Big Network Visualization Tool for iNSIdEnano

Luigi Giugliano[1], Marco Mecchia[1]

[1]Universitá degli studi di Salerno

18 febbraio 2016

Overview

Overview

iNSIdEnano

iNSIdEnano è un tool grafico che mette in evidenza le connessioni tra entità fenotipiche:

- Esposizione ai nanomateriali.
- Trattamenti farmaceutici.
- Esposizione ad agenti chimici.
- Malattie.

L'interazione tra queste entità è valutata in base al loro effetto sull'espressione dei geni.

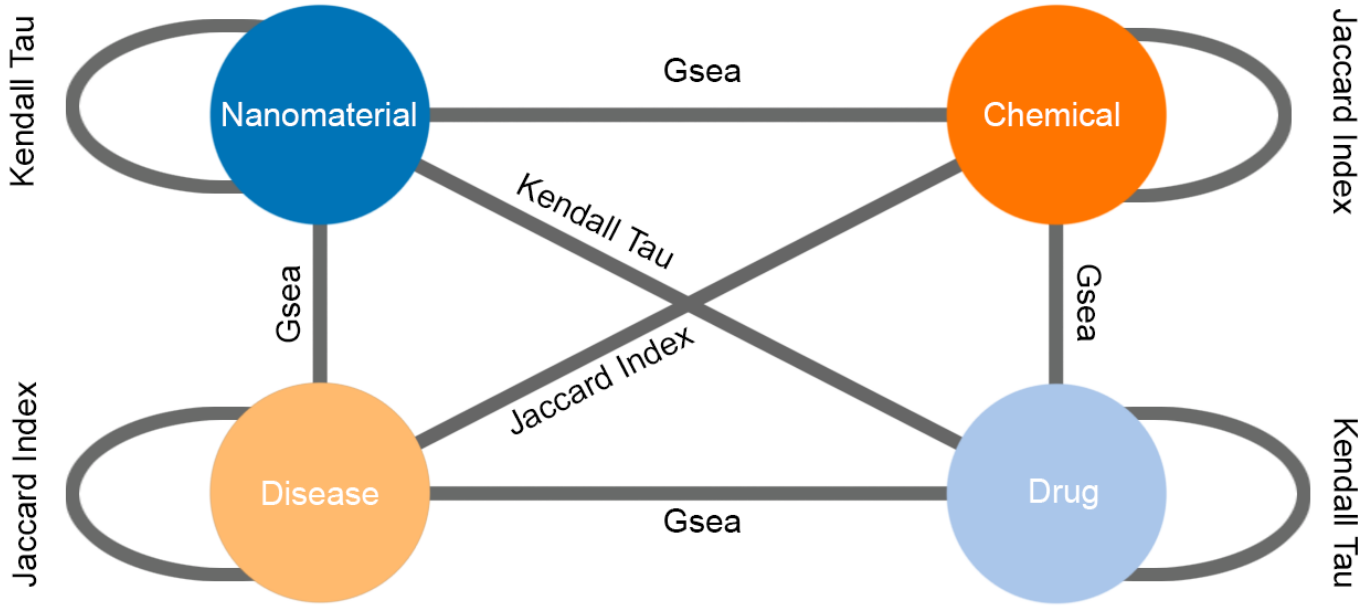
Associazione dei geni

Per ogni entità fenotipica nel dataset, è stata assegnata una lista di geni. In particolare:

- Ad un'insieme di geni vengono associati tutte le malattie e tutti gli agenti chimici.
- Ad ogni farmaco e nanomateriale vengono associate liste ordinate di geni.

Quindi per costruire una network di similarità tra entità fenotipiche è stato necessario calcolare la similarità a coppie per ogni entità.

Calcolo delle distanze: Panoramica



E' stata calcolata la distanza per ogni coppia di entità. Tali distanze sono state poi normalizzate tra -1 e 1 per renderle confrontabili.

Insieme di geni vs Insieme di geni

Il Jaccard index è stato utilizzato per calcolare la similarità tra due malattie, tra due agenti chimici o tra un agente chimico e una malattia.
Dati due insiemi A e B é definito come:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Per ogni agente chimico vengono considerati due set di geni: quelli che sono up-regolati da quell'agente chimico e quelli che sono down-regolati. Per quelli down-regolati il Jaccard index è calcolato con il segno negativo.

Geni ordinati vs Geni ordinati

La distanza Kendall Tau è stata utilizzata per calcolare la similarità tra nanomateriali e nanomateriali, tra farmaci e farmaci e tra nanomateriali e farmaci, basata sulla lista ordinata dei geni. La distanza Kendall Tau tra due liste $T1$ e $T2$ è definita come segue:

$$K(T_1, T_2) = |(i, j) : i < j, (T_1(i) < T_1(j) \wedge T_2(i) > T_2(j)) \vee \quad (2)$$

$$(T_1(i) > T_i(j) \wedge T_2(i) < T_2(J))|$$

questa distanza è compresa tra 0 e $n * (n - 1)$, dove n è la lunghezza della lista.

Geni ordinati vs insieme di geni

La Gen Set Enrichment Analysis (GSEA), basata sul test di Kolmogorov-Smirnov, è stata usata per calcolare la similarità a coppie tra nanomateriali e malattie, tra nanomateriali e agenti chimici, tra farmaci e malattie ed infine tra farmaci e agenti chimici. Il test di KolmogorovSmirnov può essere usato per confrontare elementi con una distribuzione di probabilità. La distribuzione empirica F_n per osservazioni *iid*, è definito:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I[-\inf, x](x_i) \tag{3}$$

dove:

$$I[-\inf, x](x_i)$$

è la funzione definita su X che indica l'appartenenza di un elemento in un sottoinsieme A di X che ha valore 1 per tutti gli elementi di A e 0 per tutti gli elementi di X non in A. La statistica KolmogorovSmirnov per una distribuzione cumulativa $F(x)$ è

$$D_n = sup_x[F_n(x) - F(x)]$$

La statistica KolmogorovSmirnov è stata usata non in valore assoluto per preservare il segno. Ciò aiuta a capire se un gene è up o down-regolato, ovviamente anche questi valori sono stati normalizzati tra $[-1 : 1]$

Implementazione

INSIdEnano è stato implemetato usando *R* per il back end e Javascript per il front end.

- Per far comunicare i due linguaggi, sono state usate le librerie HTMLWidgets e Shiny di R.
- Sistema con architettura client-server:
  - Il client è responsabile della gestione dell'interfaccia, del rendering della network, della formulazione e della sottomissione delle query.
  - Il server processa i dati dal database in base agli input dell'utente, e restituisce il risultato di tale computazione al client.

Overview

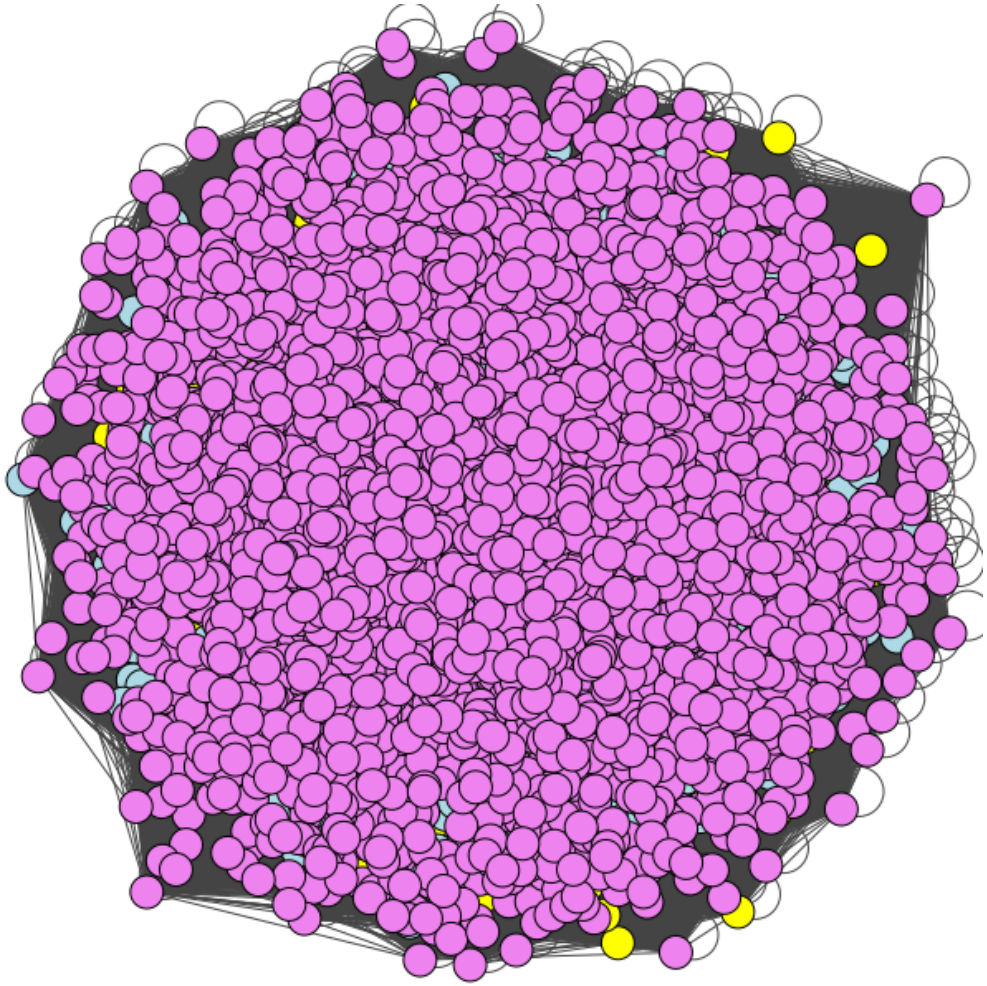
Problema

Problema

I nodi presenti nel grafo sono 3866. Le distanze sono state calcolate per ogni coppia di nodi del grafo. Il numero di archi è quindi pari a $3866 * 3865 \approx 15000000$.

L'elevato numero di archi rende questo grafo:

INVISUALIZZABILE



Soluzione

L'idea é quella di partire da un piccolo sottoinsieme di nodi, corrispondenti ai gruppi iniziali, ed organizzare l'esplorazione in maniera gerarchica.

- L'utente puo' espandere una categoria alla ricerca di un particolare nodo tramite doppioclick. Tale processo é iterabile per ogni sottocategoria.
- Trovato il nodo di interesse, é possibile visualizzare le sue connessioni con il resto del grafo cliccando sul nodo tenendo premuto alt.

In questo modo le informazioni renderizzate sono solo quelle che l'utente ha richiesto e non tutte quelle presenti nella network.

Gerarchie(1/2)

L'idea precedentemente spiegata risulta realizzabile perché sia negli agenti chimici che nei farmaci esistono delle gerarchie:
Esempio:

"Organic Chemicals :- Dichlorophen"
"Enzymes and Coenzymes :- Neopterin"

Tale gerarchia può essere aumentata con ulteriori livelli.

Gerarchie (2/2)

Nei nanomateriali e nelle malattie tali gerarchie non sono presenti.

- I nanomateriali possono essere visualizzati tutti in quanto sono solo 29.
- Le malattie sono circa 600, per cui si sconsiglia di iniziare l'esplorazione da questa categoria.

Features

La libreria progettata é stata dotata delle seguenti features:

- Disposizione dei nodi attraverso force layout.
- Espansione/Compressione dei nodi categoria tramite doppioclick.
- Messa in evidenza del vicinato di un nodo tramite shift+click.
- Visualizzazione a richiesta delle connessioni di un nodo con tutta la network tramite alt+click.
- Colorazione degli archi in base al valore. (verdi: positivi, rossi: negativi)
- Spessore degli archi in base al valore.
- Legenda in alto a sinistra per i nodi, a destra per gli archi.
- Label su gli archi.
- Label sui nodi.
- Drag and drop, release.

Implementazione

E' stato realizzato un pacchetto R contenente una generica implementazione di un HTMLWidget:

- I file R contengono le funzioni da richiamare all'interno del codice R. Tali funzioni prendono in input i dati da renderizzare e si occupano di processarle e trasferirle alla libreria di rendering.
- I file javascript contengono il codice per il rendering.

Il core della nostra implementazione è la parte di rendering. Abbiamo utilizzato una libreria di visualizzazione di grafi chiamata:

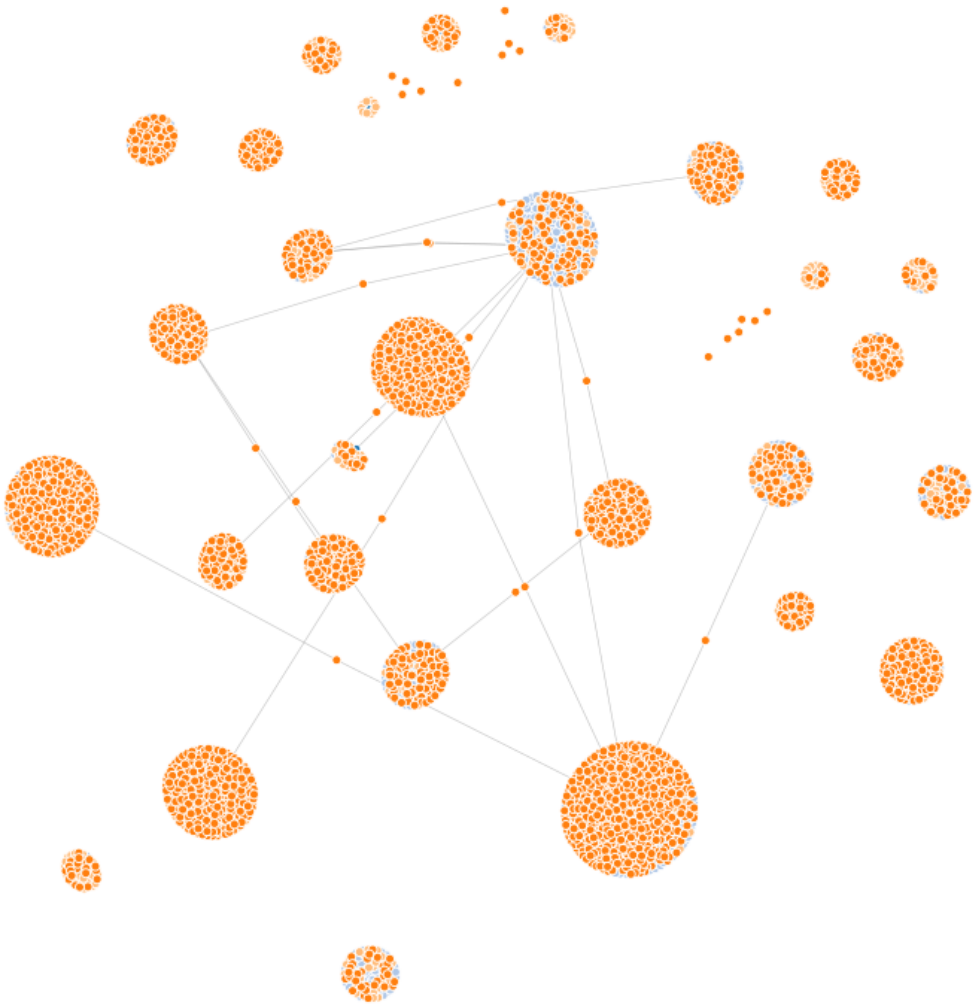
d3.js

Informazioni

Il pacchetto creato è installabile su qualsiasi sistema operativo avente R $\geq$ 3.2. Per installarlo e' sufficiente eseguire i seguenti comandi da terminale:

```
$ git clone https://github.com/Abelarm/BioInf_Project.git
$ cd BioInf_Project/graphexplorer
$ R
> devtools::install()
```

Nanocluster



Nanocluster

È stato creato un'ulteriore tool per la visualizzazione di un grafo corrispondente alla network clusterizzata base ai nanomateriali.

In questo caso, ogni nanomateriale é espandibile fino a due livelli:

- Nel primo livello, vengono visualizzati nodi fittizi corrispondenti ad ogni categoria collegata al nanomateriale.
- Nel secondo livello, vengono visualizzati i nodi effettivi di una certa categoria.

Nanocluster installation

Per installare il tool Nanocluster e' sufficiente eseguire i seguenti comandi da terminale:

```
$ git clone https://github.com/Abelarm/BioInf_Project.git
$ cd BioInf_Project/nanocluster
$ R
> devtools::install()
```

Overview

Considerazioni

- I pacchetti sviluppati non si prefiggono di essere strumenti a sé stanti.
- A partire dalla visualizzazione, risulta essere più semplice verificare esistenza e forza delle connessioni.
- Le intuizioni suggerite dalla visualizzazione grafica possono essere verificate interrogando i dati veri e propri.

Sviluppi futuri

- Rendere i tool più flessibili.
- Produrre la documentazione relativa a quanto sviluppato per sottomettere il pacchetto al CRAN.

Grazie per l’attenzione.