23 Febbraio (2025)

Corso: Strumenti formali per la Bioinformatica

Le GRN sono ipergrafi?

Rosario Pio Gnazzo
Dipartimento di Informatica, Università degli studi di Salerno

Abstract

La scoperta delle relazioni regolatorie tra geni e la ricostruzione delle reti geniche (GRN) a partire dai dati di espressione genica rappresentano una sfida computazionale di lunga data in bioinformatica. Recentemente, le reti neurali che lavorano sui grafi (GNN) hanno dimostrato il loro potenziale nella ricostruzione della struttura delle GRN, sfruttando la propagazione delle informazioni tra nodi adiacenti. In questo lavoro, generalizziamo tale approccio estendendo la rappresentazione delle GRN dai grafi agli **ipergrafi**, che permettono di modellare relazioni complesse tra più geni contemporaneamente. Abbiamo definito una regola di estrazione degli iperarchi per la costruzione della rete ipergrafica e sviluppato un modello di classificazione capace di operare su questa struttura. Il framework proposto è stato testato sui dati in-silico del benchmark DREAM5, dimostrando prestazioni promettenti. L'uso degli ipergrafi nella ricostruzione delle reti geniche apre nuove prospettive per una modellazione più accurata e profonda delle interazioni fra geni.

1 Introduzione

Le reti geniche regolatorie (GRN) rappresentano le relazioni causali tra fattori di trascrizione (TF) e i loro geni bersaglio. Modellare le GRN con strutture di rete consente di comprendere il funzionamento dei geni, interpretare processi biologici e identificare potenziali regolatori molecolari e biomarcatori per lo studio di malattie complesse. Con l'avvento delle tecnologie di sequenziamento ad alto throughput, i metodi di inferenza computazionale hanno cercato di ricostruire le GRN a partire solamente dai dati di espressione genica. Tuttavia, inferire le relazioni regolatorie tra un insieme di TF e i loro bersagli rimane una sfida aperta in bioinformatica.

1.1 Altri approcci

Nel corso degli anni, sono stati sviluppati numerosi approcci basati su metodi di apprendimento statistico e machine learning per l'inferenza delle GRN. I metodi

non supervisionati sono stati storicamente i più utilizzati, tra cui: (1) metodi basati su regressione, come TIGRESS; (2) metodi basati sull'informazione mutua, come ARACNE, CLR e MRNET; (3) metodi basati sulla correlazione, come il coefficiente di Pearson e di Spearman; e (4) reti bayesiane. Tra questi, GENIE3, basato su foreste casuali, ha ottenuto le migliori prestazioni nel benchmark DREAM5. Più recentemente, i metodi supervisionati hanno dimostrato un miglioramento significativo nella ricostruzione delle GRN, scomponendo il problema in sotto-problemi locali per ogni TF e i suoi bersagli. Diverse varianti di Support Vector Machine (SVM) sono state applicate per inferire le GRN, tra cui SIRENE, CompareSVM e GRADIS. Con il progresso del deep learning, sono stati sviluppati modelli di reti neurali profonde per prevedere le relazioni regolatorie tra geni, migliorando ulteriormente le prestazioni rispetto ai metodi tradizionali.

Tuttavia, la maggior parte di questi approcci presenta limitazioni pratiche. Le tecniche supervisionate spesso dipendono da dati eterogenei, con una ridotta capacità di generalizzazione. Inoltre, la maggior parte dei metodi si basa su modelli di completamento di matrice, che richiedono dati etichettati per ogni condizione o specie, limitandone l'applicabilità in scenari biologici reali. Infine, i meccanismi regolatori nei sistemi biologici sono intrinsecamente multi-genici, con insiemi di geni che lavorano insieme per svolgere funzioni biologiche. Il concetto di motivo di rete suggerisce che schemi regolatori ricorrenti servano da unità fondamentali delle GRN, ma molti metodi supervisionati tradizionali trattano le interazioni TF/bersaglio come coppie indipendenti, ignorando le informazioni strutturali globali della rete.

Per affrontare queste limitazioni, le reti neurali su grafi (GNN) sono emerse come una potente classe di modelli in grado di rappresentare relazioni complesse tra TF e geni bersaglio. Le GNN apprendono a partire dalla struttura della rete, propagando informazioni tra nodi e identificando pattern regolatori attraverso la classificazione di sottografi. Recentemente, è stato proposto un framework basato su GNN, GRGNN, che utilizza una combinazione di informazioni topologiche e dati di espressione genica per inferire le GRN.

1.2 Il nostro contributo

In questo lavoro, generalizziamo ulteriormente il concetto di GRN passando da una rappresentazione basata su grafi a una basata su ipergrafi. Infatti, se un grafo tradizionale permette di rappresentare relazioni binarie (TF $\rightarrow$ gene), un ipergrafo consente di modellare interazioni tra più geni contemporaneamente, fornendo una descrizione più ricca delle dinamiche regolatorie. Per fare ciò, abbiamo:

- Definito una regola di estrazione degli iperarchi dai dati di espressione genica, per costruire una rappresentazione ipergrafica delle GRN.
- Sviluppato un modello di classificazione per ipergrafi, capace di inferire relazioni regolatorie sfruttando l'informazione multi-genica fornita dagli iperarchi.
- Testato il nostro framework su dati in-silico del benchmark DREAM5.

2 Hypergraph Framework

2.1 Ipergrafi

Un **ipergrafo** è una generalizzazione di un grafo in cui un arco può connettere più di due vertici. Formalmente, un ipergrafo è definito come una coppia:

$$\mathcal{H} = (V, E) \tag{1}$$

dove:

- V è un insieme finito dei **nodi**(o vertici);
- E è un insieme di **iperarchi**, dove ogni iperarco $e \in E$ è un sottoinsieme non vuoto di V, cioè:

$$e \subseteq V.$$

A differenza di un grafo classico, in cui ogni arco connette esattamente due vertici, un iperarco di un ipergrafo può connettere un numero arbitrario di vertici. Per questo è una generalizzazione del concetto di grafi, infatti se ogni iperarco connette esattamente due vertici allora l'ipergrafo si riduce a un grafo tradizionale.

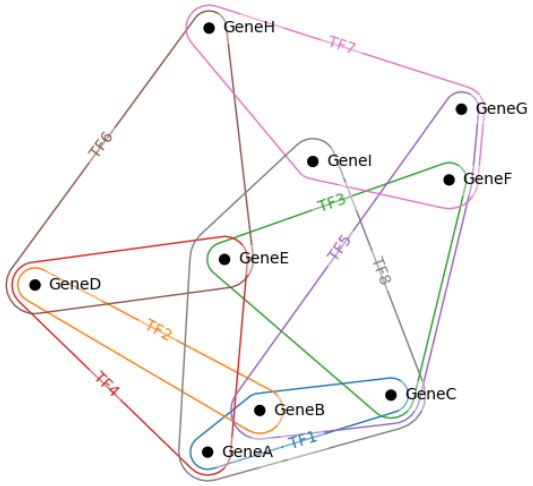

Figure 1: Es. Struttura ipergrafo

2.2 Regola di costruzione dell'ipergrafo

Sia $R \in \mathbb{R}_{m \times n}$ una matrice di espressione genica, dove m è il numero di campioni e n il numero di geni. Sia $\mathcal{T}$ un insieme di soglie per la correlazione di Pearson. Definiamo l'ipergrafo $\mathcal{H} = (V, E)$ con le seguenti proprietà:

- L'insieme dei nodi è dato dall'insieme dei geni, ovvero $V = \{G_1, G_2, \ldots, G_n\}$.
- L'insieme degli iperarchi E è costruito come segue:
  - Per ogni coppia di geni (G_i, G_j), si calcola il coefficiente di correlazione di Pearson:

$$\rho_{ij} = \frac{\sum_{k=1}^{m}(R_{ki} - \bar{R}_i)(R_{kj} - \bar{R}_j)}{\sqrt{\sum_{k=1}^{m}(R_{ki} - \bar{R}_i)^2}\sqrt{\sum_{k=1}^{m}(R_{kj} - \bar{R}_j)^2}}.$$

- Se $|\rho_{ij}| \geq t$ per una certa soglia $t \in \mathcal{T}$, allora i geni G_i e G_j vengono aggiunti allo stesso iperarco.
- Ogni iperarco $e \in E$ è definito come un sottoinsieme di geni altamente correlati rispetto a una soglia t, ossia:

$$e = \{G_i \mid \exists G_j \text{ tale che } |\rho_{ij}| \geq t\}.$$

Ne consegue che l'ipergrafo $\mathcal{H}$ è una rappresentazione delle relazioni di co-espressione genica, in cui ogni iperarco connette un gruppo di geni altamente correlati secondo una data soglia t.

2.3 Estrazione sotto-ipergrafi

Sotto-ipergrafi positivi: Dato un ipergrafo $\mathcal{H} = (V, E)$ costruito a partire da dati di espressione genica, un sotto-ipergrafo positivo viene definito a partire dalle coppie di geni (g_i, g_j) che coesistono nello stesso iperarco $e \in E$.

- Per ogni iperarco $e_k \in E$, se $|e_k| > 1$, si generano tutte le coppie possibili (g_i, g_j) con $g_i, g_j \in e_k$.
- L'insieme risultante P contiene tutte le coppie di geni che sono altamente correlate.
- Per ogni coppia $(g_a, g_b) \in P$, si individuano gli iperarchi

$$E_{ab} = e \in E \mid g_a \in e \vee g_b \in e \tag{2}$$

  che contengono almeno uno dei due geni.
- Si definisce l'insieme dei geni del sotto-ipergrafo come l'unione degli elementi di E_{ab}:

$$V_{ab} = \bigcup_{e \in E_{ab}} e. \tag{3}$$

  Si considerano solo i sotto-insiemi che contengono almeno una coppia di geni, quindi se $|V_{ab}| < 2$, il sotto-ipergrafo viene scartato. Dopodiché si costruisce la matrice di adiacenza A_{ab} in cui i nodi sono connessi se appartengono allo stesso iperarco in E_{ab}. Infine si etichetta come positivo ($y = 1$) e utilizzato per l'addestramento del modello.

Sotto-ipergrafi negativi: Per l'addestramento del modello, è necessario disporre di esempi negativi, ovvero sotto-grafi che non rappresentano interazioni biologicamente rilevanti, per fornire un insieme di controllo rispetto alle relazioni positive identificate in precedenza. Dato un ipergrafo $\mathcal{H} = (V, E)$, dove V rappresenta l'insieme di geni e E gli iperarchi, il processo di costruzione dei sotto-grafi negativi segue questi passi:

- Si considerano tutte le possibili coppie di geni $(g_i, g_j) \in V \times V$.
- Si escludono le coppie che appartengono a iperarchi esistenti, generando così il set di coppie negative $P^- = P_{tot} \setminus P^+$.
- Viene effettuata una selezione casuale di N coppie negative.
- Per ogni coppia negativa (g_i, g_j), si identificano tutti gli iperarchi contenenti almeno uno dei due geni.

- Si costruisce il sotto-grafo corrispondente considerando tutti i geni presenti negli iperarchi selezionati.
- Si costruisce la matrice delle feature **X** per i nodi e la matrice di adiacenza **A**, etichettando infine l'oggetto come negativo ($y = 0$).

2.4 Classificatore

Il classificatore proposto è una rete neurale basata su convoluzioni iper-grafiche, il cui scopo è determinare se una coppia di geni appartiene allo stesso iperarco. La rete apprende rappresentazioni strutturali dai sotto-ipergrafi costruiti a partire dai dati di espressione genica. La rete neurale, denominata *HypergraphNet*, è composta dai seguenti elementi:

- **Layer di convoluzione iper-grafica**: Tre strati di convoluzione iper-grafica basati sulla *HypergraphConv*, che propagano le informazioni tra i nodi preservando la topologia del dato iper-grafico.
- **Batch Normalization**: Dopo ogni strato convoluzionale, viene applicata la normalizzazione per ridurre il covariate shift e migliorare la stabilità della rete.
- **Funzione di attivazione ReLU**: Ogni strato convoluzionale è seguito da una funzione di attivazione non lineare $ReLU(x) = \max(0, x)$.
- **Dropout**: Viene applicato un dropout con probabilità $p = 0.3$ per prevenire l'overfitting.
- **Pooling globale**: Un'operazione di pooling medio globale (*global mean pooling*) aggrega le rappresentazioni dei nodi in un'unica rappresentazione vettoriale del sotto-ipergrafo.
- **Strato Fully Connected**: Uno strato finale lineare che riduce la rappresentazione a una singola unità di output.
- **Attivazione Sigmoide**: L'output viene trasformato mediante la funzione sigmoide $\sigma(x) = \frac{1}{1+e^{-x}}$, producendo una probabilità che rappresenta la confidenza del modello sulla classe positiva.

3 Risultati sperimentali

3.1 Addestramento e valutazione del modello

L'architettura proposta viene addestrata attraverso una validazione incrociata *Stratified K-Fold Cross-Validation*, per garantire una valutazione dei risultati più robusta. Il processo segue i seguenti passi:

- Il dataset $\mathcal{D} = \{(G_i, y_i)\}_{i=1}^{N}$ viene suddiviso in K fold mantenendo la distribuzione delle classi.
- Per ogni fold k, il modello viene addestrato sui dati di training $\mathcal{D}_{train}^{k}$ e valutato sui dati di test $\mathcal{D}_{test}^{k}$.
- L'addestramento utilizza *early stopping* per prevenire overfitting, arrestando l'ottimizzazione se la perdita di validazione non migliora per p epoche consecutive.
- Vengono calcolate metriche di valutazione come accuratezza, precisione e

AUC-ROC.
- Alla fine della cross-validation, si calcolano media e deviazione standard delle metriche per fornire una stima robusta delle prestazioni del modello.

Nello specifico la fase di train ad ogni iterazione della Cross-Validation avviene per minimizzazione della funzione di perdita di entropia binaria:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \left(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right), \tag{4}$$

dove:
- $y_i \in \{0, 1\}$ è l'etichetta reale che indica se i due geni sono nello stesso iperarco.
- $\hat{y}_i$ è la probabilità predetta dal modello.
- N è il numero totale di sotto-ipergrafi nel dataset.

Per ottimizzare i pesi della rete, viene utilizzato l'algoritmo *Adam* con un tasso di apprendimento di 0.0005 e un termine di regolarizzazione L2 (*weight decay*) pari a 5×10^{-4}.

Le performance del modello invece, sono state valutate utilizzando le come metriche:
- **Accuratezza:** misura la proporzione di predizioni corrette rispetto al totale.
- **Precisione:** misura la frazione di predizioni positive corrette rispetto a tutte le predizioni positive.
- **AUC-ROC:** valuta la capacità del modello di distinguere tra classi positive e negative analizzando la curva ROC.

3.2 Analisi dei risultati

I risultati ottenuti dal classificatore indicano una buona capacità di discriminare se due geni appartengano allo stesso iperarco e, di conseguenza, possano essere coinvolti nello stesso processo di regolazione genica. Di seguito, si riportano le metriche principali ottenute:
- **Accuracy**: $87.90\% \pm 1.51\%$
  L'accuratezza elevata indica che il modello classifica correttamente la maggior parte dei campioni, confermando una buona capacità predittiva generale. Tuttavia, questa metrica da sola non è sufficiente per valutare la qualità del modello, soprattutto in presenza di un dataset sbilanciato.
- **Precision**: $89.02\% \pm 0.79\%$
  La precisione pari a 89.02% suggerisce che il modello presenta un basso tasso di falsi positivi, garantendo un'affidabilità elevata nella predizione di coppie di geni appartenenti allo stesso iperarco. Questo aspetto è cruciale per evitare di segnalare erroneamente associazioni prive di rilevanza biologica.
- **AUC-ROC**: $78.18\% \pm 4.20\%$
  L'AUC-ROC di 78.18% indica una buona separabilità tra coppie di geni appartenenti e non appartenenti allo stesso iperarco, pur lasciando margine di miglioramento nella distinzione tra le due classi.

4 Discussione e Conclusioni

In questo lavoro, abbiamo esplorato la generalizzazione delle *Gene Regulatory Networks* (GRN) dai grafi agli *ipergrafi*, proponendo un nuovo modello per catturare interazioni complesse tra i geni. L'uso degli ipergrafi permette di infatti di rappresentare in modo più completo le relazioni *multigeniche*, superando i limiti dei modelli classici basati su grafi binari. La transizione dai grafi agli ipergrafi dunque consente di modellare **interazioni di ordine superiore**, fondamentali per comprendere processi biologici complessi. Dal punto di vista metodologico, il framework proposto introduce una **matrice di incidenza estesa**, che codifica le relazioni multi-way tra i geni che si è rivelata utile al classificatore per la generalizzazione.

4.1 Limiti e Prospettive Future

Nonostante i risultati promettenti, il modello presenta alcune limitazioni:

1. **Scalabilità computazionale** – La crescita del numero di iperarchi può rendere l'analisi computazionalmente onerosa.
2. **Validazione sperimentale** – L'efficacia del modello andrebbe verificata su dataset biologici reali più grandi per confrontare le predizioni con dati sperimentali.
3. **Interpretabilità** – Sebbene gli ipergrafi catturino interazioni complesse, è necessaria definire interpretabilità biologica per facilitare l'adozione del modello.

In sintesi, la transizione dagli approcci basati su grafi agli ipergrafi può rappresentare un passo significativo nella modellazione delle reti di regolazione genica. Questo paradigma ha il potenziale per migliorare la comprensione dei processi regolatori.