \varrho^t \, \tilde{c}_\mu(\varrho) \log \varrho \\ &= [\mathrm{M}_\mu]_0 + \begin{cases} \tilde{c}_\mu(\varrho)(\varrho^t - 1), & \varrho \neq 1, \\ c_\mu \, t, & \varrho = 1. \end{cases} \end{aligned}$$

Visto che la condizione stabilità 3.2 garantisce che per ogni metabolita $c_\mu \geq 0$, a seconda del valore di ϱ sono possibili tre regimi:

- $\varrho > 1$, tutte le concentrazioni crescono esponenzialmente nel tempo;
- $\varrho < 1$, tutte le concentrazioni si stabilizzano a $[\mathrm{M}_\mu]_0 - \tilde{c}_\mu(\varrho)$;
- $\varrho = 1$, i metaboliti con $c_\mu = 0$ rimangono costanti mentre quelli con $c_\mu > 0$ crescono linearmente

Possiamo quindi capire come ϱ sia un tasso di crescita del sistema mentre i c_μ rendano conto delle variazione locali rispetto al *trend* globale.

Anche se la rete è in uno stato stazionario (tutti i flussi sono costanti) non è garantito che a $\varrho = 1$ tutti i c_μ siano nulli (estrapolando l'andamento lineare nella figura 3.4 si ottiene $\langle\langle c_\mu \rangle\rangle \sim 0.001$); possiamo suggerire tre spiegazioni plausibili per questa situazione:

- i metaboliti in eccesso sono un prodotto di scarto delle reazioni;
- alcuni metaboliti vengono sequestrati per contribuire alla produzione di biomassa cellulare

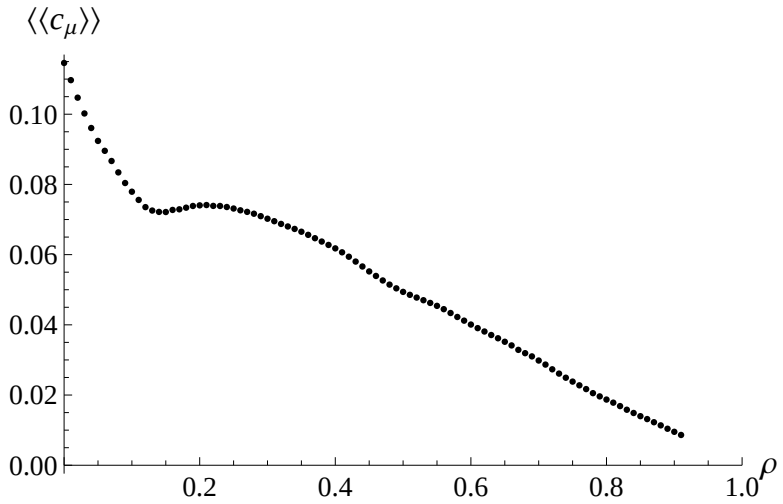


Figura 3.4: All'aumentare di ϱ i c_μ tendono a zero; la doppia parentesi indica una media sui metaboliti e sulle diverse soluzioni — Risultati da una simulazione sulla rete metabolica di *E.coli* (P=631, N=1057) per $0 < \varrho < 0.91$ e un *ensemble* di 200 soluzioni.

- il modello di metabolismo che stiamo utilizzando è incompleto.

Ulteriori analisi sono necessarie per chiarire se e quanto queste risposte diano conto dei possibili *surplus* di concentrazione.

Formulazione del problema

Sarebbe possibile portare avanti il parallelismo col modello originale ed introdurre l'analogo al vettore dei prezzi. Il significato biologico di questo oggetto può essere ricercato nella variazione in energia libera per metabolita durante il processo: una reazione che comporta un aumento netto dell'energia libera deve avere un flusso associato nullo. Sono attualmente in corso ulteriori lavori per chiarire le implicazioni di questa scelta.

Per il momento, attenendoci agli elementi sopra introdotti, possiamo porre il problema in questi termini:

> *qual è il più alto tasso di espansione ϱ compatibile con i vincoli di una rete metabolica data?*

Il tasso di espansione può essere scritto in questo modo

$$\varrho = \min_\mu \frac{\sum_i b_{i\mu} s_i}{\sum_i a_{i\mu} s_i},$$

quindi in termini matematici stiamo cercando

$$\bar{\varrho} = \max_{\vec{s}} \min_{\mu} \frac{\sum_i b_{i\mu} s_i}{\sum_i a_{i\mu} s_i},$$
$$s_i > 0, \quad \forall i.$$

Lo stato di crescita ottimale per il batterio *Escherichia coli* è stato analizzato in [12] e, significativamente, è stato fissato a $\varrho = 0.999 \pm 0.001$. Questo risultato è compatibile con l'assunzione che i flussi siano constanti nel tempo e vale in generale per le reti metaboliche.[4]
In presenza di uno stato stazionario i risultati del modello possono quindi essere usati come base di confronto con i dati sperimentali, come viene fatto con le predizioni della FBA. Bisogna però tenere conto di alcune importanti differenze:

- le differenti condizioni di stabilità ($c_\mu = 0$ o $c_\mu \geq 0$) conducono ad uno spazio delle soluzioni molto diverso, le soluzioni di FBA prima dell'ottimizzazione sono un sottoinsieme delle soluzioni di VN;
- FBA richiede l'imposizione dall'esterno di una funzione obbiettivo da massimizzare che seleziona una soluzione particolare mentre VN considera tutte le soluzioni trovate equivalenti.

Altre differenze emergeranno quando cercheremo di trovare le soluzioni a VN tramite un algoritmo numerico.

Topologia ed esempi

Per cercare di capire meglio il significato di questo modello costruiamo una rete molto semplice, senza alcuna pretesa di somiglianza biologica.

Ogni riga della matrice Ξ rappresenta un vettore-metabolita il cui prodotto scalare con il vettore-flussi deve essere maggiore di zero. Nella figura 3.5 è rappresentata una rete di cinque metaboliti collegati da due reazioni. Le soluzioni del modello sono allora i vettori-flussi $\vec{s} \geq 0$ che formano un angolo minore di $\pi/2$ con ognuno degli $\vec{\xi}_\mu$. È facile accorgersi che queste soluzioni esistono solo nel caso si possa trovare almeno un piano che lasci tutti i vettori $\vec{\xi}_\mu$ da una sola parte, i vettori dei flussi

[4] Come spiegheremo nel paragrafo successivo è dovuto alla presenza di gruppi conservati

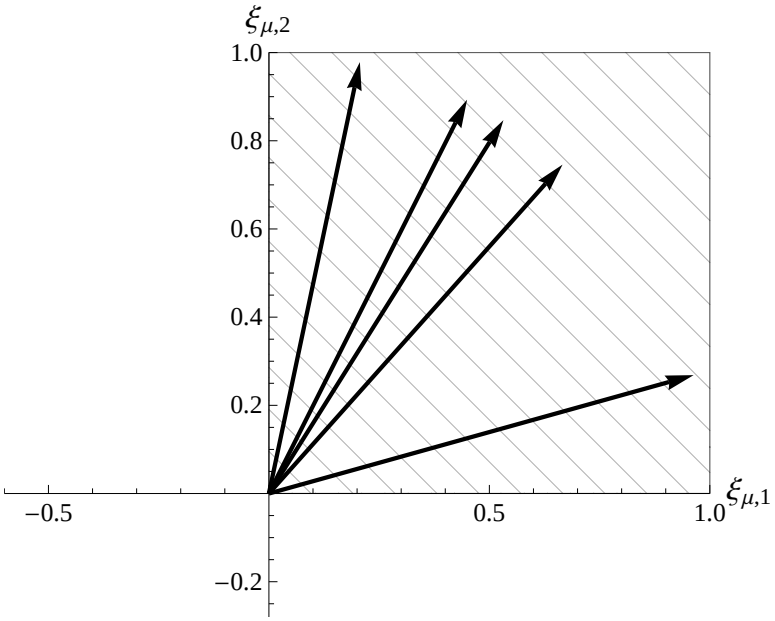


Figura 3.5: Rete metabolica con cinque metaboliti e due reazioni. A $\varrho = 0$ non ci sono coefficienti stechiometrici negativi: tutto il sottospazio $s_i > 0$ è soluzione del problema.

accettabili sono allora i versori di questi piani. La condizione limite si ha quando tra due righe della matrice stechiometrica sussiste la relazione

$$\vec{\xi}_\mu + \vec{\xi}_\nu = 0, \tag{3.9}$$

cioè i due vettori hanno la medesima direzione ma verso opposto. Per questi due vettori, infatti, deve valere

$$\begin{cases} \vec{\xi}_\mu \cdot \vec{s} \geq \vec{0} \\ \vec{\xi}_\nu \cdot \vec{s} \geq \vec{0} \end{cases} \quad \Rightarrow \quad \vec{\xi}_\mu \cdot \vec{s} = \vec{\xi}_\nu \cdot \vec{s} = 0$$

e da due condizioni lasche (diseguaglianze) passiamo ad una condizione rigida che lascia una sola possibilità per il rapporto s_2/s_1; inoltre il rango della matrice Ξ si riduce di un'unità, in quanto due righe sono linearmente dipendenti.

Nell'esempio abbiamo posto $\varrho = 0$, eliminando di fatto la parte negativa dell'equazione 3.2. L'angolo massimo formato da due $\vec{\xi}_\mu$ non può essere maggiore di 45 gradi ed è quindi possibile trovare sempre una soluzione.
Dal momento in cui ϱ è maggiore di zero le componenti negative dei vettori cominciano a crescere e il sottospazio delle soluzioni a restringersi di conseguenza – figura 3.6. Il massimo valore $\bar{\varrho}$ coincide proprio con la situazione limite in cui due

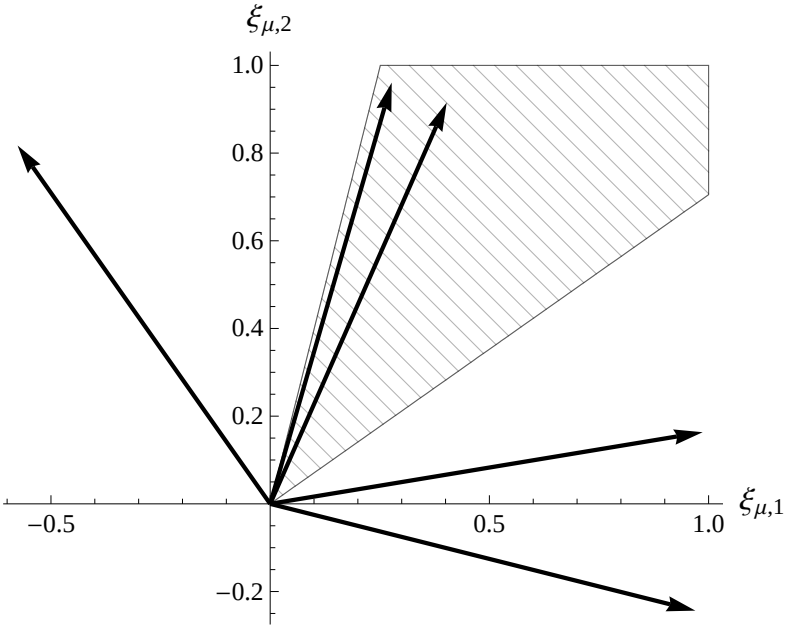


Figura 3.6: Quando $\varrho \neq 0$ diventa più difficile soddisfare tutte le condizioni: lo spazio delle soluzioni si restringe.

vettori $\vec{\xi}_\mu$ formano un angolo di π; a quel punto il sottospazio collassa in un'unica direzione, il rapporto s_2/s_1 è fissato e la soluzione è univoca a meno di una costante moltiplicativa.

La condizione limite tra i due metaboliti dell'equazione 3.9 può essere generalizzata al caso di reti con più reazioni e più metaboliti nel concetto di *gruppo conservato* (G) per cui vale

$$\sum_{\mu \in G} \vec{b}_\mu - \vec{a}_\mu = \vec{0}.$$

La presenza di gruppi conservati è una caratteristica intrinseca alla natura biologica della rete, in quanto discende dalla conservazione della massa nelle reazioni chimiche, e comporta diverse conseguenze, come si può osservare combinando la formula precedente con la 3.2:

$$\begin{aligned}\sum_{\mu \in G} c_\mu = \sum_{\mu \in G} \sum_i (b_{i\mu} - \varrho a_{i\mu}) s_i = \sum_i \sum_{\mu \in G} (b_{i\mu} - \varrho a_{i\mu}) s_i = \\ = \sum_i \left[\sum_{\mu \in G} (b_{i\mu} - a_{i\mu}) \right] s_i + (1 - \varrho) \sum_i \sum_{\mu \in G} a_{i\mu} s_i =. \\ = (1 - \varrho) \sum_i \sum_{\mu \in G} a_{i\mu} s_i\end{aligned}$$

Visto che sia i tassi di crescita sia i flussi sono quantità positive, a $\varrho = 1$ solo due situazioni sono possibili:

1. tutti i flussi delle reazioni che coinvolgono un metabolita facente parte di un gruppo conservato sono nulli;

2. i tassi di crescita di tutti i metaboliti di un gruppo conservato sono nulli.

Alcuni metaboliti molto utilizzati dal metabolismo (AMP, ADP e ATP; NAD e NADH) fanno parte di gruppi conservati: la creazioni di uno avviene sempre contemporaneamente alla distruzione dell'altro. Dobbiamo quindi escludere il primo caso e concludere che tutti i c_μ dei componenti di un gruppo conservato sono nulli.

Un'altra conseguenza evidente è che non può mai essere $\bar{\varrho} > 1$: il tasso di espansione massimo per una rete metabolica si ha in condizioni di stazionarietà. Inoltre, a $\varrho = 1$ il rango della matrice Ξ passa da P a $P' < P$. Questo aumento dei grati di libertà garantisce maggior robustezza alle soluzioni trovate[3].

3.4 Alla ricerca di un algoritmo numerico

Se non vogliamo risolvere il modello analiticamente – come vedremo in seguito non è un compito facile – serve disporre di un valido algoritmo numerico per campionare efficacemente lo spazio delle soluzioni. I lavori passati e presenti del nostro gruppo di ricerca in quest'ambito ci hanno spinto ad utilizzare l'algoritmo `minover`$^+$, che si ispira agli studi sulle reti neurali. Riteniamo sia utile quindi presentare brevemente alcune nozioni legate a questo settore di ricerca.

Relazione con le reti neurali

Il percettrone Per come lo abbiamo formulato, il problema di Von Neumann assomiglia ad un problema di classificazione molto studiato nell'ambito delle reti neurali: l'apprendimento di un *percettrone*. Il percettrone è un neurone artificiale. Nasce come schematizzazione della sua controparte biologica e come questa riceve in ingresso dei segnali binari ed in base a questi produce un segnale in uscita.
Prendendo spunto dalla biologia, i segnali in ingresso sono equiparati a delle scariche elettriche che alterano il potenziale

di membrana del neurone, alzandolo o abbassandolo in funzione del tipo di collegamento (sinapsi eccitatorie o inibitorie). Matematicamente questo equivale ad effettuare una somma pesata degli input, $\sum_{i=1}^{N} x_i w_i$; $x_i \in \{-1, 1\}$ sono gli input e $w_i \in \mathbb{R}$ i pesi sinaptici.
Se il potenziale così influenzato supera una certa soglia V_0 il neurone "spara", emette cioè un segnale

$$y = \mathrm{sgn}(\vec{x} \cdot \vec{w} - V_0).$$

Questo tipo di neurone è un classificatore binario lineare: lo spazio degli input possibili è uno spazio affine che viene diviso in due categorie da un iperpiano separatore.
L'apprendimento del percettrone consiste allora nel trovare i pesi sinaptici che diano in ogni situazione l'output desiderato, cioè un vettore $\vec{w}$ che identifichi il corretto iperpiano separatore.

In questa chiave di lettura il nostro problema equivale a classificare assieme tutti i vettori $\vec{\xi}_\mu$, tenendoli dalla stessa parte dell'iperpiano identificato dal vettore $\vec{s}$,

$$c_\mu \geq 0 \Leftrightarrow \mathrm{sgn}(\vec{\xi}_\mu \cdot \vec{s}) \geq 0.$$

Le differenze principali tra i due modelli sono costituiti dal vincolo di positività per le componenti del vettore $\vec{s}$ e da un input continuo anziché discreto.

L'apprendimento Tutta l'informazione che permette di eseguire correttamente la classificazione è codificata nei pesi sinaptici, questi pesi vanno aggiustati presentando al percettrone una serie di input di esempio di cui si conosce la giusta classificazione in un processo noto come *apprendimento*. Uno degli algoritmi più semplici è il seguente:

1. i pesi vengono inizializzati, solitamente a valori nulli;

2. si presenta al neurone l'esempio $\vec{x}_\mu$ per cui si desidera un output y^*_μ.

3. viene calcolato l'output $y_\mu = \mathrm{sgn}(\vec{x}_\mu \cdot \vec{w} - V_0)$.

4. i pesi vengono aggiornati secondo la regola:

$$w_i' = w_i + \Delta w_i = w_i + \begin{cases} 0, & y_\mu = y_\mu^*, \\ \alpha\, y_\mu^* x_{i\mu}, & y_\mu \neq y_\mu^*; \end{cases}$$

5. si procede dal punto 2 sino a che tutti gli esempi sono classificati correttamente.

La regola al punto 4 è ispirata alle ricerche di Donald Hebb, che propose una regola di aggiornamento dei pesi per cui le sinapsi di neuroni attivi contemporaneamente si rafforzano, mentre si indeboliscono se uno dei due è attivo e l'altro quiescente. In quest'ottica è chiaro perché abbiamo scelto gli input dall'insieme $\{-1, 1\}$, malgrado la somiglianza biologica imporebbe $x_i \in \{0, 1\}$; la mappatura $0 \to -1$ consente infatti una correlazione positiva anche tra due neuroni entrambi quiescenti e una correlazione negativa tra due neuroni attivi alternativamente.
Il parametro α è compreso tra zero e uno e regola la velocità dell'apprendimento.

Per semplicità possiamo riscrivere Δw in modo che abbia un'unica forma per tutti i casi

$$\Delta w_i = \alpha(1 - y_\mu^* y_\mu) x_{i\mu},$$

visto che la parentesi si annulla per $y_\mu = y_\mu^*$, o in modo più compatto

$$\begin{aligned} \Delta\vec{w} &= \alpha(1 - y_\mu^* y_\mu)\vec{x}_\mu \\ &= \alpha\delta_\mu \vec{x}_\mu. \end{aligned}$$

3.5 Minover$^+$

Ora che abbiamo una formulazione matematicamente trattabile per la regola di apprendimento possiamo cercare di adattarla a VN. Come detto nel paragrafo precedente, le differenze risiedono negli input continui e nei flussi positivi.
Per tener conto del primo problema dobbiamo modificare la definizione di δ_μ in modo che qualunque tasso di crescita positivo corrisponda ad una buona classificazione,

$$\delta_\mu = \vartheta(-c_\mu),$$

la funzione $\vartheta(x)$ vale uno se $x \geq 0$ e zero altrimenti. Inoltre per preservare i flussi positivi dobbiamo impedire che Δs_i sia minore di $-s_i$. Mettendo assieme queste due condizioni la nuova regola di apprendimento diventa

$$\vec{s}' = \min\left[0, \vec{s} + \alpha\, \vartheta(-c_\mu)\vec{\xi}_\mu\right], \tag{3.10}$$

in cui la funzione minimo scorre su tutte le componenti del vettore.

Durante l'esecuzione dell'algoritmo, va scelto ad ogni passo un vettore $\vec{\xi}_\mu$ per effettuare l'aggiornamento dei flussi. Questo procedimento può essere molto lento, specialmente nelle fasi finali dell'apprendimento dove solo pochi metaboliti anno tassi di crescita negativi. Per ovviare a questo problema vengono prima calcolati tutti i tassi di crescita, quindi si sceglie per per effettuare l'aggiornamento il metabolita di indice $\mu_0 = \arg\min_\mu c_\mu$; nel caso due metaboliti abbiano lo stesso c_μ se ne sceglie uno casualmente.

La regola di aggiornamento 3.10 e il metodo di selezione appena descritto costituiscono il cuore dell'algroitmo `minover`$^+$ (*minimum-overlap*, flussi positivi), che utilizzeremo in questa tesi [4, 11].

Riepilogando, fissato un valore per ϱ si procede come segue.

1. In mancanza di informazione aggiuntiva, i flussi di partenza vengono estratti da una distribuzione uniforme.

2. Vengono calcolati i tassi di crescita c_μ secondo la formula 3.2.

3. Si cerca la concentrazione minore, cioè la condizione maggiormente violata, $\mu_0 = \arg\min_\mu c_\mu$.

   (a) Se $c_{\mu_0} < 0$, i flussi vengono aggiornati secondo la regola
   $s_i' = \max(0, s_i + \alpha\, \xi_{i\mu_0})$, scegliendo casualmente nel caso più metaboliti abbiano $c_\mu = c_{\mu_0}$. Si torna al punto 2.

   (b) Se $c_{\mu_0} > 0$ significa che si è trovata una soluzione: l'algoritmo esce dal ciclo e restituisce il vettore $\vec{s}$.

Se lo scopo è raggiungere la condizione di maggior verosimiglianza biologica, cioè $\varrho \sim 1$, un metodo efficiente è partire da valori di ϱ più bassi, per i quali lo spazio delle soluzioni è più grande ed è quindi più facile trovare una soluzione in tempi brevi; si procede allora ad aumentare il tasso di espansione della rete sino a raggiungere il valore desiderato.

Cerchiamo di capire cosa comporta la scelta di questa regola di aggiornamento introducendo una funzione di errore (l'energia del sistema) che misuri il numero di condizioni violate; in modo che i minimi di questa funzione corrispondano alle configurazioni di flussi che soddisfano 3.2. Una scelta potrebbe essere

$$E(\vec{s}, \varrho) = -\sum_{\mu} \delta_{\mu} c_{\mu} = -\sum_{\mu} \sum_{i} \delta_{\mu} (b_{i\mu} - \varrho a_{i\mu}) s_i.$$

Il coefficiente $\delta_{\mu} = \vartheta(-c_{\mu})$ garantisce la funzione E sia definita positiva, che tutti i tassi di crescita positivi abbiano il medesimo peso e che $\min_{\vec{s}} E = 0$ quando tutti i c_{μ} sono positivi o nulli. Se calcoliamo il gradiente di questa funzione,

$$\vec{\nabla}_{\vec{s}} E = -\sum_{\mu} \delta_{\mu} \vec{\xi}_{\mu},$$

possiamo notare che è molto simile alla regola di aggiornamento che abbiamo scelto per `minover`$^+$. La differenza – importante – è che una discesa lungo il gradiente di E si ottiene eseguendo l'aggiornamento su tutte le condizioni violate contemporaneamente, mentre in `minover`$^+$ ci si avvicina di volta in volta allo $\vec{\xi}_{\mu}$ relativo alla condizione più violata. In entrambi i casi comunque si raggiunge il minimo della funzione di errore e quindi ci si ferma, dato che i coefficienti δ_{μ} producono un unico minimo comune a tutte le soluzioni. La variabilità dell'algoritmo sta dunque tutta nella scelta delle condizioni iniziali, che portano a raggiungere punti diversi dello spazio delle soluzioni.[5]

Esempi Recuperiamo la nostra rete semplificata ($P = 5$, $N = 3$) per capire all'atto pratico cosa comporta far "apprendere" il vettore dei flussi. Nella figura 3.7 possiamo vedere come

[5] La funzione E così definita è continua ma non derivabile per valori dei flussi che portano a $c_{\mu} = 0$ per qualche μ, in quanto il coefficiente δ_{μ} passa da uno a zero in modo discontinuo e la sua derivata è infinita. Questo fatto non ha particolare rilevanza nel nostro esempio ma in una trattazione matematica va prestata particolare attenzione.

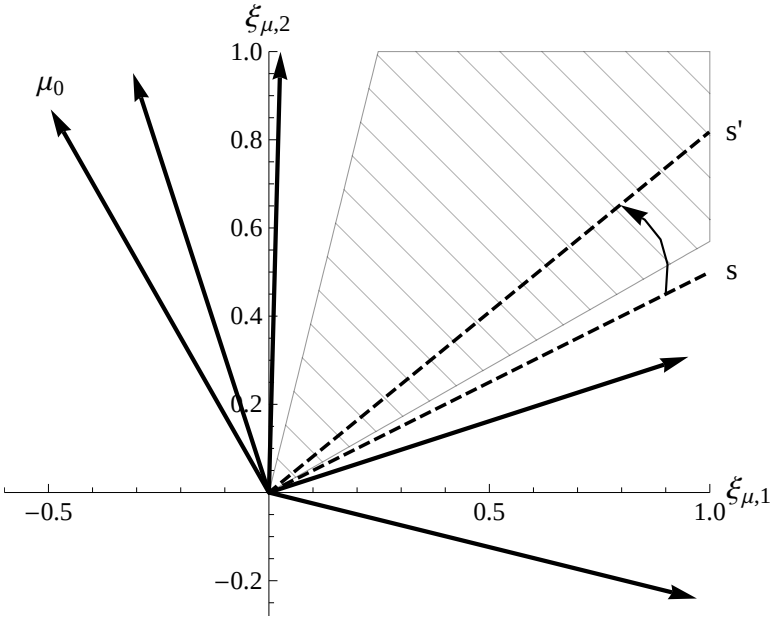


Figura 3.7: Il prodotto $\vec{\xi}_{\mu_0} \cdot \vec{s}$ non rispetta le condizioni di stabilità. Dopo la correzione, il nuovo vettore $\vec{s}' = \vec{s} + \alpha\,\vec{\xi}_{\mu_0}$ fa parte dello spazio delle soluzioni.

il metabolita M_{μ_0} abbia $c_{\mu_0} < 0$, infatti il vettore $\vec{s}$ forma con $\vec{\xi}_{\mu_0}$ un angolo maggiore di novanta gradi. L'applicazione della regola trasforma il flusso s_i in $s'_i = \min(0, s_i + \alpha\xi_{i\mu_0})$ e quindi il nuovo valore di c_{μ_0} è

$$c'_{\mu_0} = \sum_i \xi_{i\mu_0} s'_i = \sum_i (\xi_{i\mu_0} s_i + \alpha\xi^2_{i\mu_0}) = c_{\mu_0} + \alpha \sum_i \xi^2_{i\mu_0}.$$

Siccome α è una quantità positiva segue che $c'_{\mu_0} > c_{\mu_0}$: il vettore dei flussi ha ridotto l'angolo con $\vec{\xi}_{\mu_0}$. Se al passo successivo ci sarà ancora un metabolita mal classificato $\vec{s}$ si sposterà ancora, seguendo ogni volta la direzione della condizione più violata fino a formare un angolo minore di $\pi/2$ con tutti i vettori $\vec{\xi}_\mu$. La dimostrazione che qualora esista una soluzione il numero di passi richiesto per trovarla sarà finito si può trovare in [4].

La scelta della scala α è molto importante: con un passo troppo piccolo potrebbe volerci moltissimo tempo per arrivare al minimo della funzione di errore, d'altra parte un passo grande è un limite inferiore alla distanza dal minimo a cui si può arrivare. Già a questo punto possiamo intuire ciò che sarà osservato nel prossimo capitolo: il collasso lungo alcune di-

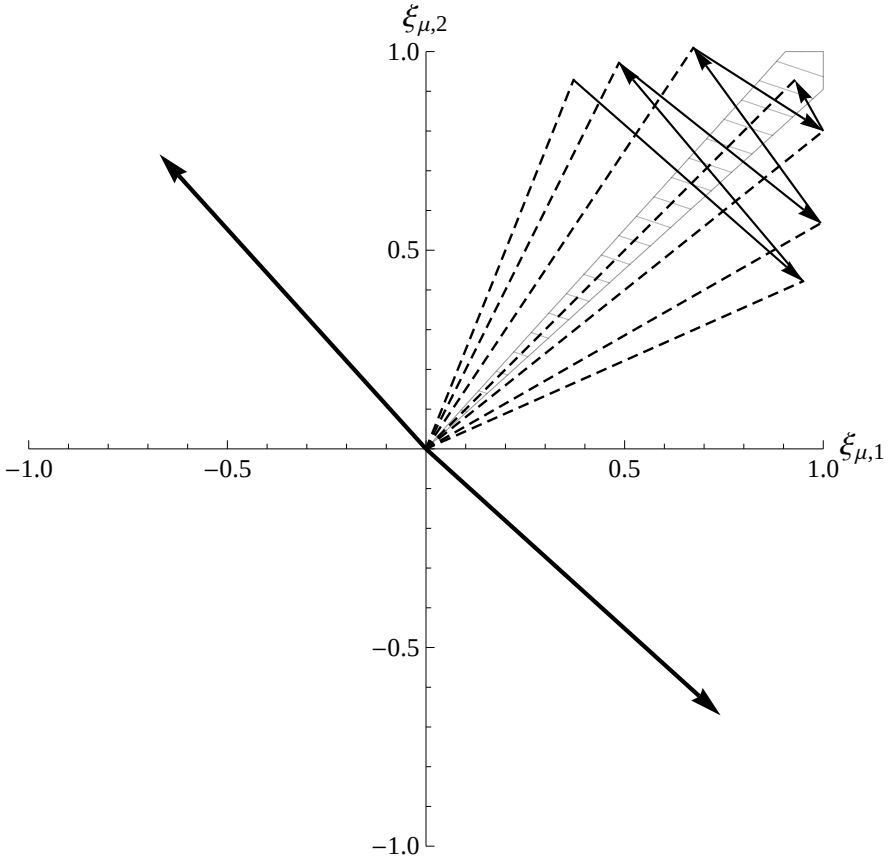


Figura 3.8: Schematizzazione dell'apprendimento in un caso particolarmente difficile per `minover`$^+$.

rezioni per $\varrho \to 1^-$ trasforma minimo da un intervallo a un punto e complica il lavoro dell'algoritmo.

Immaginiamo una situazione in cui tutte le condizioni siano soddisfatte eccetto due, relative a due metaboliti i cui coefficienti stechiometrici formino un angolo di quasi 90 gradi. L'algoritmo tenterà con tutte le sue forze di portare $\vec{s}$ esattamente sul iperpiano identificato dalla direzione dei due vettori (figura 3.8), ma ad ogni iterazione quello finirà da una parte o dall'altra fino a che α non sarà sufficientemente piccolo. Se nel frattempo le altre condizioni non sono state violate, le reazioni che coinvolgono questi due metaboliti cresceranno ad ogni passo (i Δs_i negativi sono ϱ volte più piccoli di quelli positivi) e alla fine avranno un peso sulla norma sproporzionato rispetto alla loro importanza effettiva.

Considerazioni come questa ci hanno spinto ad effettuare un'analisi più precisa del comportamento di `minover`$^+$: sarà l'argomento del prossimo capitolo.

Parte II

Analisi

Capitolo 4

Algoritmo di Von Neumann e `minover`$^+$ su reti semplici

Come suggerisce il titolo, lo scopo di questo capitolo è assumere un maggior controllo sui risultati di `minover`$^+$ lavorando su reti semplici — sebbene molto artificiose — per le quali sia possibile trovare la soluzione analitica del modello di Von Neumann. Necessariamente la notazione si farà più pesante ma cercheremo di ridurre al minimo le parti non strettamente necessarie alla comprensione.

Le condizioni di stabilità $c_\mu \geq 0$ (conservazione della massa) possono essere riscritte utilizzando la funzione ϑ di Heaviside

$$\vartheta\left(\sum_i (b_{i\mu} - \varrho a_{i\mu}) s_i\right)$$

che nel caso molto semplice di un metabolita prodotto da una sola reazione e consumato da una sola reazione diventa

$$\vartheta\left(b_{i\mu} s_i - \varrho a_{j\mu} s_j\right).$$

Questo sarà il mattone fondamentale che utilizzeremo per costruire le nostre reti.[1]

In mancanza di informazioni sulla forma dello spazio delle soluzioni è ragionevole supporre una metrica piatta nel calcolo degli integrali, cioè assegnare ad ogni configurazione dei flussi che è soluzione del modello la medesima probabilità:

$$\mu(\vec{s}) = \begin{cases} 1/V, & \text{se} \quad \Xi \cdot \vec{s} \geq \vec{0}, \\ 0, & \text{se} \quad \Xi \cdot \vec{s} < \vec{0}. \end{cases}$$

[1] La funzione $\vartheta(x)$ vale 1 se $x \geq 0$ e 0 se $x < 0$.

Nelle varie configurazioni cercheremo di calcolare il volume dello spazio delle soluzioni,

$$\int_\Omega d\mu(\vec{s}) = \int_0^\infty Ds\, \mu(\vec{s}),$$

dove con Ds intendiamo l'elemento di volume $\prod_i ds_i$, la distribuzione di probabilità marginale dei singoli flussi,

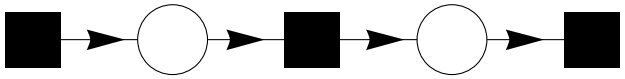
$$P(s_i = x) = \int_0^\infty \prod_{j\neq i} ds_j\, \mu(s_1, \dots, s_{i-1}, x, s_{i+1}, \dots),$$

e per finire il valore medio dei flussi,

$$\langle s_i \rangle = \int_0^\infty dx\, x\, P(s_i = x).$$

L'obiettivo di questi conti è ottenere delle quantità che possano essere confrontate facilmente con i dati provenienti dalle simulazioni.

4.1 Catene linerari



Il primo caso in esame è una semplice catena di reazioni in cui un metabolita viene consumato per produrne un altro (nel grafo di esempio qui sopra ogni quadrato nero rappresenta una reazione mentre i cerchi bianchi sono i metaboliti).
La conservazione della massa connette le reazioni due a due imponendo una gerarchia tra flussi del tipo $s_i > \varrho s_{i+1}$, quindi tanto più stringente quanto più ϱ tende ad uno.[2]

Volume Se imponiamo un limite massimo e minimo ai flussi estremi ($s_0 < A$ e $s_n > B$), l'integrale sullo spazio delle soluzioni è costituito da un prodotto dei mattoncini elementari e da due ulteriori ϑ che tengono conto delle condizioni sul bordo,

$$V_B^A(n, \varrho) = \int_0^\infty Ds \prod_{i=1}^{n-1} \vartheta(s_i - \varrho s_{i+1})\vartheta(A - s_1)\vartheta(s_n - B). \quad (4.1)$$

[2]Qui e nel seguito terremo i coefficienti stechiometrici identicamente uguali a uno: $a_{i\mu} = b_{i\mu} = 1$.

Questo primo integrale può essere eseguito facilmente in modo analitico partendo da uno degli estremi e aggiustando gli estremi di integrazione per far sparire tutte le ϑ. Il risultato è

$$V_B^A(n,\varrho) = \frac{1}{n!}\prod_{i=0}^{n-1}(\varrho^{-i}A - \varrho^{n-i-1}B)$$
$$= \frac{(\varrho^{\frac{1-n}{2}}A - \varrho^{\frac{n-1}{2}}B)^n}{n!} = \frac{[A(\varrho)-B(\varrho)]^n}{n!}.$$

Il volume dello spazio delle soluzioni è quindi un ipertriangolo in cui gli estremi dei lati sono dati dalle condizioni sui flussi: $1/\varrho$ volte più grandi dei successivi, ϱ volte più piccoli dei precedenti.

Nel caso biologico ($\varrho = 1$) questo triangolo è isoscele mentre per $\varrho \to 0$ si ha che $B(\varrho) \to 0$ e $A(\varrho) \to \infty$.

Marginale Per calcolare la probabilità marginale di un singolo flusso l'integrale (4.1) va eseguito su tutte le variabili fuorché quella relativa al flusso interessato. Nel caso di $0 < j < n$

$$P(s_j = x) = \int_0^\infty \frac{Ds}{ds_j}\prod_{i=1}^{n-1}(1-\delta_{i,j}-\delta_{i,j-1})\vartheta(s_i - \varrho s_{i+1})\times$$
$$\times\,\vartheta(A-s_1)\vartheta(s_n - B)\vartheta(s_{j-1}-\varrho x)\vartheta(x-\varrho s_{j+1}),$$

che, spezzando la produttoria all'altezza di j, si può scrivere come

$$P(s_i = x) = \int_0^\infty \prod_{i=1}^{i-1} ds_i \prod_{i=1}^{i-2}\vartheta(s_i-\varrho s_{i+1})\vartheta(A-s_1)\vartheta(s_{i-1}-\varrho x)\times$$
$$\times \int_0^\infty \prod_{i=i+1}^{n} ds_i \prod_{i=i+1}^{n-1}\vartheta(s_i-\varrho s_{i+1})\vartheta(x-\varrho s_{i+1})\vartheta(s_n - B),$$
$$= V_{\varrho x}^A(i-1,\varrho)V_B^x(n-i,\varrho).$$

Volendo una formula più simmetrica in x possiamo ripristinare s_i e inserire un'identità nella forma dell'operatore $\partial_x \int_0^x ds_i$ davanti a entrambi gli integrali. In questo moto otteniamo

$$P(s_i = x)\,dx = -\frac{\partial_x V_x^A(i,\varrho)\partial_x V_B^x(n-i+1,\varrho)}{V_B^A(n,\varrho)}dx$$
$$= \frac{\varrho^{(i-1)(n-i+1)}n!}{(i-1)!\,(n-i)!}\frac{(A-x\varrho^{i-1})^{i-1}\,(x-B\varrho^{n-i})^{n-i}}{(A-B\varrho^{n-1})^n}dx,$$

dove abbiamo aggiunto la normalizzazione ed un meno per tenere conto del segno.

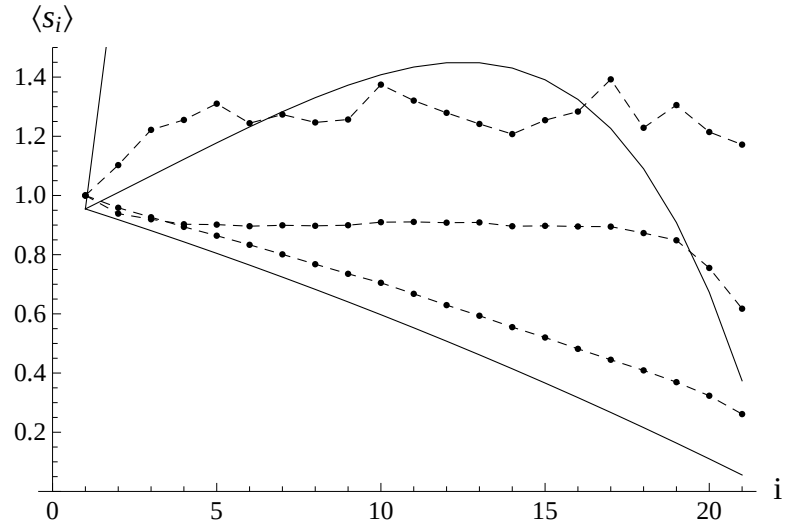


Figura 4.1: Flussi medi in una catena a $\varrho = 0.5, 0.9, 0.99$ e $n = 21$ — simulazione. Le linee continue indicano le soluzioni analitiche.

Media Calcolare il valor medio dei flussi a questo punto è immediato:

$$\langle s_i \rangle = \frac{(A\varrho^{1-i}(n-i+1) + Bi\varrho^{n-i})}{n+1}.$$

Fare il limite termodinamico ($n \to \infty$) a $\varrho \neq 1$ non ha significato perché tutti i flussi tranne il primo divergono, a $\varrho = 1$

$$\langle s_i \rangle = A - i\frac{(A-B)}{n+1} = q - i\,k.$$

Abbiamo così trovato due informazioni importanti: la condizione di massima entropia nella distribuzione dei flussi per una catena isolata è una retta e il coefficiente angolare di questa retta è legato all'inverso della lunghezza della catena ($k \propto 1/n$), quindi abbiamo una situazione in cui tutti i metaboliti vengono prodotti nella medesima quantità. Infatti

$$k = \frac{s_{i+1} - s_i}{1} = c_\mu, \quad \forall \mu.$$

Simulazione Osservando la figura 4.1, possiamo distinguere tre fasi nel comportamento di `minover`$^+$ su questa particolare realizzazione della catena:

- una prima fase a bassi ϱ in cui non risente della tipologia specifica della rete ed i flussi non sono ordinati;

- una seconda fase, compresa all'incirca tra $0.7 - 0.9$, in cui tutti i flussi centrali vengono allineati e collocati tra il primo e l'ultimo;
- una fase finale per ϱ maggiore di 0.9 in cui si apprezza appieno l'azione dei vincoli che a questo punto sono quasi rigidi.

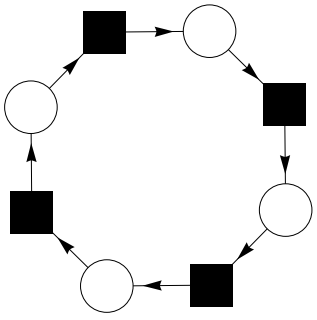
Riserviamo il raffronto con le soluzioni analitiche al paragrafo successivo.

Ricetta Avendo a disposizione il volume di una catena possiamo calcolare il volume ogni possibile rete in cui il nodo d'interazione sia una reazione (non funziona nel caso i due pezzi comunichino attraverso un metabolita). Il metodo è molto semplice:

1. si apre la catena ad un estremo con una derivata (ad esempio $\partial_x V_x^A$);
2. si apre il pezzo che si vuole agganciare (ad esempio $\partial_x V_B^x$);
3. si integra su i possibili valore della variabile di aggancio (nel nostro esempio si integra su $B\varrho^{n-m} < x < A\varrho^{1-m}$)

L'unica difficoltà — non da poco — è effettuare in modo analitico gli integrali.

4.2 Anello isolato



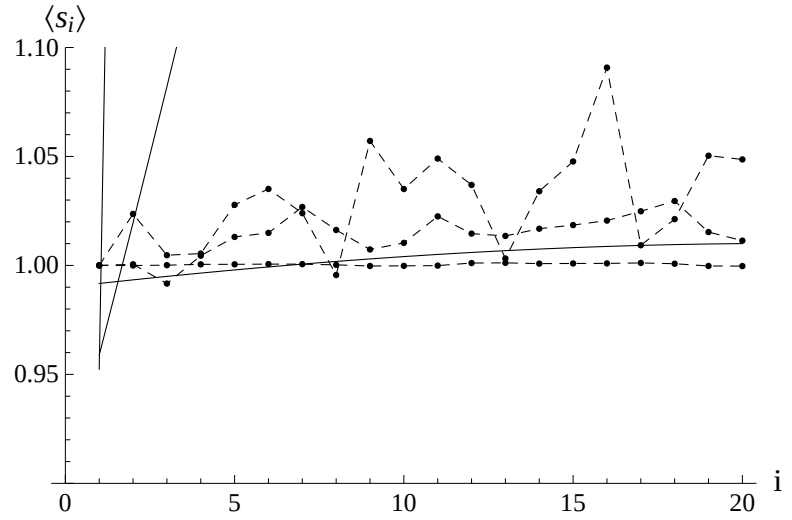


Figura 4.2: Flussi medi in un'anello a $\varrho = 0.5, 0.9, 0.99$ e $n = 20$ — simulazione. Le linee continue indicano le soluzioni analitiche.

Analizziamo ora il caso di un anello chiuso su sé stesso. A $\varrho = 1$ la conservazione della massa impone che tutti i flussi si stabilizzino ad un medesimo valore, quindi se imponiamo una norma L_p pari ad N ogni flusso varrà $N^{1-1/p}$.[3]

Per il conto analitico utilizziamo di nuovo un limite superiore ed inferiore ($A > s_1$ e $s_N > B$), in questo modo otteniamo il volume

$$V_B^A(n, \varrho) = \frac{\varrho^{-\frac{1}{2}(n-1)n} \left(A - A\varrho^{n-1}\right)^n}{n!}.$$

Come ci aspettiamo in un sistema isolato, la soluzione collassa ad un punto nel caso $\varrho = 1$ in quanto le condizioni di conservazione della massa impongono che tutte le reazioni procedano alla medesima velocità. La probabilità marginale è

$$\begin{aligned} P(s_i = x)\, dx = &- \frac{\partial_x V_x^A(n, \varrho) \partial_x V_A^x(n, \varrho)}{V_A^A(n, \varrho)} = \\ &= \frac{n! \left(A - A\varrho^{n-1}\right)^{-n} \varrho^{-(i-1)(i-n)}}{\Gamma(i)(n-i)!} \times \\ &\quad \times \left(A\varrho - x\varrho^i\right)^{i-1} \left(x - A\varrho^{n-i}\right)^{n-i}, \end{aligned}$$

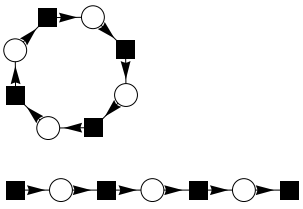
dove Γ e la funzione Gamma di Eulero.

Come atteso, sia analiticamente sia nella simulazione i flussi tendono ad equalizzarsi al crescere di ϱ (figura 4.2). A bassi

[3] La norma L_p è pari a $\sqrt[p]{\sum_{i=1}^N s_i^p}$.

ϱ, invece, l'algoritmo non esplora appieno lo spazio delle soluzioni: alcuni flussi sono divergenti mentre `minover`$^+$ incomincia l'esplorazione da una distribuzione uniforme, quindi si ferma non appena trova una soluzione, selezionando solo i flussi nella parte bassa della distribuzione. In ogni caso questo problema interessa principalmente la zona distante dalla condizione stazionaria ($\varrho = 1$) interessante dal punto di vista biologico.

4.3 Anello e catena



Questo sistema è composto dalla catena e l'anello visti precedentemente presenti come parti sconnesse di una rete. In mancanza di comunicazione diretta l'unica interazione tra i due pezzi avviene imponendo una condizione di normalizzazione. L'anello tende verso una misura nulla, quindi la massima entropia si ha quando tutta la norma è utilizzata dalla catena. Per come è congeniata la regola di aggiornamento invece, `minover`$^+$ ha bisogno di un gran numero di passi per equilibrare i flussi dell'anello, che crescono costantemente a scapito di una catena facilmente ordinabile. Alla fine della simulazione la maggior parte della norma proviene dall'anello.

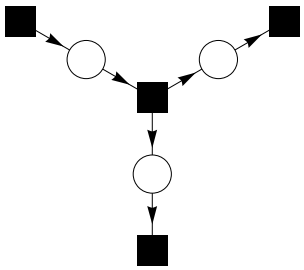
Questa è senza dubbio la situazione peggiore per `minover`$^+$ ma anche la più lontana dalla realtà. Non ha quindi grosse implicazioni nell'utilizzo dell'algoritmo per lo studio di reti biologiche, tipicamente molto connesse.

4.4 Trivio

Analizziamo ora una catena lunga n_1 collegata ad altre due di lunghezza n_2 ed n_3, chiameremo N la somma di queste lunghezze. Ci sono due modi di effettuare l'aggancio: un metabolita che viene usato come substrato da due reazioni o una

reazione che produce due metaboliti. Per semplicità di calcolo cominceremo da quest'ultimo caso.

Incrocio su una reazione



Volume Effettuare il conto analitico diventa molto complicato per questo ci limitiamo al caso $\varrho = 1$. Il volume dello spazio delle soluzioni è

$$V_B^A(n_1, n_2, n_3) = \frac{(n_2 + n_3)!\,(A - B)^N}{n_2!\,n_3!\,N!}.$$

Marginale In questo caso i flussi della prima catena seguono la distribuzione

$$P(s_i = x)\,dx = \frac{N!}{\Gamma(i)\,(N-i)!}\,x^{N-i}(1-x)^{i-1}dx,$$

mentre per la coda lunga n_2 si ha che

$$P(s_i = x)\,dx = \frac{n_2!\,N!}{(n_2+n_3)!\Gamma\,(-i+n_2+1)}\,{}_2\tilde{F}_1\left(i, -n_3; i+m; \frac{x-1}{x}\right) \times \\ \times\,(1-x)^{i+m-1}x^{-i+n_2+n_3}dx;$$

l'ultimo caso si ottiene tramite le sostituzioni $n_2 \to n_3$ e $n_3 \to n_2$.[4]

[4] ${}_2\tilde{F}_1(a, b; c; z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{z^k}{k!}$ è la funzione ipergeometrica normalizzata.

Media Considerando i flussi in sequenza, il valore medio si può scrivere

$$\langle s_i \rangle = \begin{cases} A - \frac{i}{N+1}(A-B) & 0 < i \leq n_1, \\ \langle s_{n_1} \rangle - \frac{(i-n_1)(n_2+n_3+1)}{(n_2+1)(N+1)}(A-B) & n_1 < i \leq n_1+n_2, \\ \langle s_{n_1} \rangle - \frac{(i-n_1-n_2)(n_2+n_3+1)}{(n_3+1)(N+1)}(A-B) & n_1+n_2 < i \leq N. \end{cases}$$

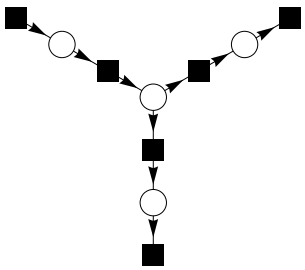
L'elemento importante è rappresentato dalla pendenza delle rette. I flussi iniziali diminuiscono ad ogni passaggio di $1/N$ e questo significa che in questa parte la rete non risente della struttura successiva ma solo del numero di flussi che ha davanti: si comporta allo stesso modo sia quando dopo l'intersezione ci sono due pezzi sia quando è presente una catena singola lunga quanto la somma di questi; il suo coefficiente angolare è infatti $1/(n_1+n_2+n_3)$.

Gli altri due pezzi partono dallo stesso punto ($(A-B) - n_1k_1$) e devono coprire a passi costanti la distanza rimanente. Questo si ottiene imponendo

$$n_2k_2 = (A-B) - n_1k_1,$$
$$n_3k_3 = n_2k_2.$$

Risolvendo si trovano esattamente i valori precedenti.[5]

Incrocio su un metabolita



Un metabolita utilizzato da due reazioni produce una struttura simile ma dalle proprietà abbastanza diverse. Nel caso

[5] I vari +1 che compaiono nelle formule sono dovuti al fatto di non aver fissato gli estremi ma imposto una maggiorazione e una minorazione. I +1 rappresentano lo spazio di variazione dei flussi più esterni.

precedente una reazione produceva due metaboliti nella medesima quantità, fissando quindi un punto di partenza comune per le successive reazioni. In questo caso un singolo metabolita viene usato in maniera concorrente da due processi, la condizione sul punto di aggancio è quindi che la somma dei flussi in uscita sia uguale al flusso in entrata

$$n_2 k_2 + n_3 k_3 = (A - B) - n_1 k_1.$$

Fissate le condizioni iniziali rimangono due parametri liberi: il valore dell'ultimo flusso prima dell'incrocio e la divisione di questo tra i due rami. Possiamo pensare di fissarli richiedendo che il pezzo iniziale continui a non dipendere dalla struttura successiva ma solamente dal numero delle reazioni, cioè $k_1 = 1/N$, e che la pendenza dei due pezzi finali sia tale da massimizzare l'entropia

$$k_2 = \max_x V_B^x \, V_B^{A - k_1 n_1 - x},$$

cioè

$$k_2 = k_3 = \frac{1}{N}.$$

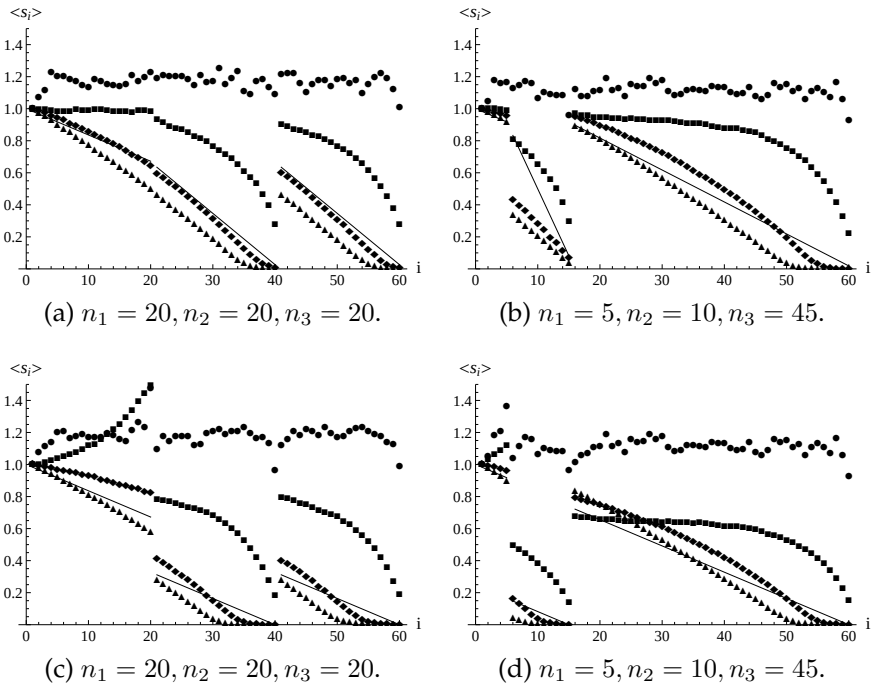
Questi risultati non sono ancora confermati da conti analitici fatti sul modello ma paiono cogliere bene le simulazioni. Considerando che le condizioni sul bordo sono identiche possiamo riassumere le differenze tra i due casi nelle condizioni di raccordo e di massima entropia mediante una forma che ne evidenzi la simmetria e complementarietà:

$$\begin{cases} n_2 k_2 = n_3 k_3 = (A - B) - n_1 k_1; \\ \dfrac{1}{k_2} + \dfrac{1}{k_3} = \dfrac{1}{k_1}, ; \end{cases}$$

$$\begin{cases} n_2 k_2 + n_3 k_3 = (A - B) - n_1 k_1; \\ \dfrac{1}{k_2} = \dfrac{1}{k_3} = \dfrac{1}{k_1}. \end{cases}$$

Simulazione Questo caso è leggermente più complesso ma sicuramente più fruttifero. Come sempre a bassi ϱ la topologia della rete perde totalmente di importanza; rafforzando il vincolo si comincia a vedere una struttura.

Nel caso di un incrocio sulla reazione c'è un buon accordo col risultato teorico (4.3 *a* e *b*) ma – come nel caso della catena e dell'anello – i pezzi soggetti a poche condizoni vengono ordinati subito e quindi perdono terreno nella spartizione della



(a) $n_1 = 20, n_2 = 20, n_3 = 20$. (b) $n_1 = 5, n_2 = 10, n_3 = 45$.

(c) $n_1 = 20, n_2 = 20, n_3 = 20$. (d) $n_1 = 5, n_2 = 10, n_3 = 45$.

Figura 4.3: Flussi medi in un trivio con incrocio su una reazione (*a* e *b*) e su un metabolita (*d* ed *e*) a $\varrho = 0.5, 0.9, 0.97, 0.99$ — le linee continue sono le soluzioni analitiche (provvisorie nei casi *c* e *d*).

norma rispetto alle parti più difficili da ordinare. Quest'effetto si vede principalmente nella parte finale delle catene e in quelle con un basso numero di reazioni.

Nell'altro caso un singolo metabolita è in competizioni tra due reazioni. Finché i vincoli laschi lo permettono (figura 4.3 *c*), `minover`$^+$ cerca di massimizzarne la produzione in modo da poter ordinare facilmente i flussi successivi.[6] Ad alto ϱ questo è impossibile ed anche in questa rete i pezzi molto corti sono molto penalizzati (figura 4.3 *c*).

Nell'utilizzare `minover`$^+$ su topologie più biologiche si deve tener presente che non esistono lunghe catene di reazioni che terminano all'esterno ma le reti sono fortemente interconnesse.

[6]Ricordiamo che il vincolo superiore sull'ultimo flusso di una catena lunga n è rilassato di un fattore $1/\varrho^n$.

4.5 Conclusioni

Oltre ad essere motivati un interesse puramente teorico, abbiamo deciso di svolgere questi conti su modellini giocattolo per avere un'idea più chiara del comportamento di `minover`$^+$ prima di utilizzarlo sistematicamente nello studio del metabolismo di *E.coli*. Nel corso dell'analisi sono emersi alcuni problemi:

1. a $\varrho \lesssim 1$ i flussi facilmente ordinabili (essenzialmente quelli con poche connessioni a valle) vengono lasciati indietro nella spartizione della norma in favore dei punti su cui l'algoritmo si concentra di più (flussi a monte e facenti parti di anelli).

2. l'algoritmo non esplora uniformemente lo spazio delle soluzioni ma è presente un *bias* legato alla distribuzione iniziale;

Parallelamente possiamo notare che:

1. la situazione $\varrho = 1$ è in ogni caso critica per la presenza di gruppi conservati e flussi congelati, che rendono difficile trovare soluzioni lungo direzioni singole con un algoritmo pensato per campionare volumi;

2. la parte di spazio delle fasi non esplorata è molto grande a bassi ϱ ma diminuisce rapidamente, a ϱ maggiori le soluzioni sono in ottimo accordo con le soluzioni analitiche del modello.

Queste considerazioni, avvalorate anche dalle prime simulazioni sull'intera rete metabolica, ci hanno spinto a non utilizzare nella nostra analisi valori di ϱ eccessivamente vini ad 1; limitandoci ad una fase in cui i vincoli legati alla topologia della rete producono soluzioni compatibili con le predizioni del modello nella zona biologicamente interessante, ma tenendoci sufficientemente lontani dalla zona critica.

Capitolo 5

Predizioni dei flussi in *E.coli*

5.1 Ricostruzione del metabolismo

Il modello di metabolismo utilizzato un questa analisi è il iJR904 GSM/GPR [16] ricavato dal ceppo di *E.coli* K-12 MG1655. Questo modello combina informazioni da fonti diverse: genomiche, trascrittomiche, proteomiche e metaboliche. Un'approccio di ottimizzazione lineare vincolata (analogo a FBA) permette di trovare gli stati ottimali di una versione della rete in diversi regimi, i risultati vengono quindi confrontati con i dati sperimentali e la rete eventualmente modificata in un processo iterativo.

Nella nostra analisi sono state effettuate alcune modifiche volte principalmente a ottenere flussi separati per le reazioni chimiche reversibili e stabilire un ambiente di coltura in cui la cellula può disporre liberamente di alcuni metaboliti (anidride carbonica, glucosio, potassio, ammoniaca, fosfato, ossigeno e solfato). Notiamo che, a differenza di quanto accade in FBA, i flussi relativi alle reazioni di *uptake* non sono limitati superiormente.
Nella formulazione finale sono presenti 1057 reazioni e 631 metaboliti.

Se esaminiamo la connettività dei nodi (figura 5.1) vediamo che sia per le reazioni sia per i metaboliti essa segue una distribuzione a potenza ma tra i secondi sono presenti specie chimiche molto connesse (come ATP, acqua, protoni, ...) che sono responsabili di un diverso comportamento nella coda della distribuzione. Questi metaboliti mettono in comunicazione parti del metabolismo anche molto distanti e non consentono di trattare come indipendenti le varie parti della rete.

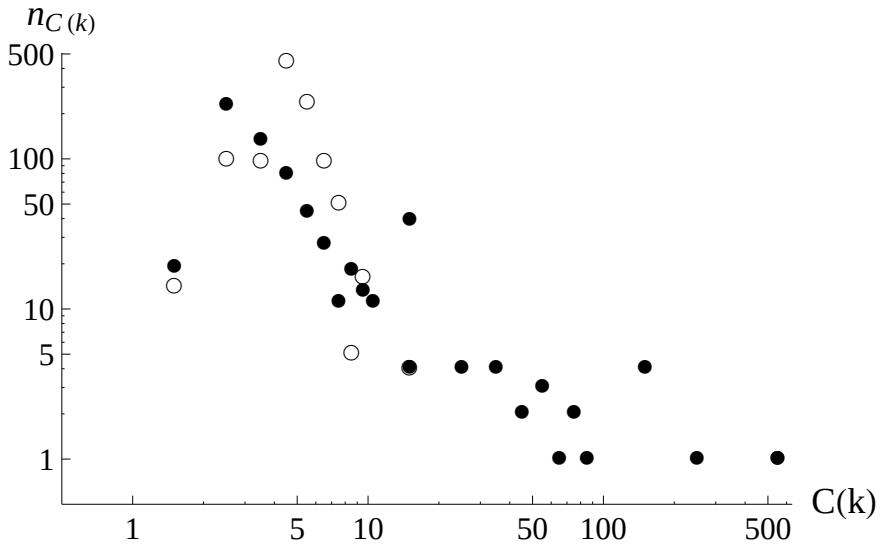


Figura 5.1: Connettività dei nodi-reazione (tondi vuoti) e dei nodi-metaboliti (pieni).

5.2 Analisi globale

Utilizzando come guida i risultati del capitolo precedente abbiamo effettuato le simulazioni per valori di ϱ compresi tra 0 e 0.91 per controllare l'evoluzione di alcune osservabili globali relative all'intera rete e tra 0.9 e 0.97 per un analisi più dettagliata dei singoli flussi nella zona biologicamente rilevante.
I valori medi di riferimento sono ricavati da un *ensemble* di 500 soluzioni in modo ottenere un campionamento il quanto più possibile uniforme dello spazio delle soluzioni.

In questo paragrafo eviteremo di soffermarci su reazioni chimiche specifiche e concentreremo l'analisi sulle proprietà globali della rete.

Volume dello spazio delle soluzioni

L'algoritmo `minover`$^+$ agisce come una mappa deterministica tra una distribuzione di flussi iniziale e un punto dello spazio delle soluzioni (che chiameremo Ω), l'elemento stocastico è quindi determinato esclusivamente dalle condizioni iniziali. Nel precedente capitolo abbiamo dimostrato come, per alcune tipologie di reti, estrarre i flussi di partenza da una distribuzione uniforme non consenta di esplorare tutto lo spazio delle soluzioni per bassi valori del tasso di crescita globale, mentre per valori più alti del parametro ϱ le soluzioni non differiscano significativamente dalla soluzione analitica.

Per analizzare questa situazione nella ben più complessa rete metabolica di *E.coli*, introduciamo una nuova osservabile: la sovrapposizione media tra due vettori dei flussi appartenenti a soluzioni diverse

$$q_{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N} q_{\alpha\beta}^{(i)},$$
$$q_{\alpha\beta}^{(i)} = \frac{s_{i\alpha}s_{i\beta}}{\|\vec{s}_\alpha\|\|\vec{s}_\beta\|}.$$

La media d'*ensemble* di questa quantità, $\langle q_{\alpha\beta}\rangle$, è legata al valore medio dell'angolo tra due soluzioni diverse e fornisce un'indicazione del volume di Ω; infatti ha valore unitario quando il modello ammette una sola soluzione (a meno di una costante moltiplicativa), mentre decresce quanto più due soluzioni possono essere differenti. Più basso è il valore di $\langle q_{\alpha\beta}\rangle$ ad un certo ϱ più è grande la variazione permessa ai singoli flussi e di conseguenza la misura di Ω.

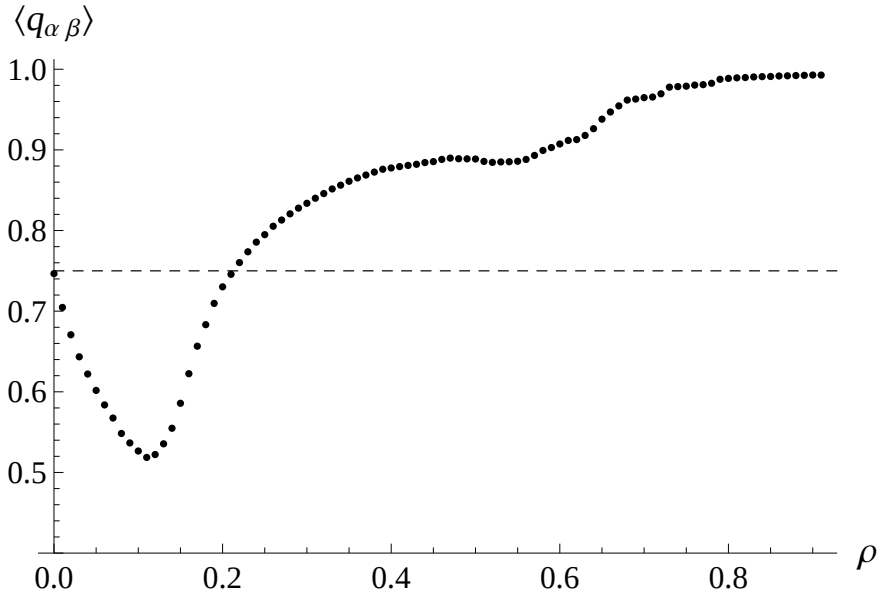


Figura 5.2: Sovrapposizione media tra i vettori di flussi di un *ensemble*; la linea tratteggiata è il valore per una coppia di vettori estratti da una distribuzione uniforme.

I risultati della nostra simulazione sono riportati nella figura 5.2. Cominciando a leggere il grafico da destra si può notare come per alti ϱ i vincoli rigidi comportino $\langle q_{\alpha\beta}\rangle \lesssim 1$ e quindi poca variabilità tra le soluzioni; diminuendo il valore del tasso di espansione la sovrapposizione media diminuisce sino a raggiungere un minimo per $\varrho \sim 0.1$, a questo punto risale al valore medio per quantità estratte da una distribuzione uniforme, $3/4$.

La spiegazione di questo fenomeno secondo noi è da imputarsi al fatto che a bassi ϱ la mappa creata da `minover`$^+$ a partire da una distribuzione uniforme corrisponde ad un sottoinsieme di Ω molto piccolo, come abbiamo visto accadere anche nel caso più semplice della catena di reazioni isolate. Evidentemente per valori di ϱ minori di 0.2 lo spazio delle soluzioni si sviluppa lungo direzioni ortogonali conducono ad un basso valore di $\langle q_{\alpha\beta}\rangle$, queste direzioni non sono esplorate a $\varrho \sim 0$ poiché qualunque punto di $\mathcal{R}_+^N$ scelto come condizione iniziale dista pochi passi dell'algoritmo da un punto di Ω.

Rapporti tra i flussi WT e PYK

L'inibizione della singola reazione

$$[\mathrm{PEP}] + [\mathrm{ADP}] \longrightarrow [\mathrm{PYR}] + [\mathrm{ATP}],$$

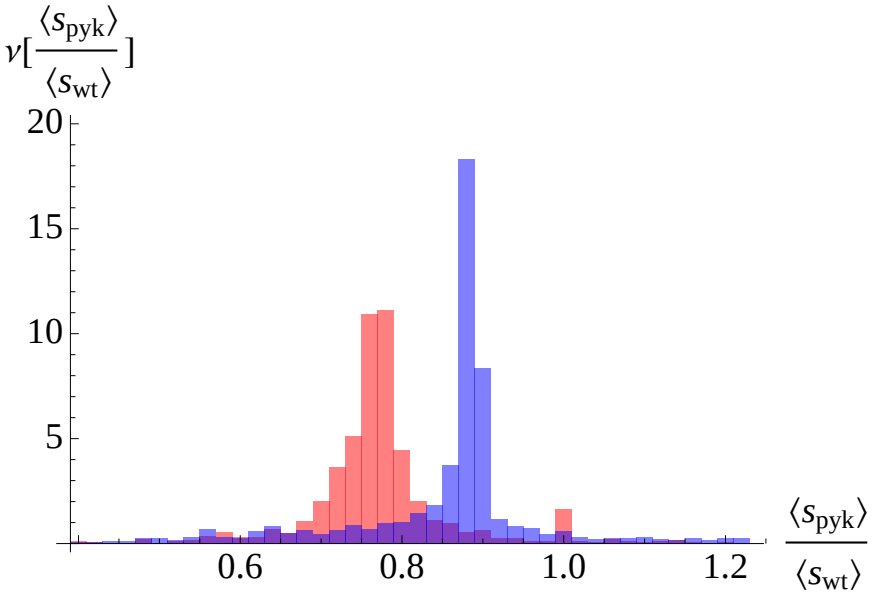


Figura 5.3: Sovrapposizione media tra i vettori di flussi di un *ensemble*.

mediata dall'enzima piruvato chinasi (PYK) è stata ottenuta ponendo uguali a zero i corrispondenti coefficienti stechiometrici nella matrice Ξ. Per poter confrontare i flussi nei diversi esperimenti (sia *in vivo* sia *in silico*) va fissata una scala comune. Abbiamo deciso quindi di fissare a uno la velocità di *uptake* del glucosio.

La risposta globale della rete corrisponde ad una diminuzione di quasi tutti i flussi (figura 5.4, istogramma rosso). Dal punto di vista biologico ci si aspetta che alcuni flussi diminuiscano per bilanciare l'aumento di alcune specifiche reazioni che devono compensare la mancata produzione del piruvato, ma sembra un argomento debole per spiegare questa diminuzione generalizzata.
La motivazione più probabile è una variazione nell'*uptake* di glucosio. Considerando i flussi normalizzati, il peso di questa reazione sulla norma è in media esattamente il valore normale della distribuzione dei rapporti. Questo fatto suggerisce di fissare come scala di riferimento l'*uptake* di glucosio medio del batterio WT anche nel caso del batterio mutato. Per ottenere questo risultato abbiamo effettuato una riscalatura in due passaggi:

- tutti i flussi sono stati divisi per l'*uptake* di glucosio cor-

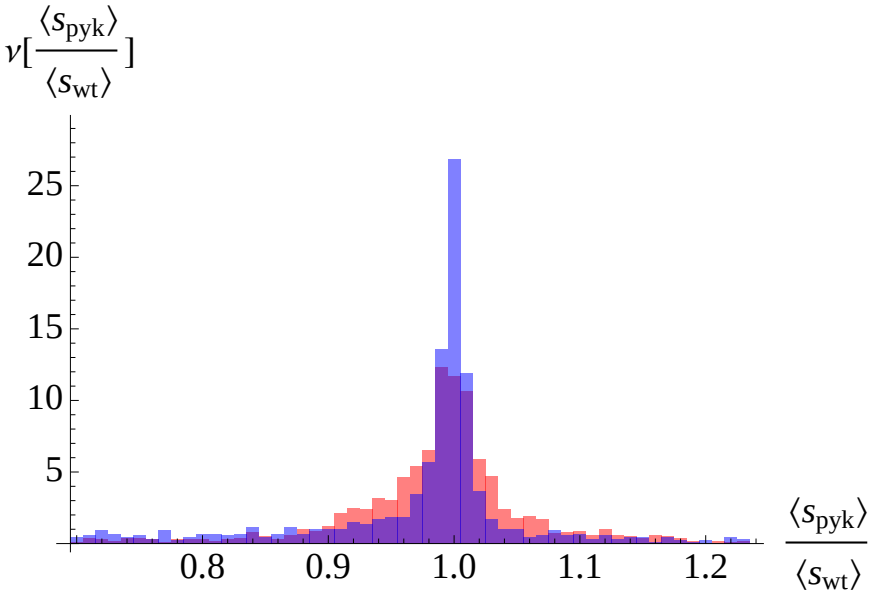


Figura 5.4: Sovrapposizione media tra i vettori di flussi di un *ensemble* — flussi riscalati.

rispondente (reazione con $i = g$),

$$s_{i\,\alpha}^{(PYK)}(\varrho) \leftarrow \frac{s_{i\,\alpha}^{(PYK)}(\varrho)}{s_{g\,\alpha}^{(WT)}(\varrho)},$$

per fissare una scala locale in ogni *ensemble*;

- i flussi vengono riscalati nuovamente tramite un coefficiente che tiene conto del diverso peso di questa reazione nella norma,

$$s_{i\,\alpha}^{(PYK)}(\varrho) \leftarrow \frac{s_{i\,\alpha}^{(PYK)}(\varrho)}{k},$$
$$k = \frac{\langle s_{g\,\alpha}^{(PYK)}(\varrho)\rangle}{\|\langle s_{\alpha}^{(PYK)}(\varrho)\rangle\|} \frac{\|\langle s_{\alpha}^{(WT)}(\varrho)\rangle\|}{\langle s_{g\,\alpha}^{(WT)}(\varrho)\rangle}.$$

Agire in questo modo permette di rimuovere questo effetto sistematico senza però alterare la significatività delle previsioni che mostreremo nel prossimo paragrafo.

Effetti delle condizioni iniziali

Abbiamo visto come la scelta delle condizioni iniziali di `minover`$^+$ costituisca un *bias* in favore di un sottoinsieme di soluzioni.

Possiamo cercare di sfruttare questa situazione per cercare dei punti di Ω_{PYK} che siano maggiormente in relazione con Ω_{WT}. Per ottenere questo risultato invece di inizializzare l'algoritmo con una distribuzione uniforme a $\varrho = 0$, siamo partiti dalle soluzioni trovate nel caso della rete completa.
L'idea alla base di questo approccio è venuta dallo studio del metodo MOMA che abbiamo presentato nel primo paragrafo del terzo capitolo. Mentre in MOMA, però, era imposta esplicitamente la distanza minima dalla soluzione WT, in questo caso si sta sfruttando una caratteristica dell'algoritmo, quindi il concetto di "vicinanza" non va inteso in senso euclideo.

Precedentemente, per stimare la distanza tra due soluzioni di uno stesso *ensemble*, abbiamo utilizzato il coefficiente di sovrapposizione $\langle q_{\alpha\beta} \rangle$; analogamente, in questo caso possiamo definire

$$q_{\alpha\alpha} = \frac{1}{N} \sum_{i=1}^{N} q_{\alpha\alpha}^{(i)},$$

$$q_{\alpha\alpha}^{(i)} = \frac{s_{i\alpha}^{(\text{WT})} s_{i\alpha}^{(\text{PYK})}}{\|\vec{s}_{\alpha}^{(\text{WT})}\| \|\vec{s}_{\alpha}^{(\text{PYK})}\|},$$

sia nel caso delle soluzioni PYK indipendenti sia in quello dei flussi ottenuti dalle controparti WT. Dalle nostre simulazioni (figura 5.5) è emerso chiaramente come questi ultimi presentino una sovrapposizione maggiore con le soluzioni WT.

Analizzando nuovamente la distribuzione dei rapporti tra i flussi (figura 5.4, istogramma blu) possiamo notare come la differenza non si esaurisca in valori in media più alti per i rapporti tra i flussi, ma la stessa distribuzione sia più piccata e dalle code più grasse. Nelle tabelle 5.1 e 5.2 mostriamo una selezione dei parametri che caratterizzano le due distribuzioni nel caso si fissi l'*uptake* del glucosio uguale a uno ovvero venga effettuata una riscalatura col sistema spiegato precedentemente.

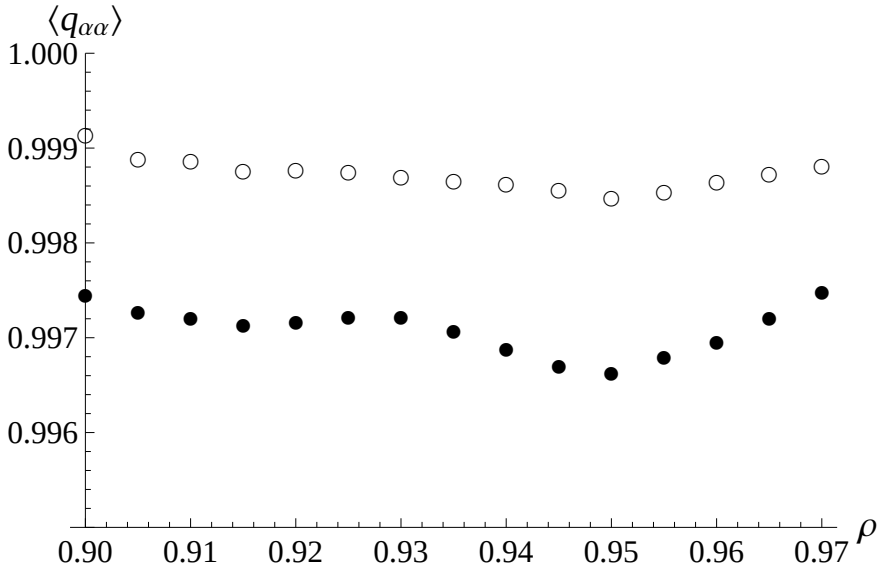


Figura 5.5: Sovrapposizione media tra i vettori di flussi di un *ensemble* nel caso di soluzioni ottenute da condizioni uniformi (tondi pieni) e in corrispondenza di soluzioni WT (tondi vuoti).

Tabella 5.1: Alcuni parametri relativi alle distribuzioni dei rapporti tra i flussi.

cond. iniziali	μ	σ	curtosi	moda
random	0.79	0.17	72	0.76
WT	1.00	1.8	462	0.88

Tabella 5.2: Parametri relativi alle distribuzioni riscalate.

cond. iniziali	μ	σ	curtosi	moda
random	1.01	0.22	80	1.00
WT	1.13	2.00	462	0.99

5.3 Analisi locale

Per poter effettuare un confronto diretto con i dati sperimentali ci siamo focalizzati sullo studio del metabolismo centrale del carbonio. Questo può essere diviso in tre vie metaboliche fondamentali: la glicolisi, la via dei pentoso fosfati e il ciclo dell'acido citrico, di cui abbiamo fornito una descrizione nel paragrafo 1.2.

Nell'ottica di limitare al minimo gli interventi sul modello che ha caratterizzato questa tesi, abbiamo deciso di effettuare la nostra simulazione sulla rete metabolica completa di K-12 MG1655 a differenza degli articoli analizzati, in cui sono state analizzate delle sotto-reti composte esclusivamente dai tre *pathway* summenzionati e poche altri collegamenti con l'esterno.

Per poter visualizzare in nostri risultati in maniera immediata e flessibile abbiamo sviluppato uno script nel linguaggio Mathematica che permette di generare un grafo bipartito metaboliti-reazioni dato un insieme di metaboliti d'interesse; le reazioni considerate sono allora tutte e sole quelle che collegano questi metaboliti. Fornendo al programma un vettore dei flussi $\vec{s}$ i nodi relativi alle reazioni vengono colorati in dipendenza dalla velocità di queste: colori tendenti al bianco indicano flussi nulli o molto piccoli, colori intensi reazioni più veloci.

Glicolisi

Il grafo 5.6 mostra la via della glicolisi a partire dal glucosio sino al piruvato. Possiamo notare delle somiglianze molto forti con le strutture che abbiamo studiato nel capitolo precedente: i flussi nella fase finale (g3p $\rightarrow$ 13dpg $\rightarrow$ 3pg $\rightarrow$ 2pg $\rightarrow$ pep $\rightarrow$ pyr) seguono un comportamento analogo a quello visto nelle catene semplici, ciò è dovuto al fatto che molti di questi metaboliti hanno bassa connettività verso l'esterno e quindi non sono presenti *pathway* alternativi.
Se scorporiamo i dati relativi ai flussi di questo pezzo a diversi ϱ (figura 5.7) e consideriamo il flusso netto nel caso delle

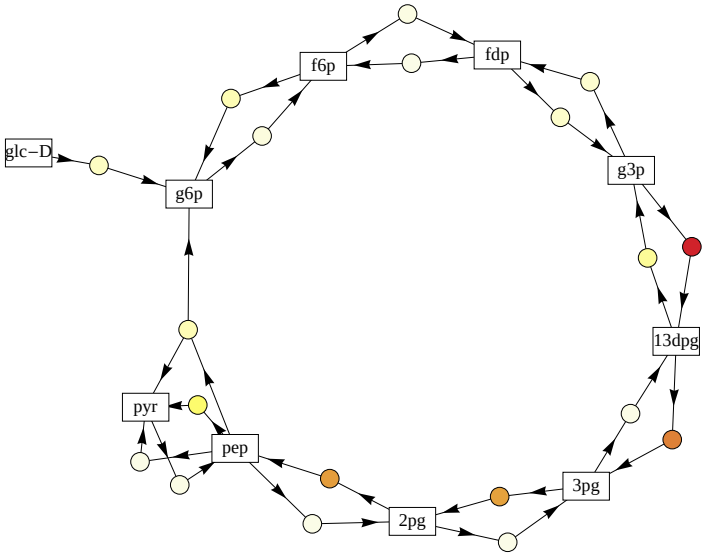


Figura 5.6: Glicolisi. Il grafo è stato costruito selezionando un insieme di metaboliti e considerando tutte e sole le reazioni tra questi. Le reazioni sono rappresentate dai pallini, colorati dal bianco al rosso secondo il valore medio del flusso corrispondente — qui è nel seguito sono mostrate le soluzione del modello a $\varrho = 0.97$.

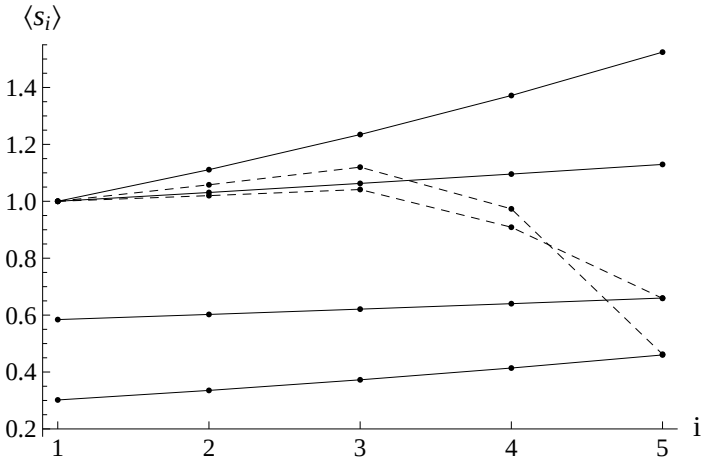


Figura 5.7: Valori medi dei flussi per le reazioni g3p $\to$ 13dpg, 13dpg $\to$ 3pg, 3pg $\to$ 2pg, 2pg $\to$ pep e pep $\to$ pyr a $\varrho = 0.9, 0.97$, le linee solide sono i massi è i minimi permessi dai vincoli. Grafico ottenuto fissando il primo flusso ad uno.

reazioni reversibili,[1]

$$s = s_+ - s_-,$$

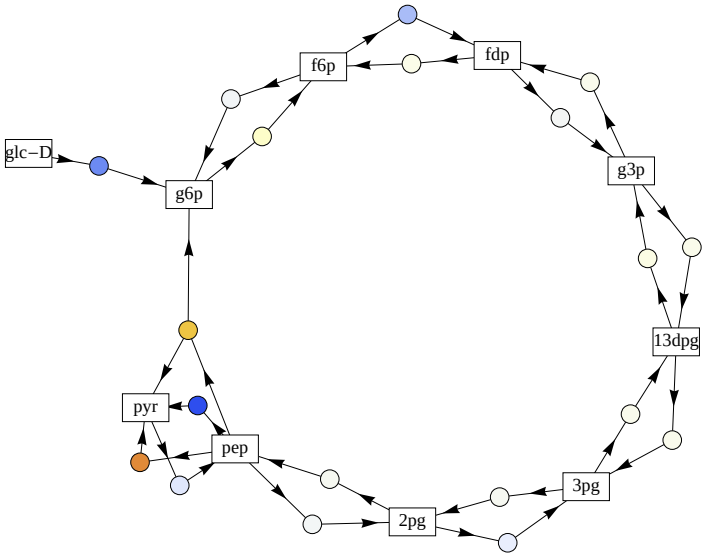
possiamo vedere come i valori medi abbiano un comportamento simile alle predizioni teoriche nel caso di una rete isolata, con una gerarchia parzialmente violata a $\varrho < 1$ che tende ad aumentare con il tasso di espanzione. Inoltre, i valori più bassi delle prime due reazioni vanno interpretati tenendo conto che i substrati corrispondenti (gliceraldeide-3-fosfato e 1,3-bisfosfoglicerato) vengono usati anche in altre reazioni.
L'ultimo passaggio di questa catena è il collegamento tra il fosfenolpiruvato e il piruvato. Nell'immagine della rete possiamo vedere che sono più d'una le reazioni chimiche che mettono in relazione questi due metaboliti ma l'unica reazione di una certa importanza è la fosforilazione che trasforma una molecola di ADP nella sua controparte ADP, proprio quella che vogliamo inibire nel nostro studio.

I rapporti tra i flussi medi della rete perturbata e di quella completa sono mostrati nella figura 5.8(a). Tutti i lavori analizzati [1, 6, 19] sono concordi nell'affermare che nel mutante il flusso di scambio tra g6p e f6p e il flusso tra f6p e g3p sono ridotti. Questo risultato sperimentale trova conferma nel nostro modello: ad alti valori di ϱ, dove ci aspettiamo maggior corrispondenza con la situazione biologica, entrambi i rapporti sono minori di uno. Inoltre disponendo di un intero *ensemble* di soluzioni siamo in grado di ricavare la deviazione media per i rapporti, come mostrato nelle figure 5.8(b-c), e dare un significato statistico più marcato a questi risultati.
Oltre a ciò il modello predice una diminuzione generale dei flussi lungo tutta la glicolisi (diminuzione dei flussi diretti e aumento dei flussi inversi) in media più marcata della riduzione globale, e una grande accelerazione nella reazione

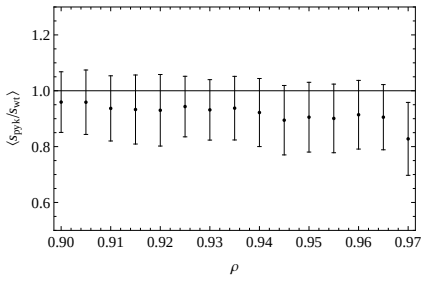
$$\mathrm{dha} + \mathrm{pep} \longrightarrow \mathrm{dhap} + \mathrm{pyr}.$$

Ulteriori dati sperimentali sono necessari per confermare o smentire questa previsione.

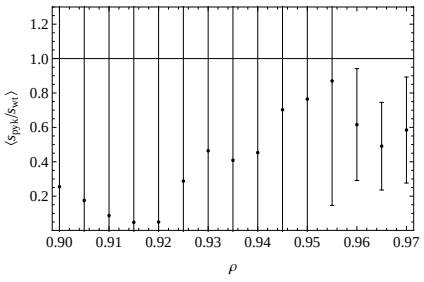
[1] Questo è un passaggio delicato. La definizione di flusso netto che abbiamo dato non tiene conto di ϱ e quindi non è una soluzione una soluzione di VN. In alternativa si può definire $s = s_+ - \varrho s_-$, in questo modo però si fissa una direzione tra metabolita consumato e metabolita prodotto, che vengono trattati in maniera asimmetrica. Questo fatto impedisce di usare questa definizione al di fuori di pezzi di rete che sono strutturati come una catena irreversibile.



(a) Media dei rapporti tra i flussi delle reazioni nel batterio mutato e in quello sano nella glicolisi. Il colore blu indica che la mutazione ha fatto diminuire i flussi, il rosso che sono aumentati. I pallini bianchi indicano che il rapporto del flussi è unitario o molto vicino a 1.



(b) **g6p** $\rightleftharpoons$ **f6p**.



(c) **f6p** + atp $\longrightarrow$ H + dhap + **g3p** + adp.

Figura 5.8: Rapporti dei flussi per i metaboliti coinvolti nella glicolisi.

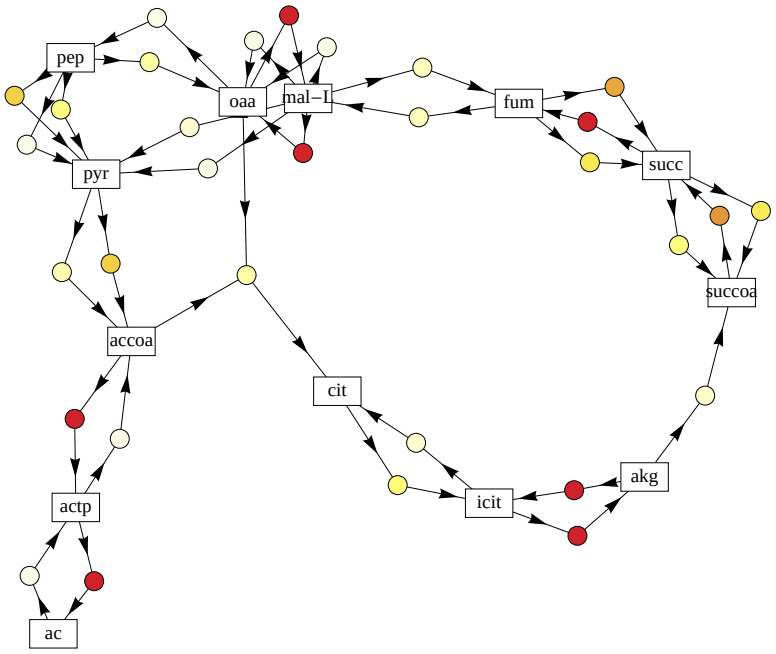


Figura 5.9: Ciclo di Krebs.

Ciclo di Krebs

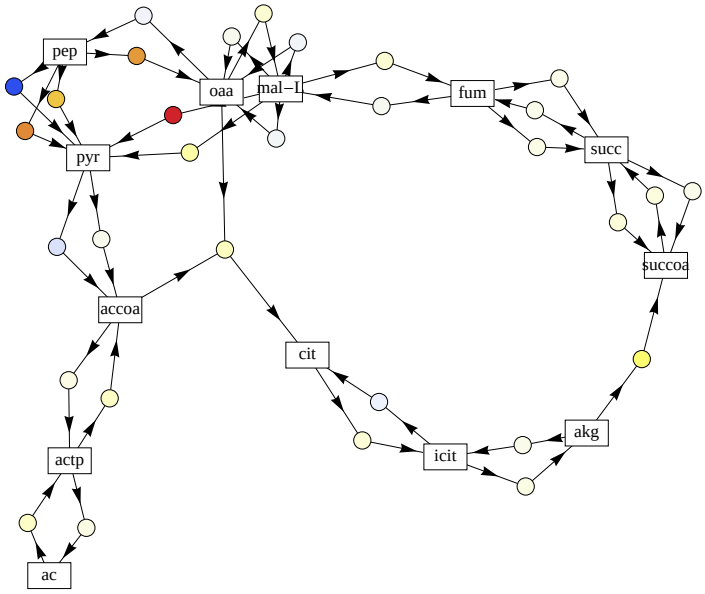
Questa parte di rete costituisce un ciclo chiuso di reazioni per la maggior parte reversibili, molto connesso con l'esterno. Ogni volta che l'azione dell'algoritmo si propaga su un metabolita legato a una reazione reversibile entrambi i flussi vanno aggiustati per mantenere la corretta proporzione. Non stupisce quindi che le coppie di flussi associate alle reazioni tra malato e ossalacetato, tra fumarato e succinato e tra isocitrato e α-chetoglutarato (akg) siano molto alte (figura 5.9), infatti l'algoritmo deve tornare ad aggiustarle ogni volta che effettua delle modifiche nei segmenti di rete collegati.
Per correggere questa anomalia è sufficiente considerare i flussi netti come abbiamo fatto nel caso della glicolisi.

Nel capitolo dedicato alla descrizione del metabolismo abbiamo visto che il piruvato è una componente fondamentale a monte del ciclo di Krebs: da esso viene prodotto l'acetil-CoA che si condensa con l'ossalacetato (oaa) per produrre citrato nel primo passaggio del ciclo. Non stupisce quindi che l'inibizione della piruvato chinasi comporti una ristrutturazione dei

flussi importante in questa parte della rete.
Gli articoli che abbiamo esaminato rilevano tutti un aumento della produzione di piruvato a partire dal malato e di ossalacetato dal fosfenolpiruvato; in più uno solo tra questi [1] indica una decisa riduzione del flusso in uscita di acetato. VN predice correttamente le modificazioni sopra elencate (figure 5.10(a) e 5.11 in alto) mentre non trova particolari differenze nei flussi delle reazioni relative all'acetato.

Mentre i risultati relativi alla glicolisi non mostrano una grande dipendenza dal tipo di condizioni iniziali scelte, in questo caso invece il confronto con i dati ottenuti inizializzando `minover`$^+$ con le soluzioni relative al metabolismo WT (figure 5.10(b) e 5.11 in basso) è più interessante: l'aumento del flusso pep $\longrightarrow$ oaa non è così marcato, mentre è presente una lieve diminuzione del flusso netto in uscita di acetato che pur non essendo della stessa entità di quella misurata in [1] sembra avvalorarne le conclusioni.

Anche in questo caso la variazione dei flussi rispetto al batterio WT è superiore alle fluttuazioni d'*ensemble*.



(a) Media dei rapporti nel ciclo di Krebs.

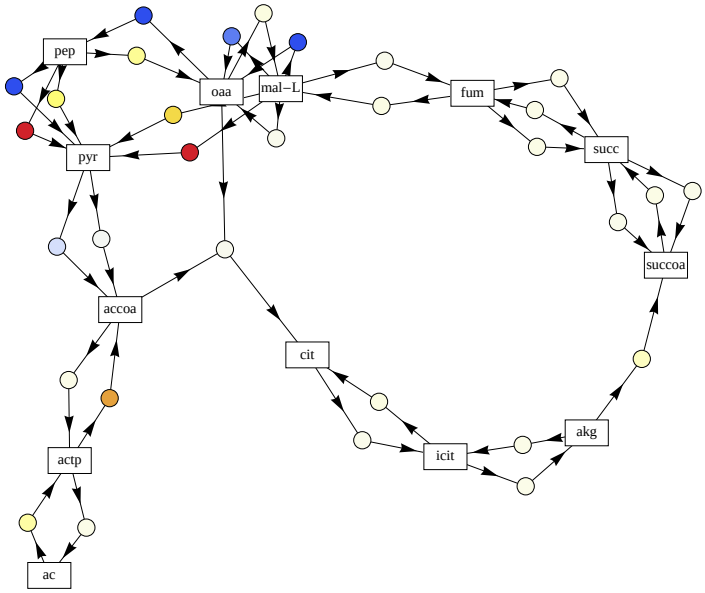
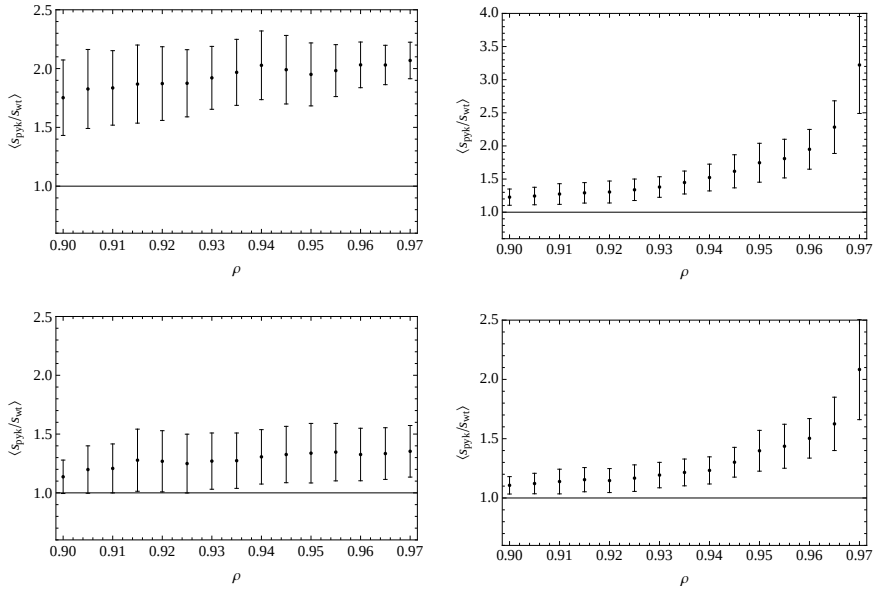


Figura 5.10: Rapporti tra i flussi ottimali per WT e PYK (sopra) e utilizzando come condizione iniziali per il mutante i flussi WT (sotto).



(a) $CO_2 + H_2O +$ **pep** $\longrightarrow$ H + **oaa** + PO_4.

(b) **mal-L** + nad(h) $\longrightarrow$ CO_2 + nadp(h) + **pyr**.

Figura 5.11: Rapporti per due reazioni in particolare in funzione di ϱ da condizioni iniziali uniformi (sopra) e WT (sotto).

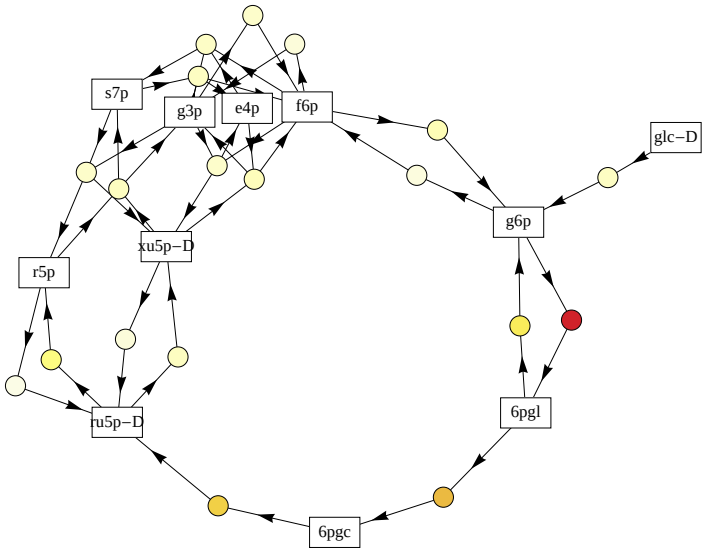


Figura 5.12: Via dei pentofosfati.

Pentoso fosfati

Secondo il nostro modello *in silico* questo *pathway* alternativo alla glicolisi è molto utilizzato, anche se questa situazione non corrisponde alle misure sperimentali.
Una possibile spiegazione ci viene fornita dall'analisi nel capitolo precedente: nel caso di una biforcazione della rete in corrispondenza di un metabolita viene privilegiata una catena lunga rispetto ad una più corta; nella via dei pentoso fosfati il percorso da g6p a f6p o g3p richiede più passaggi rispetto alla glicolisi ed occupa quindi un maggior volume nello spazio delle soluzioni (analogamente possiamo dire che ha un'entropia maggiore).
Non è possibile correggere questo comportamento nella versione del modello che stiamo utilizzando, a meno di non introdurre vincoli esterni. Questa soluzione però va contro lo spirito del modello, ossia una trattazione puramente statistica delle soluzioni permesse dalla topologia della rete. Una strada per ottenere maggior verosimiglianza è l'integrazione di questi risultati con il problema duale a quello considerato in questa tesi, ossia associare al vettore prezzi del modello di Von Neumann un vettore Δg_μ che contenga le variazioni di energia libera per ogni metabolita. Questo approccio richiede

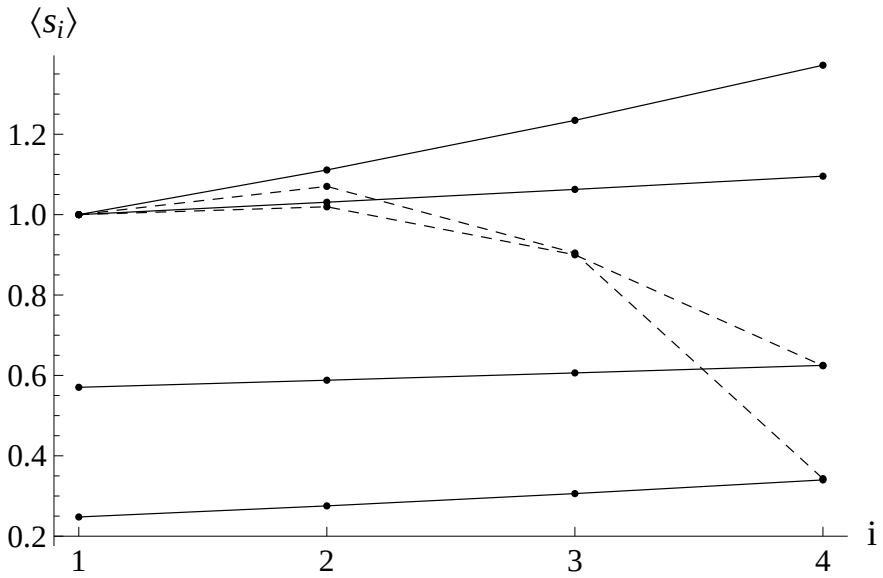


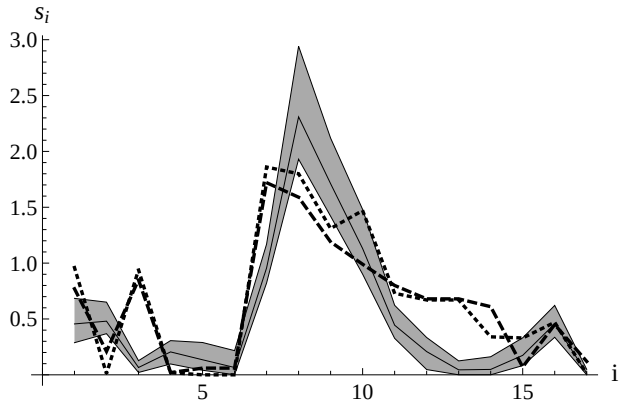
Figura 5.13: Valori medi dei flussi per le reazioni g6p → 6pgl, 6pgl → 6pgc, 6pgc → ru5p-D e il flusso combinato di ru5p-D → xu5p-D e ru5p-D → r5p — $\varrho = 0.9, 0.97$.

però una trattazione diversa, utilizzando flussi singoli per le reazioni reversibili con le conseguenze che abbiamo visto.

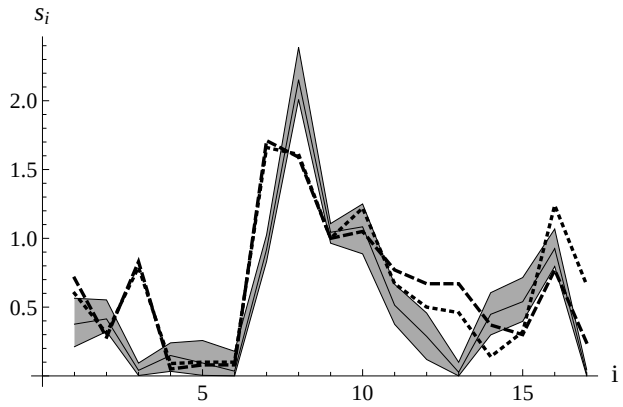
In attesa di ulteriori integrazioni del modello non ci sentiamo di trarre conclusioni biologiche da questi dati. Possiamo comunque utilizzarli come *benchmark* per risultati analitici di VN. Considerando singolarmente i flussi netti da g6p fino a ru5p-D e il totale tra ru5p-D→xu5p-D e ru5p-D→r5p otteniamo di nuovo una struttura simile alla catena semplice 5.13. Anche in questo caso si tratta di un segmento non isolato e quindi non è possibile sovrapporre questi dati alle soluzioni analitiche per una catena della stessa lunghezza. L'ordinamento gerarchico è comunque presente, anche se non è completo: i flussi centrali tendono a crescere il più possibile compatibilmente con i vincoli a $\varrho < 1$. Vediamo che anche in questo caso l'ordinamento tende a crescere all'aumentare di ϱ.

Confronti tra i valori assoluti dei flussi

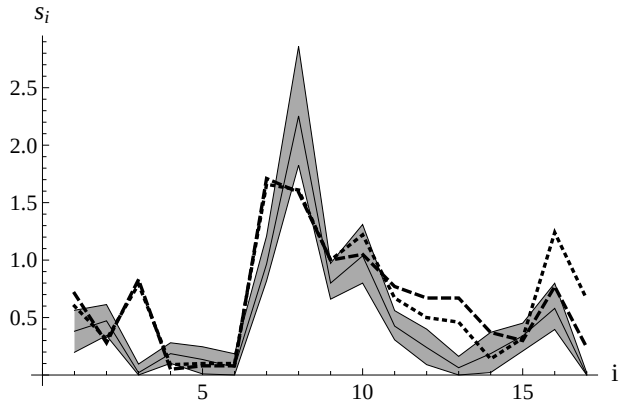
Fino a questo punto ci siamo limitati ad analizzare i risultati delle nostre simulazioni in senso tendenziale, in quando per sua natura il modello che stiamo utilizzando non fornisce un singolo valore ottimale per i flussi delle reazioni ma piuttosto un *ensemble* di soluzioni vincolate esclusivamente dalla topologia della rete e dalla conservazione della massa. Un ulte-



(a) Batterio WT



(b) Batterio PYK, condizioni iniziali uniformi



(c) Batterio PYK, condizioni iniziali WT

Figura 5.14: Confronto tra i flussi predetti in diverse condizoni di coltura (C-0.8, linea tratteggiata, e N-0.09, linea puntinata) e l'intervallo di soluzioni trovato da `minover`$^+$. La linea all'interno dell'intervallo è la soluzione media.

riore fattore da tenere in conto è il leggero *bias* introdotto da `minover`$^+$ in dipendenza dalle condizioni iniziali.

In questa sezione vogliamo fare un passo in più e mettere a confronto l'intervallo di valori trovati dall'algoritmo con i valori assoluti dei flussi trovati tramite FBA o MOMA su misure di ^{13}C-labelling.
Nella figura 5.14 mostriamo i nostri risultati per i 17 flussi considerati in nei lavori di Emmerling [6] e Segrè [19] assieme alle misure effettuate in diverse condizioni di coltura (glucosio a $D = 0.4\,\mathrm{h}^{-1}$ o ammoniaca a $D = 0.9\,\mathrm{h}^{-1}$). Per molti di questi c'è un buon accordo tra misure e simulazione ma si individuano subito alcune eccezioni che vanno contestualizzate.
Il flusso numero tre è associato alla reazione reversibile f6p $\longleftrightarrow$ g3p che solitamente avviene molto velocemente. Abbiamo visto però che per motivi entropici VN tende a preferire la via dei pentoso fosfati rispetto alla glicolisi. Questi dati mostrano come sia opportuno procedere ad un estensione del modello che permetta una più stretta aderenza ai dati biologici, pur conservando il suo carattere minimale.
Un altro flusso sottostimato è il numero sette, relativo all'ultima parte della glicolisi, a valle dei pentoso fosfati. Il problema è legato alla struttura topologica della rete: i modelli considerati nei lavori citati sono minimali e relativi esclusivamente alla zona di interesse. Inoltre, ed è quanto accade nel caso in esame, i flussi di diverse reazioni sono combinati in un unica macroreazione. Questo fatto rende problematico il confronto con la struttura molto più fine della nostra rete metabolica.
Un ultimo fattore di rumore sono le condizioni di coltura come limitazioni di glucosio o ammoniaca. All'inizio del capitolo abbiamo mostrato quali nutrienti sono assimilabili dal batterio; in un approccio VN puro una volta decisa una selezione non è possibile limitarne i flussi associati, lasciando la cellula libera di assorbire quanto serve per bilanciare i sui processi interni. Eventuali vincoli rigidi sono possibili, ma vanno al di là dello scopo di questa tesi in quanto richiederebbero una diversa e trattazione teorica.

Nonostante tutte queste considerazioni è sorprendente constatare come il modello di Von Neumann, che richiede un livello di informazione estremamente ridotto (topologia e conservazione della massa), sia in grado di produrre dei risultati che si avvicinano molto a modelli decisamente più sofisticati.

Capitolo 6

Conclusioni

La misura dei flussi metabolici è destinata ad assumere sempre maggior importanza nell'ambito delle ricerche sulle modificazioni genetiche, la bioingegneria e lo studio sempre più accurato delle connessioni tra genotopo e fenotipo cellulare.

Disporre di un modello affidabile e robusto, che sia in grado di funzionare con un numero minimo di assunzioni è quindi molto importante per eseguire degli esperimenti *in silico* che permettano di indirizzare e analizzare gli esperimenti *in vivo*. Solo questi ultimi posso infatti verificare cosa succede esattamente all'interno di un sistema complesso come una cellula, ma questa stessa complessità non permette di trovare facilmente una chiave interpretativa senza l'utilizzo di modelli adeguati.

Nel capitolo 2 abbiamo riportato una delle tecniche più promettenti per la misura dettagliata dei flussi: la marcatura con carbonio 13. Pur essendo limitata allo studio del metabolismo centrale, questa tecnica fornisce informazioni insostituibili per limitare correttamente il gran numero di gradi di libertà presenti nel problema. La possibilità di distinguere tra vie alternative permette di evidenziare più compiutamente la struttura topologica del metabolismo e di misurare direttamente le variazioni nel funzionamento dello stesso in presenza di perturbazioni.

Nel capitolo 3 abbiamo introdotto alcuni modelli che permettono di ricavare numericamente le velocità delle reazioni. Questi modelli sono molto potenti, in quanto sono in grado di fornire una gran quantità informazioni con un numero di parametri decisamente inferiore alla cinetica chimica tradizionale. Tuttavia restano limitati allo studio degli stati stazionari e richiedono una perfetta conoscenza della topologia della rete sulla quale li si vuole applicare. Non si può affermare che nel

caso del metabolismo cellulare queste condizioni siano soddisfatte appieno, la ricerca di modelli di metabolismo cellulare sempre più competi e accurati, però, procede speditamente [16] e l'assunzione di stazionarietà può essere accettata se ci si limita allo studio della cellula nella fase in cui c'è abbondanza di nutrienti e l'alta velocità delle reazioni chimiche mantiene le concentrazioni costanti una volta raggiunta la condizione di equilibrio imposta dalla rete.

I metodi analizzati richiedono l'imposizione dall'esterno dei vincoli necessari a limitare i gradi di libertà in eccesso, che vanno ricavati da misure sperimentali o da assunzioni ulteriori sugli obiettivi delle reti metaboliche.
In quest'ottica, con il presente lavoro di tesi abbiamo deciso di approfondire l'utilizzo dell'approccio che Von Neumann propose quasi cento anni fa come semplice modello di produzione economica [24]. Rispetto alle alternative come la FBA, che si basano sull'ottimizzazione vincolata e quindi richiedono necessariamente l'imposizione dall'esterno di una funzione obiettivo, il modello di Von Neumann è interamente definito dai soli vincoli di conservazione della massa e dalla massimizzazione della crescita globale del sistema. Non è richiesto quindi specificare un obiettivo di produzione per il metabolismo e questo permette di limitare al minimo le assunzioni.
In linea di principio questo modello non impone un massimo al tasso di crescita ϱ, è quindi significativo che questo massimo sia compatibile con lo stato stazionario ($\bar{\varrho} = 0.999 \pm 1$, in [12]).

Abbiamo visto come la struttura del problema di Von Neumanm sia molto simile a quella di un percettrone, è stato quindi naturale utilizzare un algoritmo come `minover` che si ispira alle regole di apprendimento delle reti neurali e ha già dato buoni risultati in passato ([2], [3], [12]).

Una volta scelto l'algoritmo per esplorare numericamente lo spazio delle soluzioni, nel capitolo 4 abbiamo tentato di porci in delle situazioni molto semplici per poter elaborare una fenomenologia delle soluzioni rispetto alle diverse topologie. Questo studio ci ha permesso di esercitare un controllo maggiore sull'algoritmo, indicando una zona ottimale (da stabilire caso per caso) in cui le soluzioni trovate sono compatibili con il modello ma non risentono delle criticità che si presentano nel caso $\varrho \lesssim 1$.

Nell'ultimo capitolo abbiamo applicato questi risultati al metabolismo di *E.coli*. I conti analitici effettuati nel capitolo precedente sono stati un aiuto insostituibile nell'affrontare l'analisi dei dati prodotti da `minover`$^+$ su una rete complessa

come quella di *E.coli*. Siamo riusciti ad individuare strutture dal comportamento simile ai semplici modelli studiati e interpretare correttamente artefatti dovuti all'algoritmo. Abbiamo visto come il *bias* sulle condizioni iniziali possa essere sfruttato per simulare un adattamento più realistico della rete ad una perturbazione esterna se assumiamo che la cellula non sia in grado di raggiungere lo stesso grado di efficienza che l'organismo non perturbato ha sviluppato nel corso di milioni di anni di selezione naturale. L'idea di utilizzare delle condizioni iniziali *biased* per implementare informazione biologica aggiuntiva e ottenere così nuovi risultati è interessante e richiede sicuramente uno studio dedicato e approfondito.
Restando nell'ambito dell'approccio tradizionale abbiamo provato a confrontare i valori dei flussi predetti nel caso di un *knockout* in uno dei punti chiave del metabolismo del carbonio, la piruvato chinase, con le misure *in vivo*. Trovando nel caso della glicolisi e del ciclo di Krebs risultati in accordo con i dati sperimentali. Inoltre, questo studio ci ha permesso di capire meglio i limiti delle assunzioni minimali del modello attraverso la comparazione di *pathway* alternativi, come la glicolisi e la via dei pentoso fosfati.

Possiamo affermare quindi di aver raggiunto una comprensione maggiore nell'utilizzare ed interpretare i risultati di un'algoritmo che si è rivelato molto promettente nello studio delle reti metaboliche.
Tra i suoi punti forza possiamo annoverare un buon potere predittivo per quanto riguarda la risposta della rete ad una perturbazione (che abbiamo verificato tramite il *knockout* di PYK). I valori dei flussi in senso assoluto sono afflitti da un problema legato alla diversa entropia nelle diverse parti della rete. Tuttavia, limitando l'analisi alle parti meno connesse, si riescono ad ottenere flussi comparabili con gli esperimenti. Se consideriamo che gli studi presi in esame utilizzano metodi ricchi di informazione aggiuntiva, sia teorica sia sperimentale, è estremamente significativo che le soluzioni identificate dal nostro modello mostrino un buon accordo con questi dati, considerando che abbiamo lavorato su una rete metabolica completa molto più articolata di quelle utilizzate negli articoli.

Durante il lavoro sono emersi spunti di approfondimento interessanti come quello già accennato sulla dipendenza dalle condizioni iniziali o la possibilità stabilire collegamento più diretto tra la topologia di queste reti (uno dei fattori chiave nel modello di Von Neumann) e i conseguenti stati ottimali del metabolismo.

Ulteriori passi in avanti possono essere compiuti estendendo il modello per comprendere il problema duale alla conservazione della massa: i vincoli termodinamici nella creazione e distruzione dei metaboliti. Il nostro gruppo è al momento impegnato anche su questo fronte attraverso uno studio che utilizza solamente la parte duale del problema sulla rete metabolica del globulo rosso, in corso di pubblicazione. La combinazione dei due approcci in un unico modello è un'impresa ardua ma potrebbe essere la strada da seguire per una miglior comprensione del fenotipo cellulare.

Bibliografia

[1] K Al Zaid Siddiquee, M J Arauzo-Bravo e K SHIMIZU. «Metabolic flux analysis of pykF gene knockout Escherichia coli based on 13 C-labeling experiments together with measurements of enzyme activities and intracellular metabolite concentrations». In: *Applied Microbiology and Biotechnology* 63.4 (gen. 2004), pp. 407–417.

[2] A De Martino e E Marinari. «The solution space of metabolic networks: producibility, robustness and fluctuations». In: *arXiv.org* q-bio.MN (feb. 2010).

[3] A De Martino et al. «Optimal flux states, reaction replaceability and response to knockouts in the human red blood cell». In: *arXiv.org* q-bio.MN (lug. 2009).

[4] A De Martino et al. «Von Neumann's expanding model on random graphs». In: *arXiv.org* cond-mat.dis-nn (mar. 2007).

[5] J.S. Edwards, R.U. Ibarra e B.O. Palsson. «In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data». In: *Nature biotechnology* 19.2 (2001).

[6] M Emmerling et al. «Metabolic flux responses to pyruvate kinase knockout in Escherichia coli». In: *Journal of Bacteriology* (2002).

[7] DA Fell. «Fat synthesis in adipose tissue. An examination of stoichiometric constraints.» In: *Biochemical Journal* (1986).

[8] Matteo Figliuzzi. «Meccanica stastistica del problema di Von Neumann». In: (2009), pp. 1–102.

[9] Eliane Fischer e Uwe Sauer. «Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS». In: *European Journal of Biochemistry* 270.5 (set. 2003), pp. 880–891.

[10] P KEDAR e R COLAH. «Proteomic investigation on the pyk-F gene knockout Escherichia coli for aromatic amino acid production». In: *Enzyme and microbial technology* (2007).

[11] W Krauth e M Mezard. «Learning algorithms with optimal stability in neural networks». In: *Journal of Physics A: Mathematical and General* 20 (1987), p. L745.

[12] C Martelli et al. «Identifying essential genes in E. coli from a metabolic optimization principle». In: *arXiv.org* q-bio.MN (feb. 2009).

[13] JD Orth e I Thiele. «What is flux balance analysis?» In: *Nature biotechnology* (2010).

[14] ET Papoutsakis. «Equations and calculations for fermentations of butyric acid bacteria». In: *Biotechnology and Bioengineering* (1984).

[15] JM Park e TY Kim. «Prediction of metabolic fluxes by incorporating genomic context and flux-converging pattern analyses». In: *PNAS*. 2010.

[16] JL Reed, TD Vo e CH Schilling. «An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)». In: *Genome Biol* (2003).

[17] U. Sauer et al. «Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism». In: *Journal of Bacteriology* 181.21 (1999), pp. 6679–6688.

[18] K Schmidt, M Carlsen e J Nielsen. «Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices». In: *Biotechnology and Bioengineering* (2000).

[19] D Segre e D Vitkup. «Analysis of optimality in natural and perturbed metabolic networks». In: *PNAS*. 2002.

[20] T. Shlomi, O. Berkman e E. Ruppin. «Regulatory on/off minimization of metabolic flux changes after genetic perturbations». In: *Proceedings of the National Academy of Sciences of the United States of America* 102.21 (2005), p. 7695.

[21] T Szyperski. «Biosynthetically Directed Fractional 13C-labeling of Proteinogenic Amino Acids». In: *European Journal of Biochemistry* (1995).

[22] T Szyperski et al. «Bioreaction network topology and metabolic flux ratio analysis by biosynthetic fractional 13C labeling and two-dimensional NMR spectroscopy». In: *Metabolic Engineering* (1999).

[23] I Thiele. «A protocol for generating a high-quality genome-scale metabolic reconstruction». In: *Nature Protocols* (2010).

[24] John Von Neumann. «A model of general economic equilibrium». In: *The Review of Economic Studies* 13.1 (1945), pp. 1–9.

Ringraziamenti

Al termine di questo lavoro mi piacerebbe ringraziare le persone che mi hanno sostenuto e lo hanno reso possibile.
In primo luogo il mio relatore, il professor Marinari, che ha avuto fiducia in me fin dal principio lasciandomi larga autonomia di giudizio su come impostare l'analisi ma è stato pronto a riportare sui binari un treno che rischiava di perdersi in troppe direzioni.
Quindi il mio correlatore, Matteo Figliuzzi, con cui ho avuto un confronto continuo e sempre stimolante. Si è preso la briga di incoraggiarmi e consigliarmi e sgridarmi ogniqualvolta ce ne fosse il bisogno.
Un ringraziamento particolare va ai miei amici, vecchi e nuovi, che mi sono stati vicini non solo durante la stesura della tesi, ma anche durante tutta questa parentesi romana.
Infine la mia famiglia, senza la quale non sarei qui ora. In tutti i sensi.