Influenza del Terreno di Coltura sul Modello di Riproduzione Cellulare di Escherichia Coli

Paolo Leoni

1. Riassunto

È stata condotta un'analisi dati relativa al ciclo di riproduzione cellulare di Escherichia Coli, studiando le relazioni tra variabili relative alla lunghezza della cellula attraverso metodi di inferenza causale. In particolare si vuole discriminare tra i modelli teorici "Chromosome-limited" e "Concurrent Cycles". Il risultato è che in due casi su sei si è attribuito con certezza il modello "Concurrent Cycles".

2. Introduzione

La divisione cellulare di Escherichia Coli è un processo fondamentale per la sopravvivenza e la replicazione di questo microrganismo. Il meccanismo esatto attraverso il quale E. coli coordina il suo ciclo cellulare e regola la divisione in risposta a variabili ambientali e interne rimane non completamente chiarito. La complessità di questo processo suggerisce l'esistenza di molteplici vie regolatorie e meccanismi di controllo.
Per esplorare ulteriormente questi meccanismi, abbiamo analizzato sei diversi dataset che documentano i tempi degli eventi chiave del ciclo cellulare di E. coli come la nascita, l'inizio e la terminazione della replicazione del DNA, l'inizio della formazione del setto e la divisione cellulare, oltre alle corrispondenti lunghezze cellulari, in sei diversi terreni di crescita: Acetato, Alanina, Glicerolo, Glicerolo + elementi traccia, Glucosio e Mannosio. Questi dati forniscono una base empirica per testare e confrontare due modelli principali di divisione cellulare.

I due modelli si chiamano "Chromosome-limited Model" e "Concurrent Cycles Model" e ci riferiremo ad essi come modello α e modello β. L'obiettivo del nostro studio è discriminare tra questi due modelli per ogni dataset utilizzando il metodo di d-separation, al fine di identificare il modello che meglio descrive il comportamento di E. coli nei diversi ambienti di crescita.

3. Metodi Utilizzati

I dati utilizzati sono stati presi da uno studio che sfrutta l'inferenza causale per chiarire i collegamenti causali nella regolazione del ciclo cellulare di E. Coli [1].
I modelli α e β sono modelli causali strutturali (SCM), che possono essere rappresentati da grafi aciclici diretti (DAG), riportati in Figura 1. Le variabili coinvolte rappresentano:

- $\boldsymbol{L_{ip}}$, la lunghezza del battere madre al momento della replicazione del DNA
- $\boldsymbol{L_b}$, la lunghezza del battere al momento della nascita;
- $\boldsymbol{L_i}$, la lunghezza del battere al momento della replicazione del DNA.
- $\boldsymbol{L_d}$, la lunghezza del battere al momento della divisione.

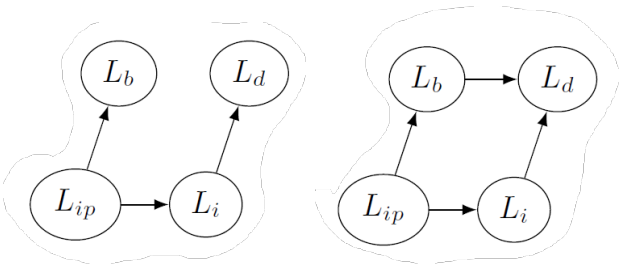

Figura 1: DAG dei modelli α e β. I nodi nel grafico sono collegati tramite delle frecce che vanno dalla causa all'effetto. Ogni nodo nel grafico rappresenta una variabile che può corrispondere a una quantità osservabile ottenuta negli esperimenti o a una variabile non osservata.

Il metodo impiegato per distinguere tra i modelli si basa sulla d-separation. Questo approccio valuta le correlazioni condizionali tra le variabili, analizzando i DAG associati a ciascun modello e applicando le regole di d-separation. Dati due nodi del DAG del modello, d-separation opera per determinare se i due sono indipendenti in seguito ad aver condizionato su un insieme di altri nodi. Un insieme di nodi S si dice che blocca un percorso p se accade almeno una delle due seguenti condizioni:

- p contiene almeno un nodo da cui esce una freccia (non-collider) che è in S
- p contiene almeno un nodo di collisione (collider) che non è in S e non ha discendenti in S

Se tutti i percorsi tra i due nodi nel DAG sono chiusi allora le variabili corrispondenti ai nodi sono indipendenti. Un collider è un nodo X che riceve frecce in entrata da due o più altri nodi (esempio: $Y \rightarrow X \leftarrow Z$).
Nel caso dei modelli α e β abbiamo guardato la relazione tra i nodi corrispondenti alle variabili L_b e L_d (vedi Figura 1), condizionando sulla variabile L_i che è quindi parte dell'insieme S.

Nel modello α, L_i è un nodo non-collider nel percorso da L_b a L_d. Dato che $S = L_i$ secondo la prima condizione, L_i chiude il percorso, bloccando qualsiasi associazione diretta tra L_b e L_d. Di conseguenza, in questo modello, L_b e L_d diventano indipendenti condizionatamente a L_i.
Nel modello β, oltre al percorso attraverso L_i come nel modello α, esiste un percorso diretto da L_b a L_d. In questo caso, condizionare su L_i chiude il percorso attraverso L_i (come nel modello α), ma il percorso diretto $L_b \rightarrow L_d$ rimane aperto poiché non passa attraverso L_i e non contiene colliders. Pertanto, L_b e L_d rimangono correlati anche quando si condiziona su L_i.
Per distinguere tra i due modelli, analizziamo la correlazione tra L_b e L_d entro un intervallo specifico $[L_{\text{imin}}, L_{\text{imax}}]$. Prevediamo che, nel modello α, la correlazione si riduca in maniera più marcata rispetto al modello β. A tal fine, abbiamo generato sinteticamente due dataset, ciascuno composto da oltre 10.000 punti: uno seguendo il modello α e l'altro il modello β. Condizionando su L_i in entrambi i casi, si osserva che nel dataset corrispondente al modello β la correlazione tra L_b e L_d diminuisce più significativamente.
Tuttavia, concentrando l'analisi sull'intervallo $[L_{\text{imin}}, L_{\text{imax}}]$ nei dati reali, si riduce notevolmente il numero di osservazioni disponibili. Questa limitazione può compromettere l'efficacia e la validità dell'analisi basata sulla d-separation.

È stato quindi utilizzato un approccio simile alla d-separation. L'idea è esaminare le relazioni tra le variabili L_b, L_i, e L_d mediante l'analisi dei residui da regressioni lineari. Per isolare gli effetti di L_i su L_b e L_d, conduciamo due regressioni lineari separate: la prima regressione tra L_i e L_b, e la seconda tra L_i e L_d. I residui di queste regressioni, indicati rispettivamente come $\text{Res}_{\text{L}_b|\text{L}_i}$ e $\text{Res}_{\text{L}_d|\text{L}_i}$, rappresentano le variazioni in L_b e L_d che non possono essere spiegate da L_i.
Successivamente, per esaminare la presenza di una possibile causa comune non rilevata tra L_b e L_d che non sia L_i, effettuiamo un'analisi della correlazione di Pearson tra i residui $\text{Res}_{\text{L}_b|\text{L}_i}$ e $\text{Res}_{\text{L}_d|\text{L}_i}$. Un coefficiente di correlazione diverso da zero suggerirebbe l'esistenza di un'influenza o di una causa comune tra L_b e L_d al di là di quella esercitata da L_i. Al contrario, un coefficiente di correlazione vicino a zero indicherebbe che, una volta rimossa l'influenza di L_i, non esistono ulteriori dipendenze significative tra L_b e L_d. Possiamo dunque affermare che se un dataset presenta un coefficiente di correlazione nullo, segue il modello α; se il coefficiente è non nullo, segue il modello β. Quindi se i residui di L_b e L_d, dopo aver regredito su L_i, non mostrano correlazione significativa, ciò supporta l'idea che non ci sono altre vie (non osservate o non misurate) che collegano L_d a L_b oltre a quelle attraverso a L_i. Questo è analogo a dire che L_b e L_d sono d-separeted da L_i.

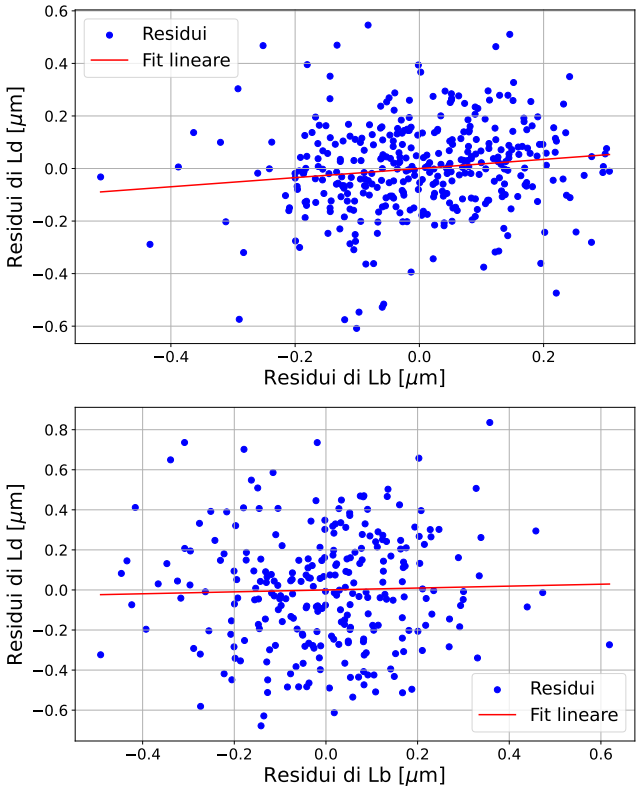

Figura 2: $\text{Res}_{\text{L}_b|\text{L}_i}$ contro $\text{Res}_{\text{L}_d|\text{L}_i}$ in Acetato (sopra) e Glucosio (sotto). È importante notare la differenza di pendenza tra le rette; $m_{acetato} = 0.26$ e $m_{glucosio} = 0.05$, possibili candidati rispettivamente ai modelli β e α.

4. Risultati

Per ogni terreno di crescita, abbiamo eseguito il test di Shapiro-Wilk per verificare la normalità

dei residui $\text{Res}_{L_b|L_i}$ e $\text{Res}_{L_d|L_i}$. In ogni caso, almeno una delle due variabili ha mostrato un p-value inferiore a 0.05, indicando una deviazione dalla normalità. Considerando anche la ridotta dimensione dei campioni, che impedisce l'applicazione del teorema del Limite Centrale, abbiamo deciso di utilizzare il coefficiente di correlazione di Spearman per analizzare la relazione tra $\text{Res}_{L_b|L_i}$ e $\text{Res}_{L_d|L_i}$. Questa decisione, pur potendo apparire restrittiva, è in realtà appropriata perché ci permette di individuare qualsiasi tipo di relazione monotona, non solo quelle lineari, tra $\text{Res}_{L_b|L_i}$ e $\text{Res}_{L_d|L_i}$. Questo approccio è utile per rivelare possibili cause sottostanti che influenzano sia L_d che L_b. Per ogni dataset sono stati eliminati gli elementi di entrambe le variabili al di fuori del 95° percentile, considerati outliers.

Tabella 1: Coefficienti di Spearman con relativi p-values di ogni terreno di crescita

	r Spearman	p value
Acetato	0.21	0.002
Alanina	0.18	0.048
Glicerolo	0.01	0.055
Glicerolo-TrEl	0.06	0.291
Glucosio	0.09	0.164
Mannosio	0.05	0.446

Dall'analisi dei dati presentati in Tabella 1, emergono come significativi solo i coefficienti di correlazione per l'Acetato e l'Alanina. I valori di r associati suggeriscono una correlazione non particolarmente robusta tra i residui; di conseguenza, la relazione tra L_b e L_d è indipendente da L_i appare debole ma non trascurabile. Possiamo dunque concludere che tali dati tendono a seguire il modello β.
Negli altri terreni di crescita la correlazione tra i residui non è statisticamente significativa, presentando un p value superiore a 0.05. Questi dati dunque non contraddicono l'ipotesi nulla secondo cui non c'è correlazione tra i residui, il che permette di non escludere il modello α. Il fatto che non ci siano abbastanza prove statistiche ($p > 0.05$) per affermare con certezza che i dati contraddicono la presenza di correlazione tra i residui (e che quindi seguono il modello α) può essere rimandato al fatto che in questi casi la rimozione della dipendenza da L_i ha già catturato la maggior parte della correlazione sistematica tra le due variabili. Tutto ciò implica che attraverso questo metodo non si osservano relazioni statisticamente significative tra L_b e L_d, il che tuttavia non esclude l'esistenza di queste relazioni. Questa argomentazione non porta ad attribuire la causa univocamente a L_i, tuttavia permette di affermare che gran parte della relazione tra L_b e L_d è spiegata in modo significativo quasi esaustivo da L_i.

5. Conclusione

L'analisi condotta aveva l'obiettivo di distinguere tra i modelli α e β, entrambi proposti per descrivere il processo di divisione cellulare in E. coli, attraverso l'utilizzo di test di indipendenza causale. Dai risultati è emerso che, in condizioni di crescita in terreni di Acetato o Alanina, i batteri tendono a seguire il modello β di divisione cellulare. Per altri terreni di crescita, invece, non è stato possibile determinare con certezza se il modello α sia predominante. È importante sottolineare che non si possono escludere altre teorie o modelli che potrebbero descrivere meglio questi processi.

Riferimenti bibliografici

[1] *Utilizzo di test di indipendenza condizionale per chiarire i collegamenti causali nella regolazione del ciclo cellulare in Escherichia coli.* URL: https://data.mendeley.com/datasets/c8fh8jy78x/1.