

UNIVERSITÀ
DEGLI STUDI
FIRENZE

Scuola di Scienze Matematiche, Fisiche e Naturali
Biologia Molecolare e Applicata
Curriculum Cellulare e Molecolare

Tesi di Laurea Magistrale

L'INTERAZIONE TRA TRASCRIZIONE E REPLICAZIONE INFLUISCE SUL METABOLISMO CELLULARE?

DOES TRANSCRIPTION-REPLICATION INTERACTION IMPACT CELLULAR METABOLISM?

ALESSIO MASONI

Relatore: *prof. Marco Fondi*

Anno Accademico 2025-2026

Alessio Masoni: *Does transcription-replication interaction impact cellular metabolism?*,
Biologia Molecolare e Applicata, © Anno Accademico 2025-2026

INDICE

1 Introduzione 5

1.1 La molteplicità genica: definizione 5

1.1.1 Effetti della molteplicità sull'espressione genica 7

1.1.2 Possibili effetti sul metabolismo 9

1.2 La Biologia dei Sistemi 11

1.2.1 Approccio teorico: la modellizzazione vincolata dei sistemi metabolici 11

1.2.2 Analisi del bilancio di flusso, FBA e pFBA 14

1.2.3 Analisi dello spazio dei flussi mediante Flux Sampling 16

1.2.4 Dati omici e integrazione nei modelli metabolici 18

1.3 Formalizzazione obiettivi e domande di ricerca 19

2 Approcci Metodologici e Modelli Utilizzati 21

2.1 Presentazione del dataset e apparato sperimentale 21

2.2 COBRApy: lo standard per l'analisi di modelli vincolati 22

2.2.1 La complessità dei sistemi batterici in un modello da meno di 12 MB: il formato SBML 23

2.3 Contestualizzazione dei modelli metabolici con RIPTiDe 25

2.3.1 Parametri di contestualizzazione 25

2.4 Energy Distance come misura di dissimilarità tra distribuzioni di flusso campionate 26

2.5 Tecniche di data visualization 28

3 Risultati e discussione 31

3.1 Pipeline per la generazione di modelli contestualizzati 31

3.1.1 Valutazione dell'affidabilità di RIPTiDe nell'integrazione dei dati trascrittomici 33
3.1.2 Costruzione del dataset di molteplicità 34
3.1.3 Utilizzo di RIPTiDe per colmare il gap tra espressione e metabolismo 37
3.2 Analisi di dissimilarità tra modelli contestualizzati 38
3.2.1 Analisi di clustering sui modelli metabolici contestualizzati 38
3.2.2 Visualizzazione del dataset nello spazio delle componenti principali 41
3.2.3 Analisi quantitativa dell'impatto della molteplicità genica sul metabolismo 45
3.3 Estensioni applicative e prospettive di ricerca 48
3.4 Materiali aggiuntivi 48

ELENCO DELLE FIGURE

Figura 1 Schema del cromosoma batterico e replicazione multiforca. 6

Figura 2 Rappresentazione geometrica di una generica matrice stechiometrica. 12

Figura 3 Vincoli metabolici e selezione dello stato ottimale nello spazio delle soluzioni. 13

Figura 4 Rappresentazione di una generica matrice di flux sampling. 17

Figura 5 Struttura di un file SBML e componenti principali. 24

Figura 6 Schema della pipeline di contestualizzazione 32

Figura 7 Prestazioni di RIPTiDe su dati trascrittomici casuali. 33

Figura 8 Distribuzione del numero medio di copie geniche in funzione del tempo di duplicazione cellulare (τ). 35

Figura 9 Relazione tra espressione genica e molteplicità cromosomica in funzione del tasso di crescita. 36

Figura 10 Analisi di clustering basata su energy distance tra modelli. 39

Figura 11 Schema concettuale del risultato atteso dal clustering. 40

Figura 12 PCA delle distribuzioni di flusso ad alta e bassa crescita. 42

Figura 13 Confronto PCA tra condizioni ad alta e bassa crescita. 44

Figura 14 Correlazione tra tasso di crescita e *energy distance*. 46

Figura 15 Relazione crescita–distanza su dati mediani aggregati. 47

1

INTRODUZIONE

1.1 LA MOLTEPLICITÀ GENICA: DEFINIZIONE

Nei procarioti, la replicazione del DNA avviene secondo un processo bidirezionale a partire da un singolo punto di origine (oriC) e termina in una regione opposta del cromosoma (ter). Quando le condizioni ambientali consentono una crescita particolarmente rapida, il tempo necessario per completare una replica del genoma (il cosiddetto C-period) può superare l'intervallo di tempo tra due divisioni cellulari successive. In tali circostanze, le cellule attivano nuovi cicli di replicazione prima che i precedenti siano terminati, dando origine al fenomeno noto come replicazione multiforca ([21]), (Figura 1). Questa modalità di replicazione implica che, in un dato momento, diverse porzioni del cromosoma possano trovarsi in differenti stati di duplicazione. Di conseguenza, i geni localizzati in prossimità dell'origine di replicazione (oriC) sono presenti in più copie rispetto a quelli situati vicino al terminatore (ter). Tale differenza nel numero di copie costituisce ciò che viene definito molteplicità genica. In *Escherichia coli*, per esempio, il rapporto medio tra le copie dei geni vicini a ori e quelli vicini a ter (rapporto oriC/ter) può variare da circa 1 durante la crescita lenta fino a valori di 4 o superiori in condizioni di crescita molto rapida ([10]).

Questa considerazione porta a ipotizzare l'esistenza di un effetto di dosaggio genico legato alla replicazione, per cui la posizione di un gene sul cromosoma influenzerebbe temporaneamente la sua disponibilità e, indirettamente, la sua espressione. Quest'ultimo però è un giudizio basato su logica deduttiva. L'effetto della molteplicità genica su un organismo procariotico è un fenomeno complesso e difficile da isolare. Lo scopo dei capitoli successivi sarà quello non solo di valutare la presenza o assenza di tale effetto, ma cercare di quantificarlo.

Alcune analisi teoriche e analisi su reti metaboliche hanno già esplorato alcune possibili conseguenze della molteplicità genica, mostrando che le fluttuazioni nel dosaggio dei geni enzimatici, indotte dalla replicazione, possono influenzare la stabilità dei flussi metabolici e modificare le concentrazioni dei metaboliti. Questi studi suggeriscono che variazioni locali nel numero di copie geniche non siano necessariamente neutre, ma possano introdurre perturbazioni significative nell'equilibrio metabolico e nelle

reti di regolazione. Da tali risultati emerge l'ipotesi che, nel corso dell'evoluzione, la vicinanza genomica dei geni funzionalmente correlati possa aver rappresentato una strategia per compensare o attenuare tali squilibri. In questa prospettiva, la molteplicità genica potrebbe aver agito non solo come conseguenza della replicazione multifork, ma anche come una delle possibili forze selettive alla base della formazione di cluster metabolici e operoni ([6]).

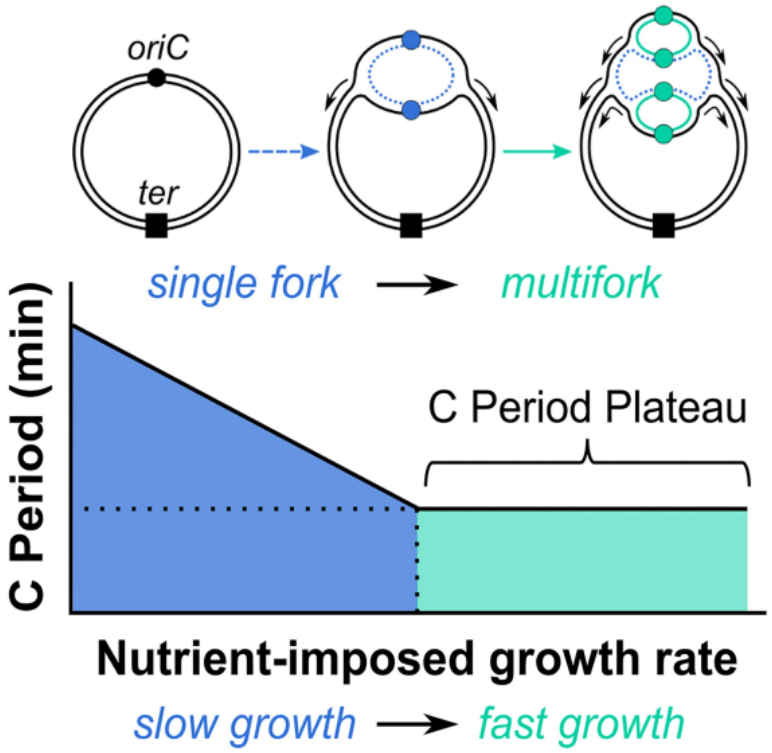


Figura 1: **A**: schema del cromosoma batterico che mostra l'origine di replicazione (oriC) e il terminatore (ter). L'inizio della replicazione genera una singola bolla replicativa, definita replicazione a singola forca. Quando un nuovo evento di inizio avviene prima che il ciclo di replicazione precedente sia completato, si instaura la replicazione multiforca.
B: secondo il modello di Cooper–Helmstetter, all'aumentare della velocità di crescita imposta dal nutriente, la durata del periodo di replicazione (C-period) diminuisce fino a raggiungere un valore minimo costante nelle condizioni di crescita più rapide ([21]).

1.1.1 *Effetti della molteplicità sull'espressione genica*

La molteplicità genica, conseguenza diretta della replicazione multiforca, comporta che geni localizzati in regioni diverse del cromosoma batterico siano presenti in numero di copie variabile durante il ciclo cellulare. In particolare, i geni prossimi all'origine di replicazione (oriC) vengono duplicati per primi e possono coesistere in più copie rispetto a quelli prossimi al terminatore (ter) ([21], [18]). Questa disomogeneità nella disponibilità genica introduce una variazione nel potenziale trascrizionale della cellula, che può tradursi in differenze nei livelli di espressione.

Jaruszewicz-Błońska et al. ([10]) descrivono il rapporto medio di copie tra due geni 1 e 2 come funzione della loro posizione relativa al cromosoma e del tempo di raddoppio cellulare T, secondo la relazione:

$$R_{1/2} = 2^{\frac{C}{T}(L_2 - L_1)} \tag{1.1}$$

dove C rappresenta la durata del periodo di replicazione, mentre L_1 e L_2 indicano le posizioni normalizzate dei geni rispetto all'origine ($L = 0$ in oriC, $L = 1$ in ter). Dall' equazione (1.1) si ricava che il rapporto di copie medio tra origine e terminatore è dato da:

$$R_{O/T} = 2^{C/T} \tag{1.2}$$

Per valori tipici di $C = 40\,\text{min}$ e $T = 20\,\text{min}$, il rapporto $R_{O/T}$ assume valori compresi tra 3 e 4, suggerendo che l'organizzazione cromosomica possa introdurre differenze sistematiche nel dosaggio genico, indipendenti dai meccanismi regolatori specifici.

In modo equivalente, il numero medio di copie di un gene x durante il ciclo cellulare può essere espresso come funzione del tempo di raddoppio τ, della posizione relativa p_x del gene lungo l'asse oriC–ter, e dei tempi C e D, secondo la relazione proposta da fondi et al. ([6]):

$$n_x = 2^{\frac{(1-p_x)C+D}{\tau}} \tag{1.3}$$

In questa formulazione, $(1 - p_x)C$ rappresenta il tempo che intercorre tra il passaggio della forcella di replicazione sul gene x e la terminazione della replicazione, mentre D è l'intervallo tra la fine della replicazione e la successiva divisione cellulare. L'esponente esprime dunque la frazione del ciclo cellulare in cui il gene si trova duplicato, e il fattore $2^{(\cdot)/\tau}$ quantifica il numero medio di copie presenti in una popolazione in crescita

esponenziale. Il modello predice che i geni situati vicino all'origine di replicazione ($p_x \approx 0$) possiedano, in media, un numero di copie pari a $n_x = 2^{(C+D)/\tau}$, mentre per i geni in prossimità del terminatore ($p_x \approx 1$) il numero medio scende a $n_x = 2^{D/\tau}$. In condizioni di crescita lenta, quando $\tau \gg C + D$, tutte le regioni cromosomiche presentano in media una sola copia, coerentemente con la replicazione non sovrapposta (*single fork*). Al contrario, durante la crescita rapida, la presenza di più cicli di replicazione sovrapposti porta a un incremento del numero di copie nelle regioni prossimali a oriC, amplificando localmente il potenziale trascrizionale.

Considerando quindi il numero medio di copie per singolo gene n_x (eq (1.3)) e assumendo costante l'attività del promotore $\Phi(\pi_x)$, è possibile modellare l'azione della molteplicità genica sull'attività trascrizionale $\langle \mathcal{A}_x \rangle$ come:

$$\langle \mathcal{A}_x \rangle \;=\; n_x \, \Phi(\pi_x) \tag{1.4}$$

Tale relazione rappresenta una semplificazione del sistema reale, poiché ingloba nella funzione $\Phi(\pi_x)$ tutti gli altri fattori che influenzano la trascrizione — come la disponibilità di RNA polimerasi, la presenza di fattori di trascrizione, le variazioni locali della struttura cromosomica e la regolazione metabolica — i quali rendono $\Phi(\pi_x)$ una funzione multiparametrica complessa che può, in parte, mascherare o modulare l'effetto diretto della molteplicità genica sull'attività trascrizionale effettiva.

1.1.2 *Possibili effetti sul metabolismo*

E' ragionevole pensare che, la molteplicità genica, alterando il numero relativo di copie dei geni codificanti per gli enzimi metabolici, possa introdurre variazioni significative nei flussi di reazione che tali enzimi catalizzano. Anche una piccola differenza nella concentrazione relativa di un enzima può infatti modificare la velocità di una reazione e, di conseguenza, l'intero bilancio dei metaboliti intermedi. Questo legame tra dosaggio genico e metabolismo è stato affrontato formalmente da fondi et al, che hanno utilizzato i principi della *Metabolic Control Analysis*, MCA) per descrivere l'impatto delle variazioni di copia sull'omeostasi metabolica.

L'MCA è un approccio quantitativo introdotto da Kacser e Burns negli anni '70 per studiare come variazioni nei livelli enzimatici influenzino i flussi e le concentrazioni dei metaboliti all'interno di reti biochimiche complesse. Essa fornisce un quadro teorico e matematico rigoroso per quantificare il grado di controllo esercitato da ciascun componente su un processo metabolico o di segnalazione, identificando i punti chiave che determinano la risposta del sistema a perturbazioni genetiche o fisiologiche ([12], [5]). In altre parole, l'MCA non si limita alla descrizione cinetica di singole reazioni, ma valuta come le variazioni locali di un enzima o di un metabolita si propaghino a livello di rete, offrendo una misura della robustezza e della distribuzione del controllo nei sistemi biologici complessi ([5]).

Nel quadro dell'MCA, l'effetto di una perturbazione nel livello di un enzima E_i sul flusso globale J di una via metabolica può essere descritto dal cosiddetto *flux amplification factor*:

$$\frac{J^{r_i}}{J^0} = \frac{1}{1 - \sum_{i=j}^{m} C_{E_i}^{J^0}(r_i - 1)}, \qquad C_{E_i}^{J} = \frac{\partial J}{\partial E_i}\frac{E_i}{J} = \frac{\partial \ln J}{\partial \ln E_i} \tag{1.5}$$

In questa formulazione, J^0 e J^{r_i} rappresentano rispettivamente i flussi nello stato di riferimento e nello stato perturbato, $r_i = [E_i]_{new}/[E_i]_{old}$ è la variazione relativa nella concentrazione dell'enzima E_i, e $C_{E_i}^{J}$ è il coefficiente di controllo del flusso, che misura quanto la variazione di un singolo enzima influenzi l'intero flusso metabolico.

Quando si considera che le variazioni di $[E_i]$ possono derivare da differenze nel numero di copie geniche (n_x) associate a diverse posizioni sul cromosoma, la relazione precedente collega direttamente la molteplicità genica alla capacità della cellula di mantenere l'omeostasi metabolica. In altre parole, un aumento locale del numero di copie di un gene enzimatico ($r_i > 1$) comporta un incremento della concentrazione dell'enzima e, di conseguenza, del flusso J, mentre una riduzione ($r_i < 1$) produce l'effetto opposto. Poiché i flussi metabolici sono il risultato dell'interazione di molte

reazioni interconnesse, anche piccole variazioni nel dosaggio di un sottoinsieme di geni possono propagarsi nel sistema, alterando le concentrazioni stazionarie dei metaboliti.

A favore di tale modello, vi sono studi di simulazione su reti metaboliche ([6]) che hanno mostrato come tali perturbazioni possono compromettere la stabilità dei flussi e spostare l'equilibrio del metabolismo verso stati subottimali. Gli autori hanno inoltre suggerito che, per minimizzare questi effetti, l'evoluzione possa aver favorito la vicinanza genomica dei geni coinvolti nella stessa via metabolica, così da mantenere proporzioni costanti nei loro livelli di espressione durante la replicazione multiforcata. Questa organizzazione spaziale dei geni rappresenterebbe una forma di compensazione naturale, in grado di attenuare le fluttuazioni introdotte dal processo replicativo.

In sintesi, la molteplicità genica può essere interpretata come un meccanismo di modulazione dinamica del metabolismo, capace di alterare transitoriamente la capacità catalitica della cellula durante la crescita rapida. Pur trattandosi di un effetto indiretto e modulato da numerosi fattori regolativi, la sua influenza sull'abbondanza genica suggerisce un possibile collegamento funzionale con il metabolismo (eq (1.5)), ma la quantificazione precisa di tale effetto rimane estremamente complessa, a causa della natura intrinsecamente multifattoriale del fenomeno e delle numerose interazioni che ne mascherano l'impatto diretto.

1.2 LA BIOLOGIA DEI SISTEMI

1.2.1 *Approccio teorico: la modellizzazione vincolata dei sistemi metabolici*

Per valutare in che modo la molteplicità genica possa influenzare il metabolismo cellulare, è stato adottato un approccio di modellizzazione basato su sistemi vincolati (*constraint-based modeling*, CBM), uno dei pilastri della biologia quantitativa. Questo approccio permette di descrivere il comportamento funzionale di una cellula o di un insieme di reazioni metaboliche non attraverso la conoscenza esplicita di tutte le costanti cinetiche (*Km*), ma imponendo vincoli stechiometrici, termodinamici e fisiologici al sistema ([15, 27, 17]).
In un modello metabolico vincolato, l'insieme delle reazioni biochimiche che avvengono in una cellula è rappresentato in forma compatta mediante una matrice stechiometrica $S_{m\times n}$, dove m è il numero dei metaboliti e n il numero delle reazioni considerate nel sistema. Ogni elemento S_{ij} della matrice indica il coefficiente stechiometrico del metabolita i nella reazione j: Considerando una generica reazione j all'interno della matrice, il segno negativo è associato ai metaboliti consumati, mentre quello positivo a quelli prodotti dalla specifica reazione. A ciascuna reazione è associata una variabile di flusso v_j, che rappresenta la velocità con cui la reazione j avviene in condizioni stazionarie. L'insieme dei flussi metabolici può quindi essere descritto da un vettore colonna che appartiene allo spazio vettoriale $\mathbb{R}^n$, (Figura 2).

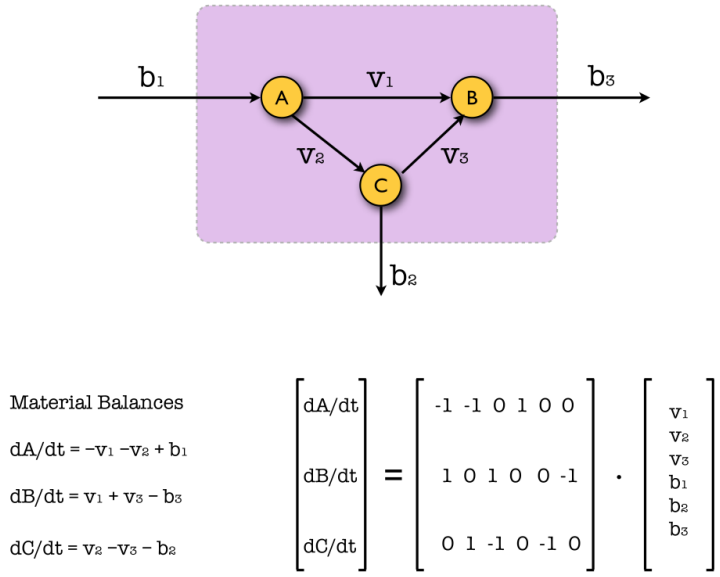


Figura 2: Schema di un sistema metabolico elementare (A, B, C) e della corrispondente matrice stechiometrica che descrive le relazioni tra le reazioni (v1, v2, v3) e cambi con l'ambiente (b1, b2, b3) secondo l'equazione dx/dt = Sv

Questa formulazione permette di tradurre la complessità del metabolismo cellulare in un problema algebrico lineare, fornendo la base matematica per le analisi vincolate come l'FBA e le sue estensioni ([15, 17]).

Rappresentando la cellula come uno spazio vettoriale appartenente a $\mathbb{R}^n$, dove n è il numero totale di reazioni che compongono il sistema metabolico, ogni punto di questo spazio corrisponde a una possibile distribuzione dei flussi metabolici $\mathbf{v} = (v_1, v_2, \ldots, v_n)$. In assenza di vincoli, esisterebbero infiniti punti nello spazio delle soluzioni, ognuno dei quali rappresenterebbe uno stato potenzialmente realizzabile del metabolismo cellulare.

L'introduzione dei vincoli biologici e fisici consente di delimitare tale spazio, definendo un insieme finito e biologicamente plausibile di soluzioni ammissibili. I principali vincoli che caratterizzano un modello metabolico vincolato sono:

- **Vincolo di bilancio di massa (steady state):** si assume che, a stato stazionario, la concentrazione di ciascun metabolita rimanga costante nel tempo, condizione formalizzata come:

  $$S \cdot \mathbf{v} = 0 \tag{1.6}$$

  dove $S_{m \times n}$ è la matrice stechiometrica che descrive la partecipazione dei m metaboliti alle n reazioni del sistema e $\mathbf{v}$ il vettore di flussi.

- **Vincoli di capacità o termodinamici:** ogni reazione possiede un limite inferiore e superiore, che riflette le condizioni termodinamiche e le capacità enzimatiche massime:

$$v_i^{min} \leqslant v_i \leqslant v_i^{max} \tag{1.7}$$

  Tali vincoli impediscono che le reazioni assumano flussi non realistici dal punto di vista fisiologico.

- **Vincoli di direzionalità:** per le reazioni irreversibili si impone che $v_i \geqslant 0$, mentre per quelle reversibili il flusso può assumere anche valori negativi ($v_i \in \mathbb{R}$), rappresentando il senso opposto della reazione.

- **Vincoli addizionali o contestuali:** in alcuni modelli è possibile introdurre ulteriori limitazioni legate a specifiche condizioni sperimentali, come la disponibilità di substrati o la regolazione genica (ad esempio, bloccando una reazione dopo una *gene knockout*).

L'insieme dei punti $\mathbf{v}$ che soddisfano i vincoli (1.6) e (1.7) definisce lo spazio delle soluzioni ammissibili o *feasible space*, che rappresenta tutte le configurazioni metabolicamente possibili della cellula coerenti con la struttura stechiometrica e le limitazioni biochimiche del sistema, (Figura 3).

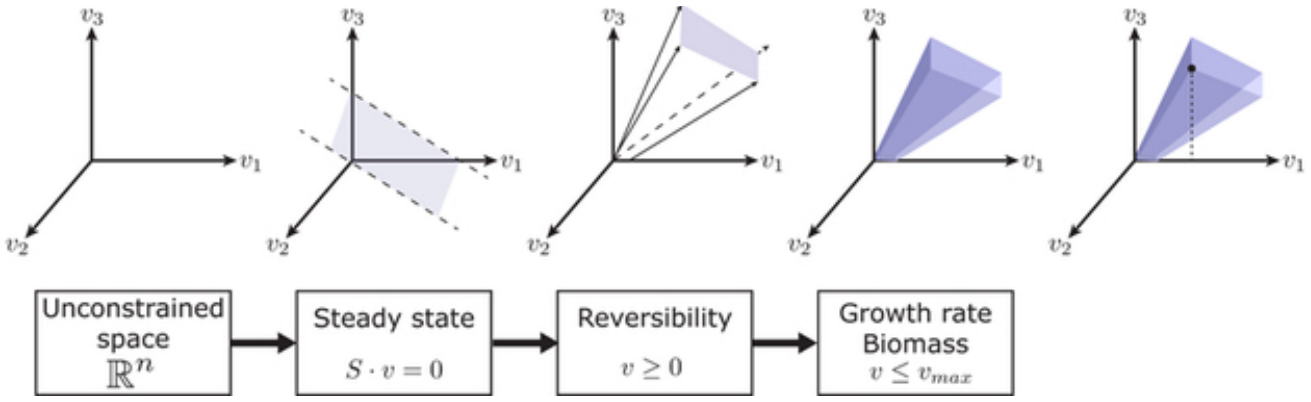


Figura 3: Rappresentazione geometrica di un sistema metabolico nello spazio vettoriale $\mathbb{R}^n$. I vincoli stechiometrici ($S\mathbf{v} = 0$) e di capacità ($a_i \leqslant v_i \leqslant b_i$) delimitano lo spazio delle soluzioni ammissibili. Ottimizzazione di una funzione obiettivo per la selezione di un'unica soluzione ottimale nello spazio delimitato dai vincoli. ([8])

1.2.2 *Analisi del bilancio di flusso, FBA e pFBA*

La *Flux Balance Analysis* (FBA), rappresenta uno degli strumenti fondamentali della biologia dei sistemi per la simulazione del metabolismo cellulare su larga scala ([15, 29]). Si tratta di un approccio matematico basato su modelli vincolati che sfrutta la programmazione lineare per andare a selezionre una soluzione ottimale nello spazio delimitato dai vincoli del modello, (Figura 3).

Come descritto in precedenza, il sistema metabolico è rappresentato dalla matrice stechiometrica $S_{m\times n}$ (Figura 2) in cui m è il numero dei metaboliti e n il numero delle reazioni. questo porta alla costruzione di un sistema di equazioni differenziali, che assumendo di trovarci in stato stazionario, quindi senza accumulo di metaboliti è possibile eguagliare a 0 come segue:

$$\begin{cases} \dfrac{dx_1}{dt} = S_{11}v_1 + S_{12}v_2 + \cdots + S_{1n}v_n, \\ \dfrac{dx_2}{dt} = S_{21}v_1 + S_{22}v_2 + \cdots + S_{2n}v_n, \\ \quad\vdots \\ \dfrac{dx_m}{dt} = S_{m1}v_1 + S_{m2}v_2 + \cdots + S_{mn}v_n, \end{cases} \qquad \Longrightarrow \quad \frac{d\mathbf{x}}{dt} = 0 \;\Rightarrow\; S\cdot\mathbf{v} = 0 \tag{1.8}$$

Dall'approssimazione a stato stazionario, si ottiene quindi un sistema di equazioni lineari, dove però analizzando contesti biologici reali, il numero di flussi sovrasta quello delle equazioni, rendendo infinite le combinazioni di flussi che soddisfano il sistema.
Per ovviare a ciò vengono aggiunti i limiti di capacità e direzionalità delle reazioni, formalizzati come:

$$v_i^{min} \leqslant v_i \leqslant v_i^{max}, \qquad i = 1, 2, \ldots, n \tag{1.9}$$

A questo punto, il modello consente di individuare una soluzione biologicamente significativa introducendo una funzione obiettivo (*objective function*). Non viene più quindi solo ricercata una soluzione tra quelle possibili, ma quella che massimizza un dato flusso, spesso come funzione obiettivo viene impostata la crescita cellulare o la produzione di un determinato metabolita ([15]). Oppure una combianzione di reazioni, tramite l'uso di un vettore di pesi:

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{bmatrix}, \qquad \text{con} \quad \lambda_i \geqslant 0, \quad \sum_{i=1}^{k} \lambda_i = 1 \tag{1.10}$$

dove $\boldsymbol{\lambda}$ è il vettore dove sono indicati i pesi da assegnare alle singole reazioni. Esso sarà impiegato per andare a risolvere il seguente problema di ottimizzazione lineare:

$$\max_{\mathbf{v}} \; Z = \boldsymbol{\lambda}^T \mathbf{v} \quad \text{soggetto a} \quad S\mathbf{v} = 0, \; v_{\min} \leqslant \mathbf{v} \leqslant v_{\max} \tag{1.11}$$

dove:

- $\mathbf{v}$ è il vettore dei flussi metabolici ($n \times 1$);
- $\boldsymbol{\lambda}$ è il vettore dei pesi ($n \times 1$), in cui ciascun elemento λ_i rappresenta il contributo della reazione i alla funzione obiettivo Z;

L'FBA consente quindi di prevedere come i flussi metabolici possano riorganizzarsi in risposta a vincoli ambientali o genetici, e si è dimostrato estremamente efficace nella predizione del tasso di crescita microbico, nella simulazione di *knockout* genici e nell'identificazione di percorsi metabolici alternativi ([29, 15, 17]). Tuttavia, una delle sue limitazioni principali risiede nel fatto che più configurazioni di flusso possono portare allo stesso valore ottimale della funzione obiettivo, rendendo la soluzione non unica e potenzialmente ridondante da un punto di vista biologico.

Per affrontare questo problema, è stata sviluppata la versione *parsimonious* dell'analisi, denominata pFBA (*parsimonious Flux Balance Analysis*) ([13]). L'idea alla base della pFBA consiste nel determinare, una volta noto il valore ottimale della funzione obiettivo primaria $Z*$ (ad esempio, la massima velocità di crescita), la soluzione che minimizza il costo complessivo dei flussi metabolici tra tutte quelle che producono lo stesso $Z*$. Tale costo è espresso come la somma dei valori assoluti dei flussi, rappresentando la quantità totale di attività enzimatica richiesta per sostenere il metabolismo:

$$\min_{\mathbf{v}} \sum_{i=1}^{n} |v_i| \quad \text{soggetto a} \quad S\mathbf{v} = 0, \quad v_{\min} \leqslant \mathbf{v} \leqslant v_{\max}, \quad Z = Z^* \tag{1.12}$$

In questa formulazione:

- $\sum |v_i|$ rappresenta la norma L_1 del vettore dei flussi $\mathbf{v}$, utilizzata come misura della *parsimonia metabolica*;
- il vincolo $Z = Z^*$ garantisce che la cellula mantenga la massima crescita (o il massimo rendimento) ottenuto nel primo livello di ottimizzazione.

Questo approccio seleziona quindi la soluzione più parsimoniosa compatibile con la crescita massima, riflettendo l'ipotesi biologica secondo cui i sistemi cellulari tendono a minimizzare l'uso di risorse catalitiche. La pFBA ha trovato ampio impiego nello studio del metabolismo microbico e nel collegamento tra flussi metabolici e livelli di espressione genica, rendendola una scelta ideale per analizzare gli effetti della molteplicità genica sui flussi cellulari.

1.2.3 *Analisi dello spazio dei flussi mediante Flux Sampling*

La *Flux Balance Analysis* (FBA), fornisce una singola soluzione ottimale del sistema, determinata massimizzando una funzione obiettivo all'interno dello spazio delle soluzioni ammissibili. Tuttavia, poiché il problema è in genere sottodeterminato, esistono molteplici configurazioni di flusso che soddisfano i vincoli del modello e producono lo stesso valore ottimale della funzione obiettivo. Per esplorare la variabilità e la struttura dello spazio delle soluzioni ammissibili, è possibile ricorrere a metodi di campionamento dello spazio dei flussi, noti come *Flux Sampling* ([24, 7]).

Il *Flux Sampling* è un approccio probabilistico che consente di ottenere una distribuzione rappresentativa dei possibili vettori di flusso $\mathbf{v}$ compatibili con i vincoli del sistema:

$$S\mathbf{v} = 0, \qquad v_{min} \leqslant \mathbf{v} \leqslant v_{max} \tag{1.13}$$

L'obiettivo non è più quello di trovare una singola soluzione ottimale, ma di generare un insieme di campioni $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ che descrivano la distribuzione dello spazio delle soluzioni ammissibili (*feasible space*). Ciascun campione rappresenta una possibile configurazione di flussi coerente con la struttura stechiometrica e i vincoli del modello.

A livello computazionale, il campionamento può essere effettuato con diversi algoritmi Monte Carlo, tra cui il metodo *Hit-and-Run* e le sue varianti (*Artificial Centered Hit-and-Run*, ACHR), esso è il principale algoritmo implementato nella libreria *"COBRA Toolbox"* e si basa sulla generazione di una sequenza di punti uniformemente distribuiti all'interno dello spazio convesso definito dai vincoli ([19]). L'insieme dei campioni ottenuti può essere organizzato in una matrice di dimensioni $k \times n$:

$$V = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1n} \\ v_{21} & v_{22} & \dots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \dots & v_{kn} \end{bmatrix} \tag{1.14}$$

dove ogni riga rappresenta un vettore di flusso $\mathbf{v}_i$ (una possibile soluzione del modello) e ogni colonna v_j rappresenta la distribuzione dei flussi associati alla reazione j. In questo modo, la matrice V può essere interpretata come un insieme di osservazioni di una variabile aleatoria multivariata $\mathbf{V} \in \mathbb{R}^n$, la cui densità di probabilità $p(\mathbf{v})$ descrive la distribuzione dei flussi metabolicamente plausibili:

$$\mathbf{V} \sim p(\mathbf{v}) \quad \text{con supporto} \quad \Omega = \{\mathbf{v} \in \mathbb{R}^n : S\mathbf{v} = 0,\ v_{min} \leqslant \mathbf{v} \leqslant v_{max}\} \tag{1.15}$$

In sintesi, il *Flux Sampling* fornisce una rappresentazione statistica dello spazio dei flussi compatibili con i vincoli del modello, offrendo una prospettiva complementare all'FBA classico: mentre quest'ultimo identifica un singolo punto ottimale, il campionamento ne descrive la distribuzione complessiva.

	A	B	C	D	E
1		SHK3Dr	DHORTS	OMPDC	G5SD
2	0	0.165138765560631	-0.142227739477932	0.142227739477932	0.0953729161214449
3	1	0.120382713187541	-0.103681051105601	0.103681051105601	0.0695248636106723
4	2	0.104973450333428	-0.0904096447989651	0.0904096447989651	0.0606255256827456
5	3	0.138516330588989	-0.119298857450738	0.119298857450738	0.0799976149035192
6	4	0.108016276478211	-0.0930303059399115	0.0930303059399115	0.0623828487583212
7	5	0.172179956145251	-0.148292045859187	0.148292045859187	0.0994394265194419
8	6	0.111512935772276	-0.0960418477426321	0.0960418477426321	0.0644022827048798
9	7	0.152803452681353	-0.131603801872297	0.131603801872297	0.0882488774842458
10	8	0.153418989948849	-0.132133943025265	0.132133943025265	0.0886043714936316
11	9	0.171827476341025	-0.147988470894169	0.147988470894169	0.0992358598452016
12	10	0.166263645132767	-0.143196558594318	0.143196558594318	0.0960225721370111
13	11	0.185411372131222	-0.159687774552899	0.159687774552899	0.107081001153286
14	12	0.184507541904309	-0.15890933828689	0.15890933828689	0.106559009183554
15	13	0.108251029465734	-0.0932325077635116	0.0932325077635116	0.062518438274102
16	14	0.170150330960364	-0.146544017814154	0.146544017814154	0.0982672604921729
17	15	0.137163594256447	-0.118133798648719	0.118133798648719	0.0792163674851066
18	16	0.138419381723743	-0.119215386209362	0.119215386209362	0.0799416421622154
19	17	0.169852902067281	-0.146287914694679	0.146287914694679	0.0980955270587555
20	18	0.109214301104915	-0.094062226011117	0.094062226011117	0.0630748185711886
21	19	0.149386174785642	-0.128660632412641	0.128660632412641	0.0862752916452068
22	20	0.104679300080432	-0.0901564230130081	0.0901564230130081	0.0604557246571369
23	21	0.123710640231447	-0.106547322757965	0.106547322757965	0.0714468847692796
24	22	0.116552544170812	-0.100382320818503	0.100382320818503	0.0673128514276665

Figura 4: Parte di una matrice di flux sampling estratta dalle analisi successive. Ogni colonna indicana una reazione metabolica e ogni riga un campionamento di flusso per tale reazione

1.2.4 *Dati omici e integrazione nei modelli metabolici*

L'integrazione di dati *omici* nei modelli metabolici rappresenta uno degli sviluppi più rilevanti della biologia dei sistemi, poiché consente di passare da rappresentazioni puramente strutturali del metabolismo a descrizioni condizionali e contestualizzate ([16]). Attraverso l'uso di dati trascrittomici, proteomici o metabolomici, è possibile costruire modelli specifici per tipo cellulare, tessuto o condizione ambientale, vincolando ulteriormente lo spazio delle soluzioni compatibili con le osservazioni sperimentali.

Negli ultimi anni sono stati sviluppati diversi algoritmi per l'integrazione sistematica dei dati omici nei modelli a scala genomica. Tra i più noti si possono citare:

- **GIMME** (*Gene Inactivity Moderated by Metabolism and Expression*) ([2]), il quale utilizza i livelli di espressione genica per penalizzare le reazioni associate a geni scarsamente espressi, selezionando una configurazione di flussi coerente con i dati trascrittomici;

- **iMAT** (*integrative Metabolic Analysis Tool*) ([30]), che bilancia in modo combinatorio le reazioni attive e inattive in base ai livelli di espressione genica, producendo uno stato metabolico compatibile con il profilo trascrizionale osservato;

- **RIPTiDe** (*Reaction Inclusion by Parsimony and Transcript Distribution*) ([11]), un approccio che integra i dati di espressione genica all'interno di un quadro parsimonioso basato sulla *parsimonious Flux Balance Analysis* (pFBA, paragrafo 1.2.2). RIPTiDe combina l'ipotesi di economia cellulare con le informazioni trascrittomiche per identificare le vie metaboliche più coerenti con lo stato trascrizionale della cellula. In pratica, il metodo assegna a ciascuna reazione un peso proporzionale all'abbondanza dei trascritti dei geni che trascrivono per gli enzimi che la catalizzano, e ricerca la distribuzione di flussi più parsimoniosa (cioè con uso minimo di risorse) che consenta di mantenere un livello ottimale di crescita. Questo consente di ottenere un modello contestuale del metabolismo, capace di riflettere la fisiologia cellulare in condizioni specifiche, anche in assenza di informazioni dettagliate sui vincoli ambientali. RIPTiDe è stato mostrato capace di migliorare le predizioni di attività metaboliche specifiche di contesto sia in esperimenti *in vitro* sia in dati metatranscrittomici *in vivo*, rispetto ad approcci precedenti basati su soglie arbitrarie di espressione o sull'integrazione diretta dei dati trascrittomici.

Questi strumenti permettono di generare modelli metabolici *context-specific*, capaci di descrivere la fisiologia di un organismo in condizioni particolari di crescita o stress, fornendo un ponte tra dati sperimentali ad alta risoluzione e predizioni computazionali. L'output comune a tutti i tools sopra citati è un modello metabolico *pruned* rispetto a

quello di partenza, ossia una versione ridotta in cui vengono eliminate le reazioni non essenziali per la funzionalità metabolica e che, in accordo con i dati omici, risultano scarsamente espresse. In molti casi, oltre alla rimozione delle reazioni, vengono anche modificati i vincoli di capacità delle stesse (v_{min} e v_{max}) per rendere le distribuzioni dei flussi più coerenti con i livelli di espressione genica e quindi con la fisiologia osservata della cellula.

1.3 FORMALIZZAZIONE OBIETTIVI E DOMANDE DI RICERCA

In questo lavoro si da per acquisito l'effetto della molteplicità genica sull'espressione genica, ossia l'aumento temporaneo del dosaggio dei geni situati in prossimità dell'origine di replicazione durante la crescita rapida. Sulla base di tale considerazione, l'obiettivo principale è valutare se, e in che misura, tale effetto si rifletta sull'attività metabolica della cellula, analizzando come le variazioni di dosaggio genico possano modificare la distribuzione dei flussi metabolici all'interno della rete.

Per indagare questo aspetto sono state impiegate tecniche di *Flux Sampling*, che consentono di esplorare la variabilità e la struttura statistica dello spazio dei flussi compatibili con i vincoli del modello. L'approccio adottato mira a confrontare le distribuzioni dei flussi metabolici in condizioni corrispondenti a diversi gradi di molteplicità genica, al fine di identificare possibili alterazioni sistematiche e quantificare l'entità con cui la replicazione può influenzare, indirettamente, l'omeostasi metabolica.

2

APPROCCI METODOLOGICI E MODELLI UTILIZZATI

2.1 PRESENTAZIONE DEL DATASET E APPARATO SPERIMENTALE

Per l'integrazione dei dati trascrittomici nei modelli metabolici è stato utilizzato il dataset di espressione genica pubblicato da sastry et al. ([23]), che rappresenta una delle più ampie e dettagliate analisi di trascrittomica condotte su *Escherichia coli*. In tale lavoro, gli autori hanno eseguito esperimenti di *RNA-Seq* su un'ampia collezione di ceppi di *E. coli K-12* sottoposti a differenti condizioni ambientali e fisiologiche, variando il substrato di crescita, la disponibilità di nutrienti e includendo anche mutanti con alterazioni specifiche nei circuiti regolatori. I dati di espressione sono forniti in unità di TPM(*Transcripts Per Million*), una misura normalizzata che consente di confrontare direttamente l'abbondanza relativa dei trascritti tra condizioni diverse. Nel dataset originale, i valori di espressione erano riportati in scala logaritmica log(TPM); ai fini di una corretta interpretazione da parte dell'algoritmo `RIPTiDe` (paragrafo 2.3), essi sono stati riconvertiti in valori di TPM lineari, così da preservare la proporzionalità tra livello trascrizionale e peso assegnato alle reazioni del modello. Ai fini di questo lavoro, sono stati selezionati esclusivamente i campioni relativi a colture cresciute su terreno minimo M9, mantenendo costante la composizione del mezzo e variando unicamente la fonte di carbonio. Le condizioni sperimentali considerate comprendono le seguenti fonti carboniose:

`glucose`, `fumarate`, `pyruvate`, `fructose`, `glycerol`, `xylose`, `sorbitol`, `D-ribose`, `glucarate`, `N-acetylglucosamine`, `galactose`, `gluconate`, `D-lyxose`, `D-arabinose`, `m-tartrate`, `acetate`.

La scelta di questo sottoinsieme di condizioni sperimentali consente di analizzare il comportamento metabolico di E. coli lungo un ampio spettro di stati fisiologici, senza introdurre variazioni nella disponibilità di altri nutrienti o in fattori di stress. Un aspetto di particolare rilievo è che le colture cresciute su differenti fonti di carbonio e con disponibilità diverse mostrano velocità di crescita sensibilmente diverse, poiché ogni substrato implica un diverso rendimento energetico. Questo gradiente di crescita

costituisce un elemento chiave nel presente studio, poiché l'effetto della molteplicità genica — ossia la variazione del numero di copie dei geni durante la replicazione — si presume che tenda ad amplificarsi con l'aumentare del tasso di crescita cellulare. L'utilizzo di dati trascrittomici raccolti in condizioni di crescita differenziate fornisce dunque un contesto sperimentale ideale per indagare la relazione tra velocità di replicazione, regolazione dell'espressione genica e attività metabolica, garantendo al contempo la coerenza biologica dei modelli ottenuti.

2.2 COBRApy: lo standard per l'analisi di modelli vincolati

`COBRApy` è la principale implementazione in linguaggio Python del framework *Constraint-Based Reconstruction and Analysis* (COBRA), progettata per la manipolazione, l'analisi e la simulazione di modelli metabolici vincolati ([4]). La libreria `COBRApy` offre diverse procedure per l'analisi vincolata dei modelli metabolici; Tra le principali figurano la Flux Balance Analysis (FBA), che identifica la distribuzione di flussi ottimale rispetto a una funzione obiettivo; la Flux Variability Analysis (FVA), che esplora l'intervallo di soluzioni compatibili con l'ottimo; e il Flux Sampling, che permette di analizzare la distribuzione probabilistica dei flussi senza la ricerca di uno stato ottimale. `COBRApy` consente inoltre di simulare knock-out genetici o enzimatici tramite analisi di eliminazione di geni o reazioni, e include varianti dell'FBA, come la parsimonious FBA (pFBA), che ricerca soluzioni a minimo costo enzimatico. Grazie a un'architettura orientata agli oggetti e all'integrazione con diversi *solver* di ottimizzazione (ad esempio `GLPK`, `CPLEX` o `Gurobi`), `COBRApy` consente di eseguire queste analisi in modo efficiente e riproducibile anche su modelli di grandi dimensioni. La sua flessibilità lo rende oggi uno strumento di riferimento per l'analisi computazionale di reti metaboliche vincolate e per lo sviluppo di *workflow* personalizzati in biologia dei sistemi.
Nel presente lavoro, `COBRApy` è stato scelto come piattaforma di riferimento per tutte le analisi successive, poiché rappresenta lo *standard de facto* per lo studio computazionale del metabolismo in biologia dei sistemi. La sua integrazione nativa in Python ha inoltre garantito una piena compatibilità con le librerie di *data science* e *data visualization* successivamente impiegate per l'elaborazione, l'analisi statistica e la rappresentazione dei risultati (paragrafo 2.5).

2.2.1 *La complessità dei sistemi batterici in un modello da meno di 12 MB: il formato SBML*

Il formato SBML (*Systems Biology Markup Language*) è stato introdotto come linguaggio indipendente dal software per la rappresentazione e lo scambio di modelli computazionali di reti biochimiche e sistemi biologici complessi ([9]). Basato su `XML`, SBML fornisce una struttura standardizzata e leggibile che descrive in modo formale le componenti di un modello — compartimenti, specie chimiche, reazioni, parametri e regole matematiche — e le relazioni tra esse. Questa formalizzazione garantisce la piena interoperabilità tra diversi strumenti di analisi, mantenendo al contempo la possibilità di modificare e personalizzare il modello a livello di singola reazione o entità molecolare. Nel presente lavoro, il modello metabolico di riferimento è stato `iML1515.xml`, una delle più recenti e complete ricostruzioni genomiche di *Escherichia coli K-12 MG1655*, disponibile nel repository pubblico *BiGG Models* ([14]). Il modello include oltre 2 700 reazioni, circa 1 500 geni e più di 1 800 metaboliti, rappresentando una delle reti metaboliche più dettagliate e validate sperimentalmente per un organismo procariote. La natura gerarchica del formato SBML consente di modificare in modo diretto la struttura del modello: è possibile aggiungere o rimuovere reazioni, introdurre nuovi metaboliti o compartimenti, e modificare i *bounds* (limiti inferiori e superiori di flusso) associati a ciascuna reazione. Queste operazioni permettono di adattare il modello a differenti condizioni sperimentali o a specifiche ipotesi biologiche, mantenendo la coerenza formale della rappresentazione. Ogni reazione nel modello è inoltre associata a una regola detta GPR(Gene–Protein–Reaction rule), che descrive il legame logico tra i geni codificanti gli enzimi e le reazioni da essi catalizzate. Le GPR sono espresse come relazioni booleane che specificano, ad esempio, se un enzima è formato da subunità codificate da geni distinti (relazione `AND`) o se più isoenzimi alternativi possono catalizzare la stessa reazione (relazione `OR`). Questa informazione è cruciale per collegare la struttura genetica del modello con le sue proprietà funzionali e per simulare l'effetto di perturbazioni geniche o di variazioni di espressione. Nel contesto di questo lavoro, il file `iML1515.xml` è stato importato e manipolato mediante l'ambiente `COBRApy`, che ne ha interpretato la struttura interna e permesso l'esecuzione di analisi vincolate di flusso e di *flux sampling*. Inoltre, attraverso l'integrazione con l'algoritmo `RIPTiDe` [11], il modello è stato ulteriormente raffinato adattando i *bounds* delle reazioni ai livelli di espressione genica derivati da dati trascrittomici. In questo modo, le simulazioni metaboliche riflettono in maniera più realistica lo stato fisiologico della cellula nelle specifiche condizioni sperimentali considerate. È riportato in Figura **??** un esempio di file in formato `SBML`, dal quale è possibile apprezzare la struttura gerarchica e l'elevato livello di dettaglio informativo che caratterizzano questo standard di rappresentazione. Ogni sezione del file è organizzata in modo modulare e descrive in maniera esplicita le entità costitutive del modello, come metaboliti e reazioni, permettendo un'interpretazione chiara e non ambigua della rete metabolica.

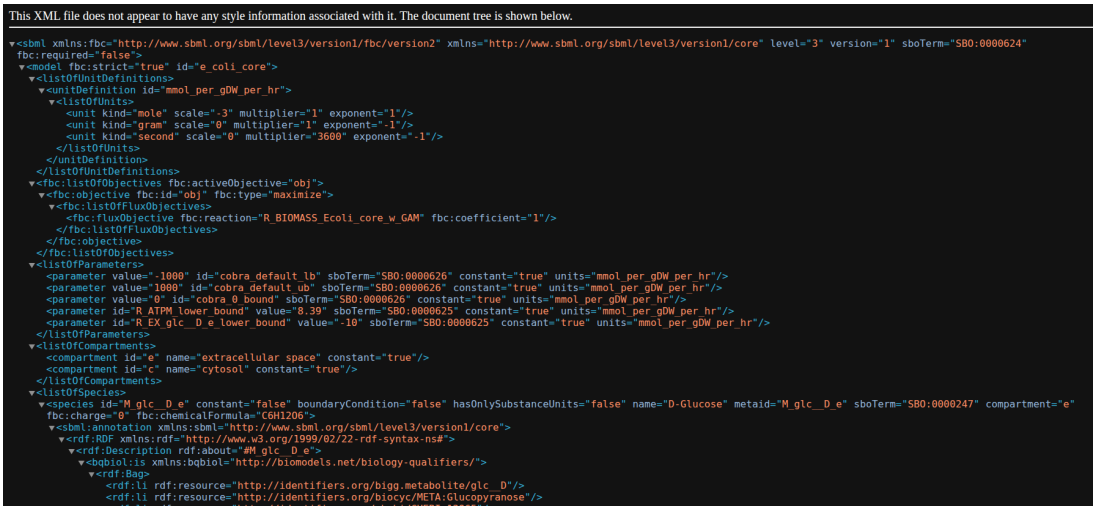


Figura 5: Esempio di struttura interna di un file in formato *SBML* (Systems Biology Markup Language). Nel frammento sono visibili le principali sezioni che definiscono l'architettura del modello: le unità di misura (`<listOfUnitDefinitions>`), gli obiettivi di flusso (`<fbc:listOfObjectives>`), i parametri globali (`<listOfParameters>`), i compartimenti (`<listOfCompartments>`) e le specie metaboliche (`<listOfSpecies>`). In questo esempio, relativo al modello del metabolismo centrale di *E. coli*, è rappresentata la molecola di D-Glucosio (`M_glc_D_e`) con le sue annotazioni esterne e i parametri di flusso associati, come i limiti inferiori e superiori (*bounds*) per la reazione di uptake.

2.3 CONTESTUALIZZAZIONE DEI MODELLI METABOLICI CON RIPTIDE

L'analisi metabolica basata su modelli vincolati può essere resa più realistica attraverso l'integrazione di dati omici (paragrafo 1.2.4), che permettono di adattare la rete metabolica alle condizioni fisiologiche effettive della cellula. A questo scopo è stato impiegato l'algoritmo `RIPTiDe` ([11]), progettato specificamente per integrare dati trascrittomici all'interno di modelli metabolici su scala genomica. RIPTiDe richiede in input due elementi principali:

- Un modello metabolico generale in formato `SBML`, come quello derivato da *iML1515*.
- Un file contenente i livelli di espressione genica, che nel presente lavoro sono stati forniti in unità di **TPM** (*Transcripts Per Million*).

A partire da queste informazioni, l'algoritmo assegna a ciascuna reazione un peso proporzionale al livello di trascrizione dei geni ad essa associati tramite le regole GPR (*Gene–Protein–Reaction*). Le reazioni catalizzate da enzimi con bassa espressione ricevono quindi un peso maggiore, penalizzandone l'utilizzo nel processo di ottimizzazione, mentre quelle con alta espressione risultano più probabili e meno costose dal punto di vista computazionale. Il cuore di RIPTiDe consiste nella risoluzione di un problema di *parsimonious flux balance analysis* (pFBA) (paragrafo 1.2.2) modificato, in cui la somma totale dei flussi viene minimizzata tenendo conto dei pesi derivati dalla trascrittomica. In questo modo, l'algoritmo seleziona la configurazione di flussi metabolicamente più parsimoniosa e coerente con l'investimento trascrizionale osservato. Le reazioni con espressione inferiore a una determinata soglia vengono automaticamente eliminate (*pruning*), generando una rete contestualizzata che riflette solo le vie potenzialmente attive nelle condizioni studiate. L'output di RIPTiDe è un nuovo modello in formato `SBML`, strutturalmente coerente con l'originale ma *prunato* e *pesato* in base ai dati di espressione. Oltre a contenere un sottoinsieme di reazioni attive, il modello risultante presenta anche una **modifica dei vincoli** di flusso, adattati in funzione dei livelli trascrittomici: le reazioni fortemente espresse mantengono ampiezze di flusso più elevate, mentre quelle debolmente espresse vengono ulteriormente vincolate. Questo approccio consente di ottenere modelli specifici per condizione, in cui la rete metabolica risulta ridotta alle sole reazioni compatibili con il profilo di espressione osservato, fornendo una rappresentazione più realistica e biologicamente informata del metabolismo cellulare.

2.3.1 *Parametri di contestualizzazione*

Per la contestualizzazione dei modelli metabolici è stata utilizzata la funzione `maxfit()` del pacchetto `RIPTiDe`, che implementa un approccio di ottimizzazione parsimoniosa

vincolato dai dati trascrittomici. La procedura di `maxfit` consiste nel ricercare il valore di frazione attiva (*fraction*) che massimizza la coerenza statistica tra i livelli di espressione genica e la struttura del modello, valutata in termini di significatività (p-value) della correlazione tra i pesi assegnati alle reazioni e i valori trascrizionali osservati. In altre parole, l'algoritmo individua il compromesso ottimale tra parsimonia del flusso e aderenza ai dati sperimentali.

Nelle analisi di questo lavoro, l'intervallo di frazione della funzione obiettivo da mantenere, è stato impostato tra `frac_min = 0.5` e `frac_max = 0.95`. Ciò ha permesso di garantire un elevato peso ai dati trascrittomici nella definizione delle reazioni attive, preservando al contempo la coerenza biologica del modello e prevenendo l'esclusione di vie metaboliche essenziali per la produzione di biomassa. Tutti gli altri parametri della funzione `maxfit()` sono stati mantenuti ai valori di default, ad eccezione del parametro `gpr`, impostato su `True` per includere le relazioni logiche di tipo `AND/OR` nelle regole gene–protein–reaction (GPR). Questo ha permesso di considerare correttamente l'effetto combinatorio dei geni nella determinazione delle capacità enzimatiche, rendendo l'assegnazione dei pesi più coerente con l'organizzazione genica sottostante.

2.4 ENERGY DISTANCE COME MISURA DI DISSIMILARITÀ TRA DISTRIBUZIONI DI FLUSSO CAMPIONATE

L'*Energy Distance* è una misura statistica non parametrica della dissimilarità tra distribuzioni di probabilità, introdotta da Gábor J Székely et al.([26]) e formalizzata successivamente come parte della famiglia delle *energy statistics*. Essa quantifica quanto due distribuzioni multivariate F e G, definite su uno spazio metrico, differiscano tra loro e risulta pari a zero se e solo se le distribuzioni coincidono. Formalmente, la distanza di energia al quadrato è definita come:

$$D^2(F, G) = 2\,\mathbb{E}\big[\|X - Y\|\big] - \mathbb{E}\big[\|X - X'\|\big] - \mathbb{E}\big[\|Y - Y'\|\big] \tag{2.1}$$

dove $X, X' \sim F$ e $Y, Y' \sim G$ sono variabili casuali indipendenti e $\|\cdot\|$ rappresenta la norma Euclidea. A differenza di molte statistiche classiche, l'Energy Distance non si limita a confrontare le medie delle distribuzioni, ma è sensibile anche alle differenze di dispersione e forma. Infatti, essa aumenta quando le due distribuzioni differiscono non solo per il centro (media), ma anche per la loro ampiezza o struttura interna, rendendola uno strumento particolarmente adatto all'analisi di dati multivariati. In termini intuitivi, può essere interpretata come una misura "di energia potenziale" tra due insiemi di punti: quanto più i campioni di F e G risultano separati nello spazio delle distanze, tanto maggiore sarà la loro dissimilarità. Maria L. Rizzo et al. ([20] e [25]) hanno dimostrato che questa proprietà conferisce all'Energy Distance una notevole

potenza nel rilevare differenze sia di localizzazione che di dispersione, superando i limiti dei test parametrici tradizionali che si focalizzano unicamente sulle medie. Nel contesto della biologia dei sistemi, questa misura risulta particolarmente utile per valutare la dissimilarità tra distribuzioni di flussi metabolici ottenute mediante *flux sampling*. Ogni condizione sperimentale o configurazione del modello genera una distribuzione distinta dei flussi; confrontarle con l'Energy Distance consente di quantificare in modo oggettivo quanto due stati metabolici differiscano tra loro, tenendo conto simultaneamente di variazioni nella media dei flussi (cioè nello stato metabolico medio) e nella loro varianza (cioè nella flessibilità o ridondanza del metabolismo). In questo modo, la misura fornisce una caratterizzazione completa delle differenze fenotipiche simulate, cogliendo variazioni strutturali che non emergerebbero da confronti basati solo su statistiche riassuntive. Dal punto di vista empirico, l'Energy Distance può essere calcolata direttamente a partire da due insiemi finiti di campioni $X = \{x_1, \ldots, x_n\}$ e $Y = \{y_1, \ldots, y_m\}$ (ad esempio due matrici di *flux sampling*) come:

$$\widehat{D}(F,G) = 2\,\overline{d}_{XY} - \overline{d}_{XX} - \overline{d}_{YY} \tag{2.2}$$

dove

$$\overline{d}_{XY} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|x_i - y_j\|, \quad \overline{d}_{XX} = \frac{1}{n(n-1)}\sum_{i\neq j}\|x_i - x_j\|, \quad \overline{d}_{YY} = \frac{1}{m(m-1)}\sum_{i\neq j}\|y_i - y_j\| \tag{2.3}$$

In questa formulazione, il termine $\overline{d}_{XY}$ rappresenta la distanza media tra i flussi appartenenti a condizioni differenti, mentre $\overline{d}_{XX}$ e $\overline{d}_{YY}$ descrivono la distanza media interna a ciascuna distribuzione. Valori elevati di $\widehat{D}(F,G)$ indicano quindi che le due distribuzioni di flusso divergono in modo significativo. Al contrario, quando si misura un'*Energy Distance* prossima a zero, ciò indica che le distribuzioni campionate coincidono sia in termini di media che di varianza. Per il calcolo numerico della *Energy Distance* è stata utilizzata la libreria `dcor` di Python, che implementa un insieme di strumenti per la misura di dipendenze e dissimilarità tra distribuzioni multivariate. In particolare, la funzione `dcor.energy_distance()` consente di stimare in modo efficiente la distanza tra insiemi di punti campionati, fornendo un'approssimazione robusta della metrica anche nel caso di distribuzioni ad alta dimensionalità, come quelle derivate dal *flux sampling* dei modelli metabolici.

2.5 TECNICHE DI DATA VISUALIZATION

Per la rappresentazione e l'analisi esplorativa dei risultati, sono state adottate diverse tecniche di *data visualization* finalizzate a descrivere in modo sintetico le relazioni tra i modelli metabolici contestualizzati e le loro distribuzioni di flusso. Le analisi di clustering sono state effettuate a partire dalle matrici di *flux sampling*, utilizzando come metrica la *energy distance* calcolata in modo *pairwise* tra tutti i modelli. Le relative matrici di distanza sono state quindi rappresentate sotto forma di dendrogrammi gerarchici, permettendo di individuare pattern di similarità e raggruppamenti coerenti con le condizioni sperimentali e con il tasso di crescita cellulare.

Parallelamente, le stesse matrici di *flux sampling* sono state proiettate in uno spazio a dimensionalità ridotta tramite analisi delle componenti principali (PCA), realizzata con la libreria `scikit-learn` di `Python`. Queste procedure hanno consentito di ottenere una visualizzazione chiara delle principali direzioni di variazione del dataset, mettendo in evidenza la separazione tra i modelli costruiti su dati di espressione grezzi e quelli normalizzati per la molteplicità genica. Nel complesso, tali tecniche di visualizzazione hanno fornito un supporto essenziale per interpretare in maniera intuitiva i risultati

delle analisi quantitative, consentendo di evidenziare trend e divergenze difficilmente osservabili nelle sole rappresentazioni numeriche.

3

RISULTATI E DISCUSSIONE

3.1 PIPELINE PER LA GENERAZIONE DI MODELLI CONTESTUALIZZATI

La pipeline sviluppata per questo lavoro è stata progettata per integrare in modo sistematico dati di trascrittomica e dati relativi alla molteplicità genica all'interno del modello genomico–metabolico di *Escherichia coli iML1515*. Come base sperimentale sono stati utilizzati i dataset di RNA-seq e i metadati pubblicati da sastry et al. ([23]), che includono un ampio spettro di condizioni di crescita su terreno minimo M9 con diverse fonti di carbonio e tassi di crescita sperimentali.

A partire da questi dati, sono stati filtrati i campioni con crescita positiva e mezzo di coltura coerente, e per ciascun gene e condizione è stata calcolata la molteplicità cromosomica attesa in funzione del tasso di crescita e della distanza genomica dall'origine di replicazione (oriC), secondo il modello di Cooper–Helmstetter ([3]). La molteplicità così ottenuta è stata successivamente riscalata all'intervallo di espressione osservato, generando un dataset parallelo di "molteplicità normalizzata". Combinando tali informazioni con i valori di espressione genica in TPM, sono stati costruiti due insiemi di profili trascrizionali: uno basato sui dati originali "raw" e uno normalizzato per la molteplicità cromosomica.

Entrambi i dataset sono stati utilizzati come input per la generazione di modelli contestualizzati mediante l'algoritmo `RIPTiDe`. Per ciascuna condizione sperimentale, `RIPTiDe` ha mappato i livelli di espressione sul modello iML1515, producendo due versioni del modello — una grezza e una normalizzata — con vincoli di flusso proporzionali all'attività trascrizionale dei geni corrispondenti. Ogni modello così ottenuto è stato poi sottoposto a campionamento dello spazio dei flussi (*flux sampling*) mediante `COBRApy`, ottenendo matrici di distribuzioni che descrivono le possibili configurazioni metaboliche compatibili con i dati di espressione.

L'intero flusso di lavoro è sintetizzato in Figura 6, che riassume graficamente le principali fasi del processo di elaborazione, dalla normalizzazione per la molteplicità genica alla contestualizzazione dei modelli e al campionamento dei flussi. Questo approccio integrato consente di confrontare, per ogni condizione, gli effetti della normalizzazione per la molteplicità genica sulle distribuzioni di flusso.

Does transcription-replication interaction impact cellular metabolism?

Michele Giovannini[1], Matteo Brilli[2], Arooba Arshad[3], Antonio Frandi[2], Silvia Buroni[3], Michael Mederer[4], Alessio Masoni[1], Francesca Vaccaro[1], Tania Alonso-Vásquez[1], Gian Luigi Garbini[1], Marco Fondi[1]

1 Department of Biology, University of Florence, Italy | **2** Department of Biosciences, University of Milan, Italy
3 Department of Biology and Biotechnology "L. Spallanzani", University of Pavia, Italy | **4** Biocenter Innsbruck, Medical University of Innsbruck, Austria

1. Background

Bacterial chromosome replication is not a neutral process: it directly influences gene expression by creating copy number variations along the oriC-ter axis[1]. During rapid growth, many bacteria (e.g. *E. coli*, *B. subtilis*) enter a mero-oligoploid state, in which multiple replication forks are simultaneously active[1,2]. This generates a dosage gradient: genes close to the origin are present in several copies, while those near the terminus lag behind[2]. This replication-dependent multiplicity effect alters transcript abundance and, in turn, may perturb metabolic homeostasis. Indeed, theoretical work has proposed that such imbalances acted as a selective force shaping bacterial genome architecture, favoring the clustering of functionally related genes and operon formation. Despite the well-established impact of replication on transcription, its consequences for cellular metabolism remain largely unexplored. By integrating gene multiplicity with genome-scale metabolic models (GEMs), we aim to assess whether replication-induced transcriptional biases propagate to the metabolic phenotype, and how they may contribute to bacterial adaptive strategies.

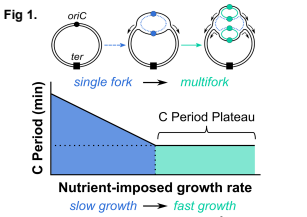


Fig 1. Replication of the *E. coli* chromosome[2]. Top: replication starts at origin (oriC, •) and ends at terminus (ter, ■), with either single forks (slow growth) or multiple forks (fast growth). Bottom: According to the Cooper–Helmstetter model, the duration of DNA replication (C period) decreases as growth rate increases, until reaching a plateau at high nutrient-imposed growth rates.

Fig 2.

$$n_x = 2^{\frac{(1-p_x)\,C+D}{\tau}}$$

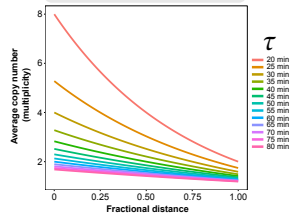
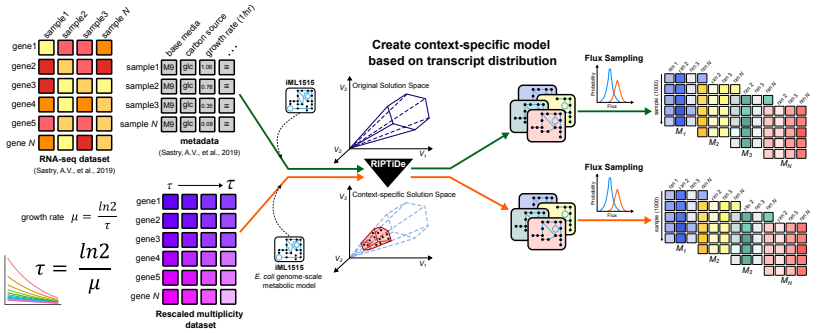


Fig 2. Average copy number of *E. coli* genome-scale metabolic model oci as a function of their position on the oriC-ter axis and division time

2. Materials and Methods



2. Published RNA-seq data from Sastry et al. (2019)[3] were used as expression input, while growth rates were converted into doubling times (τ) to calculate gene multiplicity, which was then rescaled to the observed expression range. Both TPM and rescaled multiplicity datasets were independently integrated into the *E. coli* genome-scale model (iML1515) with RIPTiDe[4], generating condition-specific constrained models that were subsequently analyzed by flux sampling to obtain flux distributions and solution matrices.

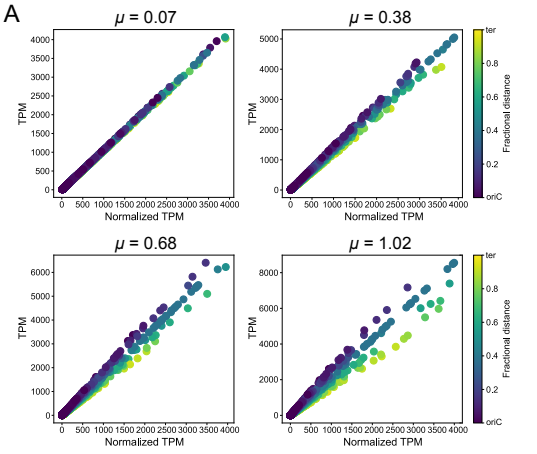
3.B At steady state, total expression can be decomposed into gene multiplicity and expression efficiency, formalized as:

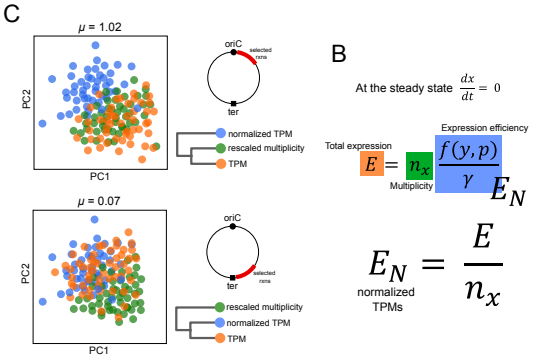
$$E_N = \frac{E}{n_x}$$

Based on this, TPM values were normalized to remove the multiplicity component. Scatter plots (**3.A**) show that the multiplicity effect increases with higher growth rates (μ), as indicated by greater dispersion of genes according to their fractional distance from oriC to ter. PCA analyses (**3.C**) suggest that differences between raw and normalized TPMs are more evident at fast growth and for genes near oriC, whereas low growth rates and genes near ter show reduced impact. Further analyses are ongoing to confirm these trends.

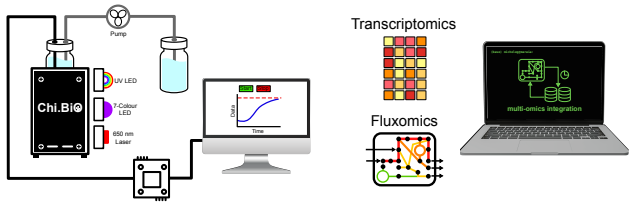
4. After method development using published RNA-seq data, ongoing experiments focus on controlled E. coli growth in a Chi.Bio bioreactor. At steady state, samples will be collected for RNA-seq and 13C-based fluxomics, and the resulting multi-omics data will be integrated into the *E. coli* GEM to dissect the metabolic impact of gene multiplicity.

3. Preliminary Results





4. What the next?



1 Januszewicz-Blohelia, J., & Lipniacki, T. (2017). Genetic toggle switch controlled by bacterial growth rate. BMC systems biology, 11(1), 117.
2 Sanders, S., Joshi, K., Levin, P. A., & Iyer-Biswas, S. (2023). Beyond the average: An updated framework for understanding the relationship between cell growth, DNA replication, and division in a bacterial system. PLoS genetics, 19(1), e1010505.
3 Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., ... & Palsson, B. O. (2019). The Escherichia coli transcriptome mostly consists of independently regulated modules. Nature communications, 10(1), 5536.
4 Jenior, M. L., Moutinho Jr, T. J., Dougherty, B. V., & Papin, J. A. (2020). Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments. PLoS computational biology, 16(4), e1007099.

Figura 6: Schema della pipeline di contestualizzazione. Il diagramma mostra le principali fasi di elaborazione: filtraggio dei dati, calcolo della molteplicità, normalizzazione, contestualizzazione tramite `RIPTiDe` e campionamento dei flussi.

3.1.1 *Valutazione dell'affidabilità di RIPTiDe nell'integrazione dei dati trascrittomici*

`RIPTiDe`, oltre a generare il modello metabolico contestualizzato, fornisce una serie di metriche quantitative utili a valutare la coerenza tra i dati trascrittomici forniti in input e la rete metabolica risultante. In particolare, l'algoritmo restituisce due indicatori principali di affidabilità: il *p-value*, che misura la significatività statistica della coerenza tra i livelli di espressione genica e i flussi predetti dal modello, e il coefficiente di correlazione r, che quantifica la relazione diretta tra distribuzione dei flussi e profilo trascrittomico. Per valutare la robustezza di queste misure e validare la coerenza biologica dei dati sperimentali, sono state eseguite oltre 3 000 simulazioni aggiuntive fornendo a `RIPTiDe` profili di espressione randomizzati. Questa procedura ha permesso di ottenere un riferimento dei valori di *p-value* e R^2 attesi in assenza di coerenza biologica, costituendo così un controllo negativo per le analisi successive. Come mostrato in Figura 7, soltanto il 27.7% delle *run* casuali ha prodotto risultati statisticamente significativi (*p-value* < 0.01), mentre la maggior parte (72.3%) è risultata non coerente con la rete metabolica. Questa differenza conferma che l'algoritmo è sensibile alla struttura del dataset di espressione e in grado di discriminare tra profili biologicamente plausibili e dati casuali. Nel presente lavoro, tutte le esecuzioni di `RIPTiDe` effettuate sui dati trascrittomici sperimentali hanno restituito *p-value* inferiori a 0.001, indicando un'elevata significatività statistica e una forte coerenza biologica tra il modello utilizzato (*iML1515*) e i profili di espressione. Questi risultati confermano l'affidabilità del processo di contestualizzazione e la capacità dell'algoritmo di integrare in modo efficace l'informazione trascrittomica, producendo modelli metabolicamente plausibili e coerenti con lo stato cellulare descritto dai dati sperimentali.

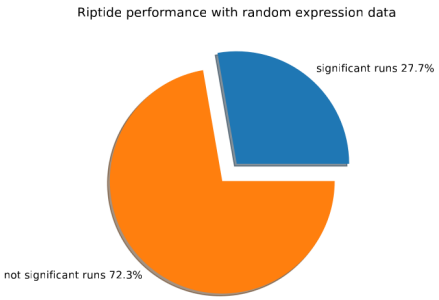


Figura 7: Prestazioni di `RIPTiDe` nell'integrazione di dati trascrittomici casuali. Il grafico mostra la percentuale di *run* considerate significative (27.7%) rispetto a quelle non significative (72.3%). Nel presente lavoro è stata adottata una soglia di significatività di ***p-value*** $<$ **0.01**: le *run* con valori superiori sono considerate non coerenti con i dati biologici, mentre quelle al di sotto della soglia indicano un'integrazione statisticamente significativa tra modello e trascrittomica.

3.1.2 *Costruzione del dataset di molteplicità*

Per costruire il dataset normalizzato per la molteplicità genica, è stato necessario stimare, per ciascun gene presente nel modello *iML1515*, il numero medio di copie atteso in funzione sia della sua posizione genomica, sia del tasso di crescita specifico della condizione sperimentale. Durante la crescita batterica, infatti, il cromosoma di *E. coli* può essere soggetto a cicli di replicazione sovrapposti: quando il tempo di duplicazione cellulare τ è inferiore al tempo necessario a completare una replicazione cromosomica ($C + D$), più origini di replicazione risultano attive simultaneamente ([10, 18]) (paragrafo 1.1). Questo fenomeno determina un gradiente di dosaggio genico che dipende sia dalla distanza dal sito d'inizio della replicazione (*oriC*), sia dal tasso di crescita μ, strettamente correlato a τ ([6]). Operativamente, per ciascun gene g_x è stata annotata la posizione genomica normalizzata $p_x \in [0, 1]$ rispetto all'origine di replicazione, a cui verrà fatto riferimento con il termine di *"Fractional Distance"*. Il numero medio di copie n_x è stato quindi calcolato secondo la formulazione del modello di Cooper–Helmstetter:

$$n_x = 2^{\frac{[(1-p_x)C+D]}{\tau}} \tag{3.1}$$

dove C rappresenta la durata della replicazione, D l'intervallo tra la fine della replicazione e la divisione cellulare, e τ il tempo di generazione.
E' facilmente osservabile la dipendenza tra valore di n_x e il tasso di crescita τ, il quale per valori alti (quindi bassi tassi di crescita) va a smorzare l'effetto che la molteplicità ha sul dosaggio genico (Figura 8).

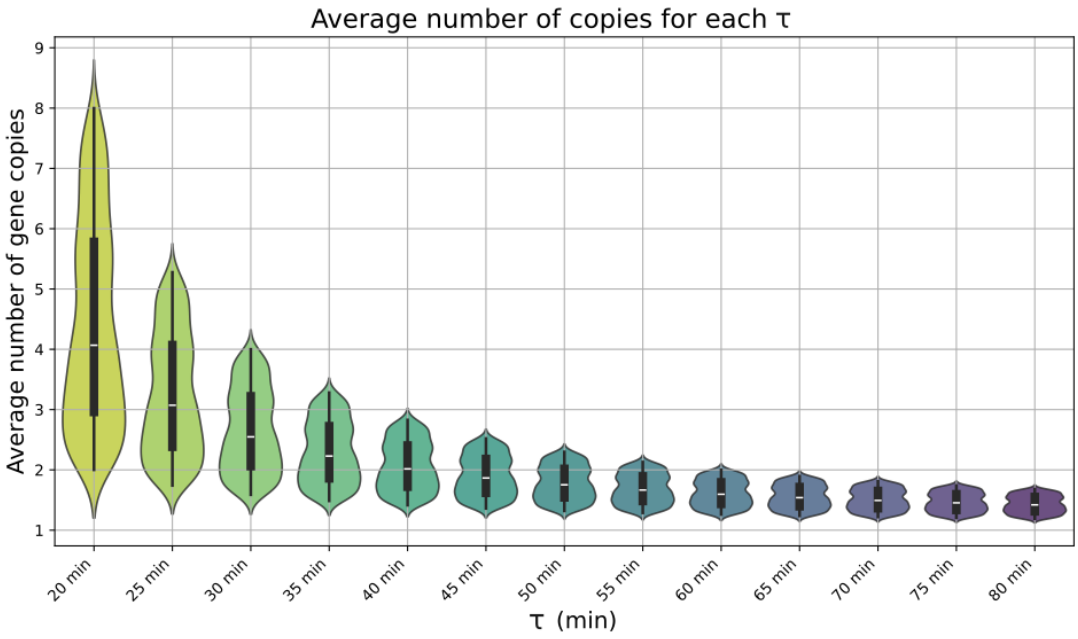


Figura 8: Distribuzione del numero medio di copie geniche in funzione del tempo di duplicazione cellulare (τ).

In questa formulazione, i geni prossimali a *oriC* (valori di p_x prossimi a 0) assumono un numero medio di copie maggiore, mentre i geni prossimali al terminatore di replicazione (*ter*, $p_x \approx 1$) tendono a un valore prossimo all'unità. All'aumentare del tasso di crescita, la diminuzione di τ comporta un incremento esponenziale di n_x per i geni vicini a *oriC*, riflettendo la presenza di più cicli di replicazione sovrapposti. Infine, per ogni condizione e per ciascun gene, il valore di espressione trascrizionale originale espresso in TPM (E_x) è stato normalizzato dividendo per il corrispondente coefficiente di molteplicità:

$$E_x = n_x f_x(y, \mathbf{p}) - \gamma E_x, \qquad E_x^{\text{norm}} = \frac{E_x}{n_x} \tag{3.2}$$

In questa formulazione, E_x rappresenta il livello di espressione del gene x, determinato dal numero medio di copie geniche n_x e da una funzione regolativa $f_x(y, \mathbf{p})$ dipendente dal fattore trascrizionale y e da un insieme di parametri $\mathbf{p}$. Il termine γE_x descrive la componente di degradazione del trascritto, mentre E_x^{norm} indica l'espressione normalizzata per la molteplicità genica, utile per isolare la componente regolatoria indipendente dal dosaggio cromosomico.
Questa trasformazione consente quindi di riportare le abbondanze trascrizionali a un livello "per cobia genica", eliminando l'effetto dovuto alle variazioni di dosaggio indotte dalla replicazione. Il risultato è un dataset di espressione normalizzato per la

molteplicità cromosomica, in cui ciascun valore E_x^{norm} rappresenta una stima dell'espressione intrinseca del gene, indipendente sia dalla posizione genomica sia dal tasso di crescita ([18, 6]).
La Figura 9 illustra l'effetto della molteplicità genica sulla distribuzione dei livelli di espressione (TPM) lungo l'asse *oriC–ter* in quattro condizioni ordinate per tasso di crescita specifico ($\mu = 0.07$, 0.38, 0.68 e 1.02 h^{-1}). All'aumentare di μ, la divergenza tra i valori di TPM grezzi e quelli normalizzati risulta progressivamente più marcata, un effetto visibile soprattutto allontanandosi dalla regione di origine della replicazione (*oriC*). Questo andamento riflette l'impatto crescente del dosaggio genico nelle condizioni di crescita rapida, dove la presenza di più origini attive amplifica localmente l'espressione dei geni prossimali, mentre la normalizzazione per la molteplicità cromosomica consente di isolare la componente regolatoria intrinseca dell'espressione.

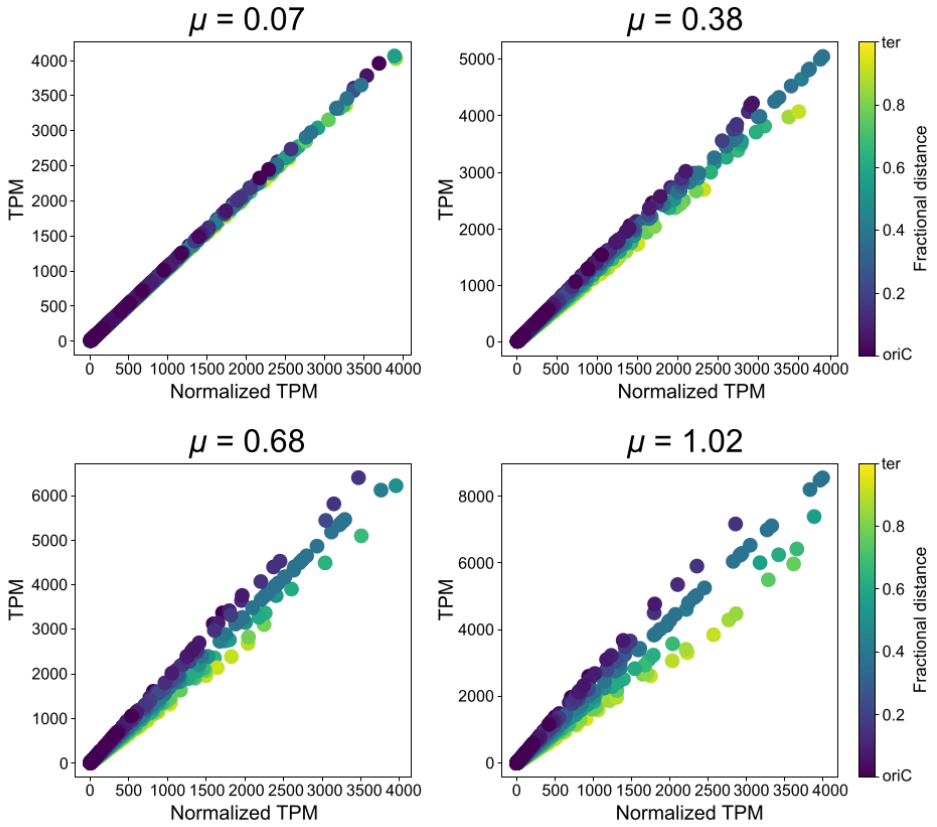


Figura 9: Figura 10. Relazione tra espressione genica e molteplicità cromosomica in funzione del tasso di crescita (μ). Confronto tra TPM grezzi e TPM normalizzati per la molteplicità genica lungo l'asse *oriC–ter* in quattro condizioni ordinate per tasso di crescita specifico. I colori rappresentano la posizione relativa dei geni sul cromosoma (da *oriC* a *ter*).

3.1.3 *Utilizzo di RIPTiDe per colmare il gap tra espressione e metabolismo*

A partire dai dataset di trascrittomica "grezzi" e da quelli "normalizzati" per la molteplicità genica, i dati di espressione sono stati mappati sul modello metabolico di riferimento di *E. coli* (*iML1515*) mediante l'algoritmo RIPTiDe [11]. In questo modo, per ciascuna condizione sperimentale descritta nel paragrafo 2.1, è stato possibile generare due modelli metabolici contestualizzati: uno basato sulle abbondanze trascrizionali originali e uno basato sui valori corretti per il dosaggio genico. Questa doppia contestualizzazione consente di valutare in che misura la normalizzazione dell'espressione influisca sulla predizione dei flussi metabolici e sull'attività complessiva della rete. Per garantire coerenza biologica con le condizioni sperimentali, prima della fase di contestualizzazione sono stati impostati nel modello i vincoli ambientali appropriati, in particolare il terreno di crescita minimo M9 e la specifica fonte di carbonio corrispondente a ciascun esperimento di sastry et al. ([23]). Questa parametrizzazione assicura che le simulazioni metaboliche riflettano realisticamente lo stato fisiologico osservato sperimentalmente e che le differenze nei flussi derivino unicamente dalle variazioni nell'espressione genica, piuttosto che da condizioni di crescita discordanti. L'applicazione di RIPTiDe in questo contesto rappresenta quindi il passaggio chiave per integrare i dati trascrittomici con la modellistica metabolica, traducendo i profili di espressione genica in distribuzioni di flusso coerenti con la fisiologia della cellula e con il suo contesto ambientale.
Questo ha permesso almeno a livello teorico di abbattere la barriera tra espressione ed effettiva variazione del fenotipo, mostrando come una variazione dell'espressione (in questo caso indotta dalla normalizzazione per la molteplicità genica), si propaghi a livello metabolico.

3.2 ANALISI DI DISSIMILARITÀ TRA MODELLI CONTESTUALIZZATI

Una volta ottenuti i modelli metabolici contestualizzati tramite *RIPTiDe* (capitolo 3.1.3), è stato necessario caratterizzarne il comportamento fenotipico dei modelli ottenuti per valutare gli effetti della normalizzazione. A tale scopo, è stata adottata la tecnica del *flux sampling* (paragrafo 1.2.3), che consente di esplorare la globalità dello spazio delle soluzioni ammissibili del modello, fornendo una descrizione statistica delle distribuzioni di flusso metabolicamente coerenti. Il *flux sampling* è stato preferito a tecniche di tipo FBA o pFBA (paragrafo 1.2.2) per la sua capacità intrinseca di descrivere in maniera più globale e realistica l'attività metabolica, considerando l'intero insieme di stati di flusso compatibili con i vincoli stechiometrici e regolativi, anziché limitarsi a un'unica soluzione subottimale. Ciò rende più semplice cogliere variazioni più sfumate nella distribuzione dei flussi, che meglio rappresentano la plasticità del metabolismo in condizioni fisiologiche differenti. Per ciascun modello contestualizzato — sia derivante dai dati di espressione grezzi sia da quelli normalizzati — è stata eseguita un'analisi di *flux sampling* mediante l'ambiente `COBRApy` ([4]) (capitolo 2.2). Il risultato di questa procedura consiste in una matrice di campionamento dei flussi, in cui ogni colonna rappresenta una reazione del modello e ogni riga una soluzione ammissibile estratta dallo spazio dei flussi, come mostrato in Figura 4. Tali matrici, ottenute per tutti i modelli, costituiscono la base per le successive analisi comparative tra le due tipologie di dataset.

3.2.1 *Analisi di clustering sui modelli metabolici contestualizzati*

Per analizzare in modo comparativo i modelli metabolici contestualizzati ottenuti a partire dai dati di espressione grezzi e da quelli normalizzati per la molteplicità genica, è stata condotta un'analisi di clustering basata sulle rispettive matrici di *flux sampling*. Ogni matrice descrive la distribuzione dei flussi metabolici campionati nello spazio delle soluzioni ammissibili di ciascun modello e fornisce pertanto una rappresentazione statistica del suo stato fenotipico complessivo (capitolo 1.2.3). Confrontare tali matrici consente di stimare quanto due modelli condividano una struttura metabolica simile o, al contrario, differiscano in termini di configurazioni di flusso accessibili. A tal fine, è stata costruita una matrice di dissimilarità *tutti contro tutti*, in cui ogni valore rappresenta la distanza tra le distribuzioni di flusso di una coppia di modelli. Come metrica di confronto è stata utilizzata l'*energy distance* (paragrafo 2.4), che può essere interpretata come una misura generale di dissimilarità tra distribuzioni multivariate. Sulla base di tale matrice, è stato quindi eseguito un clustering gerarchico per rappresentare graficamente le relazioni di similarità tra tutti i modelli considerati. Il risultato è mostrato nel dendrogramma della Figura 10, che rappresenta un'analisi di clustering *energy distance–based* sull'intero dataset di modelli contestualizzati. L'obiettivo principa-

le non era tuttavia quello di discutere la formazione dei cluster nel loro insieme, ma piuttosto di analizzare le distanze relative tra i membri di ciascuna coppia di modelli "raw–normalizzato", costruiti a partire dalla stessa condizione sperimentale. L'interesse risiede nel verificare se, e in quale misura, la normalizzazione per la molteplicità genica modifichi la distribuzione dei flussi in funzione del tasso di crescita.

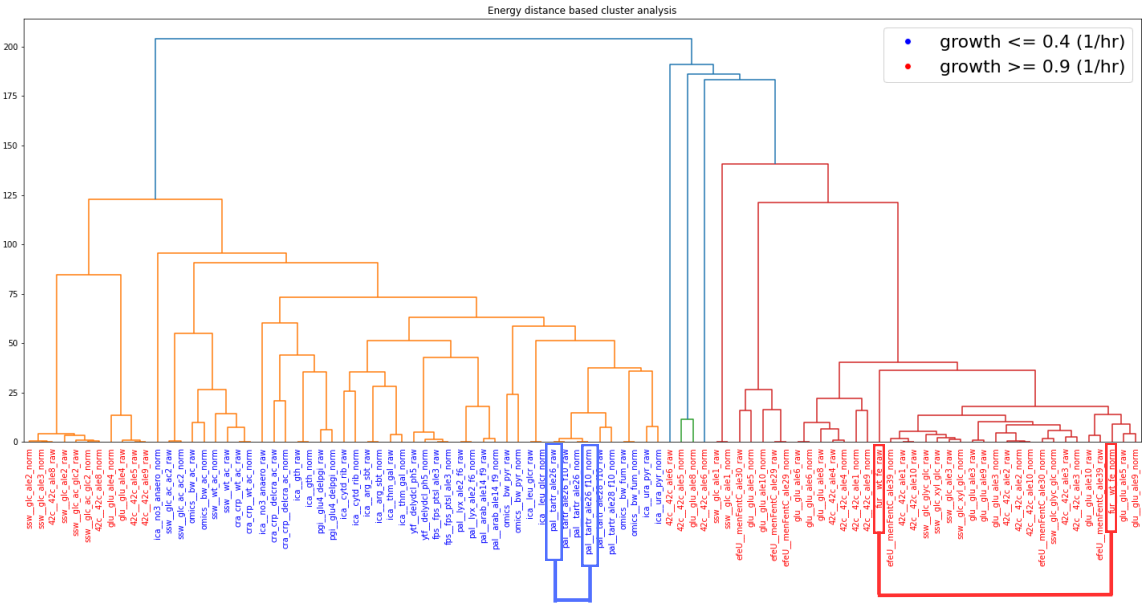


Figura 10: Dendrogramma di clustering basato su *energy distance* tra modelli metabolici contestualizzati. Le condizioni evidenziate dai riquadri colorati mostrano due esempi rappresentativi: il riquadro rosso evidenzia una coppia di modelli ad alto tasso di crescita, caratterizzata da una distanza elevata tra la versione "raw" e quella "normalizzata", mentre il riquadro blu indica una coppia a bassa crescita, in cui la distanza tra i due modelli risulta ridotta.

L'osservazione del dendrogramma suggerisce una tendenza coerente con l'ipotesi iniziale: le coppie di modelli corrispondenti a condizioni di crescita elevata presentano distanze maggiori tra le varianti grezze e normalizzate, indicando che la correzione per la molteplicità genica ha un impatto più marcato sul metabolismo in contesti in cui il dosaggio genico è fortemente sbilanciato lungo il cromosoma. Al contrario, nelle condizioni di crescita lenta, dove il numero di cicli di replicazione attivi è ridotto e la molteplicità media dei geni tende a valori prossimi all'unità, le due versioni del modello mostrano configurazioni di flusso più simili e una minore distanza nel clustering. Questo comportamento è coerente con l'atteso: la molteplicità genica, crescente al diminuire del tempo di duplicazione, introduce variazioni di espressione che si

riflettono in cambiamenti sistematici nell'attività metabolica. Quando la replicazione è rapida, i geni in prossimità dell'origine (*oriC*) risultano sovrarappresentati e possono influenzare in misura maggiore i livelli di espressione genica e i flussi metabolici corrispondenti. In condizioni di crescita lenta, invece, il contributo del dosaggio genico è trascurabile, e i due approcci di contestualizzazione convergono verso fenotipi molto simili. E' però doveroso sottolineare come l'effetto sopra citato non sia apprezzabile in tutte le condizioni sperimentali considerate. In diversi casi, infatti, le coppie di modelli "raw–normalizzato" non seguono la relazione attesa tra distanza e tasso di crescita, evidenziando come l'influenza della molteplicità genica sul metabolismo sia difficile da isolare. Ciò suggerisce che l'impatto della molteplicità cromosomica sul fenotipo rappresenti un fenomeno complesso e multiparametrico, in cui la crescita cellulare agisce come uno dei principali, ma non unici, determinanti. La Figura 11 schematizza concettualmente la relazione osservata, illustrando come la distanza nel clustering tra le coppie di modelli "raw–normalizzato" aumenti con il tasso di crescita. In sintesi, questa analisi suggerisce che l'effetto della molteplicità genica sul metabolismo non sia uniforme, ma cresca con la velocità di replicazione, e rappresneta un primo passo per definire, seppur qualitativamente l'impatto di un dosaggio genico sbilanciato sul metabolismo cellulare.

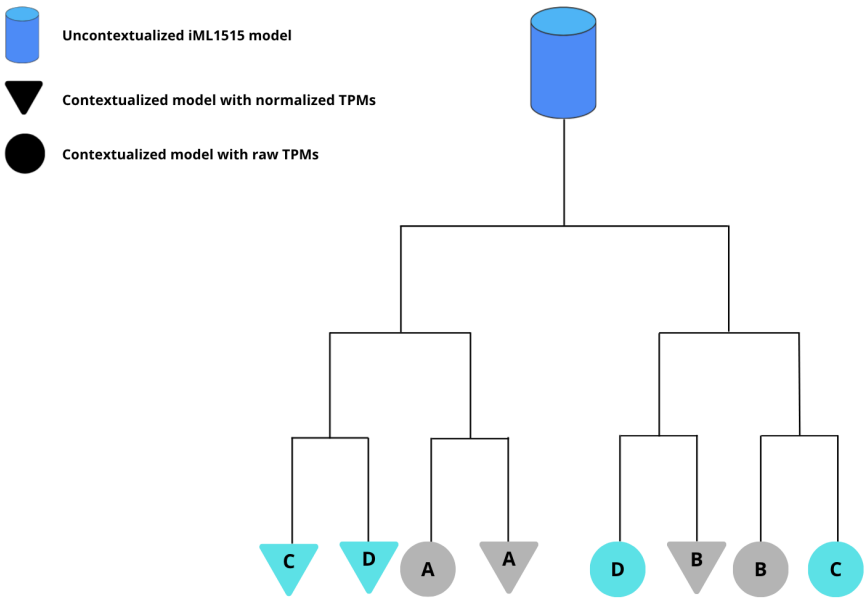


Figura 11: Schema illustrativo del comportamento atteso: nelle condizioni ad alta crescita (azzurro) le coppie di modelli "raw–normalizzato" mostrano una maggiore divergenza, mentre nelle condizioni a bassa crescita (grigio) la divergenza risulta ridotta.

3.2.2 *Visualizzazione del dataset nello spazio delle componenti principali*

In continuità con l'analisi di clustering esposta nel paragrafo 3.2.1, è stata condotta un'analisi delle componenti principali (PCA) come approccio complementare per indagare lo stesso fenomeno da una prospettiva più visuale. L'obiettivo è quello di esplorare se la separazione tra modelli generati a partire da dati di espressione grezzi e normalizzati per la molteplicità genica osservata a livello di distanza multivariata emerga in modo più evidente nella proiezione delle distribuzioni di flusso in uno spazio a bassa dimensionalità. La PCA consente infatti di ridurre la complessità del dataset, preservando la maggior parte della varianza informativa e si presenta come un valido strumento di analisi dei pattern divergenza tra modelli metabolici ([22]). Il pannello mostrato in Figura 12 riassume le PCA ottenute per tutte le matrici di *flux sampling* derivate da modelli contestualizzati in condizioni di crescita elevata ($\mu \geqslant 0.9\ h^{-1}$) e bassa ($\mu \leqslant 0.4\ h^{-1}$).

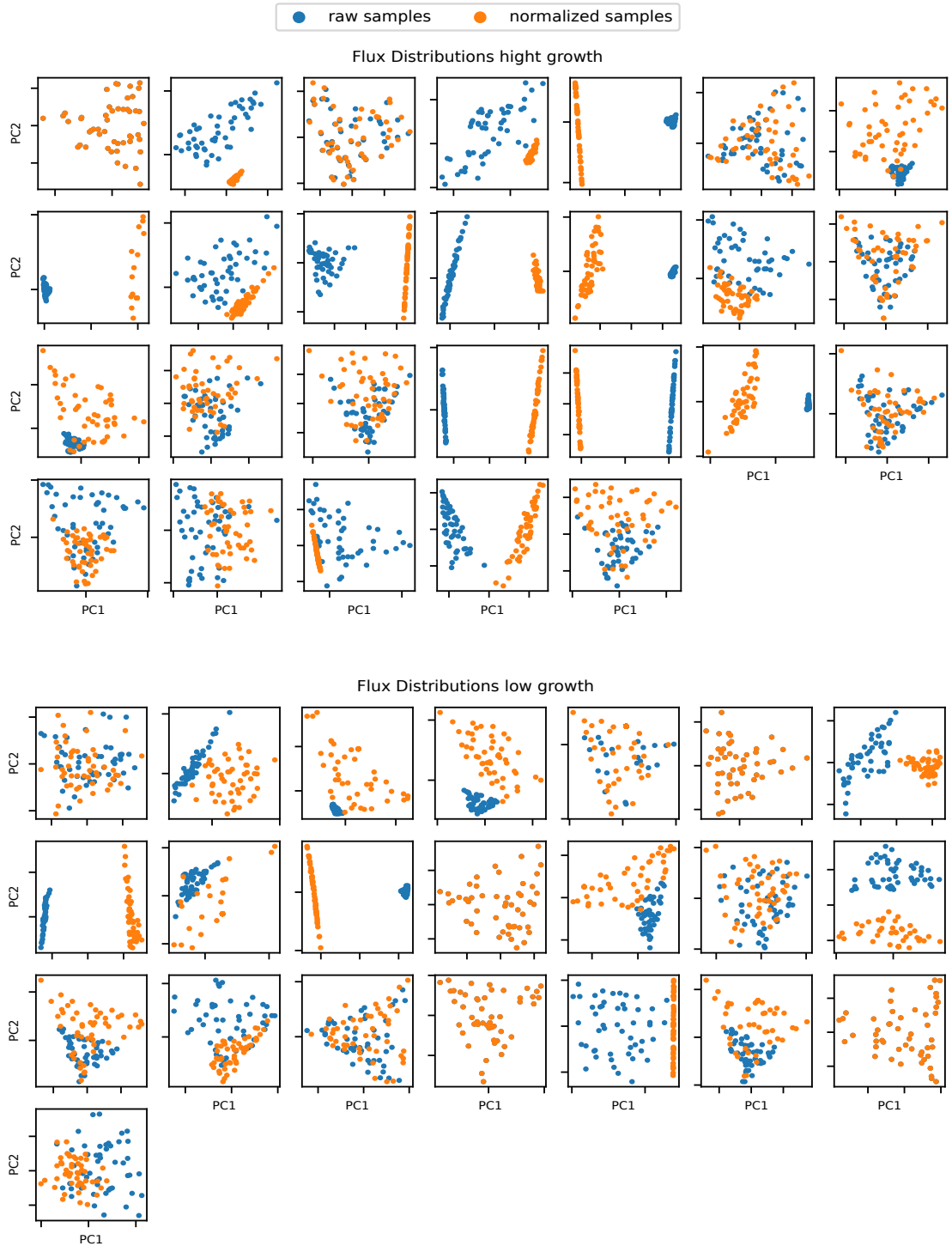


Figura 12: Proiezione PCA delle distribuzioni di flusso derivate da modelli **in alto**: a crescita alta ($\mu \geqslant 0.9\ h^{-1}$), **in basso**: a crescita bassa ($\mu \leqslant 0.4\ h^{-1}$).

Nel complesso, anche con questo approccio si può osservare una tendenza, nelle condizioni ad alta crescita, alla formazione di clusters più separati tra i campioni derivanti da dati "grezzi" e "normalizzati", suggerendo che la correzione per la molteplicità genica produca una modifica sistematica nelle distribuzioni di flusso quando il tasso di replicazione cellulare è elevato. Nelle condizioni a bassa crescita, al contrario, le due distribuzioni tendono a sovrapporsi maggiormente, indicando una minore influenza del dosaggio genico sul comportamento metabolico complessivo.

La Figura 13 mostra due esempi rappresentativi di PCA riferite a condizioni sperimentali specifiche. Nel caso ad alta crescita (pannello B), le componenti principali evidenziano una netta separazione tra i due gruppi di modelli, in linea con quanto atteso dal maggiore effetto della molteplicità cromosomica in tali contesti. Nell'esempio a bassa crescita (pannello A), invece, i punti corrispondenti ai modelli "raw" e "normalizzati" mostrano un'elevata sovrapposizione nello spazio PCA, indicando una sostanziale stabilità del metabolismo rispetto alla correzione applicata.

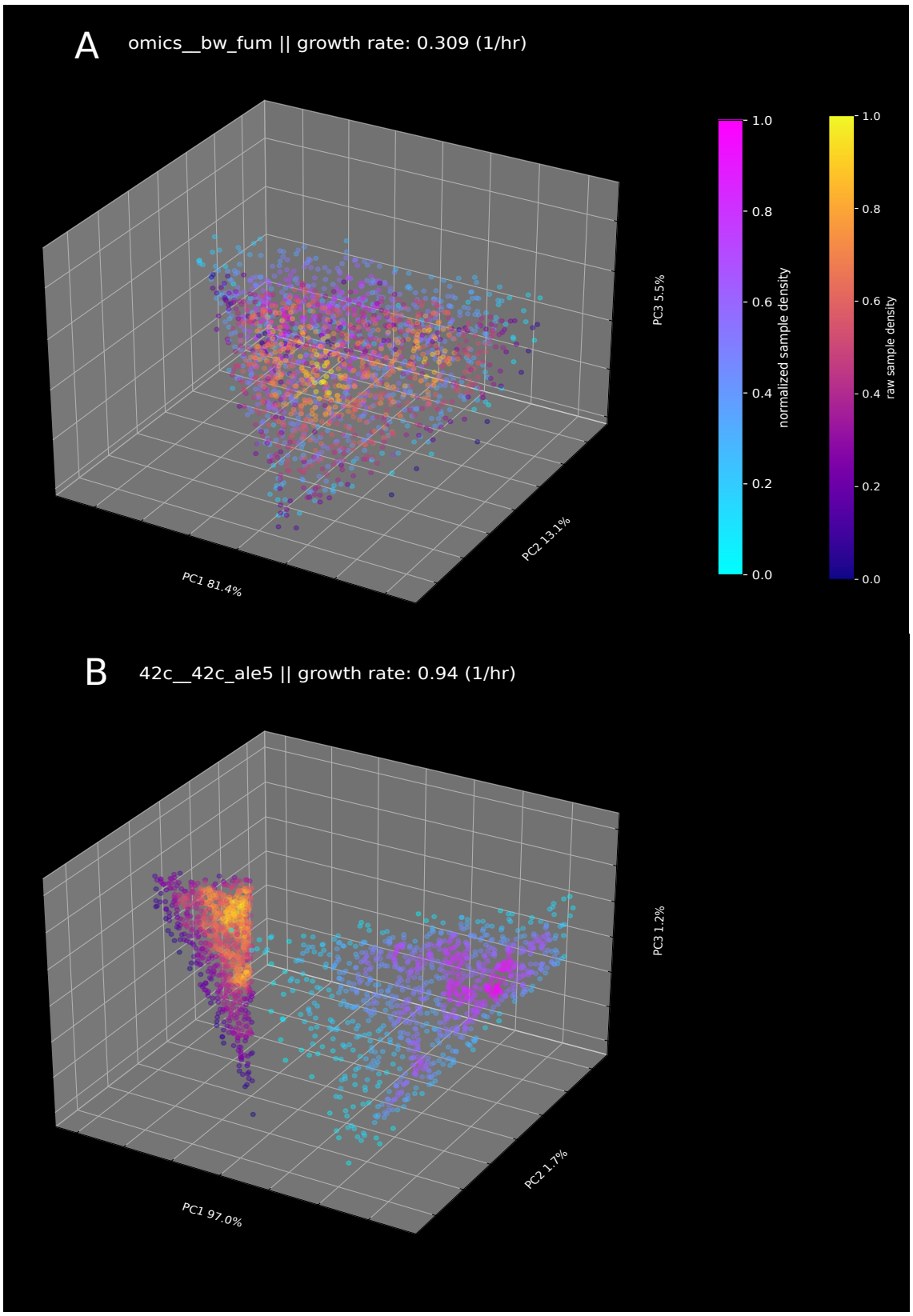


Figura 13: Confronto tra due analisi PCA rappresentative: (B) condizione ad alta crescita e (A) condizione a bassa crescita, derivate dai modelli contestualizzati con dati “grezzi” e “normalizzati”. Le separazioni più marcate nei campioni ad alta crescita riflettono un effetto più forte della molteplicità genica sulla distribuzione dei flussi metabolici.

Nonostante la tendenza generale, il livello di rumore nei dati risulta elevato e non tutte le condizioni sperimentali seguono il comportamento atteso. Questo rafforza quanto emerso nel paragrafo 3.2.1, ovvero che la risposta metabolica alla normalizzazione per la molteplicità genica possa dipendere anche da altri fattori non considerati in questo studio. Il fatto di trattare la regolazione monocopia genica come una costante rappresenta infatti una semplificazione utile ma inevitabilmente limitante: non tiene conto, ad esempio, della complessità dei meccanismi di regolazione post-trascrizionale e post-traduzionale, né delle variazioni nella disponibilità di metaboliti e cofattori che modulano dinamicamente l'attività enzimatica. Tali aspetti, contribuiscono a introdurre una variabilità aggiuntiva tra espressione genica e fenotipo metabolico, rendendo necessaria un'analisi integrata che includa più livelli di regolazione biologica ([1]).

3.2.3 *Analisi quantitativa dell'impatto della molteplicità genica sul metabolismo*

Dopo aver osservato qualitativamente, attraverso il clustering e la PCA (paragrafi 3.2.1 e 3.2.2), una possibile relazione tra la molteplicità genica e la divergenza metabolica dei modelli contestualizzati, l'attenzione si è spostata sul quantificare tale relazione in modo esplicito. L'obiettivo dell'analisi seguente è dunque quello di stimare quanto la molteplicità genica contribuisca a spiegare la varianza tra modelli generati a partire da dati di espressione grezzi e quelli normalizzati per il dosaggio genico. A tal fine, per ciascuna coppia di modelli "raw–normalizzato" è stata calcolata la *energy distance* (paragrafo 2.4), utilizzata come misura della dissimilarità tra le distribuzioni di flusso. Parallelamente, è stato costruito un vettore contenente i tassi di crescita sperimentali riportati nel dataset di sastry et al, indicato come $\boldsymbol{\mu}$, e un vettore $\mathbf{d}$ rappresentante le distanze *Energy Distance* tra le distribuzioni di flusso dei modelli "grezzi" e "normalizzati" (paragrafo 2.4):

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)^\top, \qquad \mathbf{d} = (d_1, \ldots, d_N)^\top, \qquad d_k = \mathrm{ED!}\left(\mathcal{S}^{(k)}\mathrm{raw}, , \mathcal{S}^{(k)}\mathrm{norm}\right) \quad (3.3)$$

È stata quindi calcolata la correlazione lineare di Pearson tra i due vettori, ottenendo un coefficiente di correlazione $r = 0.305$, con p-value pari a 0.001 e coefficiente di determinazione $R^2 = 0.093$. Questo risultato quantifica la componente di variabilità fenotipica imputabile alla molteplicità genica, indicando che circa il 9La relazione positiva osservata (Figura 14) suggerisce che l'impatto della molteplicità cromosomica aumenta con la velocità di crescita: i modelli associati a tassi di duplicazione più elevati mostrano una maggiore divergenza tra la versione "grezza" e quella normalizzata. Tuttavia, il contributo della molteplicità genica, pur statisticamente significativo, spiega solo una frazione limitata della variabilità complessiva, lasciando intendere che altri fattori, come la regolazione trascrizionale a livello di promotore o la disponibilità di

differenti fonti di carbonio, possano anch'essi concorrere a determinare la distanza fenotipica tra i modelli.

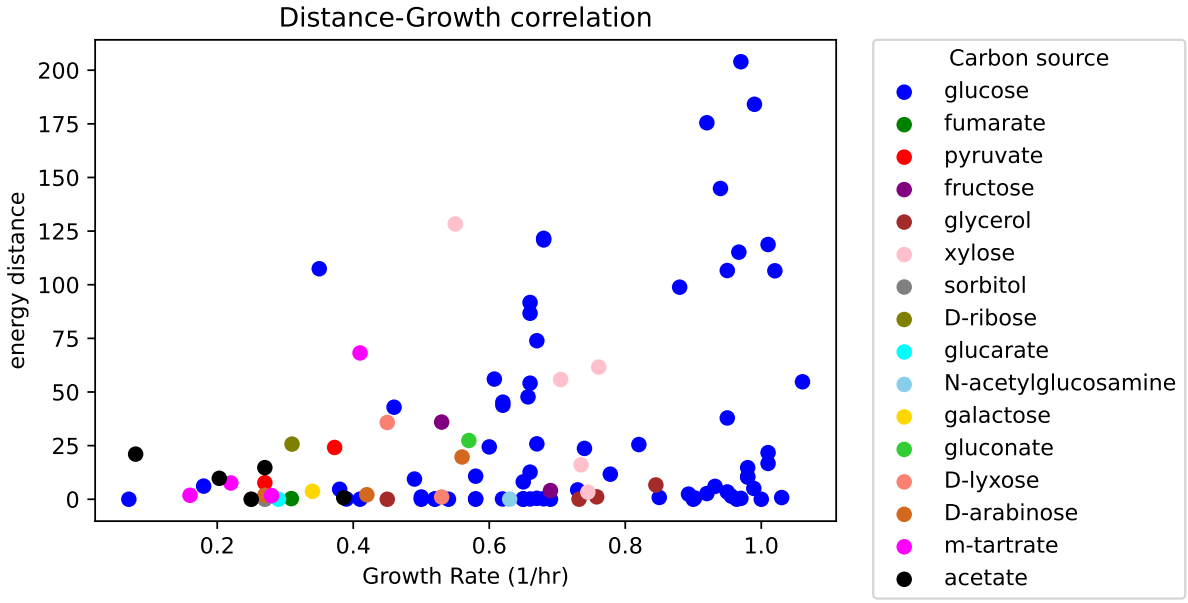


Figura 14: Relazione tra il tasso di crescita sperimentale e la *energy distance* tra coppie di modelli "grezzo–normalizzato". Le condizioni sperimentali nel grafico sono distinte in base alla fonte di carbonio utilizzata.

Per ridurre l'impatto del rumore nei dati e ottenere una rappresentazione più stabile della relazione tra tasso di crescita e distanza "grezzo–normalizzato", è stata eseguita una seconda analisi basata sul raggruppamento dei dati in intervalli (*binning*) di ampiezza 0.1 in termini di tasso di crescita (h^{-1}). All'interno di ciascun intervallo sono state calcolate la mediana dei tassi di crescita e la mediana delle corrispondenti distanze tra modelli, ottenendo così un insieme di punti rappresentativi del comportamento medio delle condizioni sperimentali (Figura 15).

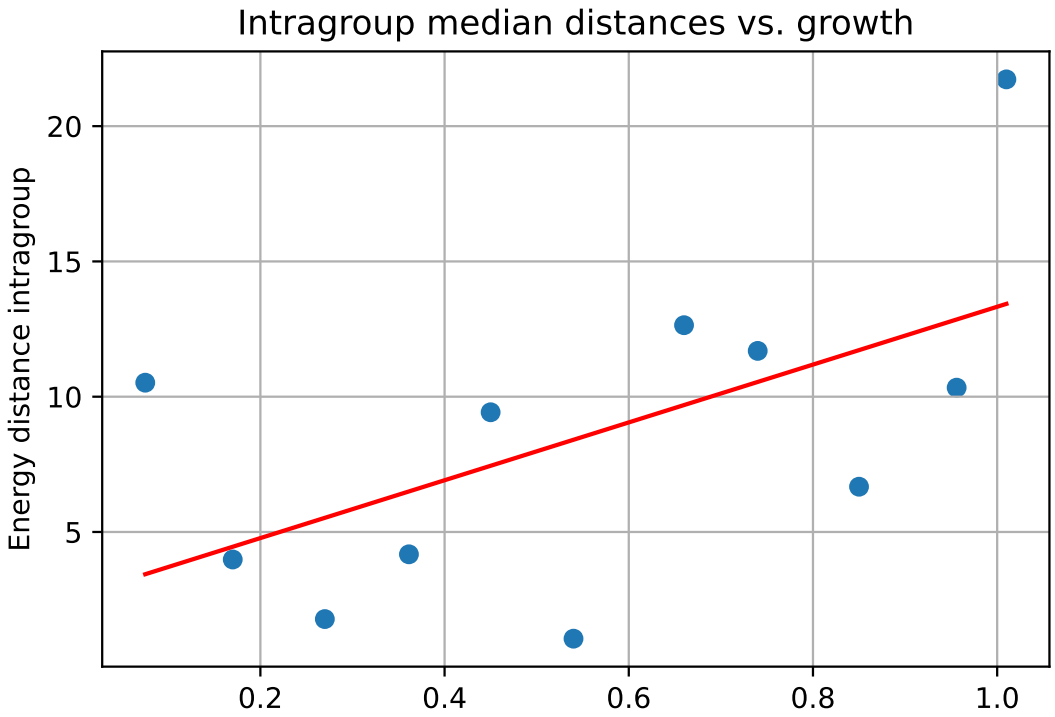


Figura 15: Relazione tra la crescita e la distanza *energy distance* calcolata su valori mediani binnati. Il segmento in rosso indica la retta di regressione lineare tra i punti

L'analisi dei valori aggregati mostra una correlazione di Pearson sensibilmente più elevata $r = 0.57$ e un coefficiente di determinazione $R^2 = 0.33$. Ciò indica una relazione più marcata tra la crescita e la distanza metabolica rispetto agli stessi parametri relativi al caso dove non è stato effettuato il *"binning"*, (Figura 14). Parte di questo incremento è sicramente attribuibile all'effetto intrinseco del processo di binning, che tende a ridurre la varianza all'interno dei gruppi ([28]), ma l'entità di tale incremento suggerisce che una componente della relazione effettiva tra i due fenomeni sia effettivamente oscurata dal rumore nei dati originali. Questo risultato supporta l'idea che, per stimare con maggiore affidabilità la forza della correlazione tra molteplicità genica e divergenza metabolica, sia necessario un numero maggiore di condizioni sperimentali, in modo da attenuare la variabilità intrinseca e ottenere valori di R^2 e r più stabili. È plausibile che i veri valori di correlazione si collochino in una fascia intermedia tra quelli ottenuti nell'analisi punto a punto (Figura 14) e quelli derivanti dai dati aggregati per intervallo.

3.3 ESTENSIONI APPLICATIVE E PROSPETTIVE DI RICERCA

Il principale risultato di questo elaborato consiste nell'aver formalizzato un approccio quantitativo per la valutazione dell'influenza della molteplicità genica sull'attività metabolica. Attraverso la costruzione di modelli contestualizzati a partire da dati trascrittomici grezzi e normalizzati per la molteplicità cromosomica, è stato possibile analizzare come il dosaggio genico, variabile con il tasso di crescita cellulare, si rifletta sulla distribuzione dei flussi metabolici predetti. L'integrazione di strumenti computazionali quali `RIPTiDe` e `COBRApy` ha permesso di ottenere una rappresentazione coerente delle variazioni fenotipiche derivanti dalle differenze di espressione, evidenziando come la molteplicità possa contribuire in maniera sensibile alla divergenza tra modelli. Le analisi di clustering, PCA e correlazione hanno complessivamente mostrato una relazione positiva tra velocità di crescita e distanza metabolica tra le coppie di modelli "raw–normalizzato", suggerendo che l'effetto del dosaggio genico sul metabolismo aumenti con la rapidità di replicazione. Nonostante il rumore intrinseco nei dati e la presenza di variabilità condizionale, l'approccio proposto fornisce una base metodologica per indagare in modo sistematico la propagazione delle variazioni di dosaggio genico e conseguentemente di trascritto fino al livello fenotipico. La possibilità di quantificare tale relazione apre la strada a studi più ampi sul ruolo della molteplicità genica come determinante dell'eterogeneità metabolica, non solo in *E. coli*, ma potenzialmente a tutti gli organismi procariotici caratterizzati da strategie di replicazione a singola OriC. Come prospettiva di ricerca, un'estensione naturale di questo lavoro consiste nell'ampliare il dataset di modelli contestualizzati, includendo un numero maggiore di condizioni sperimentali, così da migliorare la robustezza statistica delle correlazioni osservate e ridurre l'impatto del rumore. Un ulteriore sviluppo riguarda l'integrazione di dati multi-omici, in particolare di metabolomica e flussomica, che consentirebbero di validare sperimentalmente le previsioni del modello e di eliminare il possibile livello di confondimento tra trascrittomica e metabolismo. In sintesi, il lavoro presentato rappresenta un primo passo verso la definizione di un quadro quantitativo per lo studio dell'impatto della molteplicità genica sul metabolismo. Pur basandosi su un insieme di dati limitato, l'approccio dimostra la fattibilità di collegare in modo diretto le caratteristiche strutturali del genoma con il comportamento funzionale della rete metabolica, delineando un percorso di ricerca promettente verso una modellistica complementata delle dinamiche della replicazione cellulare.

3.4 MATERIALI AGGIUNTIVI

La pipeline descritta nel paragrafo 3.1.3 è stata progettata per essere completamente automatizzata, consentendo la generazione dell'intero set di modelli contestualizzati tramite un'unica esecuzione del flusso computazionale. Viene riportata di seguito:

```python
#!/usr/bin/env python3
"""
Pipeline automatizzata per la contestualizzazione di modelli metabolici di E. coli.
Integra i dati di trascrittomica grezzi e normalizzati con il modello genomico iML1515
mediante l'algoritmo RIPTiDe, mantenendo coerenza con le condizioni sperimentali
    descritte da Sastry et al. (2019).
"""
#Questo blocco definisce le librerie necessarie per  l esecuzione  dell'algoritmo.

import pandas as pd         # Gestione dei dataset tabellari
import cobra                # Manipolazione del modello genomico-metabolico
import riptide              # Contestualizzazione tramite l'algoritmo RIPTiDe
import multiprocessing as mp # Esecuzione parallela
import time, math           # Gestione temporale e operazioni matematiche


# ------------------------------------------------------------
# 1. Lettura dei dati di espressione e metadati
# ------------------------------------------------------------
def read_expression_dataset():
    """
    Carica il file Excel contenente i dati di espressione (RNA-seq) e i metadati
    relativi alle condizioni sperimentali (mezzo di crescita, fonte di carbonio, tasso di
        crescita).
    """
    metadata = pd.read_excel('41467_2019_13483_MOESM4_ESM.xlsx',
                             sheet_name='Metadata', index_col='Sample ID')
    expression = pd.read_excel('41467_2019_13483_MOESM4_ESM.xlsx',
                               sheet_name='Expression Data', index_col='log-TPM')
    return metadata, expression


# ------------------------------------------------------------
# 2. Lettura dei dati di molteplicit  genica
# ------------------------------------------------------------
def read_raw_multiplicity():
    """
    Importa il file contenente il numero medio di copie geniche per ogni gene e
        condizione.
    Questi valori saranno utilizzati per normalizzare le espressioni in TPM.
    """
    return pd.read_excel('raw_molteplicity.xlsx', sheet_name='molteplicity', index_col='
        locus_tag')


# ------------------------------------------------------------
# 3. Filtraggio e aggregazione del dataset
```

```
# ------------------------------------------------------------
def filter_dataset():
    """
    Seleziona solo le condizioni sperimentali rilevanti (mezzo M9, crescita positiva),
    rimuove repliche ridondanti e calcola la media dei replicati biologici.
    """
    carbon_sources = ['glucose', 'fumarate', 'pyruvate', 'fructose', 'glycerol', 'acetate']
    metadata, expression = read_expression_dataset()

    # Filtra i metadati mantenendo solo le condizioni valide
    metadata = metadata[
        (metadata['Growth Rate (1/hr)'] > 0) &
        (metadata['Base Media'] == 'M9') &
        (metadata['Carbon Source (g/L)'].str.extract(r'([A-Za-z0-9\-]+)', expand=False)
         .isin(carbon_sources))
    ]

    # Mantiene solo le colonne di espressione corrispondenti a condizioni valide
    expression = expression.drop([col for col in expression.columns if col not in
        metadata.index], axis='columns')

    # Elimina repliche multiple e crea indici univoci
    filtered_metadata = metadata.drop([cond for cond in metadata.index if int(cond.split(
        '_')[-1]) > 1])
    filtered_metadata.index = [idx.rsplit('_', 1)[0].rstrip('_') for idx in
        filtered_metadata.index]

    # Media delle repliche biologiche
    filtered_expression = pd.DataFrame(index=expression.index)
    exp_cols = list(expression.columns)
    i = 0
    while i < len(exp_cols):
        replicates = [exp_cols[i]]
        j = i + 1
        while j < len(exp_cols) and exp_cols[i].rsplit('_', 1)[0] == exp_cols[j].rsplit('
            _', 1)[0]:
            replicates.append(exp_cols[j])
            j += 1
        condition_root = exp_cols[i].rsplit('_', 1)[0].rstrip('_')
        filtered_expression[condition_root] = expression[replicates].mean(axis=1)
        i = j

    return filtered_metadata, filtered_expression
```

```
# ------------------------------------------------------------
# 4. Normalizzazione per la molteplicit genica
# ------------------------------------------------------------
def normalize_expression(raw_multiplicity, expression):
    """
    Converte i valori di espressione da log2(TPM) a scala lineare e li divide per
    il numero medio di copie geniche , ottenendo un dataset di espressione 'per copia
        genica'. Tale funzione viene eseguita solo per l'ottenimento del dataset '
        normalizzato'
    """
    normalized_expression = pd.DataFrame(index=raw_multiplicity.index, columns=
        raw_multiplicity.columns)
    conditions = list(expression.columns)

    for gene, _ in expression.iterrows():
        for condition in conditions:
            log_exp = float(expression.loc[gene, condition])
            exp_value = math.pow(2, log_exp)  # Conversione log2 -> valore reale
            multiplicity = float(raw_multiplicity.loc[gene, condition])
            normalized_expression.loc[gene, condition] = exp_value / multiplicity

    return normalized_expression


# ------------------------------------------------------------
# 5. Preparazione del modello metabolico
# ------------------------------------------------------------
def prepare_ecoli_model(metadata, condition_id):
    """
    Carica il modello genomico iML1515 e lo parametrizza in base alla condizione
        sperimentale.
    Imposta il mezzo M9, la fonte di carbonio corretta e regola il flusso di uptake
    fino a raggiungere il tasso di crescita sperimentale.
    """
    model = cobra.io.read_sbml_model('iML1515.xml')
    carbon_source_name = metadata.loc[condition_id, 'Carbon Source (g/L)'].split('(')[0]

    carbon_source_rxns = {
        'glucose': 'EX_glc__D_e', 'fumarate': 'EX_fum_e',
        'pyruvate': 'EX_pyr_e', 'fructose': 'EX_fru_e',
        'glycerol': 'EX_glyc_e', 'acetate': 'EX_ac_e'
        ...
    }

    # Mezzo M9 minimale
    m9_medium = {carbon_source_rxns[carbon_source_name]: 4.0,
                 'EX_o2_e': 1000.0, 'EX_nh4_e': 1000.0}
                 ...
```

```
131
    # Setting dei flussi di scambio
    for rxn in model.exchanges:
        if ex.id in m9_medium.keys():
                ex.lower_bound = -float(m9_medium[ex.id])
136     else:
            rxn.lower_bound = 0.0

    model.reactions.BIOMASS_Ec_iML1515_WT_75p37M.objective_coefficient = 1
    model.solver = 'gurobi'
141
    target_growth = metadata.loc[condition_id, 'Growth Rate (1/hr)']

    # Regola iterativamente l'uptake del carbonio fino a raggiungere il tasso di crescita
        target
    while model.optimize().objective_value < target_growth:
146     model.reactions.get_by_id(carbon_source_rxns[carbon_source_name]).lower_bound -=
            0.01

    print(f'[INFO] {carbon_source_name} uptake for model {condition_id}: '
          f'{model.reactions.get_by_id(carbon_source_rxns[carbon_source_name]).
              lower_bound}')

151 return model
# ------------------------------------------------------------
# 6. Contestualizzazione tramite RIPTiDe
# ------------------------------------------------------------
def worker(metadata, expression_dict, condition_id,
156            frac_min=0.5, frac_max=0.95, samples=1000, gpr=True):
    """
    Esegue RIPTiDe per una singola condizione sperimentale e salva il modello risultante.
    """
    print(f'[INFO] Optimizing model for condition: {condition_id}...')
161 ecoli_model = prepare_ecoli_model(metadata, condition_id)
    result = riptide.maxfit(
        model=ecoli_model,
        transcriptome=expression_dict,
        frac_min=frac_min, frac_max=frac_max,
166     gpr=gpr, samples=samples, frac_step=0.05
    )

    print(f'Objective value changed from: {ecoli_model.optimize().objective_value} '
          f'to: {result.model.optimize().objective_value}')
171
    riptide.save_output(riptide_obj=result,
                        path=f'normalized_exp_maxfit_riptide_runs_iML1515/{condition_id}'
                            ,
                        file_type='sbml')
```

```
176
# ------------------------------------------------------------
# 7. Esecuzione parallela delle simulazioni
# ------------------------------------------------------------
def run_riptide(metadata, normalized_expression, timeout=1500, num_threads=1):
181     """
    Esegue pi  simulazioni RIPTiDe in parallelo e interrompe automaticamente
    quelle che superano il tempo massimo impostato.
    """
    start, stop = 0, num_threads
186     while start < len(normalized_expression.columns):
        processes = []

        for sample_id in normalized_expression.columns[start:stop]:
            exp_dict = normalized_expression[sample_id].to_dict()
191             p = mp.Process(target=worker, args=(metadata, exp_dict, sample_id), name=
                sample_id)
            p.start()
            processes.append(p)

        # Monitoraggio dei processi con timeout
196         t1 = time.monotonic()
        for p in processes:
            elapsed = time.monotonic() - t1
            remaining = timeout - elapsed
            p.join(timeout=remaining)
201             if p.is_alive():
                p.terminate()
                p.join()
                print(f'[WARNING] Terminated hung process for sample: {p.name}')

206         start += num_threads
        stop += num_threads
```

```
# ------------------------------------------------------------
# 8. Main     organizzazione del flusso completo
# ------------------------------------------------------------
if __name__ == '__main__':
    # Lettura e filtraggio dei dati sperimentali
    metadata, expression = filter_dataset()
    raw_multiplicity = read_raw_multiplicity()

    # Selezione dei geni comuni ai due dataset
    shared_genes = [gene for gene in raw_multiplicity.index if gene in expression.index]
    expression = expression.loc[shared_genes]
    raw_multiplicity = raw_multiplicity.loc[shared_genes]

    # Normalizzazione per la molteplicit  e lancio delle simulazioni
    normalized_expression = normalize_expression(raw_multiplicity, expression)
    run_riptide(metadata, normalized_expression)
```

Il blocco di codice denominato con la dicitura `"Main"` coordina l'intera pipeline: lanciando in maniera sequenziale le funzioni di lettura dei dati, normalizzazione e generazione automatizzata dei modelli contestualizzati per ciascuna condizione sperimentale.

BIBLIOGRAFIA

[1] W. Arif and A. Kalsotra. Intersections of post-transcriptional gene regulatory mechanisms and metabolic pathways. *Biochimica et Biophysica Acta (BBA) – Molecular Basis of Disease*, 1864(3):334–343, 2017. (Cited on page 45.)

[2] Scott A Becker and Bernhard O Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, 4(5):e1000082, 2008. (Cited on page 18.)

[3] Hans Bremer and Gordon Churchward. An examination of the cooper-helmstetter theory of dna replication in bacteria and its underlying assumptions. *Journal of theoretical biology*, 69(4):645–654, 1977. (Cited on page 31.)

[4] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. Cobrapy: Constraints-based reconstruction and analysis for python. *BMC Systems Biology*, 7(1):74, 2013. (Cited on pages 22 and 38.)

[5] David A Fell. Metabolic control analysis. In *Systems biology: definitions and perspectives*, pages 69–80. Springer, 2005. (Cited on page 9.)

[6] Marco Fondi, Francesco Pini, Christopher Riccardi, Pietro Gemo, and Matteo Brilli. A new selective force driving metabolic gene clustering. *Msystems*, 9(11):e00960–24, 2024. (Cited on pages 6, 7, 10, 34, and 36.)

[7] Hulda S. Haraldsdóttir, Benjamin Cousins, Ines Thiele, and Ronan M.T. Fleming. Systematic generation of kinetic equations for genome-scale metabolic models. *PLoS Computational Biology*, 13(2):e1005396, 2017. (Cited on page 16.)

[8] Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastián N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdóttir, Jacek Wachowiak, Sarah M Keating, Vanja Vlasov, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, 14(3):639–702, 2019. (Cited on page 13.)

[9] Michael Hucka, Andrew Finney, Herbert M. Sauro, Hamid Bolouri, John C. Doyle, Hiroaki Kitano, Adam P. Arkin, Brian J. Bornstein, Don Bray, Athel Cornish-Bowden, et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003. (Cited on page 23.)

[10] Joanna Jaruszewicz-Błońska and Tomasz Lipniacki. Genetic toggle switch controlled by bacterial growth rate. *BMC systems biology*, 11(1):117, 2017. (Cited on pages 5, 7, and 34.)

[11] Matthew L Jenior, Thomas J Moutinho Jr, Bonnie V Dougherty, and Jason A Papin. Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments. *PLoS computational biology*, 16(4):e1007099, 2020. (Cited on pages 18, 23, 25, and 37.)

[12] Henrik Kacser, James A Burns, H Kacser, and DA Fell. The control of flux. *Biochemical Society Transactions*, 23(2):341–366, 1995. (Cited on page 9.)

[13] Nathan E. Lewis, Kyle K. Hixson, T. Conrad, Joshua A. Lerman, Pepa Charusanti, Ashoka D. Polpitiya, Joshua N. Adkins, Geoffrey Schramm, Samuel O. Purvine, Daniel Lopez-Ferrer, Karl K. Weitz, Roland Eils, Roland E. Kontermann, Bernhard Ø. Palsson, and Joshua N. Adkins. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6:390, 2010. (Cited on page 15.)

[14] Jonathan M. Monk, Colton J. Lloyd, Elizabeth Brunk, Nils Mih, Anand Sastry, Zachary King, Rika Takeuchi, Wataru Nomura, Zhen Zhang, Hirotada Mori, and Bernhard O. Palsson. iml1515, a knowledgebase that computes escherichia coli traits. *Nature Biotechnology*, 35(10):904–908, 2017. (Cited on page 23.)

[15] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø. Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010. (Cited on pages 11, 12, 14, and 15.)

[16] Edward J. O'Brien, Jonathan M. Monk, and Bernhard Ø. Palsson. Using transcriptomics to improve metabolic network reconstruction. *Current Opinion in Systems Biology*, 1:73–80, 2015. (Cited on page 18.)

[17] Bernhard Ø. Palsson. *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge University Press, 2015. (Cited on pages 11, 12, and 15.)

[18] Andrew W Pountain, Peien Jiang, Tianyou Yao, Ehsan Homaee, Yichao Guan, Kevin JC McDonald, Magdalena Podkowik, Bo Shopsin, Victor J Torres, Ido Golding, et al. Transcription–replication interactions reveal bacterial genome regulation. *Nature*, 626(7999):661–669, 2024. (Cited on pages 7, 34, and 36.)

[19] Nathan D. Price, Jennifer L. Reed, and Bernhard Ø. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886–897, 2004. (Cited on page 16.)

[20] Maria L. Rizzo and Gábor J. Székely. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38, 2016. (Cited on page 26.)

[21] Sara Sanders, Kunaal Joshi, Petra Anne Levin, and Srividya Iyer-Biswas. Beyond the average: An updated framework for understanding the relationship between cell growth, dna replication, and division in a bacterial system. *PLoS genetics*, 19(1):e1010505, 2023. (Cited on pages 5, 6, and 7.)

[22] Chaitra Sarathy, Marian Breuer, Martina Kutmon, Michiel E Adriaens, Chris T Evelo, and Ilja CW Arts. Comparison of metabolic states using genome-scale metabolic models. *PLoS Computational Biology*, 17(11):e1009522, 2021. (Cited on page 41.)

[23] Anand V Sastry, Ye Gao, Richard Szubin, Ying Hefner, Sibei Xu, Donghyuk Kim, Kumari Sonal Choudhary, Laurence Yang, Zachary A King, and Bernhard O Palsson. The escherichia coli transcriptome mostly consists of independently regulated modules. *Nature communications*, 10(1):5536, 2019. (Cited on pages 21, 31, and 37.)

[24] Jan Schellenberger, Nathan E. Lewis, and Bernhard Ø. Palsson. Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophysical Journal*, 100(3):544–553, 2011. (Cited on page 16.)

[25] Gábor J. Székely and Maria L. Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, 2017. (Cited on page 26.)

[26] Gábor J Székely, Maria L Rizzo, et al. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004. (Cited on page 26.)

[27] Ines Thiele and Bernhard Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, 2010. (Cited on page 11.)

[28] S Towers. Potential fitting biases resulting from grouping data into variable width bins. *Physics Letters B*, 735:146–148, 2014. (Cited on page 47.)

[29] Ajit Varma and Bernhard Ø. Palsson. Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology*, 12(10):994–998, 1994. (Cited on pages 14 and 15.)

[30] Hadas Zur, Eytan Ruppin, and Tomer Shlomi. imat: an integrative metabolic analysis tool. *Bioinformatics*, 26(24):3140–3142, 2010. (Cited on page 18.)