



# Metodi di Ensemble Gerarchici per la Predizione Strutturata della Funzione delle Proteine

Relatore

*Prof. Giorgio Valentini*

Correlatore

*Dr. Marco Notaro*

Candidato

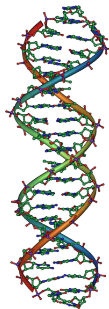
*Marco Odore*

10 Luglio 2018

# Central Dogma

---

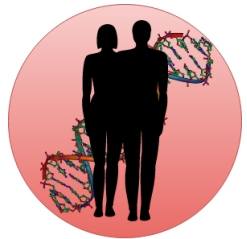
- All'interno delle molecole di DNA di ogni essere vivente esistono diverse migliaia di geni.



# Central Dogma

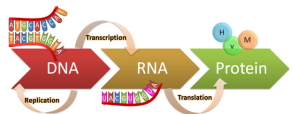
---

- All'interno delle molecole di DNA di ogni essere vivente esistono diverse migliaia di geni.
- Si stima che per l'essere umano il DNA possenga tra i 20.000 - 25.000 geni.



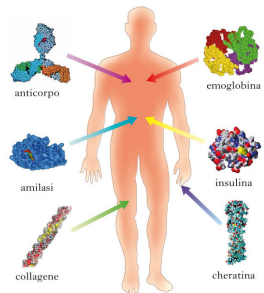
# Central Dogma

- All'interno delle molecole di DNA di ogni essere vivente esistono diverse migliaia di geni.
- Si stima che per l'essere umano il DNA possenga tra i 20.000 - 25.000 geni.
- Ogni gene all'interno del DNA è capace di codificare più proteine.



# Central Dogma

- All'interno delle molecole di DNA di ogni essere vivente esistono diverse migliaia di geni.
- Si stima che per l'essere umano il DNA possenga tra i 20.000 - 25.000 geni.
- Ogni gene all'interno del DNA è capace di codificare più proteine.
- Ogni proteina è responsabile di una o più funzioni all'interno delle cellule degli esseri viventi.



## Il problema della predizione della funzione delle proteine 1/4

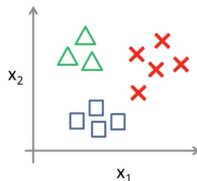
- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**



## Il problema della predizione della funzione delle proteine 1/4

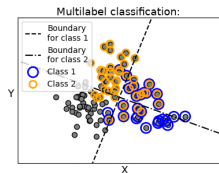
- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina (**problema multiclasse**)

Multi-class classification:



## Il problema della predizione della funzione delle proteine 1/4

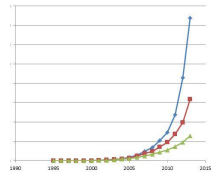
- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina (**problema multiclasse**)
- Ad ogni gene/proteina possono essere associate diverse funzioni contemporaneamente (**problema multietichetta**)





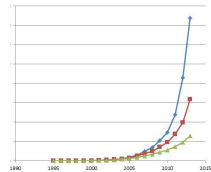
## Il problema della predizione della funzione delle proteine 1/4

- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina (**problema multiclasse**)
- Ad ogni gene/proteina possono essere associate diverse funzioni contemporaneamente (**problema multietichetta**)
- Il quantitativo di dati genomici cresce molto rapidamente.



## Il problema della predizione della funzione delle proteine 1/4

- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina (**problema multiclasse**)
- Ad ogni gene/proteina possono essere associate diverse funzioni contemporaneamente (**problema multietichetta**)
- Il quantitativo di dati genomici cresce molto rapidamente.



La **classificazione manuale** delle proteine è quindi infattibile.

## Il problema della predizione della funzione delle proteine 2/4

---

- A complicare ulteriormente il problema è il modo in cui sono *relazionate* tra loro le funzioni delle proteine.

## Il problema della predizione della funzione delle proteine 2/4

---

- A complicare ulteriormente il problema è il modo in cui sono *relazionate* tra loro le funzioni delle proteine.
- Esistono infatti due tassonomie principali per l'organizzazione delle classi:

## Il problema della predizione della funzione delle proteine 2/4

---

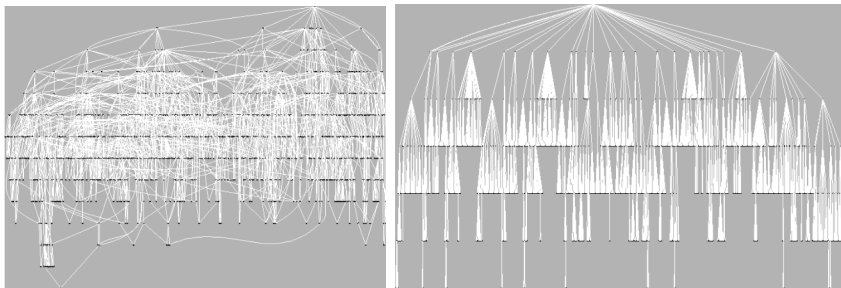
- A complicare ulteriormente il problema è il modo in cui sono *relazionate* tra loro le funzioni delle proteine.
- Esistono infatti due tassonomie principali per l'organizzazione delle classi:
  - **Gene Ontology** (GO): che organizza le funzioni come un grafo diretto aciclico (DAG), varia per ogni specie, e possiede tre ontologie differenti.

## Il problema della predizione della funzione delle proteine 2/4

---

- A complicare ulteriormente il problema è il modo in cui sono *relazionate* tra loro le funzioni delle proteine.
- Esistono infatti due tassonomie principali per l'organizzazione delle classi:
  - **Gene Ontology** (GO): che organizza le funzioni come un grafo diretto aciclico (DAG), varia per ogni specie, e possiede tre ontologie differenti.
  - **Functional Catalogue** (FunCat): che è organizzato invece come un albero, non varia in base alle specie, e descrive le funzioni in maniera più sintetica rispetto alla Gene Ontology.

## Il problema della predizione della funzione delle proteine 3/4



**Figura:** A sinistra un DAG della GO per la specie *S. cerevisiae*. A destra FunCat.

## Il problema della predizione della funzione delle proteine (GO) 4/4

---

- Data la granularità e specificità superiori della GO e il suo largo utilizzo nella comunità scientifica, all'interno della tesi ci si è soffermati sulla predizione delle sue funzioni.



## Il problema della predizione della funzione delle proteine (GO) 4/4

---

- Data la granularità e specificità superiori della GO e il suo largo utilizzo nella comunità scientifica, all'interno della tesi ci si è soffermati sulla predizione delle sue funzioni.
- Tale tassonomia presenta tre ontologie (e quindi tre DAG) principali:

## Il problema della predizione della funzione delle proteine (GO) 4/4

---

- Data la granularità e specificità superiori della GO e il suo largo utilizzo nella comunità scientifica, all'interno della tesi ci si è soffermati sulla predizione delle sue funzioni.
- Tale tassonomia presenta tre ontologie (e quindi tre DAG) principali:
  - **Processo Biologico** (BP): descrive i processi ad alto livello, come insieme di diverse attività molecolari.

## Il problema della predizione della funzione delle proteine (GO) 4/4

---

- Data la granularità e specificità superiori della GO e il suo largo utilizzo nella comunità scientifica, all'interno della tesi ci si è soffermati sulla predizione delle sue funzioni.
- Tale tassonomia presenta tre ontologie (e quindi tre DAG) principali:
  - **Processo Biologico** (BP): descrive i processi ad alto livello, come insieme di diverse attività molecolari.
  - **Funzione Molecolare** (MF): descrive le funzioni di specifici prodotti genici.

## Il problema della predizione della funzione delle proteine (GO) 4/4

---

- Data la granularità e specificità superiori della GO e il suo largo utilizzo nella comunità scientifica, all'interno della tesi ci si è soffermati sulla predizione delle sue funzioni.
- Tale tassonomia presenta tre ontologie (e quindi tre DAG) principali:
  - **Processo Biologico** (BP): descrive i processi ad alto livello, come insieme di diverse attività molecolari.
  - **Funzione Molecolare** (MF): descrive le funzioni di specifici prodotti genici.
  - **Componente Cellulare** (CC): il luogo all'interno della cellula nelle quali avviene la funzione genica.

## La predizione della funzione delle proteine tramite metodi automatici

---

Per gestire il problema della predizione della funzione delle proteine si rende quindi necessario un approccio **automatico**. I metodi più noti in letteratura sono:

## La predizione della funzione delle proteine tramite metodi automatici

---

Per gestire il problema della predizione della funzione delle proteine si rende quindi necessario un approccio **automatico**. I metodi più noti in letteratura sono:

- I metodi basati sulla **comparazione di biosequenze**: si basano sull'idea che sequenze simili condividano funzioni simili.

## La predizione della funzione delle proteine tramite metodi automatici

---

Per gestire il problema della predizione della funzione delle proteine si rende quindi necessario un approccio **automatico**. I metodi più noti in letteratura sono:

- I metodi basati sulla **comparazione di biosequenze**: si basano sull'idea che sequenze simili condividano funzioni simili.
- I metodi **basati su reti**: sono metodi applicati a dati rappresentati sotto forma di reti, che si basano sugli algoritmi di propagazione delle etichette.

## La predizione della funzione delle proteine tramite metodi automatici

---

Per gestire il problema della predizione della funzione delle proteine si rende quindi necessario un approccio **automatico**. I metodi più noti in letteratura sono:

- I metodi basati sulla **comparazione di biosequenze**: si basano sull'idea che sequenze simili condividano funzioni simili.
- I metodi **basati su reti**: sono metodi applicati a dati rappresentati sotto forma di reti, che si basano sugli algoritmi di propagazione delle etichette.
- I metodi **Kernel per spazi di output strutturato**: sono metodi che sfruttano funzioni kernel congiunte per predire in spazi di output strutturato.



## La predizione della funzione delle proteine tramite metodi automatici

---

Per gestire il problema della predizione della funzione delle proteine si rende quindi necessario un approccio **automatico**. I metodi più noti in letteratura sono:

- I metodi basati sulla **comparazione di biosequenze**: si basano sull'idea che sequenze simili condividano funzioni simili.
- I metodi **basati su reti**: sono metodi applicati a dati rappresentati sotto forma di reti, che si basano sugli algoritmi di propagazione delle etichette.
- I metodi **Kernel per spazi di output strutturato**: sono metodi che sfruttano funzioni kernel congiunte per predire in spazi di output strutturato.
- I metodi **Ensemble Gerarchici**: i metodi trattati in questa tesi.

# Metodi Ensemble Gerarchici 1/3

---

I Metodi di Ensemble Gerarchici sono metodi caratterizzati da due step principali:

# Metodi Ensemble Gerarchici 1/3

---

I Metodi di Ensemble Gerarchici sono metodi caratterizzati da due step principali:

1. *Predizione flat* delle diverse classi dell'ontologia, generando diversi predittori *indipendenti*.

# Metodi Ensemble Gerarchici 1/3

---

I Metodi di Ensemble Gerarchici sono metodi caratterizzati da due step principali:

1. *Predizione flat* delle diverse classi dell'ontologia, generando diversi predittori *indipendenti*.
2. *Combinazione e correzione gerarchica delle predizioni* sfruttando il DAG dei termini della GO.

## Metodi Ensemble Gerarchici 1/3

---

I Metodi di Ensemble Gerarchici sono metodi caratterizzati da due step principali:

1. *Predizione flat* delle diverse classi dell'ontologia, generando diversi predittori *indipendenti*.
2. *Combinazione e correzione gerarchica delle predizioni* sfruttando il DAG dei termini della GO.

Il secondo step rappresenta la componente *ensemble* del metodo. Tale step si rende necessario in quanto le predizioni flat non tengono in considerazione la struttura gerarchica dei DAG della GO, portando a risultati *inconsistenti*.

## Metodi Ensemble Gerarchici 2/3

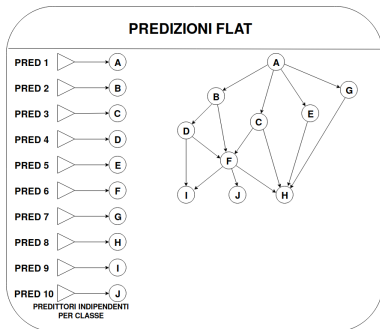
### Consistenza & True Path Rule

Un insieme di predizioni  $\hat{y} = \langle \hat{y}_1, \hat{y}_2, \dots, \hat{y}_{|N|} \rangle$ , dove  $|N|$  è la cardinalità dei termini della gerarchia, è definito *consistente*, se rispetta la *True Path Rule*, e cioè:

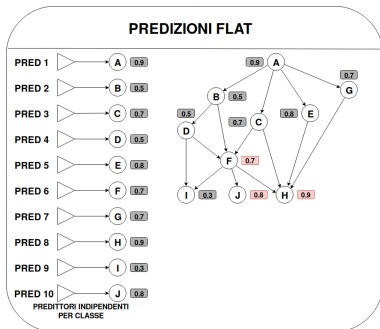
$$y \text{ consistente} \leftrightarrow \forall i \in N, j \in \text{par}(i) \rightarrow y_j \geq y_i$$

Dove  $\text{par}(i)$  indica l'insieme dei termini genitori del nodo  $i$  nella gerarchia.

## Metodi Ensemble Gerarchici (Esempio) 3/3

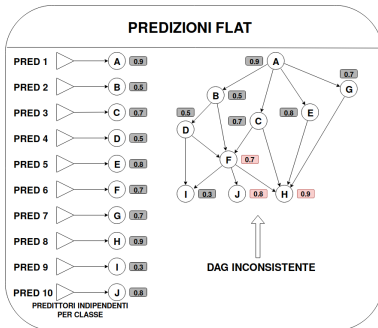


## Metodi Ensemble Gerarchici (Esempio) 3/3

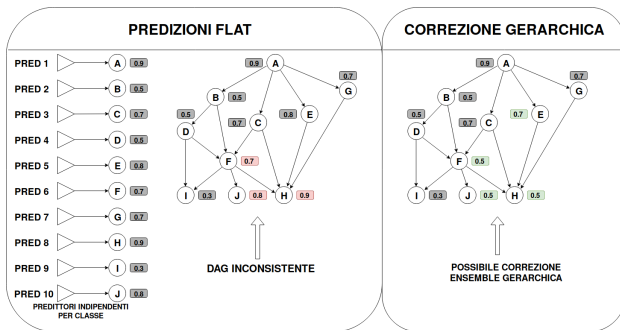




## Metodi Ensemble Gerarchici (Esempio) 3/3



## Metodi Ensemble Gerarchici (Esempio) 3/3



## Metodi Ensemble Gerarchici: Approcci

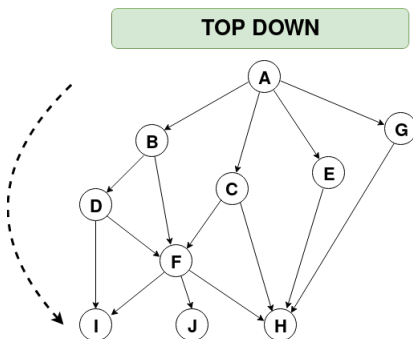
---

Esistono fondamentalmente due approcci per la correzione:

## Metodi Ensemble Gerarchici: Approcci

Esistono fondamentalmente due approcci per la correzione:

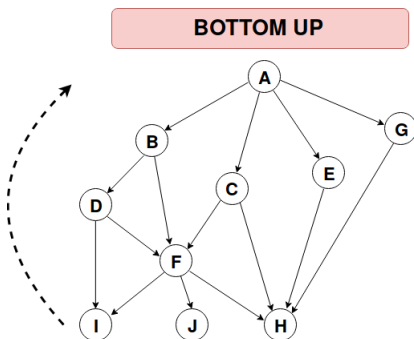
- *Top-down*: le predizioni vengono corrette dai nodi più generali a quelli più specifici.



## Metodi Ensemble Gerarchici: Approcci

Esistono fondamentalmente due approcci per la correzione:

- *Top-down*: le predizioni vengono corrette dai nodi più generali a quelli più specifici.
- *Bottom-up*: Le predizioni vengono corrette dai nodi più specifici verso quelli più generali.



## Metodo Top-Down Gerarchico (HTD-DAG)

---

- È un metodo che utilizza l'approccio *Top-Down*.

## Metodo Top-Down Gerarchico (HTD-DAG)

- È un metodo che utilizza l'approccio *Top-Down*.
- La correzione avviene ricorsivamente, percorrendo il grafo per *livelli*. Più precisamente, dato il grafo  $G = (N, E)$ , gli score flat  $f(x) = \hat{y}$  sono corretti gerarchicamente a  $\bar{y}$ , applicando la seguente regola:

### Aggiornamento con HTD-DAG

$$\bar{y}_i := \begin{cases} \hat{y}_i & \text{if } i \in \text{root}(G) \\ \min_{j \in \text{par}(i)} \bar{y}_j & \text{if } \min_{j \in \text{par}(i)} \bar{y}_j < \hat{y}_i \\ \hat{y}_i & \text{altrimenti} \end{cases}$$

Dove  $\text{par}(i)$  specifica i genitori del nodo  $i$ .

## Metodo True Path Rule per DAG (TPR-DAG)

---

- È un metodo che combina gli approcci top-down e bottom-up per la correzione delle predizioni flat.



## Metodo True Path Rule per DAG (TPR-DAG)

---

- È un metodo che combina gli approcci top-down e bottom-up per la correzione delle predizioni flat.
- È suddiviso in due step sequenziali:

## Metodo True Path Rule per DAG (TPR-DAG)

---

- È un metodo che combina gli approcci top-down e bottom-up per la correzione delle predizioni flat.
- È suddiviso in due step sequenziali:
  1. **Step bottom-up**: che partendo dai nodi più specifici del DAG, propaga quelle predizioni flat che sono considerate *positive*.

## Metodo True Path Rule per DAG (TPR-DAG)

---

- È un metodo che combina gli approcci top-down e bottom-up per la correzione delle predizioni flat.
- È suddiviso in due step sequenziali:
  1. **Step bottom-up**: che partendo dai nodi più specifici del DAG, propaga quelle predizioni flat che sono considerate *positive*.
  2. **Step top-down**: È il medesimo step utilizzato dal metodo HTD-DAG.

## Metodo True Path Rule per DAG (TPR-DAG)

---

- È un metodo che combina gli approcci top-down e bottom-up per la correzione delle predizioni flat.
- È suddiviso in due step sequenziali:
  1. **Step bottom-up**: che partendo dai nodi più specifici del DAG, propaga quelle predizioni flat che sono considerate *positive*.
  2. **Step top-down**: È il medesimo step utilizzato dal metodo HTD-DAG.
- Lo step top down si rende necessario in quanto la propagazione delle predizioni positive dal basso verso l'alto non garantisce la consistenza delle predizioni necessarie alla True Path Rule.

## Generalized Pool Adjacent Violator (GPAV) (1/4)

- Per il passo Top-Down dell'algoritmo TPR-DAG (e HTD) è possibile sfruttare gli algoritmi per la risoluzione dei problemi di **Isotonic Regression**:

### Isotonic Regression (caso generale con ordinamento parziale)

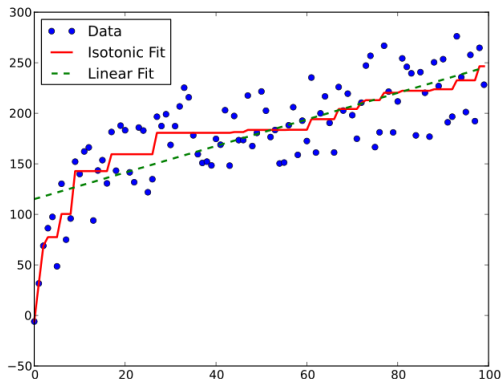
Dato un DAG,  $G(N, E)$ , con il set di nodi  $N = \{1, 2, \dots, n\}$ , si deve trovare il vettore  $x^* \in R^n$  tale che:

$$\min \sum_{i=1}^n w_i (x_i - a_i)^2$$

such that  $x_i \leq x_j \forall (i, j) \in E$

## Generalized Pool Adjacent Violator (GPAV) (2/4)

Esempio di Isotonic Regression con *ordinamento totale*:



## Generalized Pool Adjacent Violator (GPAV) (3/4)

---

Funzionamento Generale:

## Generalized Pool Adjacent Violator (GPAV) (3/4)

---

Funzionamento Generale:

- Per far funzionare l'algoritmo è necessario effettuare un *ordinamento topologico* del grafo.



## Generalized Pool Adjacent Violator (GPAV) (3/4)

Funzionamento Generale:

- Per far funzionare l'algoritmo è necessario effettuare un *ordinamento topologico* del grafo.
- L'algoritmo genera uno split del set di nodi  $N$  del DAG, in un insieme di *blocchi disgiunti definiti da  $H$*  (inizialmente  $H = N$ ).

## Generalized Pool Adjacent Violator (GPAV) (3/4)

Funzionamento Generale:

- Per far funzionare l'algoritmo è necessario effettuare un *ordinamento topologico* del grafo.
- L'algoritmo genera uno split del set di nodi  $N$  del DAG, in un insieme di *blocchi disgiunti definiti da  $H$*  (inizialmente  $H = N$ ).
- Seguendo l'ordinamento topologico del grafo, un blocco *assorbe* un suo blocco *predecessore* se si verificano determinate condizioni.

## Generalized Pool Adjacent Violator (GPAV) (3/4)

Funzionamento Generale:

- Per far funzionare l'algoritmo è necessario effettuare un *ordinamento topologico* del grafo.
- L'algoritmo genera uno split del set di nodi  $N$  del DAG, in un insieme di *blocchi disgiunti definiti da  $H$*  (inizialmente  $H = N$ ).
- Seguendo l'ordinamento topologico del grafo, un blocco *assorbe* un suo blocco *predecessore* se si verificano determinate condizioni.
- I nodi presenti nel medesimo blocco  $B_i$  *condividono il medesimo valore*  $x_i$ , e quindi a seguito dell'assorbimento sarà necessario un aggiornamento di tale valore.

## Generalized Pool Adjacent Violator (GPAV) (4/4)

---

**Algorithm 1** GPAV

---

```
1: procedure GPAV
2:    $H = N$ 
3:   for (each  $i \in N$ ) do
4:      $B_i = \{i\}$ 
5:      $B_i^- = i^-$ 
6:      $x_i = \hat{y}_i$ 
7:      $W_i = w_i$ 
8:   for  $k = 1, 2, \dots, n$  do
9:     //finché esiste un predecessore di  $B_k$  che viola la monotonicità
10:    while  $\{i \in B_k^- : x_i > x_k\} \neq \emptyset$  do
11:      // Trova l'elemento che viola maggiormente il vincolo
12:      Find  $j \in B_k^- : x_j = \max\{x_i : i \in B_k^-\}$ 
13:      Absorb( $k, j$ ) //  $j$  viene assorbito da  $B_k$ 
```

---

## Metodi GPAV e ISO-TPR

---

- Riassumendo, l'algoritmo effettua degli assorbimenti di blocchi adiacenti, finché questi violano i vincoli del problema quadratico, generando di fatto una partizione dei nodi, in cui le parti condividono lo stesso valore.

## Metodi GPAV e ISO-TPR

---

- Riassumendo, l'algoritmo effettua degli assorbimenti di blocchi adiacenti, finché questi violano i vincoli del problema quadratico, generando di fatto una partizione dei nodi, in cui le parti condividono lo stesso valore.
- L'algoritmo GPAV ha una complessità pari a  $O(n^2)$ .

## Metodi GPAV e ISO-TPR

---

- Riassumendo, l'algoritmo effettua degli assorbimenti di blocchi adiacenti, finché questi violano i vincoli del problema quadratico, generando di fatto una partizione dei nodi, in cui le parti condividono lo stesso valore.
- L'algoritmo GPAV ha una complessità pari a  $O(n^2)$ .
- Sostituendo quindi GPAV allo step Top-Down dell'algoritmo TPR-DAG visto in precedenza (invece che HTD), si ottiene l'algoritmo ISO-TPR.

## Predizione della funzione delle proteine della specie *C.elegans* (WORM) 1/4

---

Il dataset del nostro problema:



Il dataset del nostro problema:

- Si è eseguita la sperimentazione sul genoma della specie *Caenorhabditis elegans* (WORM), un organismo modello molto semplice, utilizzato frequentemente in biologia, soprattutto per studi di biologia dello sviluppo.

Il dataset del nostro problema:

- Si è eseguita la sperimentazione sul genoma della specie *Caenorhabditis elegans* (WORM), un organismo modello molto semplice, utilizzato frequentemente in biologia, soprattutto per studi di biologia dello sviluppo.
- L'insieme delle istanze utilizzato come input dai diversi predittori per il nostro problema, è dato dall'insieme dei geni della specie oggetto della sperimentazione, il quale è rappresentato da una matrice simmetrica generata dal network di interazione proteina-proteina estratto dal database *STRING* (Search Tool for the Retrieval of Interacting Genes/Proteins).

## Predizione della funzione delle proteine della specie C.elegans (WORM) 4/4

---

Un po' di info sul dataset:

Un po' di info sul dataset:

- Nel caso specifico della specie *C.elegans*, la matrice STRING ha dimensione  $15752 \times 15752$ . Il nostro problema ha quindi 15752 istanze.

Un po' di info sul dataset:

- Nel caso specifico della specie C.elegans, la matrice STRING ha dimensione  $15752 \times 15752$ . Il nostro problema ha quindi 15752 istanze.
- Per le classi funzionali/termini e annotazioni, queste cambiano in base al DAG di riferimento:

| <b>ontologia</b> | <b>numero di termini</b> | <b>numero di archi</b> |
|------------------|--------------------------|------------------------|
| BP               | 4068                     | 8066                   |
| MF               | 1163                     | 1567                   |
| CC               | 578                      | 1082                   |

**Tabella:** Statistiche relative ai DAG delle annotazioni della specie C.elegans.

## Predizione della funzione delle proteine della specie *C.elegans* (WORM) 4/4

Per evitare di avere problemi nella cross-validazione, i DAG sono stati ridotti a quei termini per cui si hanno almeno 10 annotazioni. Un po' di statistiche a seguito della selezione:

| Onto | numero di termini | media | d.std. | massimo | minimo |
|------|-------------------|-------|--------|---------|--------|
| BP   | 1335              | 71,33 | 151,68 | 2597    | 10     |
| MF   | 186               | 61,23 | 191,84 | 1806    | 10     |
| CC   | 221               | 131,9 | 302,25 | 1924    | 10     |

**Tabella:** La colonna *numero di termini* indica il numero di termini ottenuti dopo la selezione, *media* la media delle annotazioni per classe, per l'ontologia di riferimento, *d.std.* la deviazione standard delle annotazioni per l'ontologia di riferimento, *massimo* e *minimo* rispettivamente il massimo e minimo numero di annotazioni.