UNIVERSITÀ DELLA CALABRIA
DIMES
Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica

Tesi di laurea

«Ricerca di pattern su reti biologiche»

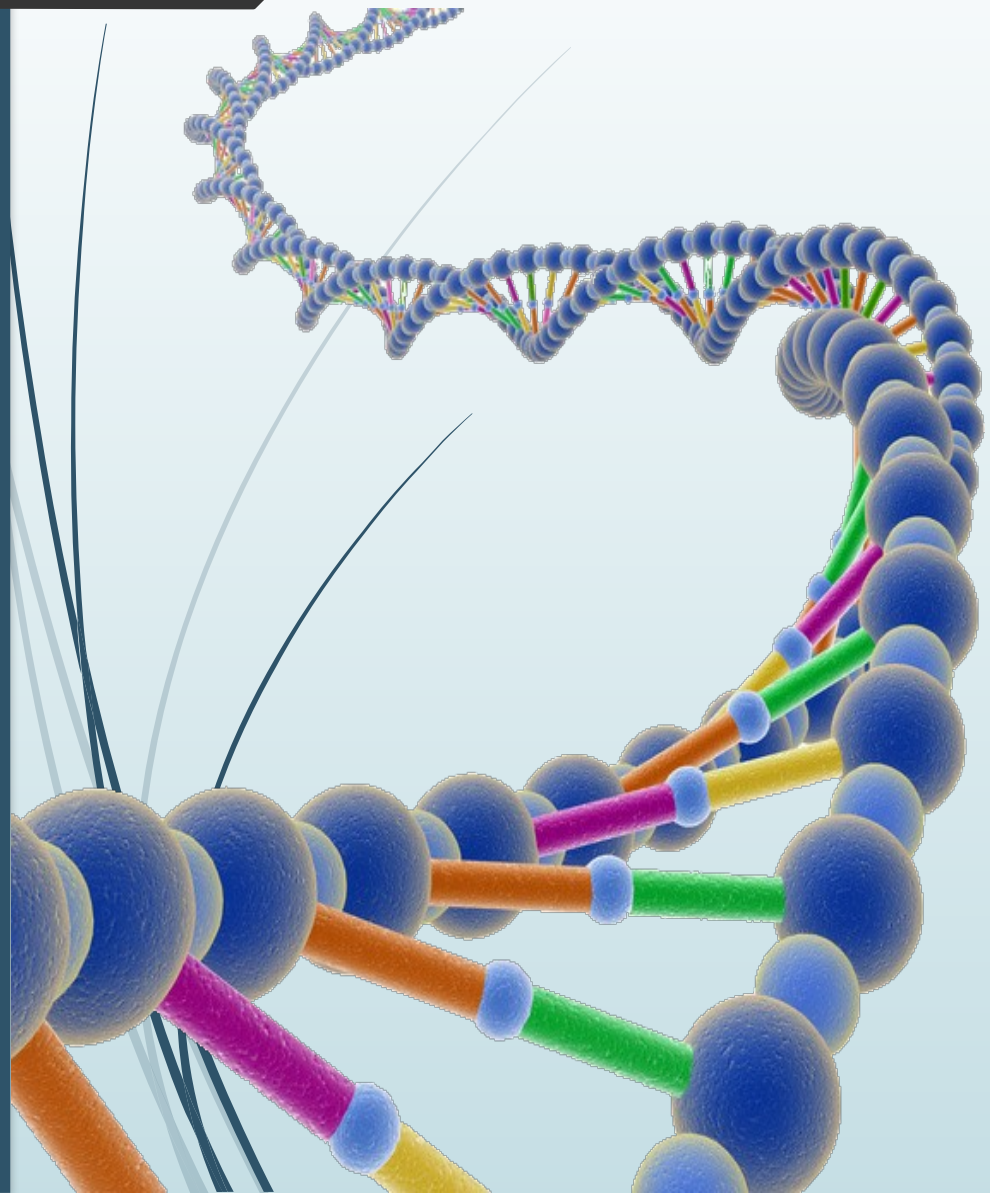
Relatore

Prof. Fabio Fassetti

Candidato

Ivonne Rizzuto
Matricola 167058

Sommario

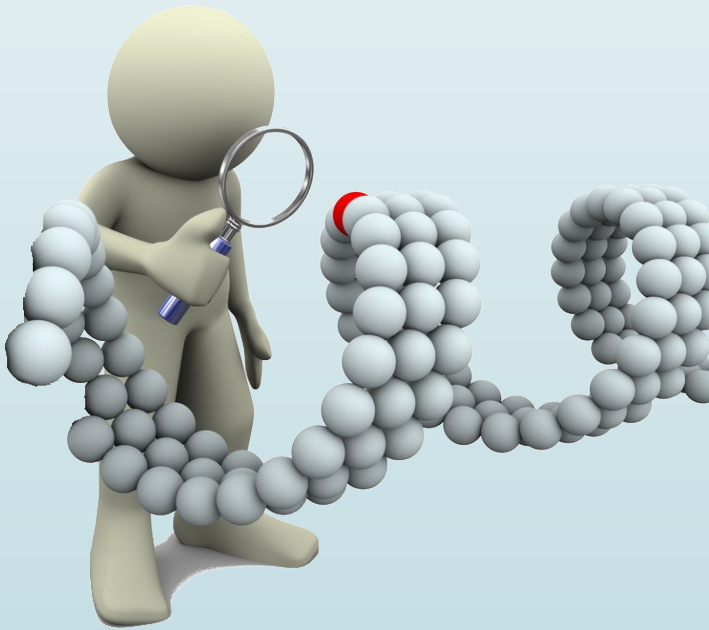


- Scopo dell’elaborato
- Definizione del modello: rete biologica
- Descrizione dei dati da utilizzare
- Tecnica adottata: Discriminative Pattern Mining
- Elaborazione dei dataset
- Analisi dei grafi associati
- Rappresentazione dei risultati

Obiettivo della tesi

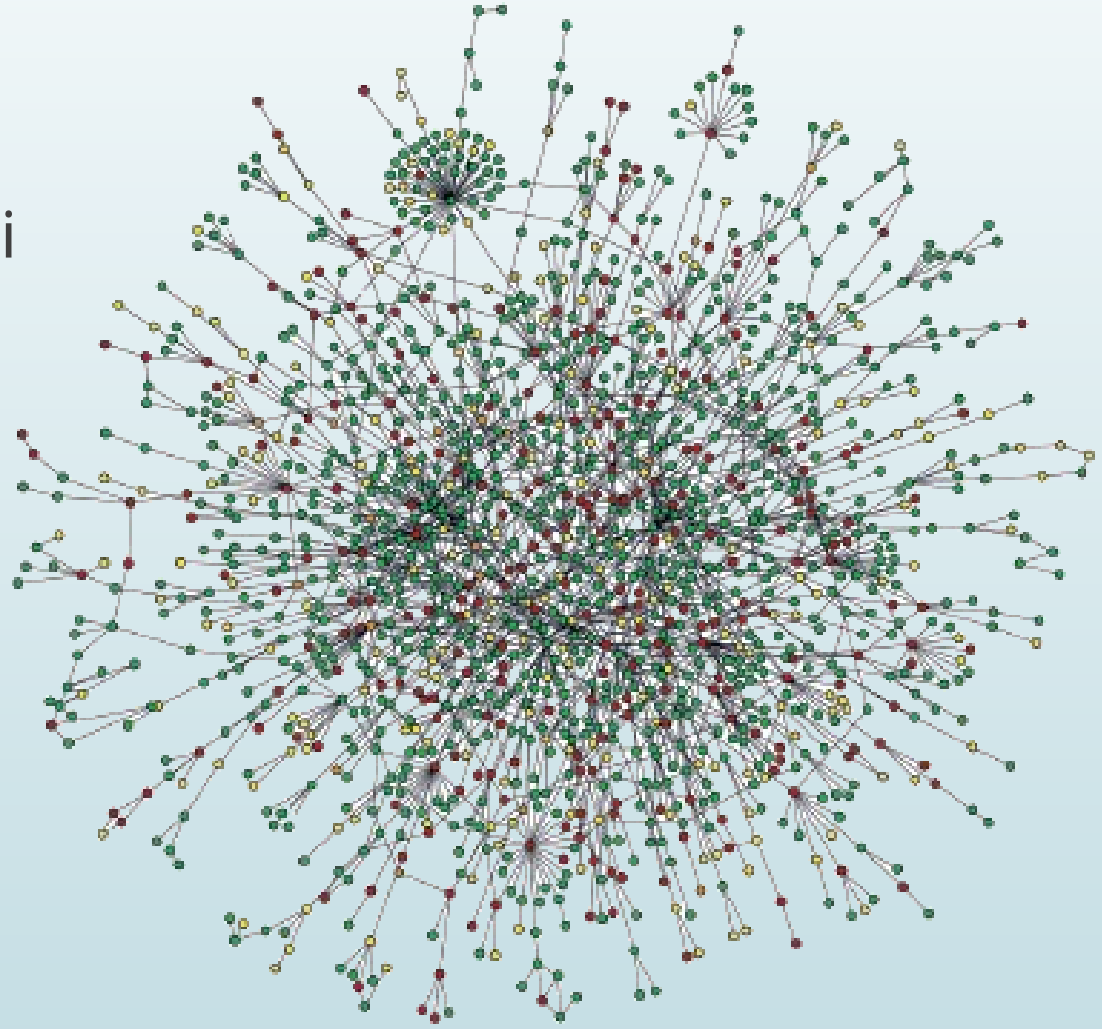
Si vuole effettuare l’analisi di dati relativi ai profili genici di alcuni campioni.

L’algoritmo utilizzato ricerca delle corrispondenze tra le informazioni, rappresentate come reti biologiche.



Rete Biologica

È oggetto di studio della Biologia dei Sistemi, per comprendere i meccanismi e le dinamiche alla base dei processi biologici che caratterizzano la vita di un organismo.

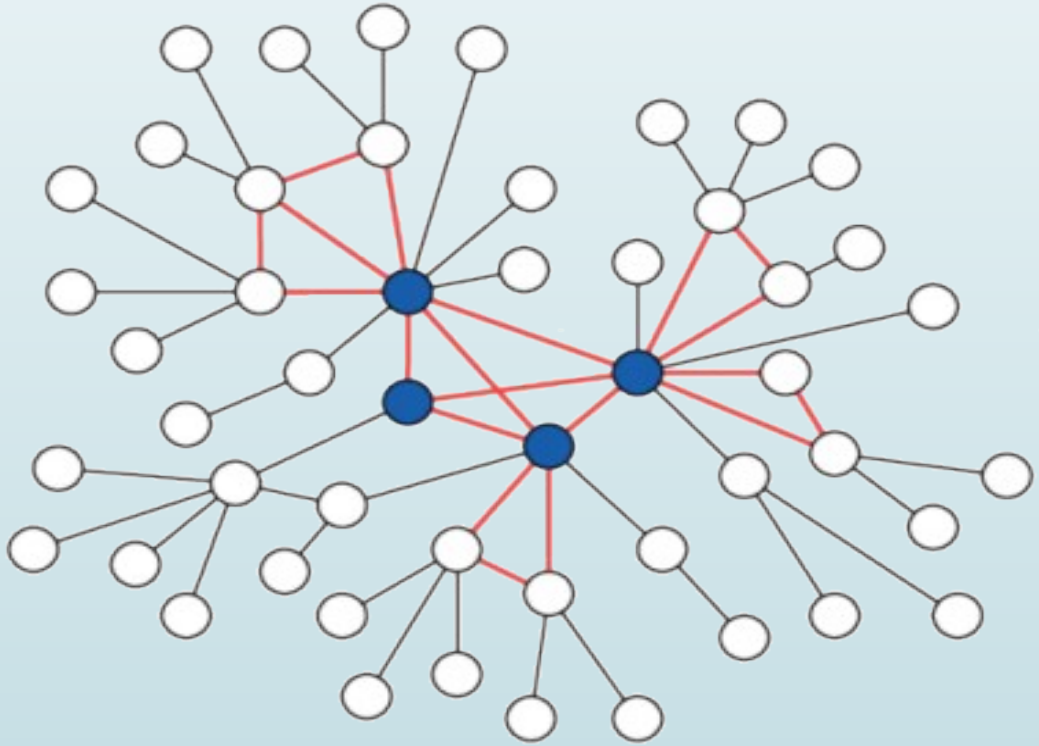


Modello: rete di interazione genica

È una rete complessa, descritta come grafo non orientato.

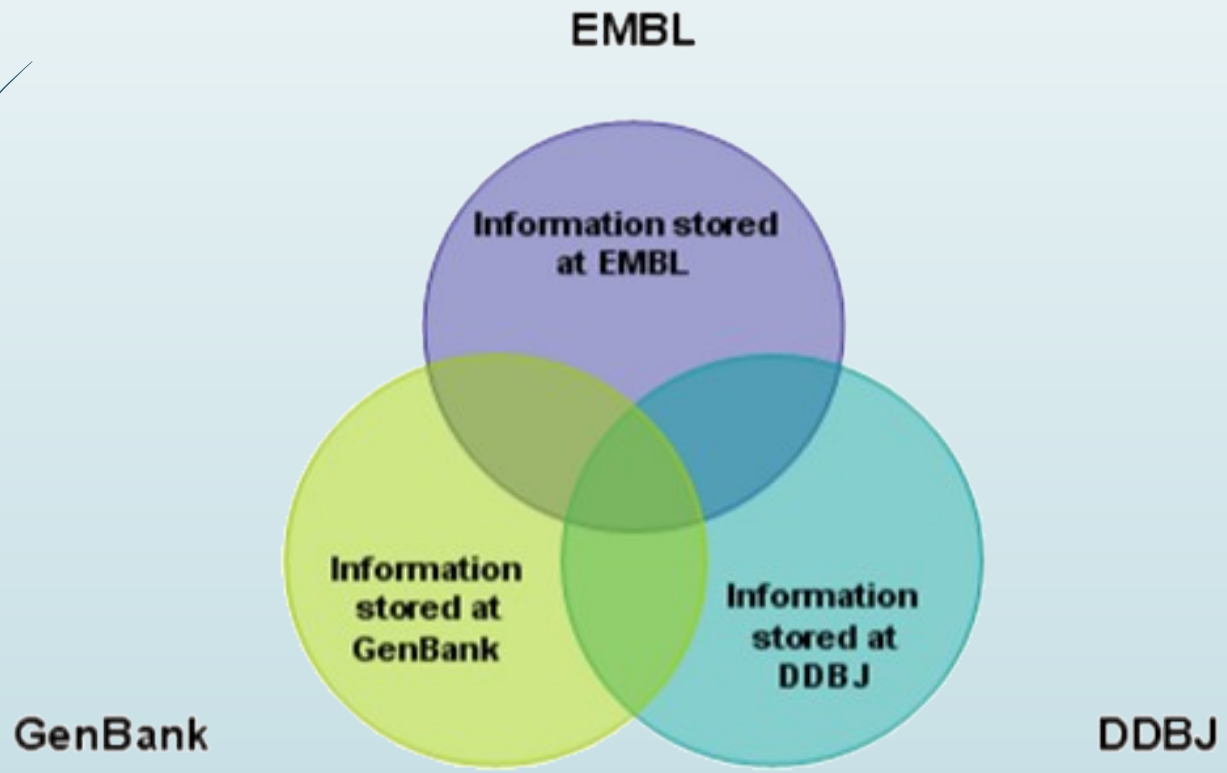
I nodi sono i geni o le proteine e gli archi esprimono le correlazioni che sussistono tra questi.

Su ogni arco ci saranno due pesi, dei valori indicativi del livello di coespressione e della significatività dell'interazione genica.



Dati analizzati

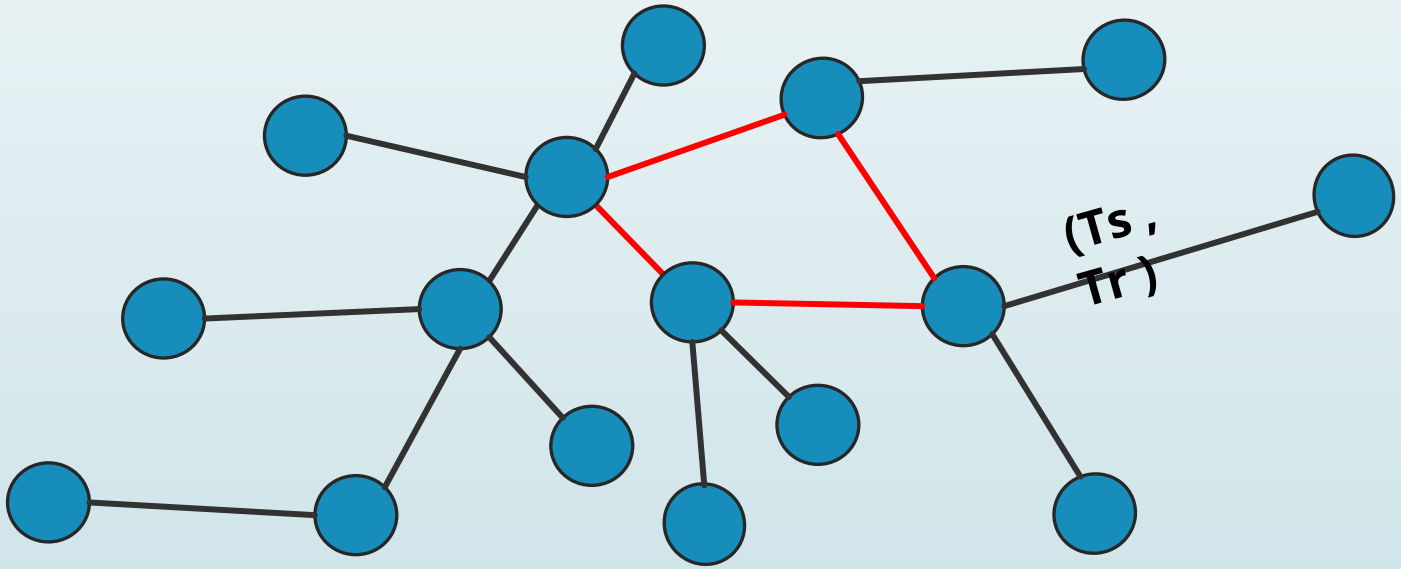
I campioni sono stati estratti dalle informazioni cliniche di pazienti, raccolte e condivise tramite tre ingenti banche dati.



Quelli utilizzati riguardano individui risultati affetti o meno da patologie come il diabete di tipo 2, la piorrea, l'obesità ed il cancro alla prostata.

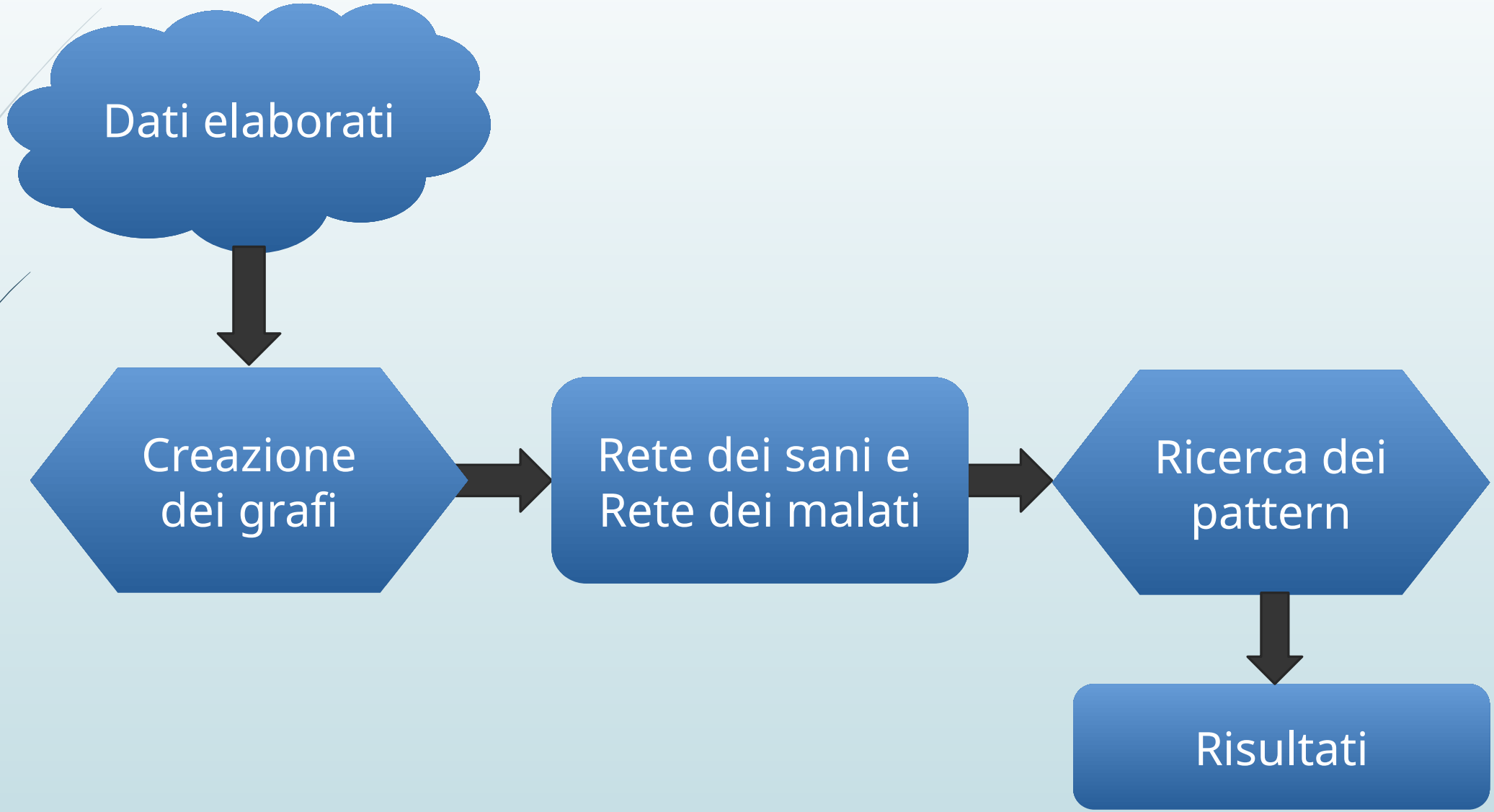
Ricerca di corrispondenze

Nei dati grezzi descritti come grafi connessi, si ricercano, usando la tecnica di Discriminating Pattern Mining, delle sottostrutture ricorrenti.



L’obiettivo è identificare delle regolarità, dei gruppi di geni uniti da legami valutati in termini di **Robustezza** e **Rilevanza**.

Algoritmo Discriminative Pattern Mining



Preparazione dei dati

Si esegue una fase di preprocessing per rendere i dati idonei ad essere elaborati dall'algoritmo.

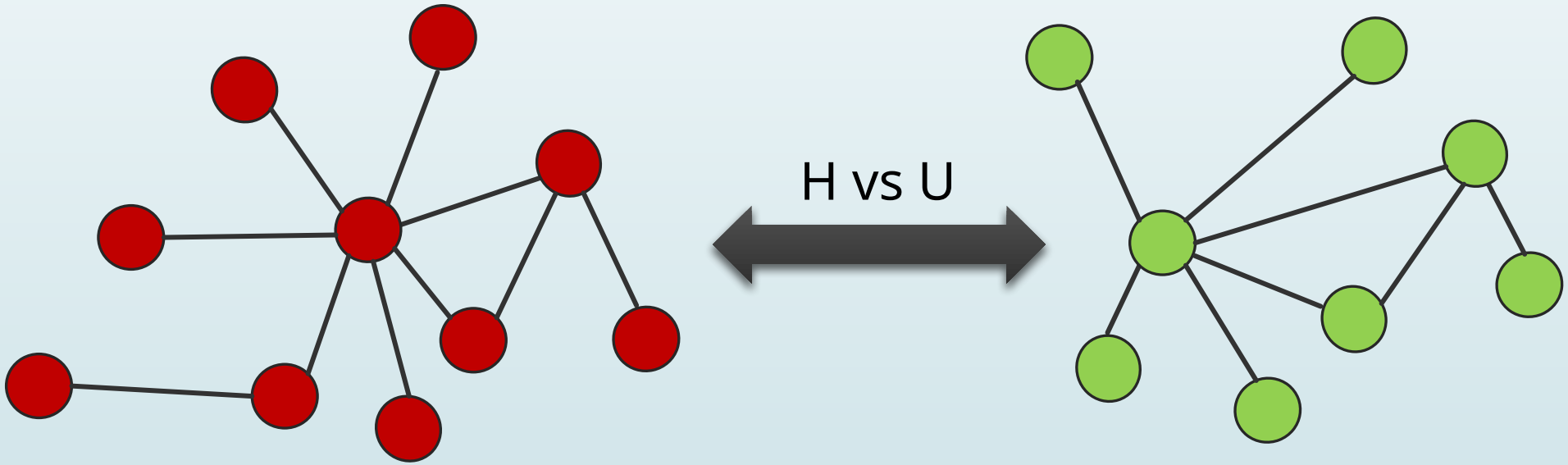
Tramite delle funzioni ad-hoc scritte in Java ed in Matlab,

si realizza:

- La conoscenza della situazione clinica dei pazienti;
- L'estrazione dei nomi dei geni identificati nei campioni;
- La suddivisione dei campioni in sani (H) e malati (U);
- La rappresentazione del livello di espressione genica in formato matriciale;

Costruzione delle reti

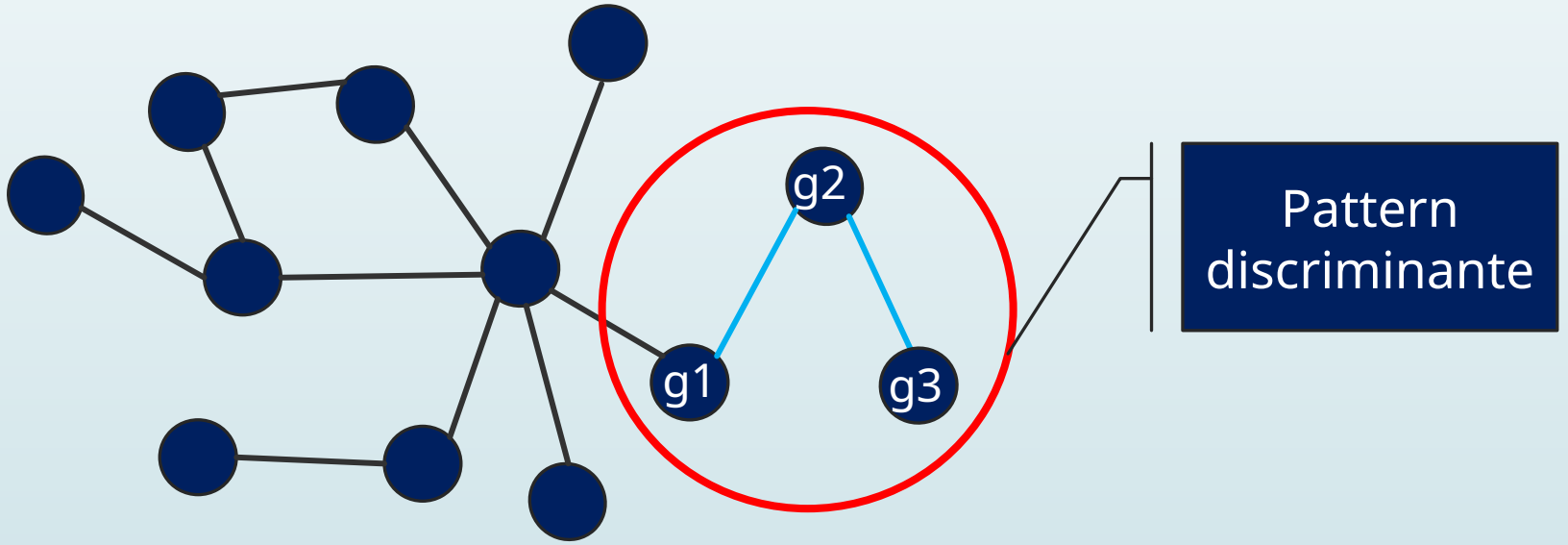
Si crea una rete per ciascuna delle due matrici, ottenendo un grafo dei pazienti malati ed uno dei pazienti sani, per ogni dataset.



L’arco tra due geni viene creato se la loro relazione soddisfa i valori di threshold specificati, cioè le soglie τs e τr fornite all’algoritmo.

Ricerca di pattern eccezionali

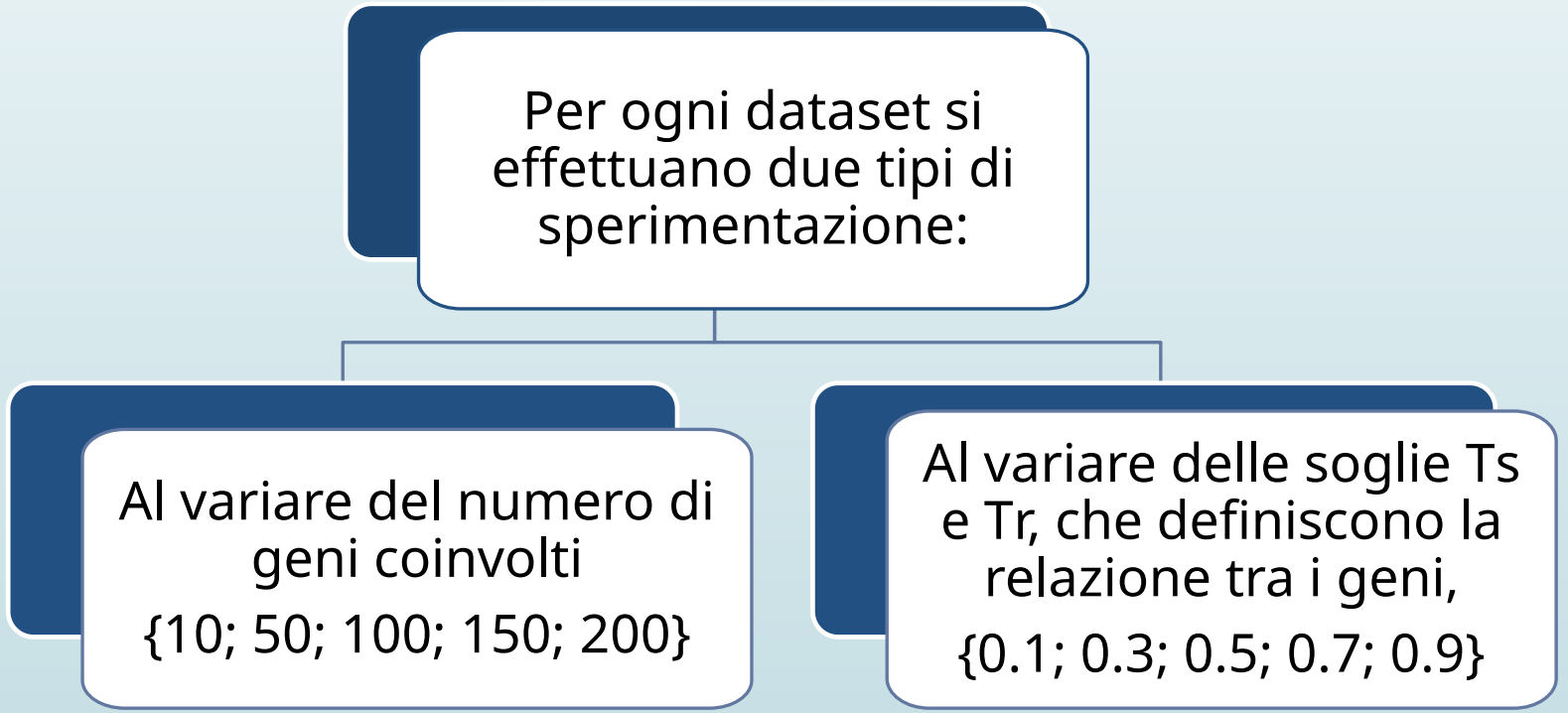
Si ricercano dei sottografi discriminanti, che siano ricorrenti e significativi in termini di livello di espressione genica .



Lo scopo è identificare le differenze fenotipiche nel codice genetico di un paziente sano rispetto a quello di un paziente malato.

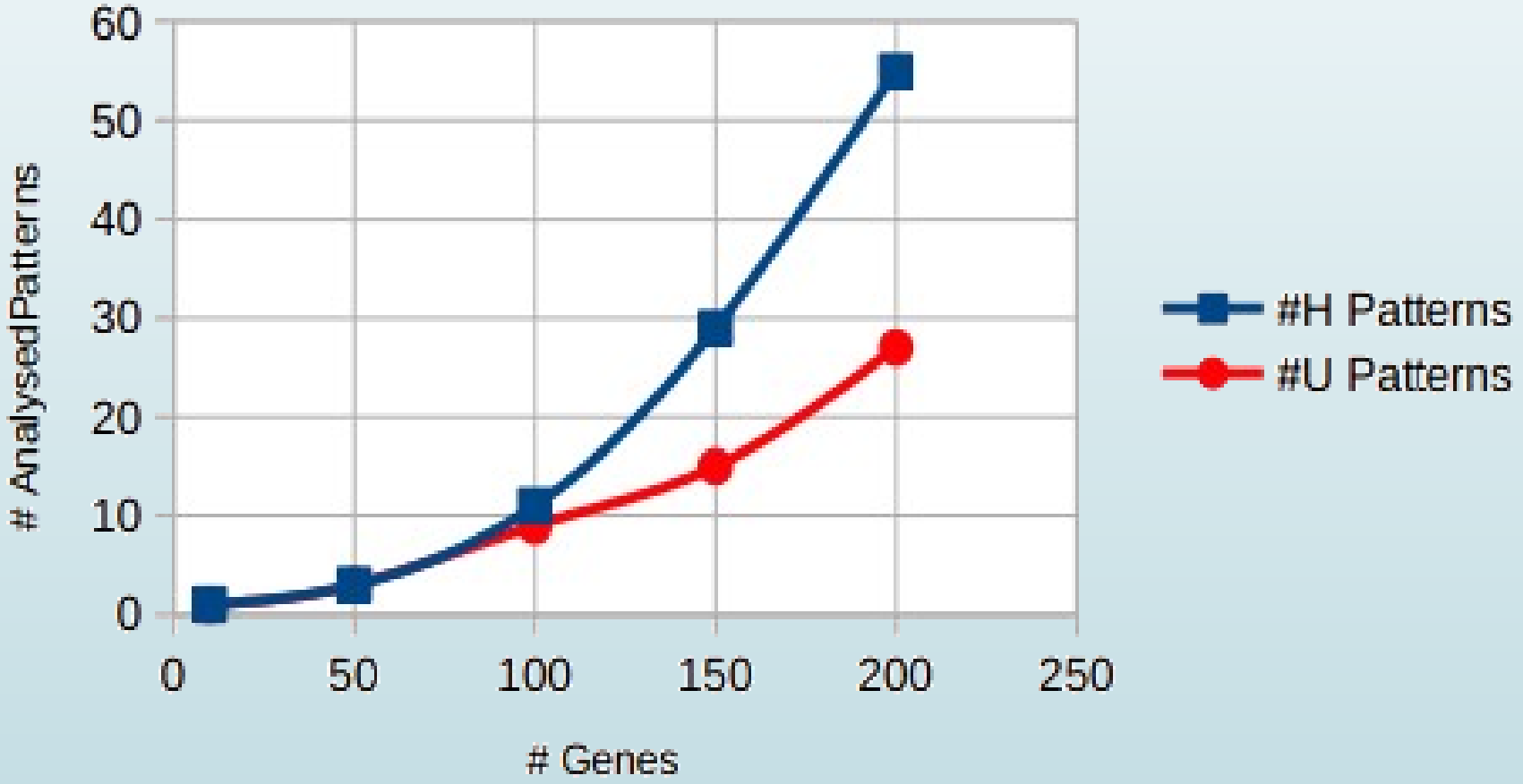
Generazione dei risultati

Si confronta il numero di pattern identificati nella rete dei sani con quello della rete dei malati, per raccogliere delle informazioni che contribuiscano ad identificare la predisposizione o l'insorgenza di una malattia.



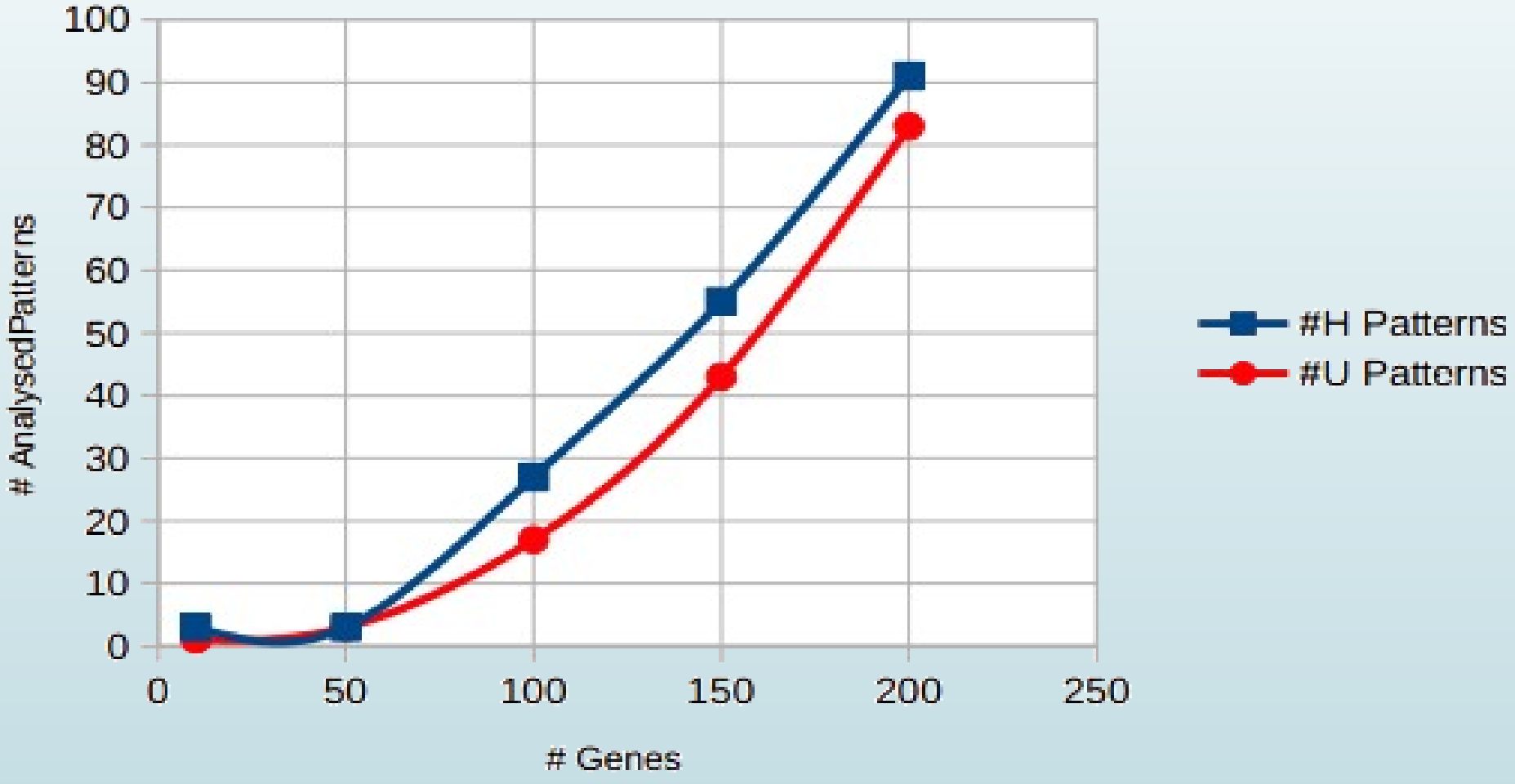
Risultati: Dataset GSE16134

Questi campioni sono stati estratti da 310 pazienti, 120 dei quali affetti da piorrea.

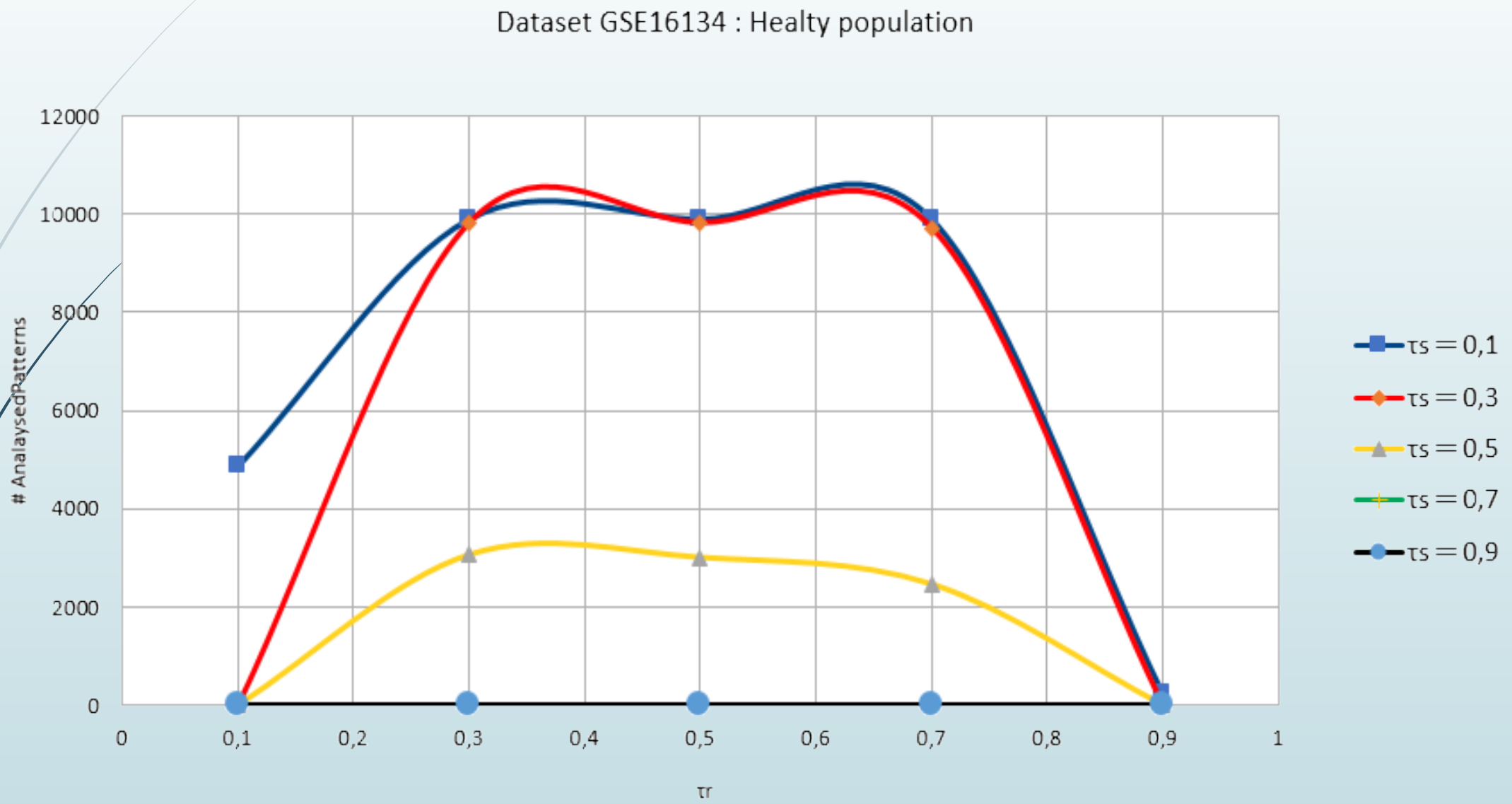


Risultati: Dataset GSE68907

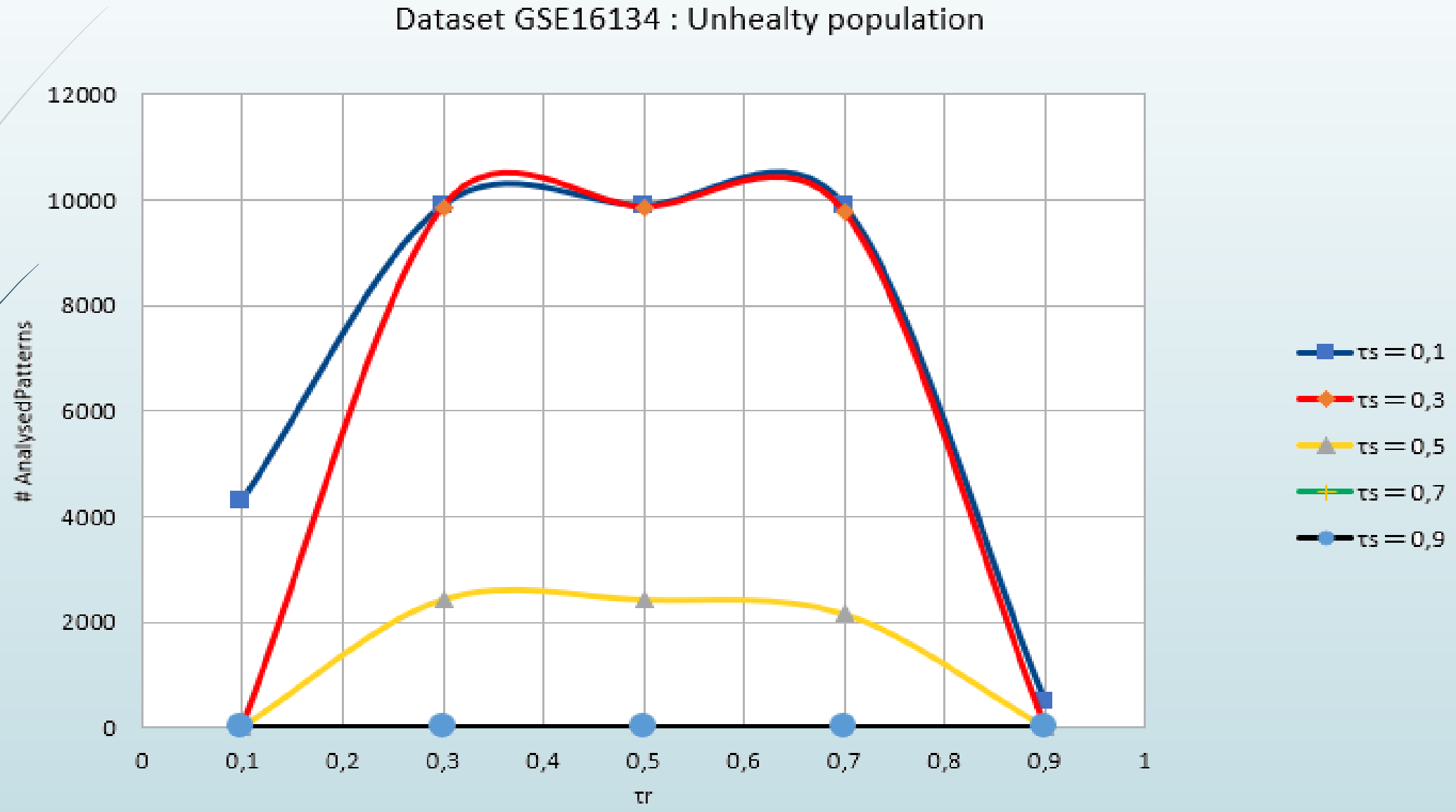
Questi campioni appartengono a pazienti in cui è stata ricercata la presenza del tumore alla prostata.



Risultati: Healty vs Unhealty



Risultati: Healty vs Unhealty



Grazie per l’attenzione!