



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

Relatore

*Prof. Giorgio Valentini*

Correlatore

*Dr. Marco Notaro*

Candidato

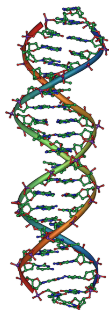
*Marco Odore*

10 Luglio 2018

# Central Dogma

---

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



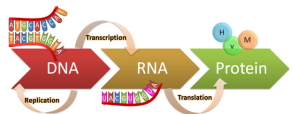
# Central Dogma

- 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.



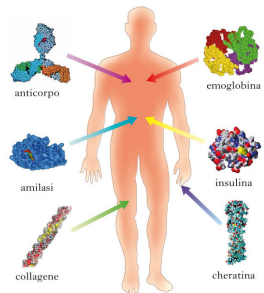
# Central Dogma

- 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.



# La funzione delle proteine

---

Le proteine sono molecole biologiche composte da amminoacidi, e le funzioni che svolgono sono molteplici:

# La funzione delle proteine

---

Le proteine sono molecole biologiche composte da amminoacidi, e le funzioni che svolgono sono molteplici:

- Metaboliche, ad esempio per la produzione di energia

# La funzione delle proteine

---

Le proteine sono molecole biologiche composte da amminoacidi, e le funzioni che svolgono sono molteplici:

- Metaboliche, ad esempio per la produzione di energia
- Di trascrizione, sintesi e processamento delle proteine stesse



# La funzione delle proteine

---

Le proteine sono molecole biologiche composte da amminoacidi, e le funzioni che svolgono sono molteplici:

- Metaboliche, ad esempio per la produzione di energia
- Di trascrizione, sintesi e processamento delle proteine stesse
- Di trasporto

# La funzione delle proteine

---

Le proteine sono molecole biologiche composte da amminoacidi, e le funzioni che svolgono sono molteplici:

- Metaboliche, ad esempio per la produzione di energia
- Di trascrizione, sintesi e processamento delle proteine stesse
- Di trasporto
- Di comunicazione intra o intercellulare

# La funzione delle proteine

---

Le proteine sono molecole biologiche composte da amminoacidi, e le funzioni che svolgono sono molteplici:

- Metaboliche, ad esempio per la produzione di energia
- Di trascrizione, sintesi e processamento delle proteine stesse
- Di trasporto
- Di comunicazione intra o intercellulare
- Di ciclo della cellula, ad esempio per la divisione e riproduzione cellulare

# La funzione delle proteine

---

Le proteine sono molecole biologiche composte da amminoacidi, e le funzioni che svolgono sono molteplici:

- Metaboliche, ad esempio per la produzione di energia
- Di trascrizione, sintesi e processamento delle proteine stesse
- Di trasporto
- Di comunicazione intra o intercellulare
- Di ciclo della cellula, ad esempio per la divisione e riproduzione cellulare

Per molte specie le funzioni di moltissimi geni (e quindi delle corrispettive proteine codificate) è **sconosciuta o parzialmente nota**.

## Il problema della predizione della funzione delle proteine 1/6

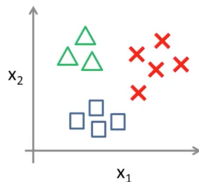
- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**



## Il problema della predizione della funzione delle proteine 1/6

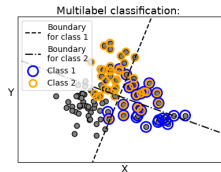
- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina (**problema multiclasse**)

Multi-class classification:



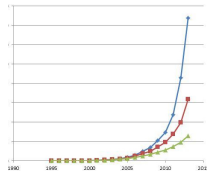
## Il problema della predizione della funzione delle proteine 1/6

- 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 1/6

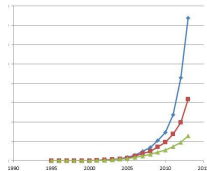
- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina (**problema multiclasse**)
- Ad ogni gene/proteina possono essere associate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





## Il problema della predizione della funzione delle proteine 1/6

- 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 2/6

---

- 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 2/6

---

- 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 2/6

---

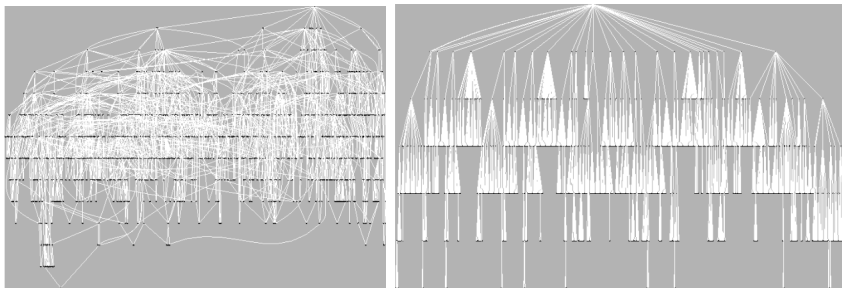
- A complicare ulteriormente il problema è il modo in cui sono *relazionate* tra loro le funzioni delle proteine.
- Esistono infatti due tassonomie principali per l'organizzazione delle classi:
  - **Gene Ontology** (GO): che organizza le funzioni come un grafo diretto aciclico (DAG), varia per ogni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

## Il problema della predizione della funzione delle proteine 2/6

---

- 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 3/6



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

## Il problema della predizione della funzione delle proteine (GO) 4/6

---

- 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) 4/6

---

- 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) 4/6

---

- 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) 4/6

---

- 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) 4/6

---

- 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.

## Il problema della predizione della funzione delle proteine (GO) 5/6

---

Gli archi dei DAG nella GO indicano inoltre 3 differenti tipi di relazione:

## Il problema della predizione della funzione delle proteine (GO) 5/6

---

Gli archi dei DAG nella GO indicano inoltre 3 differenti tipi di relazione:

- *is a*: indica una relazione di sotto-tipo.

## Il problema della predizione della funzione delle proteine (GO) 5/6

---

Gli archi dei DAG nella GO indicano inoltre 3 differenti tipi di relazione:

- *is a*: indica una relazione di sotto-tipo.
- *a part of*: indica una relazione di composizione

## Il problema della predizione della funzione delle proteine (GO) 5/6

---

Gli archi dei DAG nella GO indicano inoltre 3 differenti tipi di relazione:

- *is a*: indica una relazione di sotto-tipo.
- *a part of*: indica una relazione di composizione
- *regulates*: indica una relazione di influenza/regolazione

## Il problema della predizione della funzione delle proteine (GO) 6/6

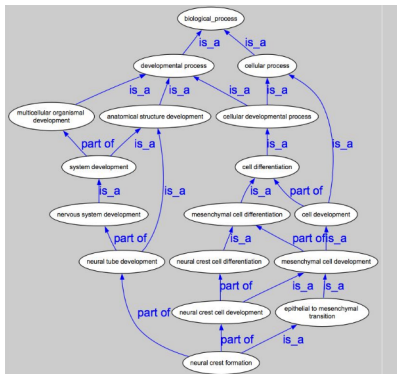


Figura: Un porzione di DAG che evidenzia i diversi tipi di relazione.



## 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

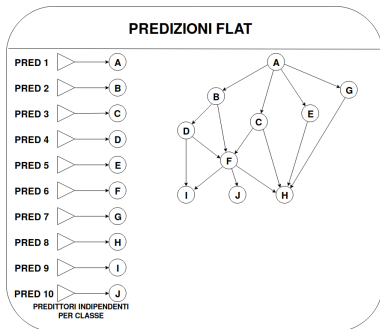
### Consistenza & True Path Rule

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:

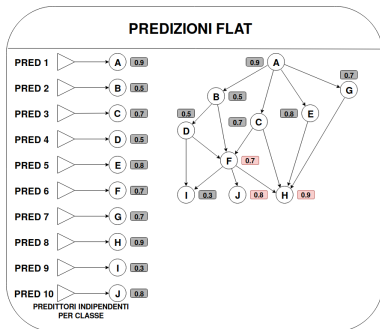
$$y \text{ consistente} \leftrightarrow \forall i \in N, j \in \text{par}(i) \rightarrow y_j \geq y_i$$

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

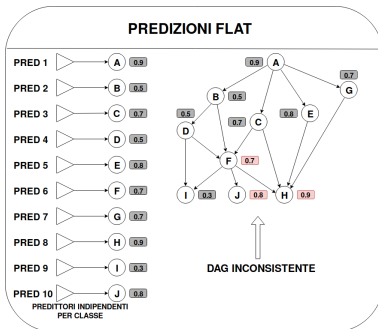
## Metodi Ensemble Gerarchici (Esempio) 3/3



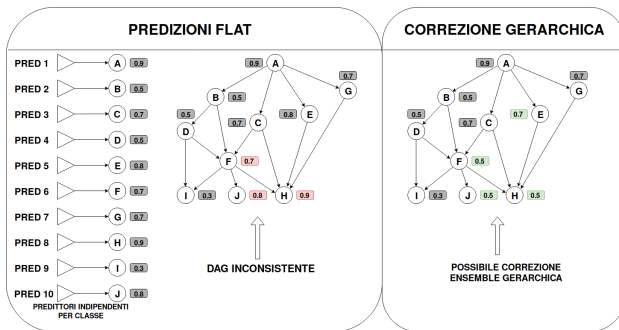
## Metodi Ensemble Gerarchici (Esempio) 3/3



## Metodi Ensemble Gerarchici (Esempio) 3/3



## Metodi Ensemble Gerarchici (Esempio) 3/3



# Metodi Ensemble Gerarchici: Approcci

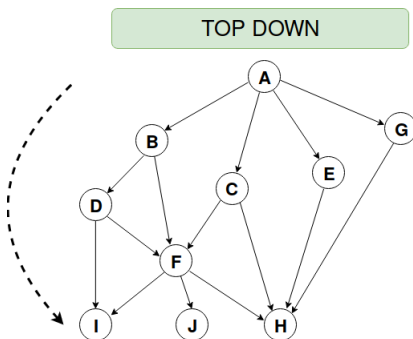
---

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.

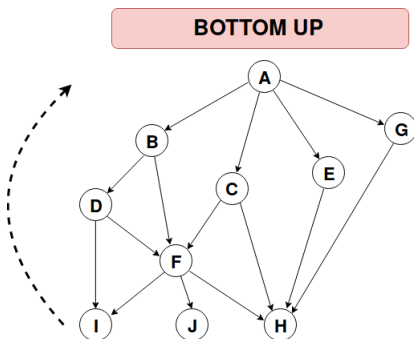




## Metodi Ensemble Gerarchici: Approcci

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) (1/2)

---

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

## Metodo Top-Down Gerarchico (HTD-DAG) (1/2)

- 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 Top-Down Gerarchico (HTD-DAG) (2/2)

---

- Per garantire la correttezza e consistenza delle correzioni, i *livelli* del grafo sono definiti come *cammino massimo dalla radice*.

## Metodo Top-Down Gerarchico (HTD-DAG) (2/2)

- Per garantire la correttezza e consistenza delle correzioni, i *livelli* del grafo sono definiti come *cammino massimo dalla radice*.
- Più formalmente, dobbiamo definire una funzione  $\psi$  che, applicata ad un nodo  $i \in N$ , restituisce il livello associato al cammino massimo, cioè:

### Livelli

$$\psi(i) = \max_{p(i,r)} l(p(r,i))$$

dove la funzione  $p(r,i)$  calcola il cammino dalla radice  $r$  al nodo  $i$  e la funzione  $l$  restituisce il livello associato ad un cammino.