Progetto d'esame di bioinformatica: Scoperta di sottotipi di malattia mediante l'integrazione di dati multi-omici

Stefano Giacoia

Il cancro alla prostata è il secondo tumore più comune negli uomini e il quarto più diffuso a livello globale. La sua prognosi e il trattamento variano a seconda del sottotipo molecolare. Esistono vari metodi per identificare i sottotipi, tra cui l'analisi dell'espressione genica, la metilazione del DNA e la struttura del genoma. A. Abeshouse et al. [1] hanno identificato sette sottotipi di cancro alla prostata, ognuno con una prognosi e una risposta al trattamento diverse. I sottotipi con una prognosi migliore mostrano una maggiore espressione di geni legati alla progressione tumorale, mentre quelli con una prognosi peggiore presentano una maggiore espressione di geni legati all'invasività e alla metastasi. Questi risultati sottolineano l'importanza dell'identificazione dei sottotipi per la diagnosi e il trattamento del cancro alla prostata.

I. INTRODUZIONE

La scoperta di sottotipi di malattia attraverso l'analisi dei dati multi-omici è un campo emergente di ricerca con il potenziale di rivoluzionare la nostra comprensione e il trattamento delle malattie. I dati multi-omici, che combinano informazioni genetiche, trascrittomiche, proteomiche e metabolomiche, possono fornire una visione più completa della biologia di una malattia e identificare sottogruppi di pazienti con caratteristiche molecolari distinte. Questi sottogruppi potrebbero avere prognosi o risposte al trattamento diverse, portando a terapie più mirate e personalizzate.

L'integrazione dei dati multi-omici è una sfida significativa, poiché i dati possono essere incoerenti, dimensionalmente complessi e gerarchici. Inoltre, i metodi di clustering tradizionali possono essere inefficaci nell'identificare sottogruppi di pazienti in base alle loro caratteristiche molecolari.

In questo studio, viene esplorato il problema della scoperta di sottotipi di malattia da dati multi-omici. Nella sezione Metodi, verranno illustrati gli approcci di integrazione dei dati e di clustering utilizzati. Verrà descritto il dataset utilizzato, analizzata la pre-elaborazione dei dati applicata, i sottotipi di malattia considerati e le metriche utilizzate per confrontare i raggruppamenti ottenuti con i sottotipi di malattia.

Nella sezione Risultati, verranno presentati e discussi i risultati dei vari approcci utilizzando tabelle e grafici. Si noti che i risultati non sono necessariamente buoni.

II. METODI

A. Dataset Utilizzato

Il dataset considerato in questo studio è stato ottenuto tramite TCGAbiolinks, un pacchetto R progettato per accedere ai dati del The Cancer Genome Atlas Research Network (TCGA). Nel caso specifico di questo studio, il focus è sulla sottotipizzazione delle malattie nel contesto del cancro alla prostata (PRAD). Per garantire la coerenza e la rilevanza dei dati, ho limitato l'analisi ai tumori solidi primari selezionando solo i campioni con dati disponibili per tutte le fonti omiche considerate (miRNA, mRNA, proteine).

B. Pre-elaborazione dei dati

La pre-elaborazione dei dati è una fase cruciale per garantire che i dati siano adatti all'analisi multi-omica e alla sottotipizzazione delle malattie.

La presenza di dati mancanti è una sfida comune nell'analisi omica. Per affrontare questo problema, ho applicato strategie di rimozione dei dati mancanti, eliminando le righe o le colonne con valori non disponibili. Questo assicura che i campioni e le caratteristiche selezionate per l'analisi siano completi. Per ridurre la complessità computazionale e concentrarci sulle caratteristiche più informative, ho inizialmente esaminato la loro varianza. Prima di tutto ho rimosso le caratteristiche con varianza prossima allo zero, successivamente mi sono concentrato sulle prime 100 caratteristiche con la varianza più elevata poiché l'idea è che le variabili che presentano la maggiore varianza tra i diversi campioni sono presumibilmente quelle che racchiudono le informazioni più rilevanti per differenziare i campioni stessi.

Dopo la selezione delle caratteristiche, ho applicato la standardizzazione usando lo z-score. Ho utilizzato questa procedura per garantire che tutte le caratteristiche avessero una distribuzione con media zero e deviazione standard unitaria. La standardizzazione è particolarmente cruciale quando si lavora con dati provenienti da diverse fonti omiche, poiché rende le diverse matrici omiche direttamente confrontabili, eliminando eventuali differenze di scala tra di esse.

C. Sottotipi di malattia

Considererò come sottotipi di malattia quelli identificati in un lavoro svolto dal Cancer Genome Atlas Research Network dove hanno utilizzato un modello di clustering integrativo (chiamato iCluster [2]) su dati multi-omici (alterazioni del numero di copie somatiche, meti-

lazione, livelli di mRNA, microRNA e proteine) e hanno scoperto tre sottotipi di malattia. Ho quindi scaricato il dataset usando la libreria TCGABiolinks prendendo in considerazione solo il tumore della prostata (PRAD) e solo i tumori solidi primari. Ho quindi allineato questo dataset con i dati multi-omici rimuovendo i campioni non comuni per garantire una corrispondenza accurata tra i dati delle sottocategorie di malattie e i dati omici. Quando si effettua l'analisi di integrazione, è cruciale che ciascun campione nel dataset delle sottocategorie di malattie abbia un corrispondente diretto nel dataset omico. Questo assicura che le etichette di sottotipo di malattia siano assegnate correttamente ai dati omici durante l'analisi.

D. Integrazione dei dati

L'integrazione dei dati è un passo cruciale per affrontare la sfida della scoperta dei sottotipi di malattie da dati multi-omici. Ho adottato due strategie di integrazioni tra i dati:

- La prima consiste nell'intregrazione dei dati utilizzando una semplice media delle matrici di somiglianza di ciascuna fonte di dati (calcolate con la distanza esponenziale euclidea scalata).
- La seconda è basata su SNF (Similarity Network Fusion [3]), con l'obiettivo di combinare le informazioni provenienti da diverse fonti omiche in un'unica rappresentazione coerente producendo una matrice di similarità integrata che tiene conto delle caratteristiche rilevanti da ciascuna fonte. Per ogni fonte omica (miRNA, mRNA, proteine), ho calcolato una matrice di similarità utilizzando la distanza euclidea scalata in modo esponenziale. Questo passaggio ha permesso di quantificare la somiglianza tra gli esempi nei diversi spazi omici. Le matrici di similarità calcolate sono state quindi integrate utilizzando l'algoritmo SNF. L'uso di SNF ha permesso di superare le sfide associate alla variabilità e alla complessità dei dati multi-omics, facilitando la creazione di una rappresentazione integrata dei campioni.

E. Eseguire la scoperta dei sottotipi di malattia

Per la scoperta dei sottotipi di malattie, è essenziale utilizzare algoritmi di clustering per raggruppare i campioni in cluster omogenei in base alle loro caratteristiche molecolari. Un approccio comune è utilizzare l'algoritmo PAM (Partitioning Around Medoids [4]) per questa fase di clustering. L'algoritmo PAM è una tecnica di clustering che, come K-means, suddivide il dataset in un numero predeterminato di cluster. Tuttavia, a differenza di K-means, PAM utilizza "medoidi", punti dati effettivi nel dataset, come rappresentanti dei cluster. Questo rende PAM più robusto agli outlier rispetto a K-means, poiché gli outlier avranno un impatto minore sulla posizione del medoide rispetto al centroide.

Nel contesto specifico della scoperta dei sottotipi di malattie, ho utilizzato l'algoritmo PAM con un numero di cluster pari al numero di sottotipi di malattie identificati da iCluster. Questo approccio consente di assegnare a ciascun campione un'etichetta di cluster, permettendo successivamente di confrontare i risultati del clustering con le sottocategorie di malattie note. Prima di eseguire l'algoritmo PAM ho convertito le matrici di similarità in matrici di distanza. In molti contesti, come l'analisi dei cluster o la visualizzazione multidimensionale, è più utile lavorare con misure di "distanza" piuttosto che di "similarità" sia perché forniscono un'interpretazione intuitiva di quanto siano "lontani" due punti dati sia perché PAM lavora con le distanze. Le misure di similarità e distanza sono concetti inversi: se due elementi sono molto simili, la loro distanza è piccola; se sono molto diversi, la loro distanza è grande.

Quindi ho eseguito PAM sulle seguenti matrici di similarità:

1. Matrici di similarità ottenute da singole fonti di dati (cioè miRNA, mRNA, proteine) utilizzando la consueta distanza euclidea esponenziale scalata. Ottenendo quindi tre diverse matrici di similarità.
2. Matrice integrata ottenuta utilizzando la media tra le matrici.
3. Matrice integrata ottenuta con SNF (Similarity Network Fusion).

Nell'ambito della nostra esplorazione completa della scoperta dei sottotipi di malattia, ho inoltre applicato lo Spectral Clustering [5] alla matrice integrata ottenuta con la Similarity Network Fusion (SNF). Il clustering spettrale è una tecnica potente che sfrutta le proprietà spettrali dei dati per scoprire le strutture sottostanti. Ho utilizzato la funzione SNFtool::spectralClustering() per eseguire il clustering spettrale sulla matrice integrata derivata da SNF. Questa funzione impiega la decomposizione spettrale per suddividere i dati in cluster in base agli autovalori e agli autovettori. I cluster risultanti rappresentano potenziali sottotipi di malattia basati su modelli sottostanti all'interno dei dati integrati. Ogni campione viene assegnato a un cluster specifico e la distribuzione dei campioni tra i cluster fornisce preziose indicazioni sull'eterogeneità della malattia.

III. RISULTATI

Nella fase finale del nostro progetto, ho condotto un'analisi comparativa completa dei vari approcci di clustering applicati nel corso dello studio. L'obiettivo è quello di valutare la coerenza e l'efficacia della scoperta dei sottotipi di malattia tra le diverse metodologie e strategie di integrazione confrontando il clustering calcolato con i

metodo di clustering	ARI	NMI	RI
PAM con miRNA	0.02467	0.02764	0.54241
PAM con mRNA	0.03766	0.05320	0.55749
PAM con proteine	0.00788	0.01965	0.55237
PAM con media tra le matrici	0.02365	0.04030	0.55981
PAM con SNF	0.17946	0.15674	0.63167
Spectral Clustering	0.11909	0.11723	0.60513

FIG. 1. Confronto delle prestazioni di diversi metodi di clustering utilizzando gli indici AI, MRI e RI.

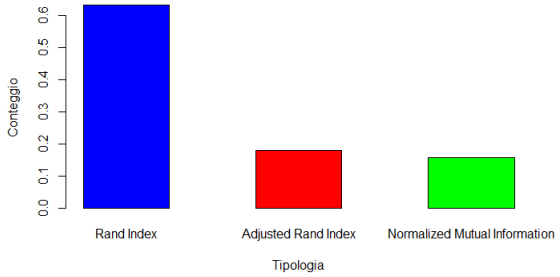

FIG. 2. barplot di confronto tra i sottotipi di iCluster e il clustering ottenuto con PAM sulla matrice integrata con SNF

sottotipi di malattia ricavate da iCluster. Questa analisi comparativa prevede diverse fasi: Per ogni metodo di clusterizzazione è stato calcolato l'Adjusted Rand Index (ARI), il Normalized Mutual Information (NMI) e il rand Index (RI) che sono tre delle misure più utilizzate per confrontare i cluster. Sono stati quindi generati grafici a barre per confrontare visivamente la distribuzione dei campioni tra i cluster per ciascun approccio di clustering. Questa visualizzazione aiuta a identificare i modelli, le somiglianze e le differenze tra i risultati ottenuti dalle diverse metodologie.

La tabella (FIG. 1) fornisce un riepilogo dell'applicazione dei tre diversi indici. È evidente che, per i raggruppamenti ottenuti da ciascun metodo considerato, rispetto ai sottotipi di malattia identificati da iCluster, la misura più rilevante è rappresentata dall'Indice di Rand. Questo indice mostra che esiste una certa sovrapposizione tra i due raggruppamenti, sebbene sia limitata per tutti gli approcci considerati. Tale risultato era prevedibile poiché iCluster è un metodo che integra diversi tipi di dati omici (come mutazioni genomiche, espressione genica, metilazione del DNA, ecc.) per identificare i sottotipi di malattia, fornendo così una visione più completa della biologia di essa, poiché considera più livelli di regolazione genica e variazioni genetiche.

Quando applichiamo l'algoritmo PAM alle singole fonti di dati (miRNA, mRNA e proteine), otteniamo ovviamente una sovrapposizione minore, poiché questi approcci considerano solo un singolo tipo di dati omici alla volta e quindi non riescono a catturare la complessità completa della biologia della malattia.

L'uso dell'algoritmo PAM sull'integrazione delle matrici di somiglianza di ciascuna fonte di dati omici, ottenuta utilizzando una semplice media delle matrici, non ha fornito un valore di Rand Index più alto rispetto all'applicazione dell'algoritmo sulle singole fonti di dati perché questo approccio non tiene conto delle possibili interazioni e correlazioni tra i diversi tipi di dati omici. Questo può portare a una perdita di informazioni importanti che potrebbero essere catturate se i dati fossero integrati in un modo che tiene conto delle loro relazioni, come avviene con SNF. Ad esempio, un gene potrebbe essere regolato a livello di mRNA, miRNA e proteine, e queste interazioni potrebbero essere perse se i dati vengono semplicemente mediati. Inoltre, la media delle matrici di somiglianza potrebbe essere influenzata da dati omici rumorosi o outlier, che potrebbero distorcere i risultati. Pertanto, è fondamentale considerare questi fattori quando si sceglie un approccio per l'integrazione dei dati omici.

Otteniamo il valore più alto dell'Indice di Rand quando eseguiamo PAM sulla matrice integrata ottenuta con SNF (vedi FIG. 2) , poiché stiamo combinando più tipi di dati omici in un'unica analisi, il che ci permette di catturare una visione più completa e complessa della biologia della malattia. Questo approccio integrato può rivelare sottotipi di malattia che potrebbero non essere evidenti quando si esaminano separatamente i dati di mRNA, miRNA o proteine.

Otteniamo un valore significativo anche quando applichiamo lo Spectral Clustering, che sfrutta la connessione tra i punti dati per creare i cluster. Questa tecnica utilizza gli autovalori e gli autovettori della matrice dei dati per proiettare i dati in uno spazio a dimensioni ridotte e raggruppare i punti dati. In termini pratici, lo Spectral Clustering risulta particolarmente efficace quando la struttura dei singoli cluster è fortemente non convessa, o più in generale, quando una misura del centro e della dispersione del cluster non fornisce una descrizione adeguata dell'intero cluster. Quando confrontiamo i cluster generati dallo Spectral Clustering con quelli identificati da iCluster, otteniamo una certa sovrapposizione perché entrambi gli algoritmi cercano di catturare la struttura sottostante dei dati. Tuttavia, mentre iCluster integra diversi tipi di dati omici per identificare i sottotipi di malattia, lo Spectral Clustering si basa sulla connettività tra i punti dati. Pertanto, se esiste una forte connettività tra i punti dati che appartengono a diversi sottotipi di malattia, lo Spectral Clustering potrebbe raggrupparli insieme.

In sintesi, sebbene abbiamo ottenuti valori di rand Index più o meno simili, l'integrazione di diversi tipi di dati omici sembra offrire la visione più completa della biologia della malattia. Questi risultati evidenziano l'importanza di utilizzare approcci integrati nell'analisi dei dati omici per una comprensione più profonda delle malattie complesse come il cancro. Questi risultati potrebbero avere un impatto significativo sulla futura ricerca scientifica e sullo sviluppo di nuove terapie.

[1] A. Abeshouse et al., "The molecular taxonomy of primary prostate cancer," Cell, vol. 163, no. 4, pp.1011–1025, 2015. [2] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," Bioinformatics, vol. 25, no. 22, pp. 2906–2912, 2009. [3] B. Wang et al., "Similarity network fusion for aggregating data types on a genomic scale," Nature methods, vol. 11, no. 3, pp. 333–337, 2014. [4] "Partitioning around medoids (program PAM)," in Finding groups in data, John Wiley & Sons, Ltd, 1990, pp. 68–125. doi: https://doi.org/10.1002/9780470316801.ch2. [5] U. Von Luxburg, "A tutorial on spectral clustering," Statistics and computing, vol. 17, pp. 395–416, 2007.