Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Applicazione di modelli grafici per l'analisi di dati omici: introduzione all'approccio CellOracle

Relatore Prof. Risso Davide
Dipartimento di Scienze Statistiche

Laureanda Panzavolta Lucia
Matricola 2045238

Anno Accademico 2023/2024

Indice

Introduzione

La teoria delle reti complesse svolge un ruolo cruciale in numerose discipline, tra cui informatica, sociologia, ingegneria, fisica, biologia molecolare e medicina. In biologia e medicina, l'analisi delle reti svolge un ruolo importante per l'identificazione di bersagli farmacologici, la determinazione della funzione di proteine o geni, la progettazione di strategie terapeutiche efficaci e la diagnosi precoce di malattie (Pavlopoulos et al., 2011). Le reti di regolazione genica (GRN) forniscono informazioni sul controllo dell'espressione genica nelle cellule, influenzata da vari fattori come i fattori di trascrizione e le loro modifiche post-traduzionali (Carninci et al., 2005), oltre all'associazione con altre biomolecole (Linding et al., 2008).

I recenti progressi tecnologici nello studio delle singole cellule hanno reso possibili le misurazioni multiomiche, permettendo di analizzare simultaneamente diverse modalità o di integrare dati omici provenienti da vari esperimenti. Sebbene stiano emergendo approcci computazionali per simulare i fenotipi cellulari post-perturbazione, molti di essi richiedono dati sperimentali di perturbazione per l'addestramento dei modelli, limitandone così la scala e l'applicazione. I modelli di rete di regolazione genetica risultano promettenti poiché ricostruiscono associazioni tra geni utilizzando dati omici unicellulari non perturbati.

È quindi evidente l'esigenza di approcci scalabili e interpretabili per comprendere le relazioni tra i meccanismi di regolazione genetica e i complessi fenotipi osservati a livello unicellulare. Questa tesi propone una strategia che combina la perturbazione computazionale con la modellazione delle GRN per superare queste limitazioni.

L'obiettivo di questo elaborato è confrontare alcuni algoritmi utilizzati per stimare un grafo a partire da un insieme di dati, con particolare attenzione a CellOracle, Lasso con *neighborhood selection*, ed Elastic Net. L'approccio CellOracle (Kamimoto et al., 2023) utilizza dati multimodali per costruire modelli personalizzati di GRN, simulando variazioni nell'identità cellulare a seguito della perturbazione dei fattori di trascrizione

(TF) e fornendo un'interpretazione chiara del ruolo contestuale dei TF nella regolazione dell'identità cellulare.

Nel primo Capitolo viene fornita una panoramica del contesto biologico, con particolare attenzione ai dati multiomici e allo studio sui dati reali, inclusa una breve definizione della cachessia, una sindrome presente in alcuni dei campioni studiati. Il secondo Capitolo introduce le diverse tipologie di grafi e i modelli grafici utilizzabili, esplorando la teoria dei grafi applicata alla biologia. Nel terzo Capitolo vengono elaborati i dati multiomici reali applicando l'approccio innovativo CellOracle, descrivendo in dettaglio come questo metodo sfrutta i dati di espressione genica per predire regolazioni geniche e costruire reti biologiche, con i risultati ottenuti. Il quarto Capitolo è dedicato all'applicazione del Lasso e dell'Elastic Net ai dati multiomici, descrivendo l'implementazione, i risultati ottenuti e un confronto delle loro performance. Infine, l'ultimo Capitolo adotta simulazioni mirate per valutare e confrontare l'efficacia dei tre approcci sopracitati, testando le performance degli algoritmi in condizioni controllate per fornire una valutazione critica delle loro capacità e limiti.

Capitolo 1

Contesto biologico

L'unità di base dell'ereditarietà trasmessa dai genitori ai figli è rappresentata dai geni. I geni sono composti da sequenze di DNA e sono disposti in posizioni specifiche sui cromosomi nel nucleo delle cellule. Essi contengono informazioni per la produzione di proteine specifiche che determinano l'espressione di una particolare caratteristica fisica o svolgono una funzione specifica all'interno di una cellula. Studiare la presenza delle proteine in una cellula fornisce una chiave di comprensione del funzionamento cellulare, dei suoi processi biologici coinvolti, del suo destino e del ruolo che svolge all'interno di un organismo. Inoltre, alcune proteine sono associate a malattie o condizioni patologiche: identificare proteine presenti nelle cellule malate e assenti in quelle sane può contribuire alla scoperta di nuovi marcatori di malattie e alla formulazione di nuove strategie diagnostiche e terapeutiche (Gonzalez & Kann, 2012).

La maggior parte del DNA si trova all'interno del nucleo di una cellula, dove forma i cromosomi. I cromosomi sono dotati di proteine chiamate istoni che si legano al DNA. Quest'ultimo è costituito da due filamenti che si attorcigliano nella forma di una scala elicoidale chiamata elica ed è composto da quattro blocchi fondamentali chiamati nucleotidi: adenina (A), timina (T), guanina (G) e citosina (C). I nucleotidi si legano tra loro (A con T, e G con C) per formare legami chimici chiamati coppie di basi, che collegano i due filamenti di DNA.

All'interno del nucleo cellulare, le molecole di DNA non si trovano in forma libera, ma sono organizzate in un complesso di proteine ed acidi nucleici chiamato cromatina. Il principale componente della cromatina è il DNA stesso, che, legandosi alle proteine istoniche, forma questa struttura organizzata. La cromatina è quindi la forma in cui

è organizzato il genoma cellulare, permettendo una compattazione e un'organizzazione funzionale del DNA all'interno del nucleo[1].

La trascrizione è il procedimento mediante il quale le informazioni contenute nel DNA vengono trasferite (trascritte) all'acido ribonucleico (RNA). L'RNA costituisce una lunga catena di basi, identica al filamento di DNA, ad eccezione della presenza della base uracile (U) che sostituisce la base timina (T). L'RNA contiene quindi un'informazione codificata in triplette, analogamente al DNA. All'inizio del processo di trascrizione, una porzione della doppia elica del DNA si apre, svolgendosi. Uno dei filamenti svolti del DNA funge da modello per la formazione di un filamento complementare di RNA. Questo filamento complementare è denominato RNA messaggero (mRNA). L'mRNA si separa dal DNA, lascia il nucleo e si sposta nel citoplasma cellulare, la parte della cellula al di fuori del nucleo. Qui, l'mRNA si lega a un ribosoma, una struttura cellulare in cui si verifica la sintesi proteica. Le proteine essenziali per l'avvio del processo di trascrizione sono comunemente chiamate fattori di trascrizione. Diversi di essi agiscono riconoscendo direttamente sequenze cis-regolatrici nei promotori o negli intensificatori. Tuttavia, non tutti si legano a sequenze specifiche nel promotore, alcuni infatti possono interagire con altre proteine già presenti o riconoscere direttamente le RNA polimerasi.

Durante la fase di traduzione, il codice dell'mRNA (derivante dal DNA) comunica al ribosoma l'ordine e il tipo di amminoacidi da associare. Gli amminoacidi vengono consegnati al ribosoma da un tipo più piccolo di RNA, chiamato RNA di trasporto (tRNA). Ciascuna molecola di tRNA trasporta un singolo amminoacido destinato a essere incorporato nella catena proteica in formazione, che assume una struttura tridimensionale complessa influenzata dalle molecole adiacenti, definite "accompagnatrici"[2].

1.1 Dati multi-omici

Le tecnologie ad alto rendimento hanno rivoluzionato la ricerca medica. L'introduzione degli array di genotipizzazione ha reso possibili studi di associazione su larga scala sull'intero genoma e metodi per esaminare i livelli globali di trascrizione, dando origine al campo della "genetica integrativa". Questi array, spesso chiamati SNP (polimorfismo a singolo nucleotide), forniscono informazioni su centinaia di migliaia di varianti del DNA a un costo ragionevole.

In confronto alle indagini su singoli omici, l'approccio multi-omico può offrire ai ricercatori una comprensione più approfondita del flusso informativo, partendo dalla causa

[1] `https://www.chimica-online.it/biologia/cromatina.htm`

[2] `https://www.msdmanuals.com/it-it/casa/aspetti-fondamentali/genetica/geni-e-cromosomi#Geni_v711445_it`

originaria della malattia (genetica, ambientale o dello sviluppo) fino alle conseguenze funzionali o alle interazioni rilevanti (Hasin et al., 2017).

La multiomica costituisce un metodo integrato per comprendere la biologia a differenti livelli e potenziare la ricerca sulle malattie umane. L'utilizzo di questo approccio, che combina dati provenienti da genomica, trascrittomica, epigenetica e proteomica, offre una visione più completa dei cambiamenti molecolari che influenzano lo sviluppo normale, le risposte cellulari e le malattie. Questa tecnica può anche impiegare dati omici raccolti da esperimenti precedenti, noti come approcci multiomici in silico, per analizzare in modo efficiente nuove relazioni biologiche.

In questa tesi sono stati integrati dati riguardanti scATAC-seq *(single-cell Assay for Transposase-Accessible Chromatin sequencing)* e scRNA-seq *(single-cell RNA sequencing)* per analizzare la relazione tra accessibilità della cromatina ed espressione genica e per studiare le reti di regolazione genica.

L'RNA-Seq è un metodo relativamente nuovo per profilare il trascrittoma, che sfrutta tecnologie avanzate di sequenziamento. Rispetto ad altri approcci, offre una misurazione notevolmente più precisa dei livelli di trascrizione e delle loro varianti isoformiche. Il trascrittoma rappresenta l'insieme completo dei trascritti (ovvero delle molecole di RNA) presenti in una cellula, in termini di quantità e tipologie, in uno specifico stadio di sviluppo o condizione fisiologica. La comprensione del trascrittoma risulta essenziale per interpretare gli elementi funzionali del genoma, per rivelare i costituenti molecolari delle cellule e dei tessuti, nonché per comprendere processi come lo sviluppo e le malattie. Gli obiettivi primari della trascrittomica includono la catalogazione di tutte le specie di trascritto, come mRNA, RNA non codificanti e RNA corti; la determinazione della struttura trascrizionale dei geni, compresi siti iniziali, estremità 5' e 3', modelli di splicing e altre modifiche post-trascrizionali; e la quantificazione dei mutevoli livelli di espressione di ciascuna trascrizione durante varie condizioni fisiologiche o stadi di sviluppo.

Negli ultimi anni, la tecnologia di sequenziamento dell'RNA ha fatto enormi progressi, migliorando significativamente la risoluzione dei dati. In passato, la raccolta dei dati di espressione genica era limitata a livello di interi gruppi di cellule, come tessuti o organismi (Wang et al., 2009). Tuttavia, successivamente, la tecnologia di sequenziamento dell'RNA a singola cellula (scRNA-seq) ha preso piede diventando il metodo più utilizzato per esplorare l'eterogeneità e la complessità delle trascrizioni di RNA all'interno delle singole cellule. Questo approccio ha anche permesso di identificare la composizione di vari tipi e funzioni cellulari all'interno di ambienti cellulari altamente

organizzati (Baek & Lee, 2020).

Per dare una breve introduzione al protocollo di RNA-seq, il primo passo consiste nella conversione della popolazione di RNA da sequenziare in frammenti di DNA complementare (cDNA). Questo viene fatto tramite trascrizione inversa e permette all'RNA di essere inserito in un flusso di lavoro di sequenziamento di nuova generazione (NGS). Il cDNA viene quindi frammentato e vengono aggiunti adattatori a ciascuna estremità dei frammenti. Dopo i processi di amplificazione, selezione dimensionale, pulizia e controllo di qualità, la libreria di cDNA viene analizzata tramite NGS, producendo brevi sequenze che corrispondono a tutto o parte del frammento da cui sono derivate. Il sequenziamento può seguire metodi di sequenziamento single-end o paired-end. Il sequenziamento single-read è una tecnica più economica e veloce che sequenzia i frammenti di cDNA da una solo estremità. I metodi paired-end sequenziano da entrambe le estremità e sono quindi più costosi, ma offrono vantaggi nella ricostruzione dei dati post-sequenziamento. È necessario fare un'ulteriore scelta tra protocolli specifici per il filamento e non specifici per il filamento. Il primo metodo consente di mantenere le informazioni su quale filamento di DNA è stato trascritto. Il valore delle informazioni aggiuntive ottenute dai protocolli specifici per il filamento li rende l'opzione preferibile. Queste *reads* (sequenze nucleotidiche), che saranno molte milioni alla fine del flusso di lavoro, possono quindi essere allineate a un genoma di riferimento, se disponibile, o assemblate per produrre una mappa della sequenza di RNA che copre il trascrittoma [3].

Il metodo ATAC-seq, introdotto nel 2013 da Buenrostro et al. (2013), rappresenta una tecnica biochimica utilizzata nella biologia molecolare e genomica per mappare le regioni del genoma accessibili alle proteine coinvolte nella regolazione dell'espressione genica. Rispetto al metodo DNase-seq, richiede un minor numero di cellule di partenza (500-50.000) e un tempo di elaborazione dei campioni più breve (Liu et al., 2018).

Il mappaggio delle alterazioni negli stati cellulari rappresenta un aspetto chiave per comprendere i sistemi biologici. Sia nello sviluppo sia nelle malattie, lo stato cellulare è governato da cambiamenti nell'espressione genica che, a loro volta, sono orchestrati da modifiche nei programmi di regolazione genica. Negli ultimi anni, è diventato sempre più evidente che questi programmi di regolazione sono stabiliti e controllati dall'attività di fattori di trascrizione (TF) che interpretano e modificano lo stato epigenetico sottostante della cromatina. Lo stato epigenetico della cromatina può essere regolato attraverso una varietà di meccanismi, compresa la modifica chimica sia del DNA che delle proteine

[3] `https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461`

istoniche che, a loro volta, alterano la dinamica e la struttura della cromatina (Grandi et al., 2021).

L'ATAC-seq utilizza un enzima chiamato trasposasi per contrassegnare le regioni di cromatina accessibili. Questo enzima è coinvolto nella trasposizione, il processo di movimento di segmenti di DNA all'interno del genoma. La trasposasi catalizza il trasferimento di segmenti di DNA, noti come elementi trasponibili o trasposoni, da una posizione del DNA a un'altra. Questo meccanismo coinvolge il taglio del DNA in punti specifici, creando estremità a singolo o doppio filamento su cui il trasposone si può legare. Queste estremità possono essere integrate in nuove posizioni nel genoma attraverso specifici meccanismi. Le trasposizioni si presentano in varie forme: alcuni trasposoni si spostano all'interno del genoma, altri vengono replicati e inseriti in nuove posizioni, mentre altri ancora possono provocare variazioni genomiche o mutazioni.

Successivamente, queste sequenze vengono amplificate e sequenziate tramite la tecnologia Next-Generation Sequencing (NGS), permettendo l'identificazione delle regioni genomiche aperte per l'attività genica (Liu et al., 2018). Questa tecnica fornisce dettagli sullo stato della cromatina, l'accessibilità del DNA e l'attività genica, ed è ampiamente utilizzata nella ricerca biologica per comprendere lo sviluppo, la differenziazione cellulare, le malattie e altri processi biologici.

L'ATAC-seq offre quindi un'analisi dettagliata della regolazione dell'espressione genica, svolgendo un ruolo cruciale nella comprensione dei meccanismi molecolari coinvolti in una vasta gamma di fenomeni biologici. Questa tecnica è fondamentale per identificare il profilo cromatinico associato a specifici tipi cellulari e per comprendere come possa essere influenzato da perturbazioni o condizioni patologiche, contribuendo così alla comprensione delle malattie e dei meccanismi coinvolti nelle stesse. Nel 2015 è stata sviluppata da Buenrostro et al. (2015) la tecnologia ATAC-seq a singola cellula (scATAC-seq) per esaminare l'accessibilità della cromatina in modo specifico alle diverse cellule presenti in campioni di tessuto contenenti una popolazione cellulare eterogenea. Tuttavia, a causa della natura intrinsecamente rumorosa e dispersa dei dati scATAC-seq, l'interpretazione accurata dei segnali biologici e la formulazione di ipotesi biologiche significative possono risultare complesse e difficoltose.

Sebbene l'RNA-seq misuri in modo più diretto l'espressione genica, l'ATAC-seq può offrire spiegazioni sui meccanismi alla base della regolazione dell'espressione genica o sulle eventuali differenze tra due tipi o condizioni cellulari.

Il protocollo Chromium Next GEM Single Cell Multiome ATAC + Gene Expression di 10x Genomics rappresenta una potente tecnologia per l'analisi integrata della cromatina

e dell'espressione genica a livello di singola cellula. Si inizia con l'isolamento dei nuclei dalle cellule. La trasposizione viene eseguita in massa utilizzando l'enzima trasposasi, che taglia preferenzialmente il DNA nucleare nelle regioni di cromatina aperta. I nuclei trasposti vengono quindi suddivisi in goccioline, o GEMs (*Gel Bead-in-Emulsions*), con una singola Gel Bead contenente un codice a barre 10x unico. All'interno della GEM, i codici a barre unici vengono attaccati all'mRNA disponibile e ai frammenti di DNA trasposto in un singolo nucleo. Dopo questa incubazione, le GEM vengono rotte e raggruppate prima della pulizia, preamplificazione e costruzione della libreria. Da un singolo pool di GEMs vengono create due librerie, una per il sequenziamento dell'RNA e una per l'ATAC [4].

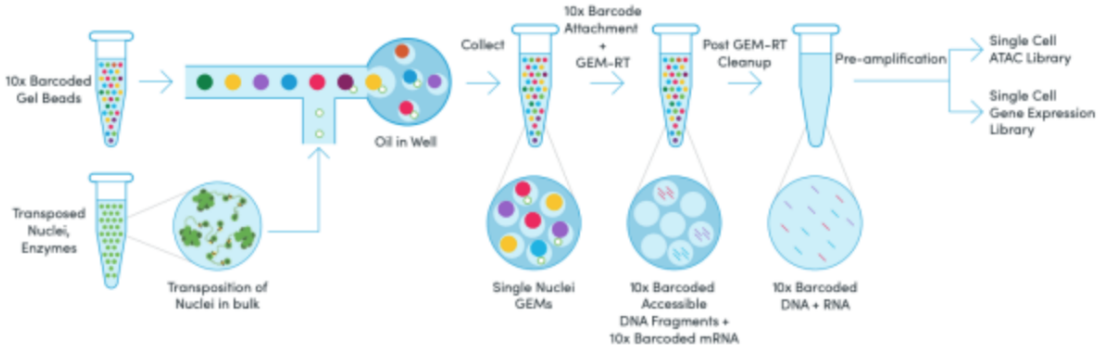

FIGURA 1.1: Protocollo del Single Cell Multiome ATAC + Gene Expression. [4]

1.2 Cachessia

Nel Capitolo 3 si tratteranno dati reali riguardanti un caso di studio eseguito su topi da laboratorio. Questo studio ha lo scopo di studiare la cachessia e per questo motivo, di seguito, verrà introdotta questa patologia in ambito umano.

Circa l'80 % dei pazienti oncologici è affetto da cachessia, una sindrome multifattoriale caratterizzata da una grave perdita di peso, principalmente causata dal deperimento muscolare, con o senza perdita di massa grassa, che può essere parzialmente mitigata mediante supporto nutrizionale (Fearon et al., 2011). Questo disturbo è guidato da una combinazione eterogenea di fattori come la ridotta assunzione di cibo e i cambiamenti metabolici (apparentemente indotti da fattori derivati dal tumore e dal soggetto), un'aumentata spesa energetica, un catabolismo e un'infiammazione eccessivi.

È ben noto quindi come la condizione cachettica sia caratterizzata anche da una componente anoressica che può portare a una ridotta assunzione di cibo e, alla fine, a

[4] `https://www.10xgenomics.com/blog/introducing-chromium-single-cell-multiome-atac-gene-expression`

una perdita di peso corporeo (Laviano et al., 2008). Quando la spesa energetica a riposo supera l'assunzione di cibo, si instaura uno squilibrio energetico negativo, in parte connesso al metabolismo tumorale. Per questo motivo, la massa tumorale e il suo metabolismo anaerobico contribuiscono all'influenza del tumore sul corpo, inducendo una spesa energetica eccessiva per sostenere i meccanismi correlati al tumore, potenzialmente portando direttamente a una diminuzione delle proteine muscolari e al conseguente deperimento muscolare (Friesen et al., 2015). Oltre a impattare sulla qualità della vita, la cachessia ostacola anche la possibilità per i pazienti di seguire in modo ottimale i trattamenti farmacologici in corso. Attualmente, non esiste ancora una cura per questa malattia, ma l'approfondimento degli studi sui meccanismi molecolari alla base della perdita di massa muscolare potrebbe contribuire a individuare soluzioni terapeutiche (Leduc-Gaudet et al., 2023).

1.2.1 Lo studio

In questo studio si vuole studiare l'effetto della cachessia sulla composizione cellulare del tessuto muscolo-scheletrico in topi con tumore al colon. Per l'esperimento sono stati arruolati sei topi: tre affetti dal carcinoma colon-26 (C26), organismo modello per studiare la cachessia, e tre controlli sani (CTRL).

- Su quattro topi (due cachetici e due sani) sono stati misurati simultaneamente l'RNA e l'ATAC dallo stesso nucleo. Questi campioni verranno chiamati *multiome* in quanto sono composti da due misurazioni.
- Su due topi (uno cachetico e uno sano) è stato misurato solo l'RNA.

Per ogni replica multiome è stato effettuato il controllo della qualità, ovvero sono stati rimossi i nuclei di bassa qualità o danneggiati. Per verificare se un nucleo è stato danneggiato durante il suo isolamento, si calcola la percentuale di reads mitocondriale in ogni cellula. Un'alta percentuale di RNA mitocondriale può indicare cellule in apoptosi o danneggiate.

Prima di identificare i tipi cellulari, è stata effettuata l'integrazione. Questo è uno step importante per eliminare gli effetti di batch, ovvero le differenze tecniche tra i dataset. Per questa correzione è stato usato il metodo di Seurat (Versione 3). Per identificare gli anchors, coppie di cellule che si ritiene siano simili tra i diversi dataset, è stata usata la reciprocal PCA ('RPCA') (Stuart et al., 2019).

L'identificazione cellulare è stata effettuata con un metodo semi-supervisionato. I collaboratori hanno fornito una lista di geni marcatori dei tipi cellulari che si aspettano

di trovare. Quindi, è stato effettuato il clustering: per prima cosa sono stati identificati i k-nearest neighbors per ogni cellula utilizzando il grafo dei Shared Nearest Neighbors (SNN), e poi è stato calcolato il clustering con l'algoritmo di Louvain (Stuart et al., 2019).

Successivamente, si è visualizzata l'espressione media dei marcatori in ogni cluster mediante una heatmap. Ad ogni cluster viene assegnata l'etichetta del tipo cellulare corrispondente al marcatore espresso. Successivamente è stato effettuato nuovamente il clustering su un gruppo cellulare, i Myonuclei, per identificare i sottogruppi. E' stato effettuato lo stesso procedimento, precedentemente descritto.

Nelle successive analisi verranno considerati solo i dati relativi ai topi per i quali sono stati misurati simultaneamente l'RNA e l'ATAC dallo stesso nucleo, fondamentali per la costruzione e l'analisi delle reti di regolazione genica.

Lo scopo di questo elaborato è la costruzione e l'analisi delle reti di regolazione genica. Come primo approccio è stato utilizzato CellOracle[5], per costruire reti di regolazione genica basate su machine learning e dati multi-omici. Successivamente, sono stati introdotti altri due approcci, *neighborhood selection* con il Lasso e Elastic Net, noti per la loro capacità di fornire predizioni robuste e soluzioni sparse. Per il confronto delle reti risultanti è stata impiegata la teoria dei grafi per esplorare la struttura e l'organizzazione delle interazioni geniche identificate. Questa procedura ha permesso di valutare la coerenza biologica delle predizioni e di identificare pattern comuni e distintivi tra i diversi metodi.

[5] `https://morris-lab.github.io/CellOracle.documentation/`

Capitolo 2

Introduzione ai modelli grafici

Per introdurre i concetti fondamentali della teoria dei grafi si riporta sia la descrizione empirica che quella matematica come sono originariamente definiti nella letteratura (Pavlopoulos et al., 2011).

Un modello grafico è un modello statistico associato a un grafo. I nodi del grafo corrispondono alle variabili casuali di interesse, e gli archi codificano le dipendenze condizionali ammesse tra le variabili. Le proprietà di fattorizzazione sottostanti ai modelli grafici facilitano il calcolo trattabile con distribuzioni multivariate, rendendo i modelli uno strumento prezioso in una moltitudine di applicazioni. Inoltre, i modelli grafici diretti ammettono interpretazioni causali intuitive e sono diventati un pilastro per l'inferenza causale. L'idea di utilizzare grafi i cui nodi corrispondono a variabili casuali al fine di descrivere strutture di indipendenza condizionata era emersa in statistica prima che Pearl & Paz (1986) suggerissero questo approccio nel contesto dell'informatica. Il primo lavoro statistico in cui sono stati formulati è di Speed (1978). Si possono distinguere due tendenze tradizionali di base, ossia l'uso di grafi non diretti e diretti (aciclici). Si noti che i modelli statistici descritti da tali grafi possono essere compresi come modelli di (speciali) strutture di indipendenza condizionata. I grafi non diretti (UGs) sono apparsi negli anni '70 in fisica statistica come strumenti per descrivere le relazioni tra variabili casuali discrete. Successivamente, negli anni '80, vennero introdotti i grafi diretti aciclici, denominati DAGs (Maathuis et al., 2018).

2.1 Indipendenza e indipendenza condizionata

Si ricorda che praticamente qualsiasi concetto definito per la probabilità può anche essere esteso alla probabilità condizionata. Due eventi A e B sono indipendenti se $Pr(A \cap B) = Pr(A)Pr(B)$, o equivalentemente, $Pr(A|B) = Pr(A)$ e $Pr(B|A) = Pr(B)$.

Difatti, in presenza di indipendenza, la realizzazione di A non va ad influenzare in alcun modo la probabilità di manifestarsi di B e viceversa.

Possiamo estendere questo concetto agli eventi condizionatamente indipendenti. In particolare, due eventi A e B sono condizionatamente indipendenti dato un evento C con $Pr(C) > 0$ se $Pr(A \cap B|C) = Pr(A|C)Pr(B|C)$. Dalla definizione di probabilità condizionata,

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

se $Pr(B) > 0$. Condizionando su C, otteniamo

$$Pr(A|B, C) = \frac{Pr(A \cap B|C)}{Pr(B|C)}$$

se $Pr(B|C), Pr(C) \neq 0$. Se A e B sono condizionatamente indipendenti dato C, otteniamo

$$\begin{aligned} Pr(A|B, C) &= \frac{Pr(A \cap B|C)}{Pr(B|C)} \\ &= \frac{Pr(A|C)Pr(B|C)}{Pr(B|C)} \\ &= Pr(A|C) \end{aligned}$$

Quindi, se A e B sono condizionatamente indipendenti dato C, allora $Pr(A|B, C) = Pr(A|C)$. Va sottolineato che l'indipendenza condizionata non implica l'indipendenza marginale [1].

2.2 Grafo non diretto

Un grafo G può essere definito come una coppia (V, E) dove V è un insieme di vertici che rappresentano i nodi ed E è un insieme di archi che rappresentano le connessioni tra i nodi. Definiamo come $E = \{(i, j)|i, j \in V\}$ la singola connessione tra i nodi i e j. In questo caso, diciamo che i e j sono vicini. Una connessione multipla consiste in due o più archi che hanno gli stessi nodi terminali. Tali archi multipli sono particolarmente importanti per le reti in cui due elementi possono essere collegati da più di una connessione. In tali casi, ogni connessione indica un diverso tipo di informazione.

[1] `https://www.probabilitycourse.com/chapter1/1_4_4_conditional_independence.php`

2.3 Grafo diretto

Un grafo diretto è definito come una tripla ordinata $G = (V,\ E,\ f)$, dove f è una funzione che mappa ciascun elemento in E a una coppia ordinata di vertici in V. Le coppie ordinate di vertici sono chiamate archi diretti o frecce. Un arco $E = (i,\ j)$ è considerato con direzione da i a j.

Se c'è una freccia da i a j, allora scriviamo questo come $i \to j$, o equivalentemente come $[ij] \in E$. Nota che $[ij]$ non è lo stesso di $[ji]$. Se $i \to j$ o $j \to i$, diciamo che i e j sono adiacenti e si scrive $i \sim j$. Per percorso si intende una sequenza di vertici $\{V_1, \ldots, V_k\}$ tale che $V_i \sim V_{i+1}$ per ogni $i = 1, \ldots, k-1$. In contrasto, per un percorso diretto da V_1 a V_k viene richiesto che $V_i \to V_{i+1}$ per ogni $i = 1, \ldots, k-1$. Quando i primi e gli ultimi vertici coincidono, cioè $V_1 = V_k$, il percorso diretto è chiamato ciclo diretto (Edwards, 2000).

I grafi diretti sono principalmente adatti per la rappresentazione di schemi che descrivono *pathways* biologici o procedure che mostrano l'interazione sequenziale degli elementi in uno o più punti temporali e il flusso di informazioni attraverso la rete. Queste sono principalmente reti metaboliche, di trasduzione del segnale o regolatorie.

2.3.1 Grafo diretto aciclico

Ci si concentra sui grafi diretti senza cicli diretti. Questi sono generalmente conosciuti come grafi aciclici diretti, o DAGs per abbreviare. Se $i \to j$, allora i è chiamato un "genitore" di j e j è chiamato un "figlio" di i. L'insieme dei "genitori" di j è indicato come pa(j) e l'insieme dei figli come ch(j).

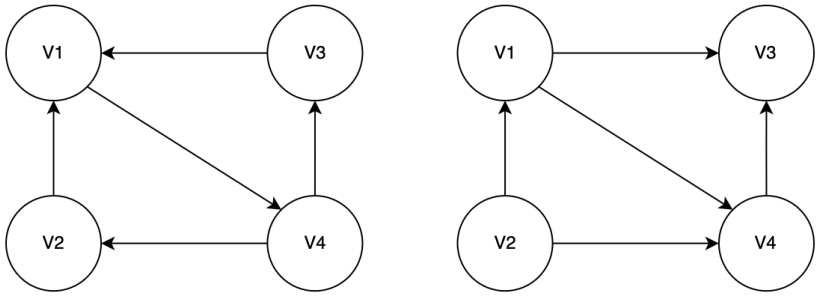

FIGURA 2.1: Due grafi diretti. A sinistra un grafo diretto ciclico, a destra un DAG. Fonte: `https://hyperskill.org/learn/step/44126`

Nel costruire un DAG, viene disegnata una freccia da V_i a V_j, dove $i < j$, a meno che $f(V_j|V_{j-1}...V_1)$ non dipenda da V_i. Questa è la differenza chiave tra i DAGs e i grafi non diretti. In entrambi i tipi di grafi, l'assenza di un arco tra V_i e V_j è equivalente

a una relazione di indipendenza condizionale tra V_i e V_j. Nei grafi non diretti, sono indipendenti condizionatamente date tutte le variabili rimanenti, mentre nei DAGs sono indipendenti condizionatamente date tutte le variabili precedenti (Edwards, 2000).

2.4 Grafo pesato

Un grafo pesato è definito come un grafo $G = (V, E)$ dove V è un insieme di vertici e E è un insieme di archi tra i vertici $E = \{(u, v)|u, v \in V\}$. Associati ad esso vi è una funzione di peso $w : E \to \mathbb{R}$, dove $\mathbb{R}$ denota l'insieme di tutti i numeri reali. Nella maggior parte dei casi, il peso $w_{i,j}$ dell'arco tra i nodi i e j rappresenta la rilevanza della connessione. Di solito, un peso maggiore corrisponde a una maggiore affidabilità di una connessione. I grafi ponderati sono attualmente le reti più ampiamente utilizzate in tutto il campo della bioinformatica.

Il grafo bipartito è un grafo non diretto $G = (V, E)$ in cui V può essere suddiviso in due insiemi V_1 e V_2 in modo che $(u, v) \in E$ implichi che $u \in V_1$ e $v \in V_2$ oppure $v \in V_1$ e $u \in V_2$. Le applicazioni di questo tipo di grafo alla visualizzazione o alla modellizzazione delle reti biologiche vanno dalla rappresentazione dei collegamenti enzima-reazione nei percorsi metabolici alle ontologie o connessioni ecologiche, come discusso in Burgos et al. (2008) o Picard et al. (2008).

Se $G = (V, E)$ è un grafo, allora $G_1 = (V_1, E_1)$ è chiamato sottografo se $V_1 \subseteq V$ ed $E_1 \subseteq E$, dove ogni arco in E_1 è incidente con i vertici in V_1.

Esempi e forme che descrivono i tipi di grafi sopra menzionati vengono riportati in Figura 2.2. Le strutture dati più comuni utilizzate per rendere queste reti leggibili dai computer sono le matrici di adiacenza o le liste di adiacenza.

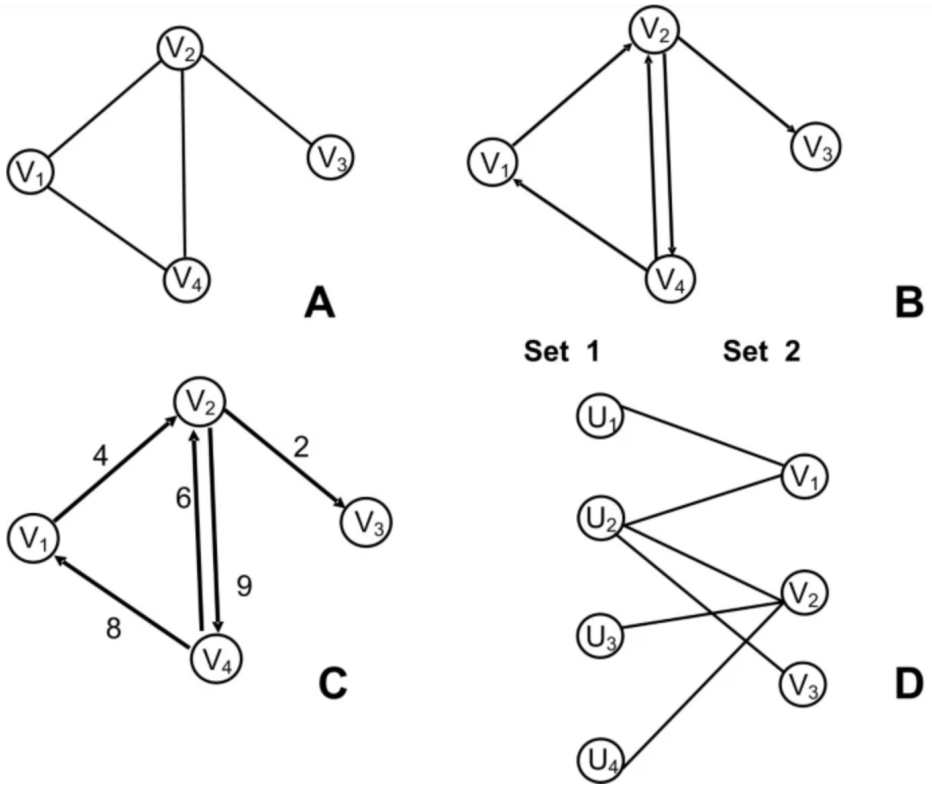

FIGURA 2.2: **A**. Grafo indiretto: $V = V_1, V_2, V_3, V_4$, $|V| = 4$, $E = (V_1, V_2), (V_2, V_3), (V_2, V_4), (V_4, V_1)$, $|E| = 4$. **B**. Grafo diretto: $V = V_1, V_2, V_3, V_4$, $|V| = 4$, $E = (V_1, V_2), (V_2, V_3), (V_2, V_4), (V_4, V_1), (V_4, V_2)$, $|E| = 5$. **C**. Grafo pesato: $V = V_1, V_2, V_3, V_4$, $|V| = 4$, $E = (V_1, V_2, 4), (V_2, V_3, 2), (V_2, V_4, 9), (V_4, V_1, 8), (V_4, V_2, 6)$, $|E| = 5$. **D**. Grafo bipartito: $V = U_1, U_2, U_3, U_4, V_1, V_2, V_3$, $|V| = 7$, $E = (U_1, V_1), (U_2, V_1), (U_2, V_2), (U_2, V_3), (U_3, V_2), (U_4, V_2)$, $|E| = 6$. (Pavlopoulos et al., 2011)

2.5 Struttura dei dati

Come anticipato in precedenza, le due principali strutture dati utilizzate per memorizzare le rappresentazioni dei grafici di rete sono la matrice di adiacenza e la lista di adiacenza.

2.5.1 Matrice di adiacenza

Dato un grafo $G = (V, E)$, la rappresentazione tramite matrice di adiacenza consiste in una matrice $A = (a_{ij})$ di dimensione $n \times n = |V| \times |V|$ tale che $a_{ij} = 1$ se $(i, j) \in E$ o $a_{ij} = 0$ o $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$, dove $n = |V|$. Nel caso di grafi pesati, $a_{ij} = w_{ij}$ se $(i, j) \in E$, altrimenti $a_{ij} = 0$. Per i grafi non diretti, la matrice è simmetrica perché $a_{ij} = a_{ji}$. Questa regola non si applica ai grafi diretti, poiché in quel caso le parti triangolari superiore e inferiore della matrice rivelano la direzione degli archi. Le matrici di adiacenza richiedono uno spazio di $\Theta(|V|^2)$ e sono più adatte per grafi densi e non per grafi sparsi. Questa struttura dati è quindi più efficiente per reti "affollate", dove la densità delle connessioni tra gli elementi è relativamente alta. Nel caso di un grafo completamente connesso in cui tutti i nodi sono collegati tra loro, le matrici di adiacenza sono altamente consigliate.

2.5.2 Lista di adiacenza

Dato un grafo $G = (V, E)$, la rappresentazione mediante lista di adiacenza consiste in un array Adj di $|E|$ elementi, dove per ogni $e \in E$ si ha $Adj(0, e) = i \in V$. Le liste di adiacenza richiedono uno spazio $\Theta(|V| + |E|)$ e sono preferibili per grafi sparsi con una bassa densità di connessioni. Un esempio di come queste strutture dati rappresentano un grafo è mostrato in Figura 2.3.

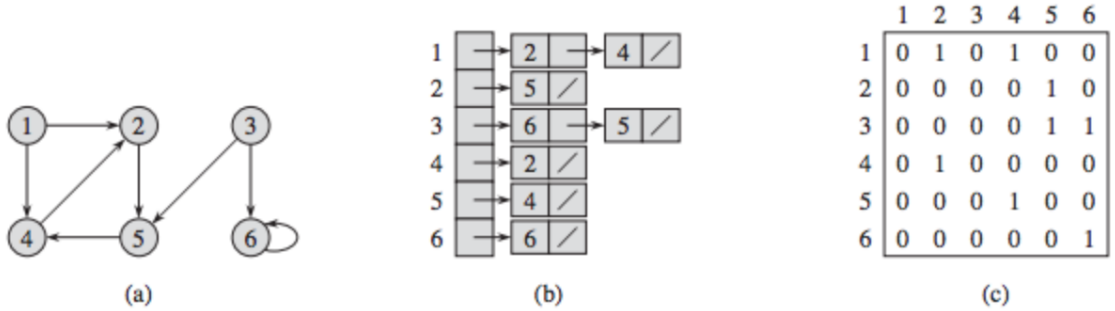

FIGURA 2.3: **(a)**. Un Grafo Diretto: Un grafo casuale composto da sei nodi e otto archi diretti. **(b)**. Lista di Adiacenza: La struttura dati che rappresenta il grafo diretto utilizzando liste. **(c)**. Matrice di Adiacenza: La struttura dati che rappresenta il grafo diretto utilizzando una matrice 2D. Lo zero rappresenta l'assenza della connessione mentre l'uno rappresenta l'esistenza della connessione tra due nodi. La matrice non è simmetrica poiché il grafo è diretto. Fonte: `https://cis.temple.edu/~pwang/9615-AA/Lecture/09-Graph-1.htm`

2.6 Centralità della rete

Esaminare diverse proprietà di una rete può fornire preziose informazioni sull'organizzazione interna di una rete biologica, sulla ripartizione delle molecole tra i processi cellulari, così come sui vincoli evolutivi che hanno plasmato la struttura funzionale di una rete proteica, metabolica o regolatoria di un organismo.

Il grado di un nodo in un grafo non orientato è il numero di connessioni o archi che il nodo ha con altri nodi ed è definito come $\deg(i) = k(i) = |N(i)|$, dove $N(i)$ è il numero dei vicini del nodo i. Se una rete è orientata, allora ogni nodo ha due gradi differenti: il grado entrante $\deg_{\text{in}}(i)$, che è il numero di archi arrivano nel nodo i, e il grado uscente $\deg_{\text{out}}(i)$, che è il numero di archi che escono dal nodo i (Pavlopoulos et al., 2011).

La connettività totale di una rete è definita come $C = \frac{E}{N(N-1)}$, dove E è il numero di archi e N è il numero totale di nodi. La struttura di connettività delle reti biologiche è spesso informativa per quanto riguarda l'interazione e la reversibilità delle reazioni, i composti che strutturano la rete, come nel metabolismo, o le relazioni trofiche, come nelle reti alimentari.

La densità del grafo mostra quanto sparsa o densa sia una rete in base al numero di connessioni per insieme di nodi ed è definita come $density = \frac{2|E|}{|V|(|V|-1)}$. Le reti biologiche sono solitamente scarsamente connesse, poiché ciò conferisce un vantaggio evolutivo per preservare la robustezza.

Come possono essere classificati o ordinati i nodi in base alle loro proprietà? Nelle reti biologiche, è importante individuare nodi centrali o nodi intermedi che influenzano la topologia della rete, a seconda naturalmente della domanda biologica. Una tale domanda potrebbe essere trovare le molecole in una via biologica che non sono necessariamente centrali ma hanno un ruolo biologico cruciale nella trasduzione del segnale, per individuare molecole che sono cruciali per stimolare l'espressione dei geni.

Le misure di centralità più utilizzate in ambito biologico sono:

- **Centralità di grado** (Degree Centrality): Un nodo importante è coinvolto in un grande numero di interazioni. Per un nodo i, la centralità di grado è calcolata come $C_d(i) = \deg(i)$. Per i grafi diretti, ogni nodo è ovviamente caratterizzato da due centralità di grado. Queste sono $C_{d_{\text{in}}}(i) = \deg_{\text{in}}(i)$ e $C_{d_{\text{out}}}(i) = \deg_{\text{out}}(i)$. I nodi con centralità di grado molto alta sono chiamati *hub* poiché sono connessi a molti vicini. La rimozione di tali nodi centrali ha un grande impatto sulla topologia della rete. È stato dimostrato che le reti biologiche tendono ad essere robuste contro le perturbazioni casuali, ma la rimozione degli *hub* spesso porta al fallimento del sistema (Zotenko et al., 2008).

- **Centralità betweenness** (Betweeness Centrality): I nodi che sono intermedi tra vicini hanno un rango più elevato. Senza questi nodi, non ci sarebbe modo per due vicini di comunicare tra loro. Pertanto, la centralità betweenness mostra i nodi importanti che si trovano su una grande proporzione di percorsi tra altri nodi nella rete. Per nodi distinti $i, j, w \in V(G)$, sia σ_{ij} il numero totale di percorsi più brevi tra i e j. Con percorso più breve si intende il percorso con il minor numero di archi necessari per connettere due nodi. Allora, la centralità di betweenness è calcolata come $C_b(w) = \sum_{i \neq j \neq w} \frac{\sigma_{ij}(w)}{\sigma_{ij}}$, dove $\sigma_{ij}(w)$ rappresenta il numero di percorsi più brevi da i a j che passano attraverso il nodo w. Il calcolo di questa misura di centralità è discusso in Rong et al. (2009).

- **Centralità dell'autovettore** (Eigenvector Centrality): I nodi che sono connessi a vicini importanti hanno un rango più alto. Sia $G = (V,\ E)$ un grafo non orientato e A la matrice di adiacenza della rete G. La centralità dell'autovettore è l'autovettore C_{eiv} dell'autovalore più grande λ_{max} in valore assoluto tale che $\lambda C_{eiv} = AC_{eiv}$. Formalmente, se A è la matrice di adiacenza di una rete G con $V(G) = v_1, ..., v_n$, e $\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$, allora la centralità dell'autovettore $C_{eiv}(v_i)$ del nodo v_i è data dalla i-esima coordinata x_i di un autovettore normalizzato che soddisfa la condizione $Ax = \rho(A)x$ (Özgür et al., 2008). In biologia, questa misura

di centralità è stata utilizzata, per esempio, per identificare interazioni genetiche sintetiche (Paladugu et al., 2008), associazioni gene-malattia (Özgür et al., 2008) o *hubs* di rete (Zotenko et al., 2008).

- **Centralità di vicinanza** (Closeness Centrality): indica i nodi importanti che possono comunicare rapidamente con altri nodi della rete. Sia $G = (V, E)$ un grafo non orientato. Allora, la centralità è definita come $C_{clo}(i) = \frac{1}{\sum_{t \in V}^{|V|} dist(i,j)}$, dove $dist(i, j)$ indica la distanza o il percorso più breve ρ tra i nodi i e j (Hansen et al., 2020).

- **Centralità dell'eccentricità** (Eccentricity Centrality): La centralità dell'eccentricità è la misura che indica quanto facilmente accessibile è un nodo rispetto agli altri nodi. Sia $G = (V, E)$ un grafo non orientato. La centralità dell'eccentricità è calcolata come $C_{ecc} = \frac{1}{\max(dist(i,j))}$, dove $dist(i, j)$ è il percorso più breve tra i nodi i e j. L'eccentricità C_{ecc} di un vertice V è la massima distanza tra v e qualsiasi altro vertice. Nelle reti biologiche, le proteine o altre bioentità con un'elevata eccentricità sono facilmente raggiungibili funzionalmente da altri componenti della rete, e quindi possono facilmente percepire cambiamenti nella concentrazione di altri enzimi o molecole a cui sono collegati. Al contrario, quelle proteine che hanno eccentricità più basse spesso svolgono un ruolo funzionale marginale nel sistema (Chavali et al., 2010).

2.7 Modelli Grafici per grafi indiretti

I modelli grafici sono strumenti che vengono regolarmente utilizzati per investigare le strutture di dipendenza complesse nei dataset biomedici ad alto rendimento. Consentono una visione olistica a livello di sistema dei vari processi biologici, permettendo un'interpretazione intuitiva e rigorosa (Ni et al., 2022).

Con l'avvento delle tecnologie omiche ad alto rendimento è emersa una crescente necessità di strumenti di analisi dati per esplorare le relazioni tra diversi output biologici. Questi output possono essere, ad esempio, i livelli di espressione di diversi geni, la presenza di varianti genomiche, l'accessibilità di regioni genomiche individuali (struttura del cromatina), o le abbondanze di singoli metaboliti. La modellazione grafica con reti biologiche è uno dei modi più popolari per analizzare questo tipo di dati. Le reti biologiche sono strutture in cui i nodi rappresentano singole unità come geni, proteine o metaboliti e i collegamenti riflettono una relazione tra coppie di nodi come co-espressione

o interazione fisica. Questo tipo di rete sono state cruciali nell'identificare geni di suscettibilità alle malattie e nell'individuare potenziali bersagli terapeutici farmacologici (Shutta et al., 2022).

2.7.1 Modelli Grafici Gaussiani

Molto utilizzati sono i modelli grafici gaussiani (GGMs), strumenti per inferire le dipendenze tra variabili biologiche. Applicazioni popolari sono la ricostruzione di reti di associazione tra geni, proteine e metaboliti. I GGMs sono uno strumento di ricerca esplorativa che può essere utile per scoprire relazioni interessanti tra geni o per identificare geni di interesse terapeutico, ma non necessariamente inferiscono una rete nel senso meccanicistico. Anche se i GGMs sono ben investigati dal punto di vista teorico e applicativo, importanti estensioni non sono ben conosciute all'interno della comunità biologica. Il GGM è un caso speciale di un grafo di indipendenza condizionata. In un GGM, assumiamo che X segua una distribuzione normale multivariata:

$$X = (X_1, \ldots, X_p) \sim \text{MVN}(\mu, \Sigma),$$

dove $\mu = (\mu_1, \ldots, \mu_p)$ è un vettore p-dimensionale di medie e Σ è la matrice di covarianza $p \times p$ di X. Assumendo che la matrice di covarianza Σ sia non singolare, la struttura di indipendenza condizionale della distribuzione può essere comodamente rappresentata da un modello grafico $G = (V, E)$, dove $V = \{1, \ldots, p\}$ è l'insieme dei nodi ed E l'insieme degli archi in $V \times V$. Una coppia (i, j) è contenuta nell'insieme degli archi E se e solo se X_i è condizionalmente dipendente da X_j, date tutte le rimanenti variabili $X_{V\setminus\{i,j\}} = \{X_k; k \in V \setminus \{i, j\}\}$ (Proprietà di Markov). Ogni coppia di variabili non contenuta nell'insieme degli archi è condizionatamente indipendente, date tutte le rimanenti variabili, e corrisponde a un elemento zero nella matrice di covarianza inversa (matrice di precisione).

I pesi degli archi di un GGM corrispondono alle correlazioni parziali tra le variabili in X. La correlazione parziale tra le variabili X_i e X_j è una misura della loro associazione condizionale, data la presenza degli altri elementi di X (Altenbuchinger et al., 2020).

Un approccio naturale per stimare la struttura del grafo è la selezione in avanti o all'indietro. Nella ricerca in avanti, la stima iniziale dell'insieme degli archi è l'insieme vuoto e gli archi vengono aggiunti iterativamente fino a quando non viene soddisfatto un opportuno criterio di arresto. La selezione o l'eliminazione di un singolo arco in questa strategia richiede una stima di massima verosimiglianza (MLE) per $O(p^2)$ modelli differenti. La procedura non è adatta per grafi ad alta dimensionalità e l'esistenza della

stima di massima verosimiglianza non è garantita se il numero di osservazioni è inferiore al numero di nodi.

Un modo preferibile per procedere è la *neighborhood selection* con il Lasso, basata sull'ottimizzazione di una funzione convessa, applicata consecutivamente a ciascun nodo del grafo. Il metodo è computazionalmente molto efficiente ed è consistente anche per contesti ad alta dimensione. Questa selezione è un sottoproblema della selezione della covarianza. Il *neighborhood* (ne_i) di un nodo $i \in V$ è il più piccolo sottoinsieme di $V \setminus \{i\}$ tale che, date tutte le variabili X_{ne_i} nel *neighborhood*, X_i è condizionalmente indipendente da tutte le rimanenti variabili. Il *neighborhood* di un nodo $i \in V$ consiste di tutti i nodi $j \in V \setminus \{i\}$ tali che $(i, j) \in E$. Date n osservazioni i.i.d. di X, la *neighborhood selection* mira a stimare individualmente il vicinato di qualsiasi variabile data o di un nodo dato. Questa selezione può essere formulata come un problema di regressione standard e può essere risolta in modo efficiente con il Lasso (Tibshirani, 1996), come vedremo nel successivo Capitolo 4.

La *neighborhood selection* con il Lasso è quindi un'alternativa computazionalmente attraente alla selezione standard della covarianza per grafi sparsi ad alta dimensionalità. Questo genere di selezione stima le restrizioni di indipendenza condizionale separatamente per ciascun nodo nel grafo e quindi è equivalente alla selezione di variabili per modelli lineari gaussiani. Lo schema di selezione *neighborhood* è consistente per grafi sparsi ad alta dimensionalità. La consistenza dipende dalla scelta del parametro di penalizzazione. Questa soluzione potrebbe includere un numero illimitato di variabili rumorose nel modello (Meinshausen & Bühlmann, 2006).

L'accuratezza della stima di MLE tramite selezione in avanti è particolarmente scarsa se il numero di nodi nel grafo è comparabile con il numero di osservazioni. Al contrario, la *neighorhood selection* con tramite Lasso, proposta da Meinshausen & Bühlmann (2006) è ragionevolmente accurata per stimare grafi con diverse migliaia di nodi, utilizzando solo alcune centinaia di osservazioni. In questo caso gli archi tra variabili vengono identificati a partire dal vicinato di ciascun nodo, ottenuto applicando un'unica regressione penalizzata.

2.8 Neighborhood selection

Invece di assumere un modello fisso, un approccio più flessibile è assumere che sia il numero di nodi nei grafi (numero di variabili), indicato da $p(n) = |V(n)|$, sia la distribuzione (la matrice di covarianza) dipendano in generale dal numero di osservazioni, in modo che $V = V(n)$ e $\Sigma = \Sigma(n)$. In aggiunta, anche il *neighborhood* dipende in

generale anche da n. In altre parole, regredendo la variabile X_i su tutte le altre variabili si ottiene il vettore dei coefficienti per la previsione ottimale,

$$\hat{\beta}^i = \arg\min_{\beta_i \in \mathbb{R}^{p-1}} \mathbb{E}(X_i - \sum_{j \in V(n)\setminus\{i\}} \beta_{ij} X_t)^2,$$

da cui si ottiene il set di vicini di un nodo $i \in V(n)$ che può essere quindi definito come

$$\text{ne}_i = \{j \in V(n) \setminus \{i\} : \beta_{ij} \neq 0\}.$$

L'introduzione della regolarizzazione Lasso porta ad una contrazione dei coefficienti verso lo zero, risultando in alcuni coefficienti stimati esattamente come zero. Questo processo migliora l'identificazione dei vicini di X_i, con l'uso del parametro di regolarizzazione λ, il quale svolge un ruolo fondamentale. Valori più elevati di λ tendono a ridurre le dimensioni dell'insieme stimato, mentre più variabili sono generalmente incluse in $\hat{ne}_i^{\lambda}$ se il valore di λ viene diminuito.

La stima Lasso $\hat{\beta}_i^{\lambda}$ di β_i è ottenuta da

$$\hat{\beta_i^{\lambda}} = \arg\min_{\beta_i \in \mathbb{R}^{p-1}} (\frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \sum_{j \neq i} \beta_{ij} x_{kj})^2 + \lambda \|\beta_i\|_1)$$

tramite i minimi quadrati, dove $\|\beta_i\|_1 = \sum_{j \in V(n)\setminus\{i\}} |\beta_{ij}|$.

La stima di *neighborhood* è definita dagli elementi non nulli dei coefficienti stimati dalla regressione con penalizzazione L1,

$$\hat{ne}_i^{\lambda} = j \in V(n) \setminus \{i\} : \hat{\beta}_{ij}^{\lambda} \neq 0.$$

Secondo Meinshausen & Bühlmann (2006), l'impiego della convalida incrociata per la scelta del parametro λ non è la soluzione ottimale in quanto può portare all'inclusione di un numero considerevole di variabili spurie nel vicinato, generando così falsi positivi. Come alternativa, si suggerisce l'adozione di un parametro di regolazione variabile specifico per ciascun nodo:

$$\lambda(\alpha) = \frac{2\hat{\sigma}_i}{\sqrt{n}} \widetilde{\Phi}^{-1}(\frac{\alpha}{2p(n)^2})$$

dove $\widetilde{\Phi} = 1 - \Phi$ (Φ è la funzione di ripartizione di una N(0, 1)), $0 < \alpha < 1$, $\hat{\sigma}_i^2 = n^{-1} \mathbf{X}_i^T \mathbf{X}_i$. La probabilità di unire erroneamente due componenti di connettività distinte con la stima dell'insieme di archi è limitata dal livello α sotto la scelta di $\lambda = \lambda(\alpha)$ del parametro di penalizzazione. Il valore ottimale di λ è determinato quando

i dati possono essere considerati generati da una distribuzione Gaussiana multivariata e sotto alcune assunzioni (si veda Meinshausen & Bühlmann (2006)).

Un livello $\alpha = 0.05$ è stato utilizzato per il calcolo di λ.

2.9 Modelli di regressione regolarizzati

Nel campo dell'analisi statistica e del machine learning, i modelli di regressione sono ampiamente utilizzati per comprendere le relazioni tra variabili e fare previsioni su dati futuri. Tuttavia, in presenza di molte variabili predittive o quando queste sono fortemente correlate, i modelli di regressione tradizionali possono incorrere in problemi di instabilità e sovradattamento, portando a previsioni poco affidabili.

Per affrontare queste problematiche, sono stati sviluppati i modelli di regressione regolarizzata, che aggiungono un termine di regolarizzazione alla funzione obiettivo del modello. Questo termine di regolarizzazione impone vincoli sulla complessità del modello, limitando l'importanza delle singole variabili o promuovendo la sparsità dei coefficienti.

Tra i principali metodi di regressione regolarizzata troviamo il Lasso (*Least Absolute Shrinkage and Selection Operator*), il Ridge e l'Elastic Net, tutti ampiamente utilizzati per gestire problemi di multicollinearità e per selezionare un sottoinsieme significativo di variabili predittive. Il Lasso è particolarmente efficace nel promuovere la sparsità dei coefficienti, mentre l'Elastic Net combina la regolarizzazione del Lasso con quella della regressione Ridge, offrendo un compromesso tra distorsione e varianza.

2.9.1 Modello Lasso

Il modello Lasso è un metodo di regressione lineare che combina la selezione delle variabili e la regolarizzazione. Introdotto da Tibshirani (1996), il Lasso è particolarmente utile nei contesti in cui il numero di variabili predittive è alto rispetto al numero di osservazioni, o quando si desidera ottenere un modello più interpretabile eliminando le variabili non rilevanti.

Sia $\mathbf{y} = (y_1, \ldots, y_n)^T$ la risposta e $\mathbf{X} = (x_1 | \ldots | x_p)$ la matrice del modello, dove $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$, sono i predittori. Nello specifico si tratta di risolvere il seguente problema di ottimizzazione:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{k=1}^{n} \left(y_k - \sum_{j=1}^{p} X_{kj} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

dove il primo termine rappresenta la somma dei residui quadrati (*loss function*), il secondo termine la penalizzazione L1.

Come già accennato precedentemente, il parametro di penalizzazione λ gioca un ruolo cruciale nel modello Lasso. Difatti, un suo valore molto grande riduce tutti i coefficienti β a zero; un valore molto piccolo rende il modello simile a una regressione lineare ordinaria, senza penalizzazione. Un valore ottimale di λ bilancia la bontà di adattamento del modello e la sua complessità, permettendo la selezione automatica delle variabili più rilevanti.

Gli algoritmi `glmnet` utilizzano la discesa coordinata ciclica (*Pathwise Coordinate Descent*), che ottimizza successivamente la funzione obiettivo su ciascun parametro, mantenendo fissi gli altri, e ciclando ripetutamente fino alla convergenza.

2.9.2 Modello Elastic Net

Elastic Net è una generalizzazione dei modelli Lasso e Ridge Regression. Introdotto da Zou & Hastie (2005), Elastic Net combina i termini di penalizzazione L1 e L2, risultando in una soluzione che può selezionare variabili e gestire collinearità.

La penalità del modello di regressione *Elastic Net* è controllata da α e colma il divario tra la regressione Lasso ($\alpha = 1$) e la regressione Ridge ($\alpha = 0$). Il parametro di regolazione λ controlla la forza complessiva della penalità.

Come già accennato, la penalità Ridge riduce i coefficienti dei predittori correlati tra loro, mentre il Lasso tende a sceglierne uno e a scartare gli altri. La penalità *Elastic Net*, invece, procede nel seguente modo: se i predittori sono correlati in gruppi, un $\alpha = 0.5$ tende a selezionare o escludere l'intero gruppo di caratteristiche. Questo è un parametro di livello superiore, per questo motivo si può scegliere un valore a priori o sperimentare con diversi valori. L'*Elastic Net* con $\alpha = 1 - \epsilon$ per qualche piccolo $\epsilon > 0$ si comporta in modo molto simile al Lasso, ma elimina qualsiasi degenerazione e comportamento anomalo causato da correlazioni estreme [2].

Sia $\mathbf{y} = (y_1, \ldots, y_n)^T$ la risposta e $\mathbf{X} = (x_1 | \ldots | x_p)$ la matrice del modello, dove $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$, sono i predittori. Dopo una trasformazione di posizione e scala, possiamo assumere che la risposta sia centrata e i predittori siano standardizzati,

$$\sum_{k=1}^{n} y_k = 0, \quad \sum_{k=1}^{n} x_{kj} = 0 \quad \text{e} \quad \sum_{k=1}^{n} x_{kj}^2 = 1, \quad \text{per} \quad j = 1, 2, \ldots, p.$$

[2] https://glmnet.stanford.edu/articles/glmnet.html

Per ogni λ_1 e λ_2 non negativi fissati, il problema di ottimizzazione dell'Elastic Net naïve è:

$$\hat{\beta}^{elasticnet} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{k=1}^{n} \left(y_k - \sum_{j=1}^{p} X_{kj}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

dove λ_1 e λ_2 sono i parametri di penalizzazione per i termini L1 e L2 rispettivamente.

Tuttavia, il naïve Elastic Net non si comporta in modo soddisfacente a meno che non sia molto vicino alla regressione Ridge o al Lasso. Questo è il motivo per cui viene chiamato naïve, la cui traduzione in italiano è "ingenuo". Nel contesto della previsione, un metodo di penalizzazione accurato ottiene buone prestazioni di previsione attraverso il compromesso tra distorsione e varianza. Lo stimatore naïve Elastic Net è una procedura a due fasi: per ogni λ_2 fissato, si trovano prima i coefficienti della regressione Ridge e poi si esegue la riduzione di tipo Lasso lungo i percorsi di soluzione dei coefficienti Lasso. Questo approccio sembra comportare una doppia quantità di riduzione che non aiuta molto a ridurre le varianze e introduce una distorsione extra non necessaria.

Un possibile modo, proposto da Zou & Hastie (2005), per risolvere il problema della naïve Elastic Net è quello di utilizzare

$$\hat{\beta}^{rescaled} = (1 + \lambda(1 - \alpha))\hat{\beta}^{elasticnet}$$

Si noti che questo comporta un ridimensionamento non banale per la Ridge quando α=0, il che probabilmente non ha molto senso. D'altra parte, questo non comporta alcun ridimensionamento per il Lasso quando α=1, nonostante varie affermazioni nella letteratura secondo cui lo stimatore Lasso potrebbe beneficiare di un certo ridimensionamento.

Nel contesto di questo elaborato, le regressioni regolarizzate vengono impiegate come strumento di stima della struttura del grafo, riducendo il problema ad una serie di regressioni in cui, a turno, un nodo svolge il ruolo di risposta e gli altri di predittori. Si noti che quando si utilizza il Lasso, questo approccio consiste nel *neighborhood selection* proposto da Meinshausen & Bühlmann (2006). Qui vengono considerati anche approcci simili, basati su Elastic Net e regressione Ridge (si veda il Capitolo 4).

Capitolo 3

Elaborazione dei dati e approccio CellOracle

Le tecnologie a singola cellula stanno avanzando, consentendo lo sviluppo di metodi per caratterizzare l'identità cellulare attraverso la regolazione genica. Reti di regolazione genica (GRN, *Gene Regulatory Network*), dedotte da dati multiomici a livello di singola cellula, vengono utilizzate per eseguire perturbazioni in silico dei fattori di trascrizione, simulando così i cambiamenti nell'identità cellulare.

Per definizione, una rete di regolazione genica è un insieme di relazioni regolatorie tra fattori di trascrizione (TF) e siti di legame dei TF su specifici mRNA, che governano determinati livelli di espressione degli mRNA e delle proteine risultanti (Yachie-Kinoshita & Kaizu, 2019).

Esistono approcci computazionali per simulare i fenotipi a singola cellula dopo perturbazioni, ma molti richiedono dati sperimentali, limitandone la scala (Ji et al., 2021). Sono necessari approcci scalabili e interpretabili per comprendere la correlazione tra i meccanismi di regolazione genica e i fenotipi complessi osservati. I modelli basati sull'apprendimento profondo rappresentano una "scatola nera", che limita l'interpretazione dei meccanismi di regolazione genica alla base degli eventi biologici simulati (Kamimoto et al., 2023).

Si presenta qui una strategia che supera tali limitazioni, combinando perturbazioni computazionali con la modellazione delle reti geniche. In questo contesto, gli approcci di modellazione delle reti di regolazione genica sono promettenti poiché ricostruiscono associazioni sistematiche gene-gene dai dati omici a singola cellula non perturbati. Si

applica un approccio basato sull'apprendimento automatico, CellOracle[1], dedicato all'analisi delle perturbazioni genetiche in silico, impiegando dati omici a livello di singola cellula e modelli di rete di regolazione genica.

Il metodo si struttura in due passi chiave. Innanzitutto, utilizza i dati di espressione genica a singola cellula per costruire un grafo delle reti di regolazione genica (GRN) di base, identificando le relazioni regolatorie tra i geni in modo da poter comprendere come i fattori di trascrizione influenzino l'espressione genica in contesti cellulari specifici. Successivamente, Celloracle raffina il GRN di base attraverso un processo iterativo che tiene conto della specificità cellulare e delle interazioni predittive tra i geni. Il modello del GRN finale viene quindi ottimizzato, integrando ulteriori informazioni biologiche e migliorando la precisione delle previsioni delle interazioni geniche.

Ci sono diverse opzioni per la costruzione della base GRN di CellOracle:

- Utilizzando dati scATAC-seq o dati bulk ATAC-seq.
- Utilizzando dati provenienti da un database di promotori.
- Partendo da un elenco di geni bersaglio di fattori di trascrizione.

In questo elaborato le GRN di base vengono costruite basandosi su dati scATAC-seq. E' quindi possibile utilizzare i dati scATAC-seq per ottenere la regione di DNA del promotore/potenziatore accessibile. I dati del promotore/potenziatore specifici del campione verranno successivamente convertiti in GRN base.

3.1 Preelaborazione dei dati scATAC-seq

Come primo passo le cellule vengono selezionate in base al numero di picchi, considerando una soglia minima di 2000 e una massima di 13000. I picchi sono regioni del genoma che rappresentano una regione di cromatina aperta.

Poiché i dati di accessibilità alla cromatina a livello cellulare singolo sono estremamente sparsi, una stima accurata dei punteggi di co-accessibilità richiede l'aggregazione di cellule simili per creare dati di conteggio più densi. Il pacchetto *Cicero*[2] fa ciò utilizzando un approccio dei k-nearest-neighbors, creando così insiemi sovrapposti di cellule. Vengono costruiti questi insiemi basandosi sull'embedding UMAP (McInnes et al., 2018), una mappa di coordinate a dimensioni ridotte che rappresenta la similarità delle cellule.

Successivamente, si estraggono i picchi promotori/potenziatori attivi dai dati scATAC-seq. La funzione principale di *Cicero* è utilizzare i dati di accessibilità alla cromatina a

[1] `https://morris-lab.github.io/CellOracle.documentation/`
[2] `https://cole-trapnell-lab.github.io/cicero-release/docs_m3/`

livello di singola cellula per predire interazioni cis-regolatorie (come quelle tra enhancer e promotore) nel genoma, esaminando la co-accessibilità che assume valori tra -1 e 1.

Le regioni genomiche co-accessibili si riferiscono a regioni del genoma che mostrano una tendenza a essere accessibili contemporaneamente in una cellula o in un gruppo di cellule. Più nello specifico, questa co-accessibilità implica che queste regioni geniche condividano uno stato simile di apertura della cromatina in risposta a determinate condizioni o in uno specifico contesto cellulare.

3.2 Annotazione TSS

Prima di costruire la rete genica di base (GRN), è necessario annotare i picchi coaccessibili e filtrare gli elementi attivi del promotore/enhancer. Questo processo è essenziale per definire le regioni in cui cercare motivi conservati. Inizialmente, si identificano i picchi intorno ai siti di inizio della trascrizione (TSS, *Transcription Start Sites*) all'interno del genoma. La sede di inizio della trascrizione è il luogo in cui il primo nucleotide del DNA viene trascritto in RNA. L'annotazione dei TSS è cruciale per comprendere l'inizio della trascrizione genica e identificare le regioni promotori coinvolte nella regolazione genica, ma è più impegnativa rispetto a quella delle regioni codificanti. Ciò è dovuto al fatto che le regioni non tradotte 5' e 3' (UTR) evolvono più rapidamente rispetto alla regione codificante, e meno evidenza esterna è disponibile per supportare le annotazioni (Leung, 2023).

Al termine dell'annotazione, si uniscono i dati di *Cicero* con le informazioni sui picchi TSS e si filtrano tutti i picchi con connessioni deboli ai picchi TSS. Questi elementi identificati costituiscono i componenti attivi del promotore/enhancer per la costruzione della rete genica di base (GRN).

Concludendo, con il processo di integrazione si ottiene quindi un file contenente i picchi TSS o i picchi che hanno una connessione robusta con un picco TSS, il nome del gene associato al sito TSS e il punteggio di co-accessibilità tra il picco e un picco TSS. Il punteggio di co-accessibilità tra questi due picchi fornisce una misura della vicinanza o dell'interazione tra questi due elementi.

3.2.1 Interpretazione punteggi di co-accessibilità

Si fornisce qui una breve interpretazione di questa misura:

- Valori positivi ($0 <$ co-accessibilità ≤ 1): maggiore probabilità di interazioni fisiche tra le regioni genomiche considerate. Valori vicini a uno indicano una forte

co-accessibilità, suggerendo che le regioni sono fisicamente vicine nel nucleo e tendono a interagire attivamente. Questo può indicare la presenza di complessi di regolazione genica o interazioni funzionali. Se il punteggio è esattamente uno, significa che il picco è esso stesso un TSS.

- Valore neutro (co-accessibilità = 0): le regioni genomiche considerate non mostrano interazioni fisiche significative o non ci sono prove di co-accessibilità tra di loro.

- Valori negativi (-1 $\leq$ co-accessibilità $<$ 0): tendenza alla separazione o alla repulsione tra le regioni genomiche considerate. Valori vicini a -1 indicano una forte co-accessibilità negativa, il che può implicare che le regioni siano spazialmente separate nel nucleo e tendono a evitare l'interazione.

3.3 Scansione del motivo di legame del fattore di trascrizione

Nei paragrafi precedenti sono state identificate le regioni accessibili del DNA del promotore/enhancer, utilizzando i dati di scATAC-seq. Di seguito, si costruisce la rete genica di base scandendo le sequenze genomiche regolatorie per i motivi di legame dei fattori di trascrizione. La GRN di base viene generata combinando i picchi scATAC-seq e le informazioni sui motivi. La rete genica di base, ossia la lista delle connessioni potenziali tra i fattori di trascrizione e i geni bersaglio, verrà poi utilizzata nella fase di inferenza della rete genica.

Per proseguire con l'analisi, il file relativo ai dati scATAC-seq deve essere convertito in un formato CSV con tre colonne: indice, picco e nome del gene.

Si può scegliere di utilizzare un riferimento personalizzato per il legame del fattore di trascrizione o i motivi predefiniti di CellOracle durante l'analisi dei motivi. In questo caso si è scelto di proseguire con la seconda opzione, ossia utilizzando il pacchetto Python *Gimmemotifs*[3]. *Gimmemotifs* fornisce molti set di dati di motivi generati da database pubblici, tra cui CisDB, ENCODE, HOMER e JASPAR[4].

Specificatamente, in questa circostanza, verrà utilizzata l'impostazione predefinita **gimme.vertebrate.v5.0**, un database non ridondante e clusterizzato di motivi noti nei vertebrati. Questi motivi provengono da CIS-BP[5] e da altre fonti.

[3] `https://gimmemotifs.readthedocs.io/en/master/`
[4] `https://gimmemotifs.readthedocs.io/en/master/overview.html`
[5] `https://cisbp.ccbr.utoronto.ca/`

3.3.1 Creazione dell'oggetto TFinfo

Il modulo di analisi dei motivi ha una classe personalizzata, TFinfo. In questa fase dell'analisi vengono eseguiti i seguenti passaggi:

1. I dati di picco vengono convertiti in sequenze di DNA.

2. Le sequenze di DNA vengono esaminate alla ricerca di motivi di legame TF, ottenendo così il punteggio di legame per ogni coppia di picco e motivo. Punteggi di legame bassi indicano che la predizione o la misura dell'interazione tra il fattore di trascrizione e la sequenza del DNA è debole. In altre parole, ciò potrebbe indicare una minore probabilità o una minore forza legame del fattore di trascrizione a quella particolare regione genomica.

3. I risultati della scansione del motivo vengono postelaborati. Si filtrano i motivi con punteggi di legame bassi, impostando un threshold pari a 10.

4. Si convertono i dati in rete di regolazione genica di base. L'output risultante sarà la base della GRN finale utilizzata nella fase di costruzione del modello e presenta i picchi sulle righe, una colonna contenente i nomi dei geni e le colonne rimanenti relative ai fattori di trascrizione.

3.4 Preparazione dei dati scRNA-seq

Prima di avviare un'analisi con CellOracle, non basta avere a disposizione i dati scATAQ-seq ma è necessario preelaborare anche i dati di scRNA-seq. Si preparano i dati di scRNA-seq come un oggetto AnnData, utilizzando Scanpy[6]. Scanpy è un toolkit Python per l'analisi di dati scRNA-seq.

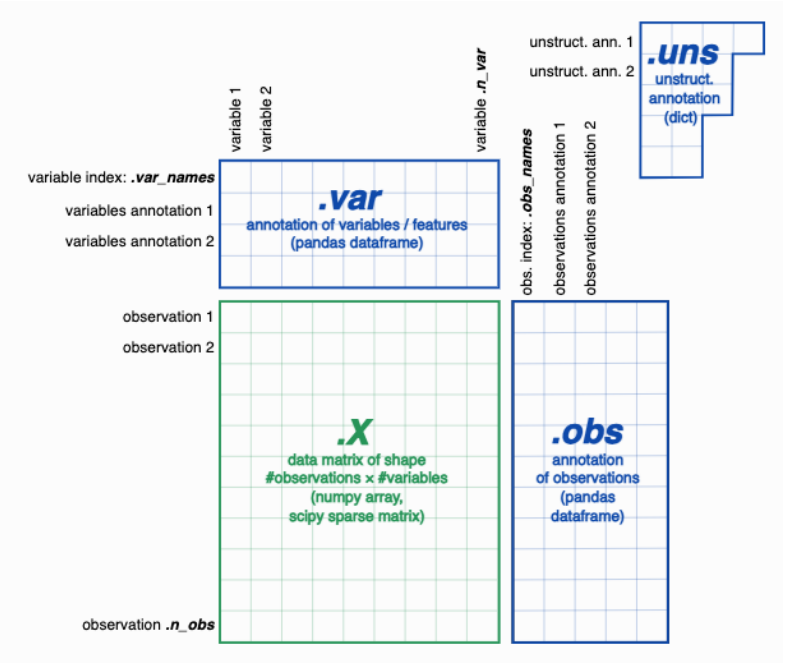

FIGURA 3.1: Struttura di un oggetto **AnnData**

Un oggetto AnnData memorizza una matrice di dati, le annotazioni delle osservazioni e delle variabili come un DataFrame, e le annotazioni non strutturate come un dizionario. In sintesi, è una raccolta di contenitori dati più semplici: array, matrici sparse, dataframe.

Per preprocessare i dati di scRNA-seq corretti per gli effetti di batch, si considera il clustering di cellule fornito e si prosegue con la riduzione della dimensionalità (si veda il Capitolo 1). Il passaggio di riduzione della dimensionalità richiede una attenta considerazione durante la preparazione dei dati per un'analisi con CellOracle. Per un'analisi di successo, la rappresentazione dovrebbe riflettere la transizione cellulare di interesse.

Si procede quindi con una combinazione di tre algoritmi: diffusion map (Coifman & Lafon, 2005), grafo diretto forzato (Jacomy et al., 2014) e PAGA[7] (Partition-based Graph Abstraction). Per eliminare il rumore dal grafo, lo si rappresenta nello spazio

[6] `https://scanpy.readthedocs.io/en/stable/`
[7] `https://github.com/theislab/paga`

della diffusion map (e non nello spazio delle componenti principali). Calcolare le distanze entro alcune componenti di diffusione equivale a ripulire il grafo dal rumore, prendendo semplicemente alcune delle prime componenti spettrali. È molto simile a pre-processare una matrice di dati utilizzando la PCA. Questo passaggio non è obbligatorio per PAGA; tuttavia, il suo utilizzo può migliorare la qualità dei risultati. Successivamente, si calcola il grafo PAGA. I dati PAGA determinano le posizioni iniziali dei cluster per il calcolo del grafo diretto forzato.

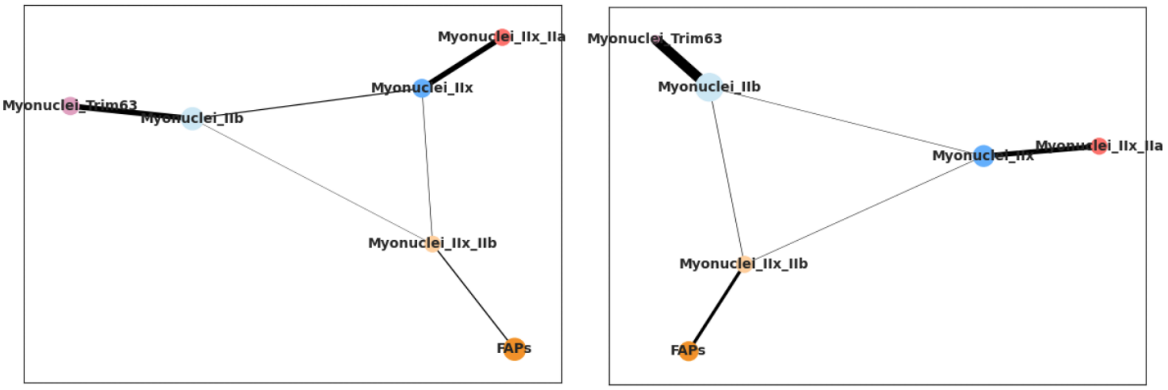

FIGURA 3.2: A sinistra: grafico PAGA per i topi C26. A destra: grafico PAGA per i topi CTRL.

Un grafico PAGA si ottiene associando un nodo a ciascun cluster cellulare e collegando ciascun nodo tramite archi ponderati che rappresentano una misura statistica della connettività tra i cluster. Le connessioni più marcate possono rappresentare relazioni più forti o significative tra i gruppi cellulari. La forza di una connessione può essere valutata in base a vari parametri, tra cui la probabilità di transizione, la similarità genetica o altre misure di relazione tra i cluster.

Dopodichè, si calcola il grafo diretto forzato delle cellule. Rispetto a alcuni metodi di riduzione dimensionale, i grafi diretti dalle forze possono catturare strutture più dettagliate di ramificazione. Tuttavia, potrebbero risultare instabili in presenza di dati con molte ramificazioni sovrapposte. Per evitare questa sovrapposizione, le informazioni sulla traiettoria cellulare PAGA possono essere utilizzate per inizializzare il calcolo del grafo diretto dalle forze. Questo tipo di grafi è preferito in quanto generalmente produce una struttura di linea ad alta risoluzione.

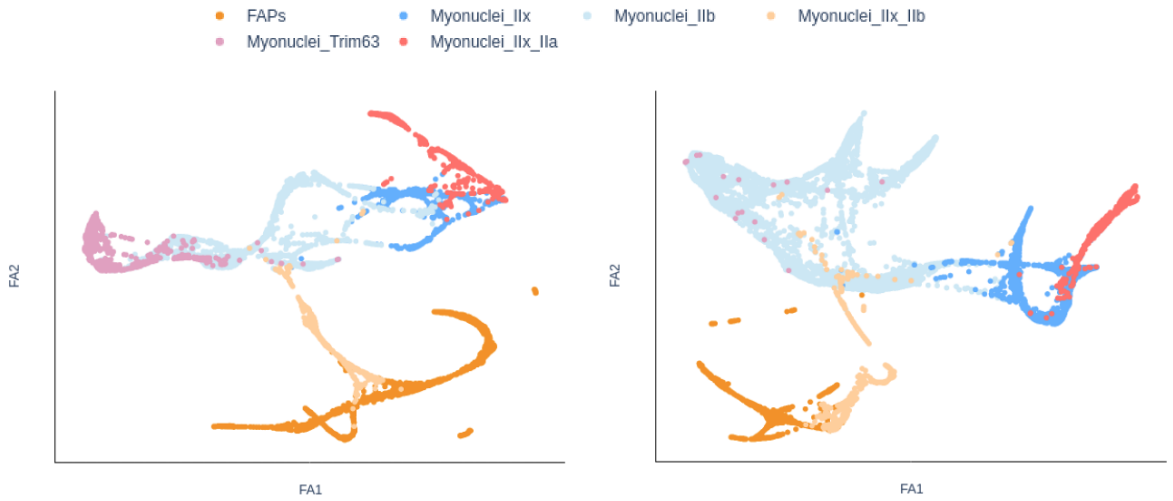

FIGURA 3.3: A sinistra: grafo diretto forzato per i topi C26. A destra: grafo diretto forzato per i topi CTRL.

Le coordinate FA1 e FA2 non sono direttamente legate a una riduzione della dimensionalità come in algoritmi specifici come UMAP (McInnes et al., 2018) o t-SNE (van der Maaten & Hinton, 2008), ma svolgono un ruolo simile nella disposizione spaziale dei nodi in un grafo diretto dalle forze. Possono essere considerate come le coordinate risultanti dalla simulazione fisica del sistema. La loro interpretazione è legata alla disposizione relativa dei nodi nella visualizzazione grafica del grafo.

3.5 Costruzione del modello GRN

Fino a questo punto dell'analisi non era ancora stato utilizzato CellOracle. Introduciamo quindi ora questo approccio.

CellOracle utilizza due tipi di dati di input durante la costruzione del modello di rete di regolazione genica (GRN):

- Dati scRNA-seq grezzi.
- Rete di regolazione genica di base. La GRN di base rappresenta le connessioni tra i fattori di trascrizione (TF) e i geni target. La struttura dei dati è una matrice binaria. La GRN, come già detto precedentemente, viene costruita partendo dai dati scATAC-seq.

CellOracle costruisce modelli GRN suddivisi per cluster. È possibile confrontare la struttura dei modelli GRN tra i cluster, consentendo di indagare sulle configurazioni GRN specifiche per il tipo di cellula ed esplorare i cambiamenti strutturali mentre le

GRN vengono riconfigurate lungo la traiettoria di differenziazione cellulare. Questo approccio utilizza due classi personalizzate: Oracle e Links.

Oracle è responsabile per quasi tutta la costruzione del modello GRN e la simulazione delle perturbazioni dei fattori di trascrizione (TF). Essa esegue i seguenti passaggi in sequenza: importa ed elabora i dati grezzi della scRNA-seq, importa i dati di base della GRN e successivamente costruisce il modello della rete di regolazione genica.

La classe Links è utilizzata per memorizzare i dati della GRN finale. Contiene anche numerose funzioni per l'analisi e la visualizzazione della rete.

Per quanto riguarda i dati grezzi relativi all'scRNA-seq, si considerano solo i geni con più di un conteggio e si normalizza la matrice di espressione genica con il conteggio totale di UMI per cellula. La normalizzazione è un passaggio critico che corregge le differenze tecniche tra cellule. In seguito, si rimuovono i geni poco variabili, riducendo così il tempo di calcolo durante le fasi di ricostruzione e simulazione della GRN. Questo passaggio, inoltre, migliora l'accuratezza complessiva dell'inferenza GRN rimuovendo i geni rumorosi. Si decide di utilizzare i primi 2000 geni più variabili.

3.5.1 Creazione di un oggetto Oracle

Si utilizza un oggetto Oracle durante le fasi di preelaborazione dei dati e di inferenza GRN. Per questo motivo, per iniziare, si crea un'istanza di un nuovo oggetto Oracle in cui si inseriscono i dati di espressione genetica (in formato AnnData) e le informazioni sui fattori di trascrizione (base GRN).

Si ricorda che per l'analisi CellOracle, i dati relativi a scRNA-seq devono includere i conteggi grezzi di espressione genica, le informazioni di clustering e i dati di traiettoria corretti per gli effetti di batch.

CellOracle utilizza la stessa strategia di *velocyto*[8] per visualizzare le transizioni cellulari. Questo processo richiede preventivamente l'imputazione KNN (k-nearest neighbors), per la quale dobbiamo prima calcolare e selezionare le componenti principali.

3.5.2 Calcolo della rete di regolazione genica

Si crea un GRN specifico per tutti i cluster cellulari. Utilizzando la GRN di base, CellOracle costruisce i modelli GRN come una lista di archi diretti tra un fattore di trascrizione e i suoi geni bersaglio. È necessario rimuovere gli archi deboli o non significativi in base al *p-value*, prima di effettuare l'analisi della struttura di rete.

[8] `https://velocyto.org/`

CellOracle utilizza la tecnica di bagging ridge o la regressione Bayesian ridge per dedurre le reti di regolazione genica. In questo caso si decide di proseguire con la regressione bagging ridge in quanto tende a produrre risultati di inferenza migliori (Kamimoto et al., 2023). Si costruisce un modello che predice l'espressione di un gene bersaglio sulla base dell'espressione di geni candidati regolatori:

$$x_j = \sum_{i=0}^{n} b_{i,j} x_i + c_j \tag{3.1}$$

dove x_j è l'espressione di un singolo gene bersaglio e $x_i, i = 0, \ldots, n, i \neq j$, è il valore dell'espressione genica del gene candidato regolatore (fattore di trascrizione) che regola il gene x_j. $b_{i,j}$ è il valore del coefficiente del modello lineare e c_j è l'intercetta per questo modello. L'elenco dei potenziali geni regolatori per ogni gene target è quello generato dalla costruzione della rete di regolazione genica di base.

Il calcolo di regressione viene eseguito per ogni cluster cellulare in parallelo dopo che la matrice di espressione genica dei dati scRNA-seq è stata divisa in diversi cluster. Il modello di regressione specifico per ogni cluster può catturare relazioni regolatorie non lineari o miste. Inoltre, viene applicata una regolarizzazione L2 tramite il modello Ridge. La regolarizzazione aiuta non solo a distinguere connessioni regolatorie attive da connessioni casuali, inattive o false nella GRN di base, ma riduce anche il sovradattamento in campioni più piccoli.

Il modello di Bagging Ridge fornisce in aggiunta al valore dei coefficienti la loro distribuzione per determinare la significatività della connessione gene-gene dedotta. Si considera il seguente modello:

$$x_j | b \sim \mathcal{N}(\sum_{i=0}^{n} b_{i,j} x_i + c_j, \epsilon), \quad b \sim \mathcal{N}(\mu_b, \sigma_b) \tag{3.2}$$

dove μ_b è il centro della distribuzione di b e σ_b è la sua deviazione standard.

Utilizzando la distribuzione a posteriori, possiamo stimare i coefficienti b. Si utilizza poi un bootstrap aggregating (bagging) per stimare la variabilità delle stime Ridge di b, utilizzando un t-test per testare l'ipotesi nulla che $b = 0$. Il *p-value* aiuta a identificare connessioni robuste minimizzando le connessioni derivanti da rumore casuale. La regolarizzazione viene applicata al coefficiente b per evitare che il coefficiente b diventi estremamente grande a causa del sovradattamento e per identificare variabili informative. Utilizzando Bagging Ridge come modello, si imposta manualmente la forza di regolarizzazione (λ) a dieci. Se si stabilisce un valore più basso, aumenta la sensibilità e è possibile rilevare connessioni di rete più deboli. Tuttavia, potrebbe esserci più rumore.

Se, invece, si seleziona un valore più alto, si riduce la possibilità di sovradattamento.

Successivamente, si calcolano diversi punteggi di rete: centralità di grado, centralità di betweenness e centralità dell'autovettore.

Ci si concentra maggiormente sulla centralità dell'autovalore, una misura di centralità che assegna importanza a un nodo in base alla sua connettività con altri nodi importanti nella rete (si veda Capitolo 2). I nodi con centralità dell'autovettore elevata sono quelli che sono collegati ad altri nodi importanti. Se un fattore di trascrizione non ha un collegamento nella rete in un cluster, i punteggi di rete non possono essere calcolati e non saranno mostrati.

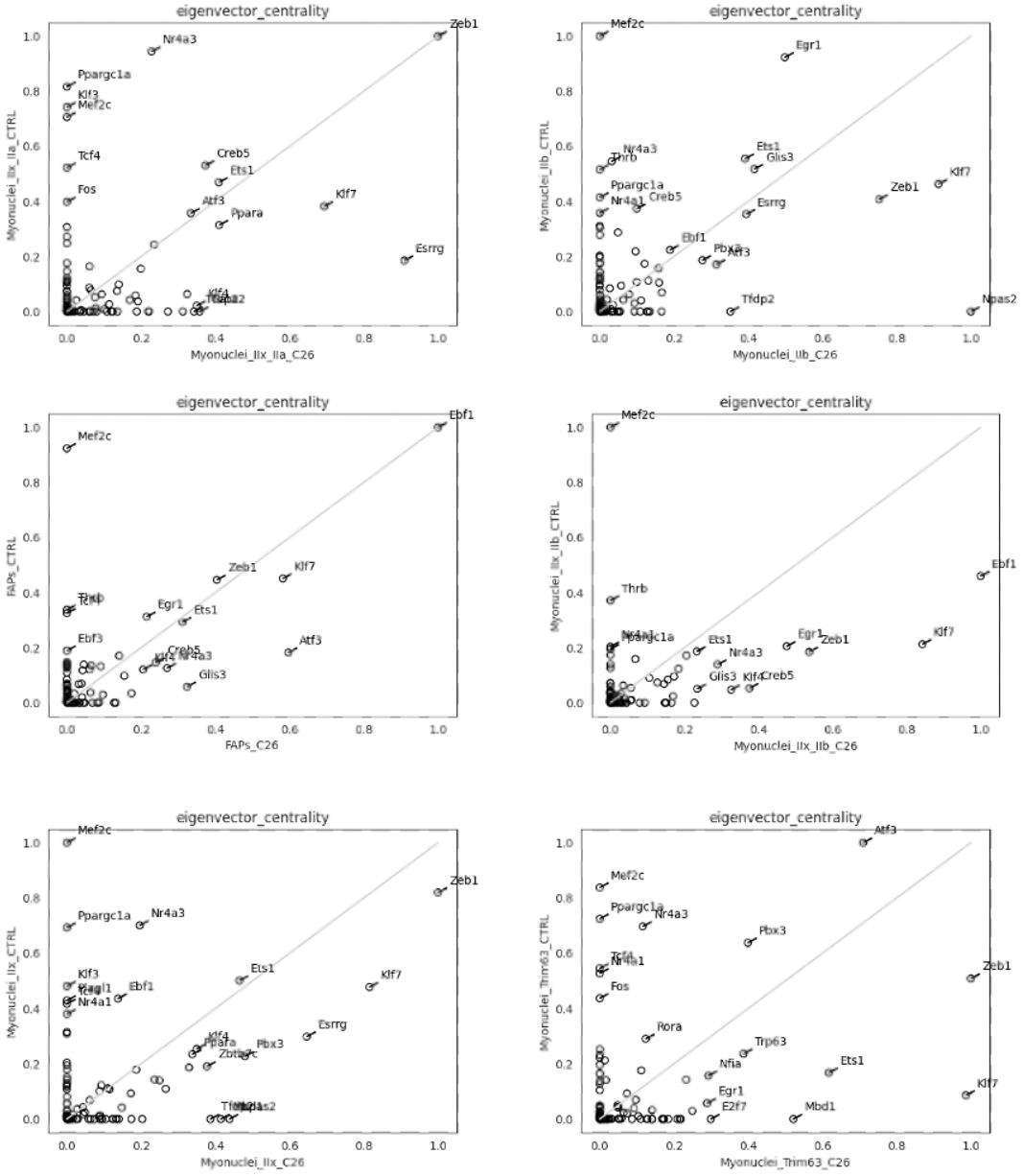

FIGURA 3.4: Confronto della misura di centralità dell'autovettore tra i cluster dei topi affetti da cachessia e di quelli sani.

Confrontando questi punteggi della rete dei topi C26 e CTRL relativa ad ogni singolo cluster cellulare, possiamo analizzare le disparità nella struttura della rete di regolazione genica. Il cluster cellulare FAPs sembra presentare misure di centralità simili tra topi sani e malati, rispetto agli altri cluster.

Per tutti i cluster cellulari il fattore di trascrizione Mef2c emerge come nodo centrale nelle reti dei controlli. Presente in abbondanza nel tessuto muscolare scheletrico, Mef2c gioca un ruolo cruciale nel dirigere il differenziamento terminale dei precursori miogenici (mioblasti) (Ferrari et al., 2013). Questo implica che Mef2c è coinvolto in molteplici interazioni con altri geni e proteine, influenzando direttamente o indirettamente una serie di processi biologici critici per lo sviluppo e la funzione muscolare.

3.6 Analisi delle perturbazioni in silico dei TF

Dopo aver costruito i modelli GRN di CellOracle, è possibile effettuare l'analisi delle perturbazioni in silico dei fattori di trascrizione (TF). CellOracle utilizza i modelli GRN per simulare cambiamenti nell'identità cellulare in risposta a perturbazioni dei fattori di trascrizione. Le Figure relative a questa analisi sono riportate in Appendice, in quanto la loro interpretazione non risulta di particolare interesse statistico. Sono presenti solo quattro grafici in quanto vengono considerate esclusivamente le perturbazioni dei geni che differiscono maggiormente tra controlli e cachetici, basandosi sulla Figura 3.4.

Come discusso in questo Capitolo, il secondo passaggio del metodo CellOracle implica una regressione Ridge simile a *neighborhood selection*, che usa una penalità L2 invece che L1. Si proverà quindi a sostituire questo passaggio con Lasso e Elastic Net.

Capitolo 4

Applicazione dei modelli ai dati multi-omici

Nel presente Capitolo si approfondisce l'implementazione delle tecniche di regressione Lasso e Elastic Net applicate ai dati multi-omici. Si ricorda che la regressione Lasso e l'Elastic Net sono metodi di penalizzazione che si utilizzano per migliorare la predizione e la selezione delle variabili in modelli di regressione, particolarmente efficaci quando ci si trova di fronte a dati ad alta dimensionalità e a collinearità tra le variabili.

L'analisi si focalizza sull'applicazione di queste tecniche a due tipologie di dati: l'espressione genica e l'integrazione di quest'ultima con l'informazione derivata dalla rete di regolazione genica (GRN). L'espressione genica rappresenta il livello di attività di vari geni in condizioni specifiche ed è misurata attraverso scRNA-Seq. Questi dati, se analizzati singolarmente, possono fornire informazioni cruciali sui processi biologici sottostanti.

Tuttavia, per ottenere una comprensione più completa e dettagliata dei meccanismi biologici, è spesso utile integrare i dati di espressione genica con altre fonti di informazione, come le reti di regolazione genica. Le GRN rappresentano le interazioni tra geni, delineando come l'espressione di un gene possa influenzare l'espressione di altri geni. Questa integrazione consente di sfruttare la conoscenza delle interazioni regolatorie per migliorare la predittività e l'interpretabilità dei modelli di regressione, fornendo nuove intuizioni sui meccanismi molecolari e sui potenziali bersagli terapeutici.

E' di rilevante importanza specificare che tutto ciò è stato eseguito separatamente per il campione dei topi C26 e per quello dei topi controlli. Inoltre, visto l'elevato numero di informazioni a disposizione, si è deciso di estrarre dalla matrice delle espressioni geniche una specifica popolazione cellulare, chiamata *Myonuclei IIb*.

I Myonuclei sono nuclei cellulari presenti nelle fibre muscolari scheletriche. Le cellule muscolari scheletriche sono uniche in quanto sono cellule estremamente grandi che contengono migliaia di nuclei. Questi nuclei devono lavorare insieme per mantenere la funzione del muscolo scheletrico. Il nucleo è uno dei siti principali di integrazione dei segnali e di regolazione dell'espressione genica. Ad ogni modo, esaminare i processi nucleari nel muscolo scheletrico può essere difficile perché i Myonuclei sono difficili da isolare (Cutler et al., 2017).

All'interno del tessuto muscolare scheletrico, esistono diverse tipologie di fibre muscolari, ciascuna con caratteristiche e funzioni specifiche. Una di queste tipologie è rappresentata dalle fibre muscolari di tipo IIb. Queste specifiche fibre muscolari, anche conosciute come fibre a contrazione rapida e glicolitiche, sono specializzate nei movimenti esplosivi e di breve durata. Esse sono caratterizzate da un alto contenuto di enzimi glicolitici, che consentono loro di produrre rapidamente energia tramite la glicolisi anaerobica.

4.1 Preparazione dei dati

I dati relativi all'espressione genica vengono incorporati in un oggetto SingleCellExperiment, che è una struttura dati flessibile progettata per contenere e manipolare dati di singole cellule. In seguito, viene eseguito un clustering delle singole cellule basato sui profili di espressione genica di quest'ultime. Vengono inoltre calcolati i fattori di normalizzazione dei fattori di dimensione per tener conto delle differenze nella profondità di copertura e nella varianza tra le singole cellule. Per stabilizzare la varianza e ridurre la distorsione nei dati viene eseguita una trasformazione log(x+1) dei dati di conteggio delle singole cellule (Risso et al., 2014).

Successivamente, ove necessario, la matrice di espressione genica viene integreata con la rete di espressione genica.

I dati da qui ottenuti, assumendo che si distribuiscano in modo Gaussiano, sono pronti per essere modellati. Come già detto in precedenza verranno implementati il modello di regressione penalizzato Lasso e Elastic Net nei seguenti modi.

Considerando solo la matrice di espressione genica:

- **Lasso 1**, **Elastic Net 1**: sia per la variabile risposta sia per le covariate si selezionano tutti i geni.
- **Lasso 2**, **Elastic Net 2**: per la variabile risposta si considerano tutti i geni, mentre per le covariate solo i geni che sono anche fattori di trascrizione.

Matrice di espressione genica + GRN:

- **Lasso 3**, **Elastic Net 3**: per la variabile risposta si considerano tutti i geni, mentre per le covariate solo i geni che sono anche fattori di trascrizione.
- **Lasso 4**, **Elastic Net 4**: sia per la variabile risposta sia per le covariate si selezionano i geni che sono anche fattori di trascrizione.

4.2 Selezione del λ ottimale

A differenza dell'approccio CellOracle, il quale considera un unico parametro di regolazione per modellare (pari a dieci), si è scelto di utilizzare un λ specifico per ciascun nodo, come suggerito da Meinshausen & Bühlmann (2006):

$$\lambda_i(\alpha) = \frac{2\hat{\sigma}_i}{\sqrt{n}} \widetilde{\Phi}^{-1}(\frac{\alpha}{2p^2})$$

dove $\hat{\sigma}_i$ è la stima della deviazione standard di y_i, n è il numero di righe della matrice e p è il numero di variabili di ogni singolo modello.

Questo metodo permette di determinare un valore di λ che riflette l'importanza e la struttura relativa a ciascuna variabile, tenendo conto delle loro specifiche proprietà.

I vantaggi di questo procedimento sono:

- Adattabilità: Ogni modello riceve un valore di λ che tiene conto della specifica varianza e distribuzione della variabile risposta, migliorando potenzialmente l'accuratezza della selezione delle variabili.
- Flessibilità: L'approccio è flessibile e può essere facilmente adattato a diverse configurazioni di dati e livelli di significatività.

Mentre gli svantaggi sono:

- Complessità Computazionale: Il calcolo di un valore di λ ottimale per ciascun modello aumenta la complessità computazionale rispetto all'uso di un singolo valore di λ per tutte le variabili.
- Interpretabilità: I risultati possono essere meno interpretabili, poiché ogni modello ha un proprio λ.

L'approccio adottato ci permette di costruire una rete basata sulle proprietà individuali di ciascuna variabile, con un'attenzione particolare alla loro varianza e distribuzione. Questo ha migliorato la precisione dell'analisi, permettendo di identificare relazioni interessanti tra le variabili nel dataset studiato.

4.3 Matrice di espressione genica

Di seguito vengono riportati i risultati della modellazione in termini di misura di centralità dell'autovettore, considerando come dati di partenza solo la matrice di espressione genica. Vengono rappresentati solo i nomi dei fattori di trascrizione che sono risultati di particolare interesse osservando la centralità degli autovettori dei nodi relativi alla rete ottenuta con l'approccio CellOracle.

Il modello Lasso impone un vincolo di sparsità, riducendo molti coefficienti a zero per evitare il sovradattamento. Come si può notare in Figura 4.1 questo si riflette nella rete, dove molte connessioni tra nodi possono essere ridotte o eliminate, portando a valori di centralità degli autovettori più bassi.

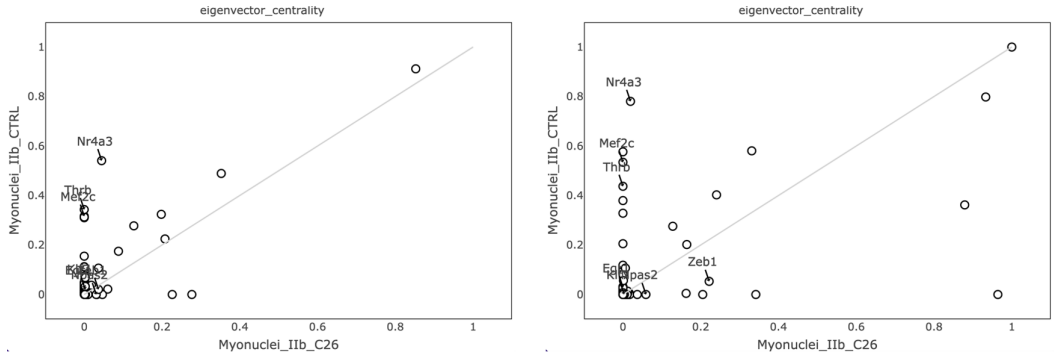

FIGURA 4.1: Confronto della misura di centralità dell'autovettore tra cachetici e sani, relativo al gruppo cellulare Myonuclei IIb. A sinistra: Lasso 1. A destra: Lasso 2.

In maniera contraria, i valori dell'Elastic Net in Figura 4.2 sono meno sparsi, poichè questo modello può trovare un equilibrio tra riduzione dei coefficienti e selezione delle caratteristiche.

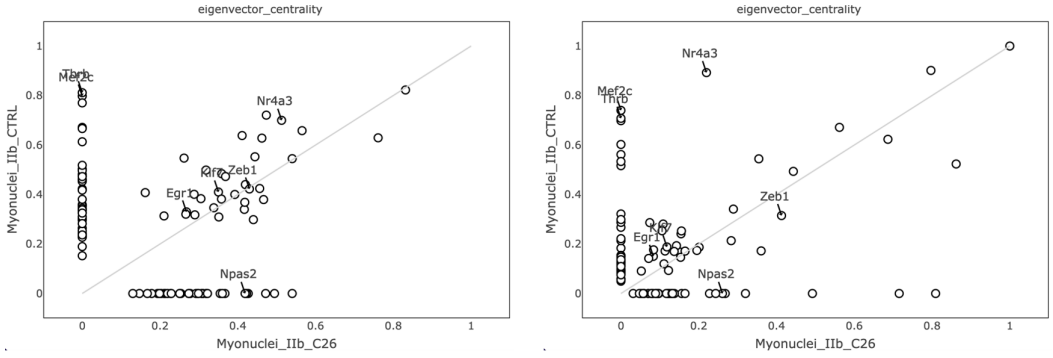

FIGURA 4.2: Confronto della misura di centralità dell'autovettore tra cachetici e sani, relativo al gruppo cellulare Myonuclei IIb. A sinistra: Elastic Net 1. A destra: Elastic Net 2.

4.4 Matrice di espressione genica + GRN

In Figura 4.3 e 4.4 vengono riportati i risultati della modellazione in termini di misura di centralità dell'autovettore, considerando come dati di partenza la matrice di espressione genica integrata con le informazioni relative alla rete di espressione. Come per il paragrafo precedente, vengono rappresentati solo i nomi dei fattori di trascrizione che sono risultati di particolare interesse osservando la centralità degli autovettori dei nodi relativi alla rete ottenuta con l'approccio CellOracle.

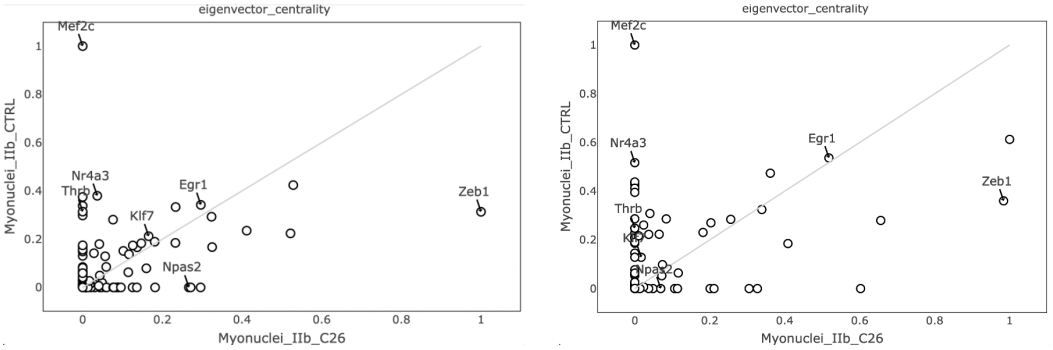

FIGURA 4.3: Confronto della misura di centralità dell'autovettore tra cachetici e sani, relativo al gruppo cellulare Myonuclei IIb. A sinistra: Lasso 3. A destra: Lasso 4.

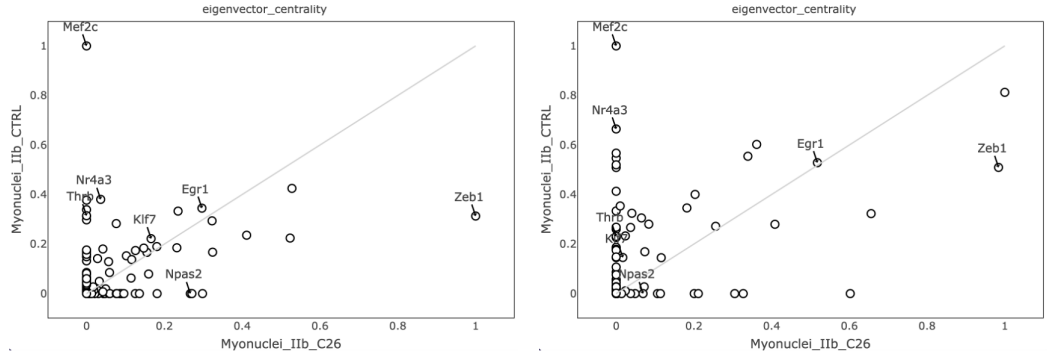

FIGURA 4.4: Confronto della misura di centralità dell'autovettore tra cachetici e sani, relativo al gruppo cellulare Myonuclei IIb. A sinistra: Elastic Net 3. A destra: Elastic Net 4.

In Figura 4.5, si confrontano le misure di centralità dell'autovettore tra individui cachetici e sani, secondo l'approccio CellOracle. Questa rappresentazione differisce da quella presentata nel Capitolo 3 (Figura 3.4) poiché, al fine di garantire una comparazione più accurata, la centralità è stata ricalcolata senza considerare il peso dei coefficienti. Ciò è dovuto al fatto che, per i modelli Lasso ed Elastic Net, la centralità calcolata non è ponderata.

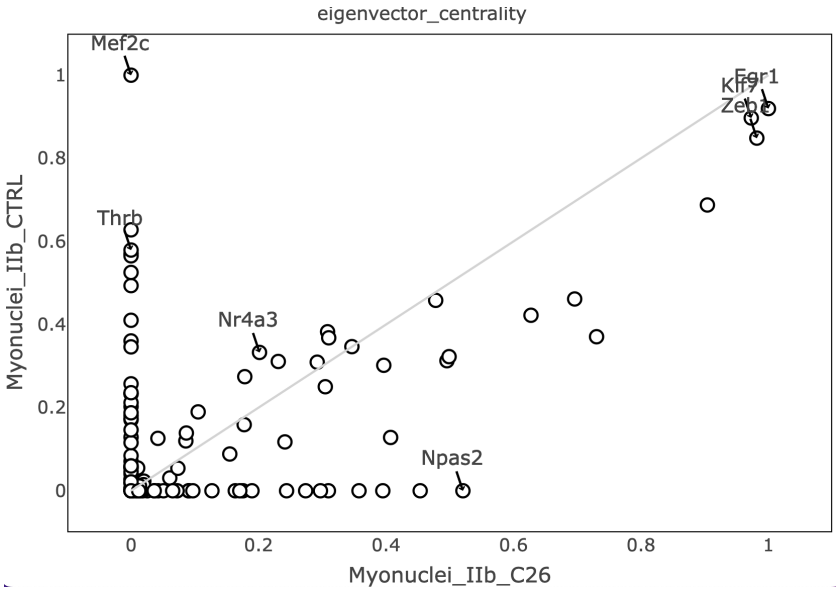

FIGURA 4.5: Confronto della misura di centralità dell'autovettore tra cachetici e sani, relativo al gruppo cellulare Myonuclei IIb. Approccio CellOracle

Le Tabelle relative alle misure di centralità di CellOracle, Lasso 3 e Elastic Net 3 sono presenti in Appendice. Sono stati selezionati solo questi tre approcci in quanto tutti e tre considerano gli stessi dati di partenza. Questa scelta ci consente di garantire che i tre metodi principali abbiano approssimativamente lo stesso numero di connessioni, consentendoci di valutarli in modo più accurato.

Per quanto riguarda i topi C26, geni come Zeb1, Egr1, Klf7 mostrano valori di centralità elevati in almeno una delle reti, suggerendo un ruolo potenzialmente cruciale in queste reti di regolazione genica. Risulta quindi di particolare interesse approfondire il loro ruolo:

- **Zeb1**: noto per il suo ruolo nella transizione dell'epitelio-mesenchimale (EMT), un processo fondamentale per lo sviluppo embrionale, per il mantenimento dell'omeostasi dei tessuti, per la fibrosi tissutale e per il cancro [1]. Zeb1 è implicato nella metastatizzazione del cancro al colon attraverso la promozione dell'EMT (Gregory et al., 2008).

- **Egr1**: fattore di trascrizione coinvolto principalmente nei processi di lesione tissutale, risposte immunitarie e fibrosi. Studi recenti hanno mostrato che Egr1 è strettamente correlato all'inizio e alla progressione del cancro e potrebbe partecipare alla proliferazione, invasione e metastasi delle cellule tumorali, nonché all'angiogenesi tumorale (Wang et al., 2021).

[1] `https://irinsubria.uninsubria.it/retrieve/handle/11383/2090460/107510/PhD_thesis_rizzosamantha_completa.pdf`

- **Klf7**: esperimenti funzionali hanno dimostrato che il silenziamento di questo fattore di trascrizione ha portato a una riduzione della proliferazione, migrazione e invasione cellulare, indicando il suo coinvolgimento nella promozione della crescita tumorale e della metastasi (Li & Liu, 2023).

I risultati delle reti sono più simili tra Lasso 3 ed Elastic Net 3 rispetto a CellOracle. In Tabella A.1 si può notare che alcuni fattori di trascrizione presenti nella rete CellOracle non sono presenti nelle altre due reti.

Relativamente ai topi CTRL, un gene emerge particolarmente centrale in tutte e tre le reti analizzate. Si parla del fattore di trascrizione **Mef2c** che regola la differenziazione e la crescita del cuore e del muscolo scheletrico (Piasecka et al., 2021). Anche **Zeb1** e **Egr1** mostrano valori rilevanti di centralità. Come detto precedentemente, i risultati delle reti sono più simili tra Lasso 3 ed Elastic Net 3 rispetto a CellOracle. In Tabella A.2 si può notare che alcuni fattori di trascrizione presenti nella rete CellOracle non sono presenti nelle altre due reti.

4.5 Confronto tra i modelli

Ci si concentra ora sul confronto delle singole reti generate dai differenti modelli.

L'obiettivo è individuare le caratteristiche uniche di ciascun modello e comprendere come questi influenzino la struttura delle reti che producono. In sostanza, cerchiamo di rispondere alla domanda: "Che cosa rende ogni modello unico e come questo si riflette nelle reti che genera?"

In Tabella 4.1 si sintetizzano le caratteristiche salienti delle reti ottenute attraverso le varie metodologie. Le dimensioni, la densità e il numero di connessioni delineano il profilo strutturale di ciascuna rete, offrendo una base per il confronto. Per quanto riguarda l'approccio CellOracle, sono state considerate due reti. La prima, esaminata anche nel Capitolo 3, è stata ottenuta mantenendo solo 2000 connessioni con un *p-value* ≤ 0.001. Questo approccio ha permesso di concentrare l'analisi sulle interazioni più robuste e statisticamente significative. Successivamente, è stata stimata una seconda rete utilizzando solo il filtro basato sui *p-value*.

Metodo	**Campione**	**Densità della rete**	**Dimensione matrice rete**	**N° connessioni**
Celloracle 2000	C26	0.001	1275*1275	2000
	CTRL	0.001	1383*1383	2000
Celloracle filtro pvalue	C26	0.009	1275*1275	14398
	CTRL	0.01	1383*1383	18887
Lasso 1	C26	0.006	1504*1504	13949
	CTRL	0.009	1595*1595	22738
Lasso 2	C26	0.001	1504*1504	1252
	CTRL	0.001	1595*1595	2158
Lasso 3	C26	0.001	1135*1135	1018
	CTRL	0.001	1266*1266	2125
Lasso 4	C26	0.017	61*61	61
	CTRL	0.033	69*69	153
Elastic Net 1	C26	0.03	1504*1504	74464
	CTRL	0.04	1595*1595	103049
Elastic Net 2	C26	0.003	1504*1504	5980
	CTRL	0.004	1595*1595	8993
Elastic Net 3	C26	0.001	1135*1135	1018
	CTRL	0.001	1266*1266	2130
Elastic Net 4	C26	0.02	61*61	61
	CTRL	0.05	69*69	242

TABELLA 4.1: Confronto tra le reti ottenute dai differenti modelli (approcci), considerando anche i differenti dati di partenza

Talvolta, per cogliere appieno la complessità di tali reti, ci si focalizza anche su strumenti più visivi e intuitivi, i *CATplot* (Irizarry et al., 2005). Quest'ultimo è un tipo di grafico utilizzato per visualizzare come varia il valore di concordanza tra due insiemi di dati man mano che si considera un numero crescente di osservazioni. Tipicamente viene rappresentato come un grafico a linee, sull'asse x sono presenti i rank delle osservazioni considerate, mentre sull'asse y viene rappresentata la misura di concordanza, che può variare da 0 a 1. Questo tipo di grafico è utile per esaminare la consistenza o la relazione tra due serie di dati mentre il numero di osservazioni tenute in considerazione aumenta.

Il confronto basato sulla concordanza dei modelli, avviene attraverso i gradi IN (presenti in Appendice), i gradi OUT e la betweenness. Molte volte la linea relativa al Lasso 3 non si vede in quanto è posizionata sotto alla linea di Elastic Net 3, e lo stesso succede per Lasso 4 e Elastic Net 4.

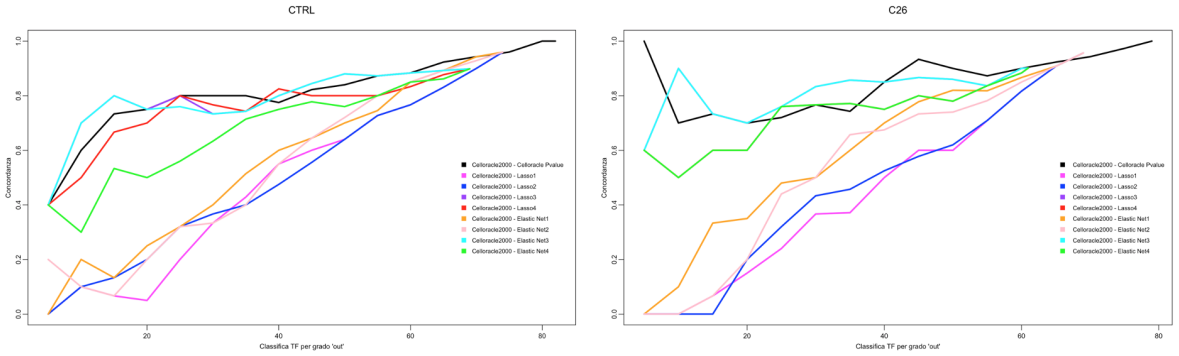

FIGURA 4.6: Confronto dei gradi OUT dei differenti modelli (approcci). A sinistra: topi controllo. A destra: topi cachetici

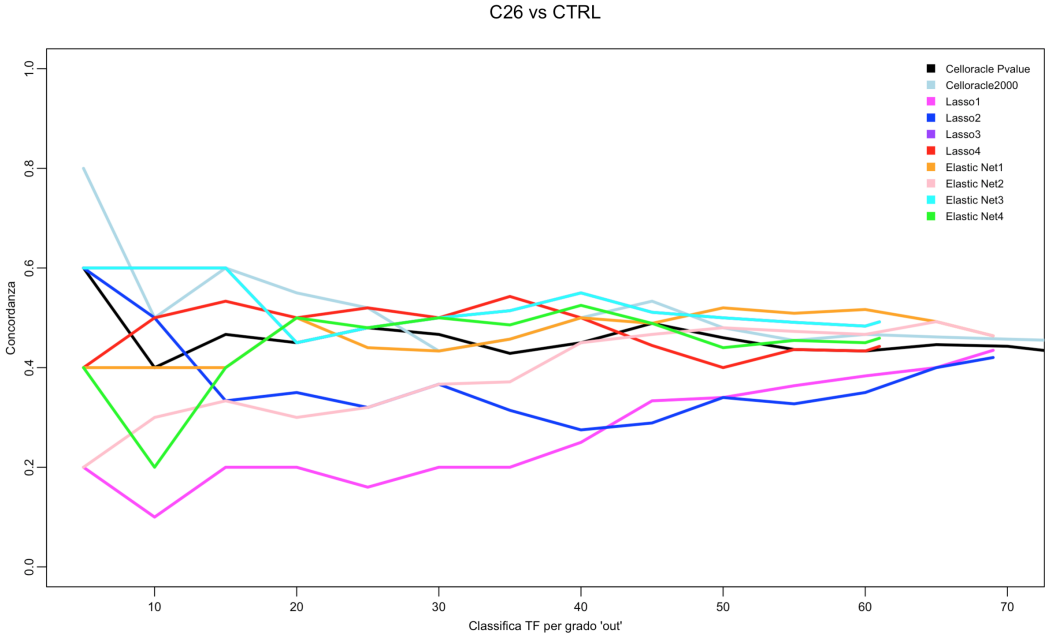

FIGURA 4.7: Cachetici vs sani. Confronto dei gradi OUT dei differenti modelli (approcci).

Analizzando il *grado "out"* del Lasso e dell'Elastic Net, che utilizzano come input la matrice di espressione genica integrata con la rete genica regolatoria (GRN), si può notare come questi due metodi si allineino maggiormente alle caratteristiche di CellOracle poiché si basano sugli stessi dati di partenza. Considerando invece le rimanenti metriche di confronto, non risultano differenze importanti tra l'Elastic Net e il Lasso. Le uniche differenze più evidenti emergono per il Lasso 1 e il Lasso 2 in quanto, oltre ad eseguire una selezione più approfondita dei geni, utilizzano come dati di partenza esclusivamente la matrice di espressione genica.

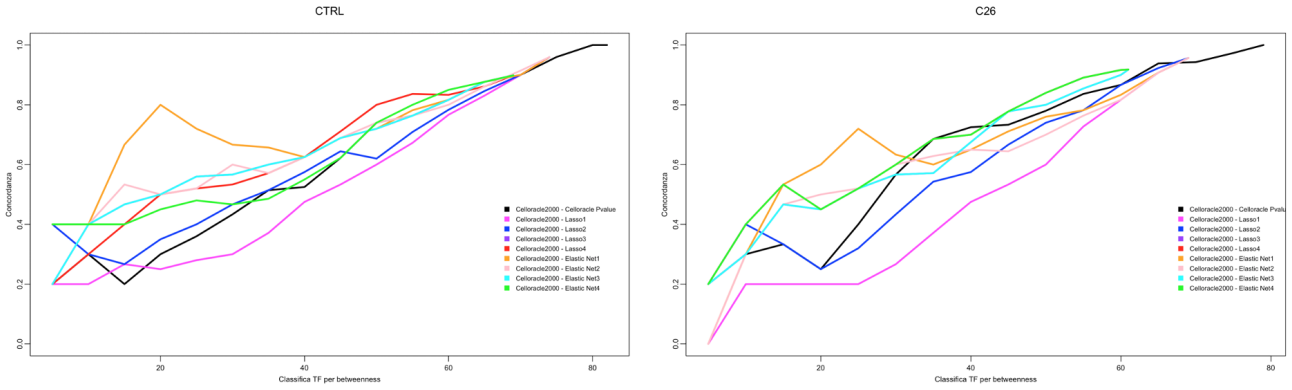

FIGURA 4.8: Confronto della *betweenness centrality* dei differenti modelli (approcci). A sinistra: topi controllo. A destra: topi cachetici

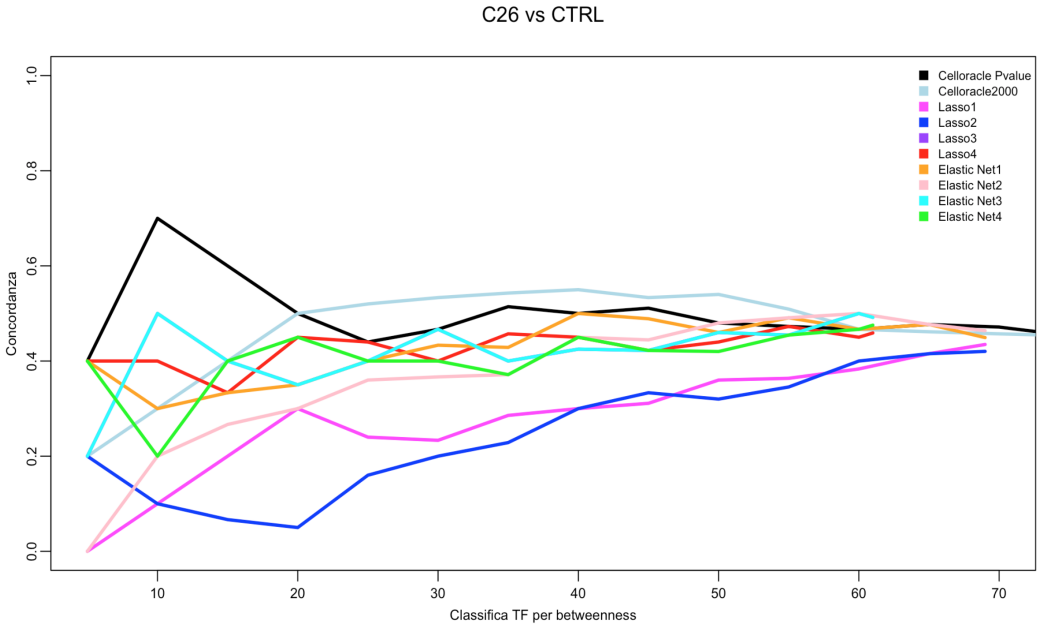

FIGURA 4.9: Cachetici vs sani. Confronto della *betweenness centrality* dei differenti modelli (approcci).

Per un ulteriore confronto si procede con un'analisi approfondita dei set di geni arricchiti (GSEA) (Subramanian et al., 2005), la quale viene molto utilizzata nell'ambito della bioinformatica in quanto consente di valutare se un insieme specifico di geni sia sovra o sottorappresentato nell'espressione genica. Questo tipo di analisi fornisce una comprensione dettagliata dei processi biologici coinvolti e delle vie metaboliche attive all'interno di un determinato campione.

Dopo una preliminare GSEA, le reti vengono visualizzate tramite gli Hive Plot (Hanson et al. (2014), Krzywinski et al. (2012)). Le comunità geniche sono state stimate

utilizzando l'algoritmo *Leiden* (Traag et al., 2019) e sono rappresentate sugli assi dei grafici mediante colori differenti degli archi. La lunghezza di ciascun asse è proporzionale alla dimensione della rispettiva comunità. Gli archi tra due nodi nella stessa comunità sono disegnati con il colore specifico della comunità, mentre le connessioni tra due nodi in due comunità diverse sono colorate in grigio. I nodi hub, definiti come nodi con più di venti vicini, vengono rappresentati come cerchi neri pieni e sono specificati a partire dalla Tabella A.3 in Appendice. Ciascun asse (comunità) è stato annotato con l'insieme di geni più arricchito. Come detto nel Paragrafo precedente, ci si concentra sul confronto tra Celloracle, Lasso 3 e Elastic Net 3.

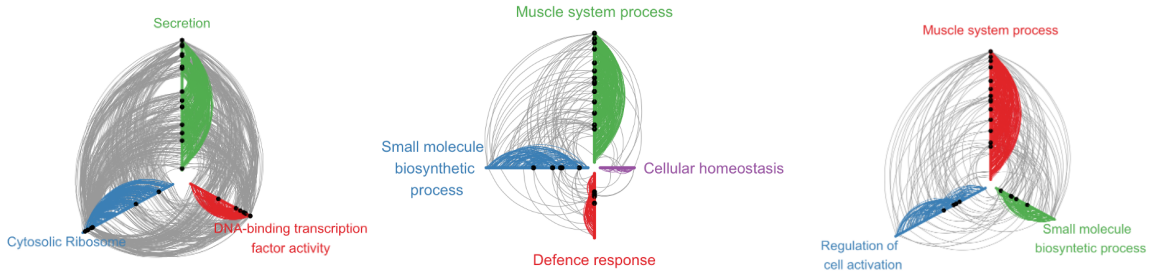

FIGURA 4.10: Hive plot topi C26. A sinistra: CellOracle. In centro: Lasso 3. A destra: Elastic Net 3.

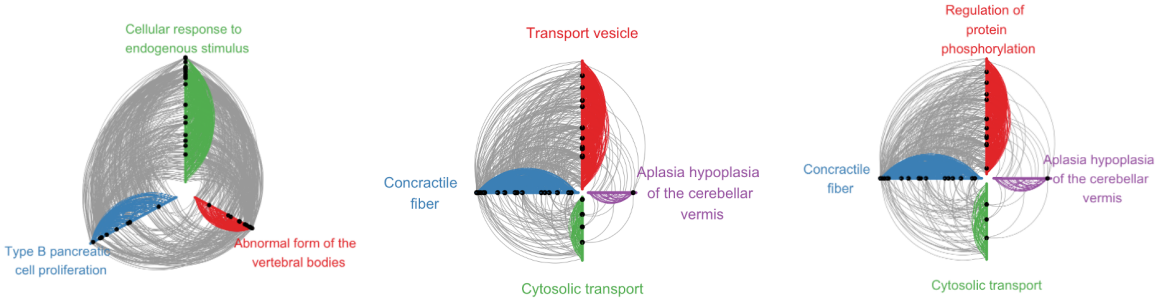

FIGURA 4.11: Hive plot topi CTRL. A sinistra: CellOracle. In centro: Lasso 3. A destra: Elastic Net 3.

In Tabella A.3, CellOracle ha evidenziato geni centrali molto associati a funzioni come la secrezione e il ribosoma citosolico, entrambi noti per il loro ruolo nella risposta

fisiologica e nella sintesi proteica essenziale per l'omeostasi cellulare e la funzione muscolare. Questi risultati sono coerenti con la letteratura che collega questi processi alla patogenesi della cachessia muscolare, dove l'alterata secrezione e la disfunzione del ribosoma giocano un ruolo critico (Fearon et al. (2011), Mannelli et al. (2020)). La ricerca di Tsoli & Robertson (2013) evidenzia che la cachessia muscolare è caratterizzata da infiammazione maligna e alterazioni metaboliche, contesto che supporta ulteriormente l'interpretazione dei risultati ottenuti attraverso CellOracle.

D'altro canto, Lasso3 C26 (Tabella A.5) e Elastic Net3 C26 (Tabella A.7), pur mostrando una certa sovrapposizione nei geni identificati, sembrano focalizzarsi su aspetti più specifici come il processo biosintetico di piccole molecole e la regolazione dell'attivazione cellulare. Questi aspetti potrebbero essere rilevanti per comprendere come la cachessia muscolare influenzi processi metabolici e l'attività cellulare in modo differenziato rispetto ai controlli sani.

Nel confronto tra i topi sani, l'analisi con CellOracle (Tabella A.8) ha evidenziato alcune funzioni cellulari, inclusa la proliferazione delle cellule pancreatiche di tipo B e le risposte agli stimoli endogeni, indicando la sua sensibilità nel rilevare cambiamenti in risposta a stimoli fisiologici normali. Questo è importante per distinguere i profili genici normali da quelli patologici.

In conclusione, sebbene ogni metodo abbia rivelato informazioni preziose nei diversi contesti biologici, è difficile valutare quale metodo funzioni meglio sui dati reali. Lasso3 ed Elastic Net3 identificano termini di Gene Ontology (GO)[2] che riguardano il muscolo, mentre CellOracle trova termini che possono essere interessanti per la cachessia.

Tuttavia, per affrontare questa sfida, nel capitolo successivo si farà ricorso alle simulazioni.

[2] `https://geneontology.org/`

Capitolo 5

Studio di simulazione

Nel contesto dell'analisi dei dati ad alta dimensioniltà, la selezione di variabili e la costruzione di modelli predittivi robusti è di rilevante importanza. Diversi algoritmi sono stati sviluppati per affrontare queste problematiche, ciascuno con caratteristiche specifiche.

Nel corso di questo Capitolo vengono valutate le prestazioni dei tre algoritmi precedentemente presentati, tramite la simulazione di tre grafi con strutture differenti:

1. il grafo di Erdös-Renyi (Erdős & Rényi, 1959), noto anche come grafo casuale, è caratterizzato dal fatto che, per ogni coppia di nodi, la presenza di un arco che li collega è determinata da una variabile casuale di Bernoulli. La sua dimensione in questo caso viene fissata a $p = 10$ (numero di nodi) e 20 archi;

2. il grafo Scale-free, la cui distribuzione delle connessioni per nodo segue una legge di potenza. Il modello più noto per la generazione di questo grafo è quello di Barabási-Albert (Barabási & Albert, 1999). Come dimensione vengono considerati 10 nodi e $m = 2$ (numero di nuovi collegamenti per ogni nuovo nodo);

3. il grafo Hub in cui alcuni nodi, chiamati "hub", sono collegati a molti altri nodi nella rete. In questa analisi il numero di hub è posto a 2.

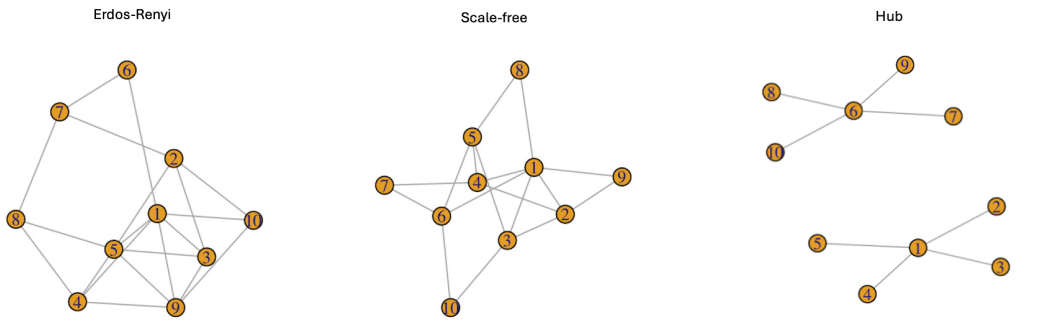

FIGURA 5.1: Grafi per la simulazione con 10 nodi. A sinistra: il grafo Erdos-Renyi. In centro: il grafo Scale-free. A destra: grafo Hub.

Successivamente, sulla base di questi grafi sono stati generati 100 datasets di numerosità $n = 50, 100, 200$ per $p = 10$.

Assumendo di avere dati Gaussiani, è stata impiegata la funzione **rmvnorm** della libreria **mvtnorm**. Come vettore per la media μ, a tutti gli elementi è stato assegnato il valore cinque. La matrice di varianza-covarianza è stata ottenuta invertendo la matrice di precisione Ω, ricavata dalla matrice di adiacenza di ciascun grafo. Facendo riferimento a Meinshausen & Bühlmann (2006), tutte le variabili presentano una varianza condizionale identica e la correlazione parziale tra i vicini è impostata a 0.245 (valori assoluti inferiori a 0.25 garantiscono la positività della matrice di covarianza inversa); ossia, $\Omega_{ii} = 1$ per tutti i nodi $i \in V$, $\Omega_{ij} = 0.245$ se esiste un arco che collega i e j e $\Omega_{ij} = 0$ altrimenti. Gli elementi diagonali della matrice di covarianza corrispondente sono generalmente superiori a 1 ma, per ottenere una varianza costante, tutte le variabili vengono riscalate in modo che gli elementi diagonali di Ω siano tutti unitari. Questa matrice deve essere definita positiva per poter essere invertita. Il fatto che il determinante di una matrice sia positivo non implica necessariamente che la matrice sia definita positiva o semidefinita positiva.

Si controllano quindi gli autovalori della matrice. Siccome nel nostro caso tutti gli autovalori sono positivi, la matrice è definita positiva.

Per la scelta del parametro di regolazione dei tre algoritmi si procede come nei Capitoli precedenti: per l'algoritmo di CellOracle si usa un λ pari a dieci; per il Lasso e Elastic Net $\lambda_i(\alpha) = \frac{2\hat{\sigma}_i}{\sqrt{n}}\widetilde{\Phi}^{-1}(\frac{\alpha}{2p(n)^2})$, pensato appositamente per i dati Gaussiani.

Per replicare l'algoritmo Bagging Ridge introdotto da Celloracle sono stati eseguiti venti bootstrap su un sottoinsieme casuale di righe della matrice di espressione, ossia sulle cellule. Successivamente, si calcola la media dei coefficienti ottenuti dalla regressione Ridge per ogni gene nei differenti bootstrap, e si esegue il test t a un campione.

Questo processo è stato iterato p volte, poiché ad ogni iterazione la variabile di risposta cambia.

Per valutare la capacità degli algoritmi impiegati nel riprodurre correttamente la struttura dei grafi, sono stati utilizzati come metriche la precisione (P), il recupero (R) e l'*F1-score* (media armonica tra P e R):

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F_1 = 2\frac{P \cdot R}{P + R}$$

dove TP (*True Positives*) rappresenta il numero di archi correttamente identificati dall'algoritmo, FP (*False Positives*) indica il numero di archi individuati dall'algoritmo che non sono presenti nel grafo reale, TN (*True Negatives*) è il numero di archi che sono correttamente assenti e FN (*False Negatives*) conta il numero di archi del grafo originale che non sono stati rilevati dall'algoritmo.

5.1 Risultati

Nelle seguenti Figure vengono rapprentate le metrice sopra menzionate. Considerando i grafi di Erdös-Renyi e Scale-free, il Lasso sembrebbe l'algoritmo migliore in termini di F_1-score e precisione.

In generale il Lasso mostra un miglioramento all'aumentare delle osservazioni, come atteso dalla consistenza del *neighborhood selection*.

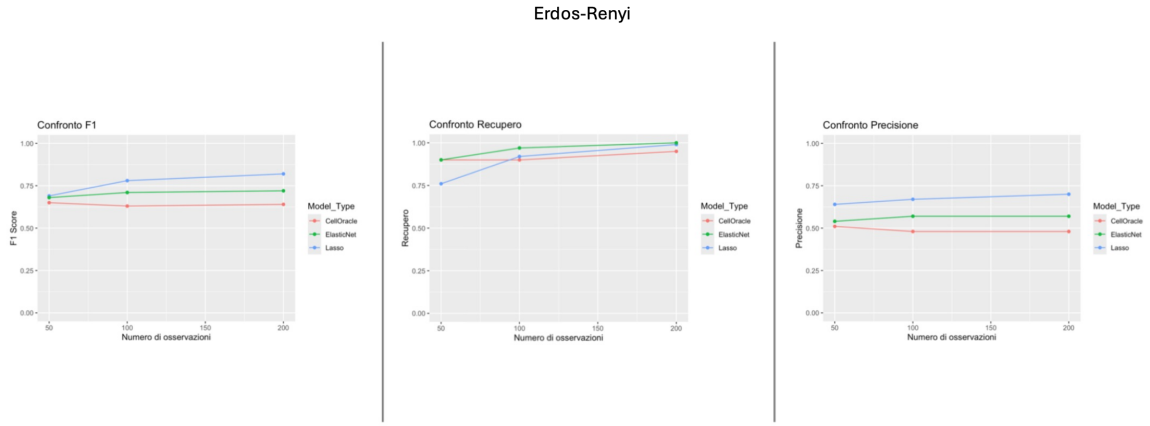

FIGURA 5.2: F_1 score, recupero (R), precisione (P) degli algoritmi considerati per il grafo di Erdös-Renyi.

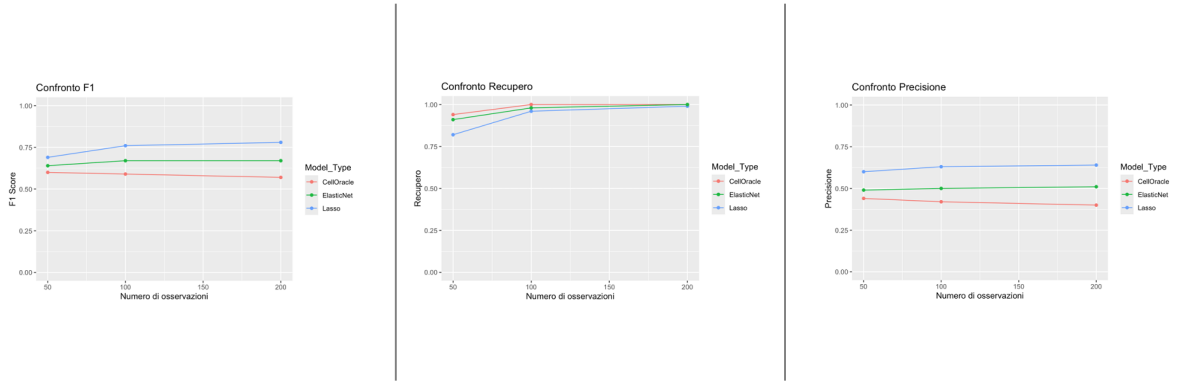

FIGURA 5.3: F_1 score, recupero (R), precisione (P) degli algoritmi considerati per il grafo Scale-free.

Per quanto riguarda il grafo hub i valori di F_1 e della precisione sono molto bassi per tutti e tre gli algoritmi. Mentre, considerando il recupero, tutti e tre iniziano con valori elevati e all'aumentare del numero di osservazioni questa metrica tende a stabilizzarsi, indicando una buona capacità di identificare tutte le istanze positive. In sintesi, sebbene tutti e tre gli algoritmi mostrino buone capacità di recupero, le loro basse prestazioni in termini di F1 Score e precisione richiederebbero ulteriori indagini per interpretare le differenze rispetto alle simulazioni precedenti.

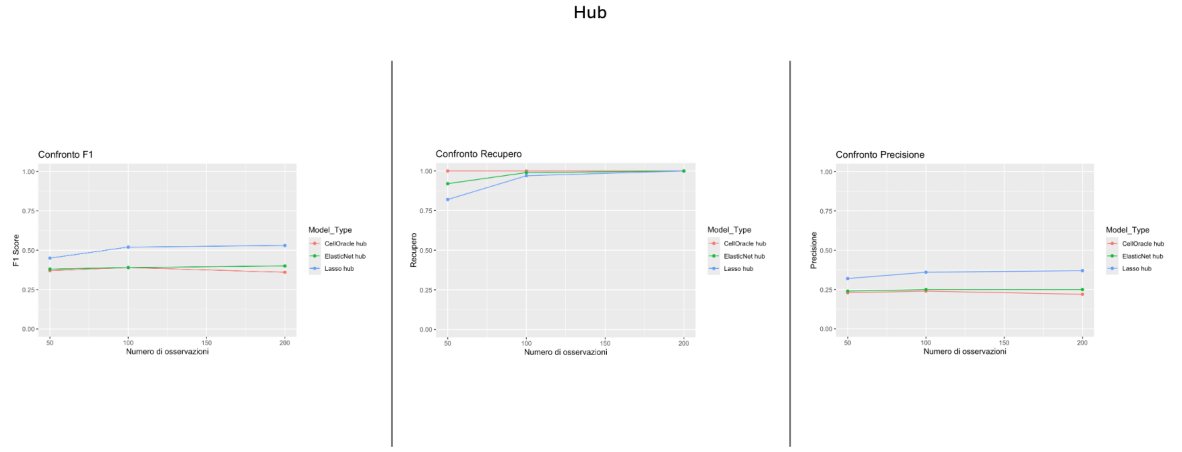

FIGURA 5.4: F_1 score, recupero (R), precisione (P) degli algoritmi considerati per il grafo Hub.

Concludendo, confrontando i risultati tra i tre tipi di strutture di grafo, il recupero è elevato e simile, indicando che tutti gli algoritmi riescono a identificare bene le istanze positive indipendentemente dalla topologia del grafo. Inoltre, i risultati migliori in

termini di F_1-score e precisione sono stati ottenuti con i grafi Erdos-Renyi, seguiti dai grafi Scale-Free, mentre i grafi Hub mostrano le prestazioni peggiori.

Questo suggerisce che le caratteristiche topologiche dei dati influenzano significativamente le prestazioni degli algoritmi. I grafi Erdös-Renyi, con la loro distribuzione casuale dei collegamenti, sembrano favorire una migliore identificazione e classificazione rispetto ai grafi Scale-Free e Hub. D'altra parte, i grafi Hub, con nodi centralizzati e meno distribuiti, rappresentano una sfida maggiore per gli algoritmi.

Argomentando invece i tre algoritmi, l'Elastic Net emerge come il miglior algoritmo complessivo, mostrando buone performance in F_1, Precisione e Recupero nei grafi Scale-Free e Erdos-Renyi. Ha una performance stabile e elevata, rendendolo una scelta affidabile. Il Lasso è molto simile, con ottime performance soprattutto in Precisione, ma leggermente inferiore a ElasticNet in Recupero nei grafi Erdos-Renyi. CellOracle è il meno performante dei tre, specialmente nei grafi Hub, ma ha comunque buoni risultati in Recupero.

Conclusioni

Dal presente elaborato vengono tratte due principali conclusioni, ciascuna considerando diverse prospettive: le simulazioni di dati e i dati reali. Le simulazioni sono state eseguite in quanto con i dati reali è difficile valutare quale metodo funzioni meglio.

Considerando le simulazioni di dati, l'Elastic Net risulta essere il metodo più efficace. Questo approccio mostra buone prestazioni in termini di F1, Precisione e Recupero. La sua performance è stabile ed elevata, rendendolo una scelta affidabile per la stima delle reti di regolazione genica. Il Lasso, sebbene molto simile all'Elastic Net, offre ottimi risultati soprattutto in Precisione, ma risulta leggermente inferiore in termini di Recupero. Infine, CellOracle, pur essendo il meno performante dei tre metodi nelle simulazioni, mostra comunque buoni risultati in termini di Recupero.

Considerando i dati reali, Lasso3 ed Elastic Net3 identificano termini di Gene Ontology (GO) che riguardano il muscolo, mentre CellOracle trova termini che possono essere interessanti per la cachessia. Da letteratura, CellOracle è riconosciuto per la sua capacità di sfruttare dati multimodali per costruire modelli personalizzati di reti di regolazione genica (GRN), simulando variazioni nell'identità cellulare a seguito della perturbazione dei fattori di trascrizione (TF) e fornendo interpretazioni sistematiche del ruolo contestuale dei TF nella regolazione dell'identità cellulare (Kamimoto et al., 2023).

I risultati ottenuti suggeriscono diverse implicazioni future e applicazioni pratiche. L'approccio CellOracle potrebbe essere ulteriormente sviluppato per migliorare la sua efficacia anche nelle simulazioni di dati, contribuendo alla comprensione delle interazioni geniche e allo sviluppo di nuove strategie terapeutiche. L'efficacia dell'Elastic Net nelle simulazioni sottolinea l'importanza di considerare metodi robusti per la modellazione delle reti geniche, applicabili in vari contesti, dalla diagnosi alla personalizzazione delle terapie.

Appendice

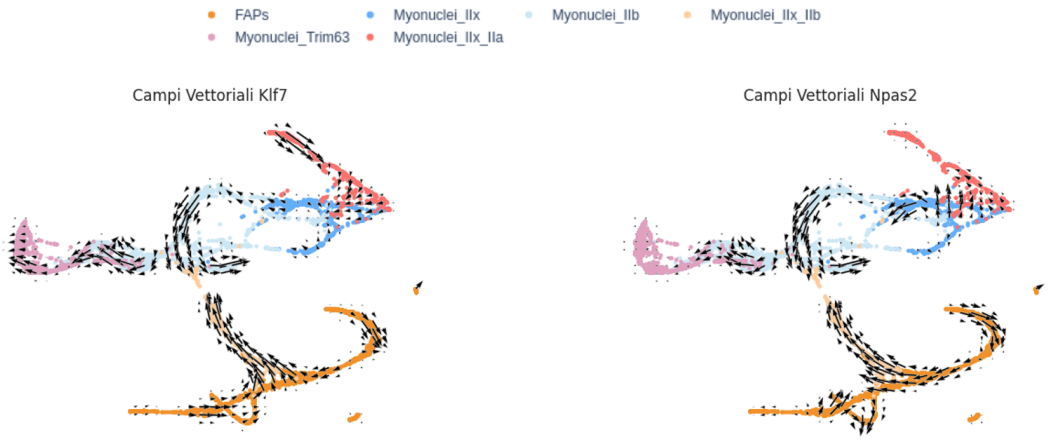

FIGURA A.1: Analisi delle perturbazioni in silico di alcuni fattori di trascrizione (TF), relativo ai topi cachetici. A sinistra: campi vettoriali del fattore di trascrizione Klf7. A destra: campi vettoriali del fattore di trascrizione Npas2.

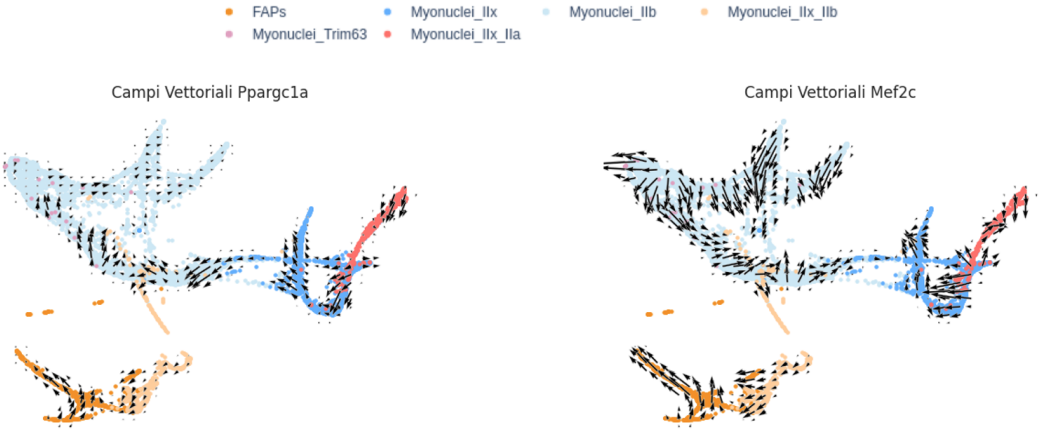

FIGURA A.2: Analisi delle perturbazioni in silico di alcuni fattori di trascrizione (TF), relativo ai topi sani. A sinistra: campi vettoriali del fattore di trascrizione Ppargc1a. A destra: campi vettoriali del fattore di trascrizione Mef2c.

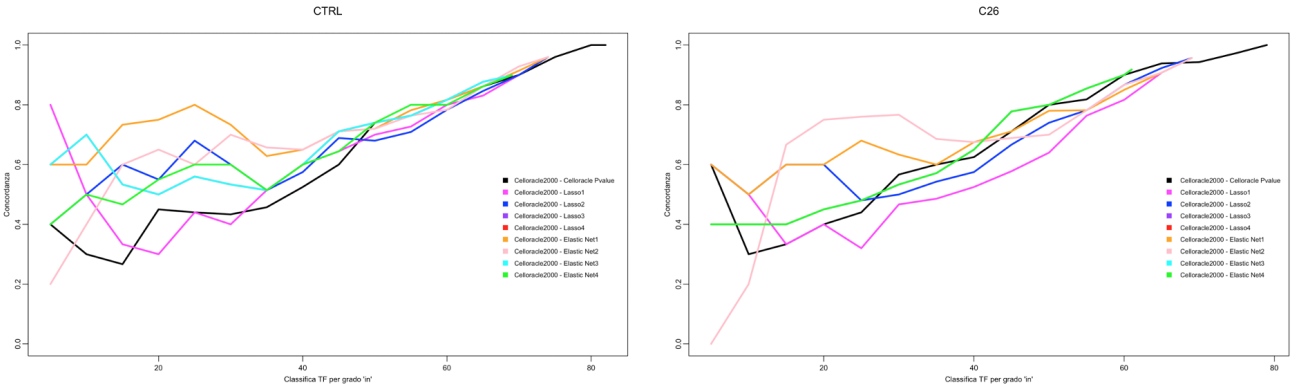

FIGURA A.3: Confronto dei gradi IN dei differenti modelli (approcci). A sinistra: topi controllo. A destra: topi cachetici

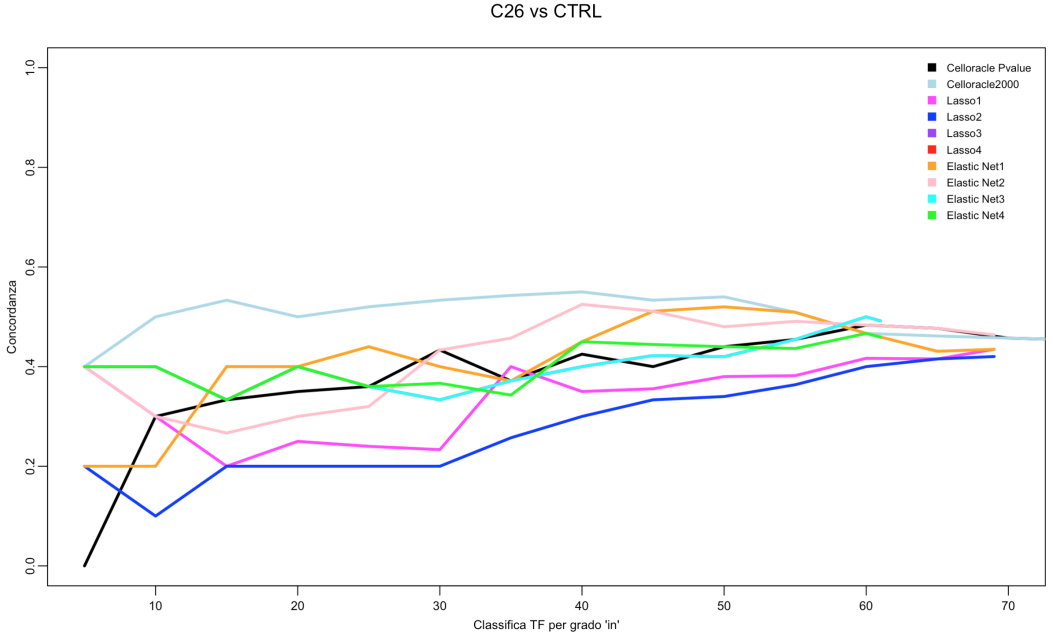

FIGURA A.4: Cachetici vs sani. Confronto dei gradi IN dei differenti modelli (approcci).

TABELLA A.1: Confronto *eigenvector centrality* tra CellOracle, Lasso 3 e Elastic Net 3, considerando solo il campione dei topi cachetici.

Gene	**Celloracle C26**	**Lasso 3**	**Elastic Net 3**	**Media**
Zeb1	0,982	1	1	0,994
Atf3	0,73	0,529	0,528	0,596
Pbx3	0,696	0,522	0,523	0,58
Egr1	1	0,297	0,297	0,531
Klf7	0,973	0,166	0,166	0,435
Esrrg	0,397	0,412	0,411	0,407
Npas2	0,521	0,267	0,267	0,352
Ets1	0,904	0,059	0,059	0,341
Tfdp2	0,395	0,296	0,298	0,33
Ebf1	0,478	0,233	0,235	0,315
Rora	0,292	0,324	0,323	0,313
Nfia	0,231	0,325	0,324	0,293
Tcfl5	0,274	0,271	0,271	0,272
Ppara	0,347	0,233	0,232	0,271
Glis3	0,499	0,137	0,155	0,264
Klf4	0,627	0,076	0,076	0,26
Mitf	0,358	0,182	0,181	0,24
Trp63	0,308	0,181	0,181	0,223
Fli1	0,407	0,116	0,116	0,213
Mbd1	0,454	0,077	0,076	0,202
Zbtb7c	0,31	0,147	0,147	0,201
E2f7	0,496	0,043	0,043	0,194
Pbx1	0,297	0,136	0,136	0,19
Creb5	0,305	0,102	0,102	0,17
Sox6	0,178	0,16	0,16	0,166
Nr1h4	0,31	0,087	0,087	0,161
Epas1	0,164	0,125	0,125	0,138
Runx2	0,155	0,114	0,114	0,128
Spi1	0,171	0,096	0,096	0,121
Vdr	0,087	0,126	0,126	0,113
Npas3	0,241	0,043	0,033	0,106

Gene	Celloracle C26	Lasso 3	Elastic Net 3	Media
Foxf2	0,176	0,058	0,058	0,097
Wt1	0,097	NA	NA	0,097
Nr4a3	0,202	0,037	0,036	0,092
Hes7	0,09	NA	NA	0,09
Hoxb4	0,091	0,09	0,09	0,09
Fosb	0,178	0,042	0,042	0,087
Foxp2	0,244	0,008	0,008	0,087
Id2	0,19	0,03	0,03	0,083
Foxa3	0,127	0,048	0,048	0,074
Gata2	0,051	0,085	0,085	0,074
Sim1	0,065	0,078	0,078	0,074
Tcf7l1	0,105	0,04	0,042	0,062
Gli2	0,042	0,049	0,049	0,047
Itgb2	0,018	0,06	0,06	0,046
Mecom	0,074	0,03	0,03	0,045
Gli3	0,019	0,056	0,057	0,044
Egr3	0,043	NA	NA	0,043
Prdm5	0,086	0,019	0,019	0,041
Erg	0,061	0,028	0,028	0,039
Bcl11a	0,037	NA	NA	0,037
Runx3	0,036	NA	NA	0,036
Osr1	0,072	0,008	0,008	0,029
Hesx1	0,026	NA	NA	0,026
Hlf	0,01	0,024	0,024	0,019
Sp100	0,013	0,017	0,017	0,016
Meox1	0,002	0,021	0,021	0,015
Sox7	0,011	0,013	0,013	0,012
Zfp957	0,012	NA	NA	0,012
Sox4	0,015	0,007	0,007	0,01
Gata3	0,009	0,009	0,009	0,009
Myb	0,01	0,009	0,009	0,009
Gata4	0,011	0,007	0,007	0,008
Rorb	0,019	0,003	0,003	0,008
Hmga1	0,004	0,008	0,008	0,007

Gene	**Celloracle C26**	**Lasso 3**	**Elastic Net 3**	**Media**
Lef1	0,02	0	0	0,007
Ikzf4	0	0,005	0,005	0,003
Alx1	0	NA	NA	0
Hoxb2	0	NA	NA	0
Hoxd11	0	0	0	0
Irx5	0	NA	NA	0
Neurod2	0	NA	NA	0
Nr1i2	0	NA	NA	0
Otx1	0	NA	NA	0
Pdx1	0	NA	NA	0
Pou3f4	0	NA	NA	0
Prop1	0	NA	NA	0
Rfx8	0	NA	NA	0
Sp7	0	NA	NA	0

Tabella A.2: Confronto *eigenvector centrality* tra CellOracle, Lasso 3 e Elastic Net 3, considerando solo il campione dei topi sani.

Gene	**Celloracle CTRL**	**Lasso 3**	**Elastic Net 3**	**Media**
Mef2c	1	1	1	1
Egr1	0,92	0,342	0,345	0,536
Zeb1	0,849	0,313	0,313	0,492
Klf7	0,897	0,212	0,222	0,444
Fos	0,571	0,375	0,377	0,441
Atf3	0,371	0,424	0,425	0,407
Thrb	0,579	0,314	0,314	0,402
Ebf1	0,458	0,333	0,333	0,375
Tcf4	0,525	0,298	0,298	0,374
Nr4a3	0,333	0,38	0,38	0,364
Ppargc1a	0,361	0,338	0,339	0,346
Nr4a1	0,41	0,311	0,313	0,345
Klf4	0,422	0,281	0,283	0,329
E2f7	0,313	NA	NA	0,313
Klf3	0,628	0,149	0,149	0,309
Pbx3	0,461	0,224	0,224	0,303
Rora	0,309	0,293	0,294	0,299
Klf5	0,565	0,156	0,158	0,293
Ets1	0,688	0,085	0,085	0,286
Esrrg	0,302	0,235	0,236	0,258
Trp63	0,383	0,189	0,19	0,254
Zbtb7c	0,368	0,183	0,184	0,245
Ppara	0,347	0,184	0,185	0,239
Glis3	0,323	0,166	0,166	0,218
Nfia	0,311	0,167	0,167	0,215
Fosb	0,275	0,18	0,18	0,212
Plagl1	0,494	0,055	0,055	0,201
Stat2	0,236	0,175	0,176	0,196
Ar	0,233	0,169	0,169	0,19
Creb5	0,25	0,151	0,153	0,185
Hic1	0,233	0,131	0,132	0,165

Gene	Celloracle CTRL	Lasso 3	Elastic Net 3	Media
Vdr	0,139	0,174	0,174	0,162
Prdm1	0,346	0,045	0,046	0,146
Creb3l2	0,257	0,078	0,078	0,138
Fli1	0,128	0,137	0,137	0,134
Ebf3	0,202	0,085	0,085	0,124
Hnf4g	0,116	NA	NA	0,116
Sox6	0,159	0,079	0,079	0,106
Erg	0,031	0,142	0,142	0,105
Thra	0,188	0,059	0,06	0,102
Hivep3	0,212	0,036	0,036	0,095
Gli3	0,019	0,129	0,129	0,092
Tbx15	0,146	0,036	0,036	0,073
Npas3	0,118	0,049	0,049	0,072
Runx2	0,088	0,063	0,063	0,071
Rfx2	0,174	0,017	0,017	0,069
Tcf7l1	0,19	0,007	0,007	0,068
Myef2	0,181	0,007	0,007	0,065
Hoxd3	0,06	0,066	0,066	0,064
Meox2	0,057	0,065	0,065	0,062
Rarg	0,171	0,008	0,008	0,062
Gli1	0,129	0,027	0,027	0,061
Runx1	0,018	0,081	0,081	0,06
Sp7	0,06	NA	NA	0,06
Gli2	0,126	0,019	0,019	0,055
Lmx1b	0,055	NA	NA	0,055
Prdm5	0,12	0,014	0,014	0,049
Ahr	0	0,058	0,058	0,039
Etv1	0,084	0,015	0,015	0,038
Nr2f2	0,04	0,03	0,03	0,033
Elf4	0,071	0,012	0,012	0,032
Tfap2a	0	0,042	0,042	0,028
Hlf	0,055	0,013	0,013	0,027
Lmx1a	0,055	0,011	0,011	0,026
Nr1d1	0,027	0,026	0,026	0,026

Gene	Celloracle CTRL	Lasso 3	Elastic Net 3	Media
Mecom	0,054	0,011	0,011	0,025
Lef1	0,023	NA	NA	0,023
Rorb	0,015	0,026	0,026	0,022
Sp100	0,011	0,027	0,027	0,022
Sp6	0,021	NA	NA	0,021
Shox2	0,046	0,006	0,006	0,019
Foxc1	0,028	0,004	0,004	0,012
Hltf	0,019	0,004	0,004	0,009
Bcl6b	0,007	0,008	0,008	0,008
Arid3c	0,005	NA	NA	0,005
Nrl	0,004	NA	NA	0,004
Dmrta1	0	NA	NA	0
Foxc2	0	NA	NA	0
Gata1	0	NA	NA	0
Id4	0	NA	NA	0
Npas1	0	NA	NA	0
Sox8	0	0	0	0

Gene Set CellOracle C26	Nomi dei geni hub
Secretion	Actn2, Esrrg, Glis3, Klf7, Pbx1, Pbx3, Ppara, Rora, Syne1, Tfdp2, Zeb1, E2f7, Mbd1, Npas3, Nr1h4, Zbtb7c
Cytosolic Ribosome	Creb5, Mitf, Ebf1, Egr1, Ets1, Fli1, Id2, Klf4, Runx2
DNA-binding transcription factor activity	Npas2, Sorbs1, Trp63, Atf3, Fosb, Foxf2, Spi1, Tcfl5

Tabella A.3: Gene Set enrichment analysis dei geni tenuti da CellOracle, considerando il campione di topi cachetici

Gene Set Celloracle CTRL	Nomi dei geni hub
Cellular response to endogenous stimulus	Ebf1, Fos, Hivep3, Itga4, Mef2c, Nfia, Ppara, Thrb, Zeb1, Atf3, E2f7, Fosb, Hic1, Klf3, Klf4, Klf5, Klf7, Rarg, Thra, Zbtb7c
Type B pancreatic cell proliferation	Ctnna3, Nr4a1, Nr4a3, Pbx3, Rora, Trp63, Creb3l2, Creb5, Ets1, Plagl1
Abnormal form of the vertebral bodies	Esrrg, Peg3, Ppargc1a, Tbc1d1, Tcf4, Egr1, Glis3, Myef2, Prdm1, Prdm5, Rfx2, Stat2

Tabella A.4: Gene Set enrichment analysis dei geni tenuti da CellOracle, considerando il campione di topi sani

Gene Set Lasso3 C26	**Nomi dei geni hub**
Muscle system process	Ebf1, Esrrg, Glis3, Klf7, Mitf, Nfia, Npas2, Pbx3, Ppara, Rora, Spi1, Tfdp2, Trp63, Zeb1
Small molecule biosynthetic process	Atf3, E2f7, Egr1, Foxf2, Mbd1
Defence response	Ets1, Fli1, Gata2, Klf4
Cellular homeostasis	Nessun Hub

Tabella A.5: Gene Set enrichment analysis dei geni tenuti dal Lasso 3, considerando il campione di topi cachetici

Gene Set Lasso3 CTRL	**Nomi dei geni hub**
Regulation of protein phosphorylation	Ahr, Ebf1, Ebf3, Egr1, Fli1, Fosb, Hic1, Klf3, Klf4, Klf7, Pbx3, Plagl1, Runx2, Stat2
Concractile fiber	Ar, Atf3, Creb3l2, Creb5, Esrrg, Fos, Glis3, Mef2c, Nfia, Nr4a1, Nr4a3, Ppara, Ppargc1a, Rora, Tcf4, Thrb, Trp63, Vdr, Zbtb7c, Zeb1
Cytosolic transport	Erg, Ets1, Klf5, Prdm1
Aplasia hypoplasia of the cerebellar vermis	Tfap2a

Tabella A.6: Gene Set enrichment analysis dei geni tenuti dal Lasso 3, considerando il campione di topi sani

Gene Set Elastic Net3 C26	**Nomi dei geni hub**
Muscle system process	Ebf1, Esrrg, Glis3, Klf7, Mitf, Nfia, Npas2, Pbx3, Ppara, Rora, Spi1, Tfdp2, Trp63, Zeb1
Regulation of cell activation	Ets1, Fli1, Gata2, Klf4
Small molecule biosyntetic process	Atf3, E2f7, Egr1, Foxf2, Mbd1

Tabella A.7: Gene Set enrichment analysis dei geni tenuti da Elastic Net 3, considerando il campione di topi cachetici

Gene Set Elastic Net3 CTRL	**Nomi dei geni hub**
Transport vesicle	Ebf1, Ebf3, Egr1, Fli1, Fosb, Hic1, Klf3, Klf4, Klf7, Pbx3, Plagl1, Runx2, Stat2
Concractile fiber	Ar, Atf3, Creb3l2, Creb5, Esrrg, Fos, Glis3, Mef2c, Nfia, Nr4a1, Nr4a3, Ppara, Ppargc1a, Rora, Tcf4, Thrb, Trp63, Vdr, Zbtb7c, Zeb1
Cytosolic transport	Ahr, Erg, Ets1, Klf5, Prdm1
Aplasia hypoplasia of the cerebellar vermis	Tfap2a

Tabella A.8: Gene Set enrichment analysis dei geni tenuti da Elastic Net 3, considerando il campione di topi sani

Bibliografia

Altenbuchinger, M., Weihs, A., Quackenbush, J., Grabe, H. J. & Zacharias, H. U. (2020). Gaussian and mixed graphical models as (multi-) omics data analysis tools. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1863**, 194418.

Baek, S. & Lee, I. (2020). Single-cell atac sequencing analysis: From data preprocessing to hypothesis generation. *Computational and Structural Biotechnology* **18**, 1429–1439.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *science* **286**, 509–512.

Buenrostro, J., Giresi, P., Zaba, L. et al. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y. & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490.

Burgos, E., Ceva, H., Laura Hernández, R. P. J. P., Devoto, M. & Medan, D. (2008). Two classes of bipartite networks: Nested biological and social systems. *Physical Review E* **78**.

Carninci, P. et al. (2005). The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563.

Chavali, S., Barrenas, F., Kanduri, K. & Benson, M. (2010). Network properties of human disease genes with pleiotropic effects. *BMC systems biology* **4**, 1–11.

Coifman, R. R. & Lafon, S. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* **102**, 7426–7431.

Cutler, A. A., Corbett, A. H. & Pavlath, G. K. (2017). Biochemical isolation of myonuclei from mouse skeletal muscle tissue. *Bio-protocol* **7**, e2654–e2654.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer Science & Business Media.

Erdős, P. & Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae Debrecen* **6**, 290–297.

Fearon, K., Strasser, F., Anker, S., Bosaeus, I., Bruera, E., Fainsinger, R., Jatoi, A., Loprinzi, C., MacDonald, N., Mantovani, G. & Davis, M. (2011). Definition and classification of cancer cachexia: An international consensus. *The Lancet Oncology* **12**, 489–495.

Ferrari, S., Battini, R., Magli, A., Angelelli, C., Badodi, M. G. S., Baruffaldi, F., Molinari, S. et al. (2013). Il fattore di trascrizione mef2c ai crocevia tra proliferazione, sviluppo e differenziamento muscolare. In *IL FATTORE DI TRASCRIZIONE MEF2C AI CROCEVIA TRA PROLIFERAZIONE, SVILUPPO E DIFFERENZIAMENTO MUSCOLARE*. pp. 1–10.

Friesen, D., Baracos, V. & Tuszynski, J. (2015). Modeling the energetic cost of cancer as a result of altered energy metabolism: implications for cachexia. *Theoretical Biology and Medical Modelling* **12**, 1–18.

Gonzalez, M. W. & Kann, M. G. (2012). Chapter 4: Protein interactions and disease. *PLoS Computational Biology* **8**, e1002819.

Grandi, F., Modi, H., Kampman, L. et al. (2021). Chromatin accessibility profiling by atac-seq. *Nature Protocols* **17**, 1518–1552.

Gregory, P. A., Bert, A. G., Paterson, E. L., Barry, S. C., Tsykin, A., Farshid, G., Vadas, M. A., Khew-Goodall, Y. & Goodall, G. J. (2008). The mir-200 family and mir-205 regulate epithelial to mesenchymal transition by targeting zeb1 and sip1. *Nature cell biology* **10**, 593–601.

Hansen, D. L., Shneiderman, B., Smith, M. A. & Himelboim, I. (2020). Chapter 6 - calculating and visualizing network metrics. In *Analyzing Social Media Networks with NodeXL (Second Edition)*. Morgan Kaufmann, second edition ed., pp. 79–94.

Hanson, B. A., Memisevic, V. & Chung, J. (2014). Hiver: 2d and 3d hive plots for r. *R/CRAN pkg. ver. 0.2-27* .

Hasin, Y., Seldin, M. & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology* **18**.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G., Geoghegan, J., Germino, G. et al. (2005). Multiple-laboratory comparison of microarray platforms. *Nature methods* **2**, 345–350.

Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one* **9**, e98679.

Ji, Y., Lotfollahi, M., Wolf, F. A. & Theis, F. J. (2021). Machine learning for perturbational single-cell omics. *Cell Systems* **12**, 522–537.

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L. & Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751.

Krzywinski, M., Birol, I., Jones, S. J. & Marra, M. A. (2012). Hive plots—rational approach to visualizing networks. *Briefings in bioinformatics* **13**, 627–644.

Laviano, A., Inui, A., Marks, D., Meguid, M., Pichard, C., Rossi Fanelli, F. & Seelaender, M. (2008). Neural control of the anorexia-cachexia syndrome. *American Journal of Physiology- Endocrinology and Metabolism* **295**, E1000–E1008.

Leduc-Gaudet, J., Franco-Romero, A. & Cefis, M. e. a. (2023). Mytho is a novel regulator of skeletal muscle autophagy and integrity. *Nature Communications* **14**.

Leung, W. (2023). Annotation of transcription start sites in drosophila .

Li, Z. & Liu, Q. (2023). The oncogenic role of klf7 in colon adenocarcinoma and therapeutic perspectives. *International Journal of Genomics* **2023**, 5520926.

Linding, R., Jensen, L. J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M. B. & Pawson, T. (2008). Networkin: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Research* **36**, D695–D699.

Liu, C., Wang, M., Wei, X. et al. (2018). An atac-seq atlas of chromatin accessibility in mouse tissues. *Scientific Data* **6**.

Maathuis, M., Drton, M., Lauritzen, S. & Wainwright, M. (2018). *Handbook of Graphical Models*. Boca Raton: CRC Press, 1st ed.

Mannelli, M., Gamberi, T., Magherini, F. & Fiaschi, T. (2020). The adipokines in cancer cachexia. *International journal of molecular sciences* **21**, 4860.

McInnes, L., Healy, J. & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* .

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso .

Ni, Y., Baladandayuthapani, V., Vannucci, M. & Stingo, F. C. (2022). Bayesian graphical models for modern biological applications. *Statistical Methods & Applications* **31**, 197–225.

Özgür, A., Vu, T., Erkan, G. & Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* **24**, i277–i285.

Paladugu, S. R., Zhao, S., Ray, A. & Raval, A. (2008). Mining protein networks for synthetic genetic interactions. *Bmc Bioinformatics* **9**, 1–14.

Pavlopoulos, G., Secrier, M., Moschopoulos, C. et al. (2011). Using graph theory to analyze biological networks. *BioData Mining* **4**.

Pearl, J. & Paz, A. (1986). GRAPHOIDS: A graph-based logic for reasoning about relevance relations. In *Proceedings of the 8th European Conference on Artificial Intelligence (ECAI-86)*. Brighton, United Kingdom.

Piasecka, A., Sekrecki, M., Szcześniak, M. W. & Sobczak, K. (2021). Mef2c shapes the microtranscriptome during differentiation of skeletal muscles. *Scientific Reports* **11**, 3476.

Picard, F., Miele, V., Daudin, J. et al. (2008). Deciphering the connectivity structure of biological networks using mixnet. *BMC Bioinformatics* **10**.

Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology* **32**, 896–902.

Rong, Z. H., Li, X. & Lu, W. L. (2009). Pinning a complex network through the betweenness centrality strategy. In *2009 IEEE International Symposium on Circuits and Systems*. IEEE.

Shutta, K. H., De Vito, R., Scholtens, D. M. & Balasubramanian, R. (2022). Gaussian graphical models with applications to omics analyses. *Statistics in medicine* **41**, 5150–5187.

Speed, T. (1978). Relations between models for spatial data, contingency tables and markov fields on graphs. *Advances in Applied Probability* **10**, 111–122.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. (2019). Comprehensive integration of single-cell data. *cell* **177**, 1888–1902.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**, 267–288.

Traag, V. A., Waltman, L. & Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 5233.

Tsoli, M. & Robertson, G. (2013). Cancer cachexia: malignant inflammation, tumorkines, and metabolic mayhem. *Trends in Endocrinology & Metabolism* **24**, 174–183.

van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605.

Wang, B., Guo, H., Yu, H., Chen, Y., Xu, H. & Zhao, G. (2021). The role of the transcription factor egr1 in cancer. *Frontiers in oncology* **11**, 642547.

Wang, Z., Gerstein, M. & Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63.

Yachie-Kinoshita, A. & Kaizu, K. (2019). Cell modeling and simulation. *Encyclopedia of Bioinformatics and Computational Biology* **2**, 864–873.

Zotenko, E., Mestre, J., O'Leary, D. P. & Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS computational biology* **4**, e1000140.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301–320.