Università Degli Studi Di Milano

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

Relatore

Prof. Giorgio Valentini

Correlatore

Dr. Marco Notaro

Candidato

Marco Odore

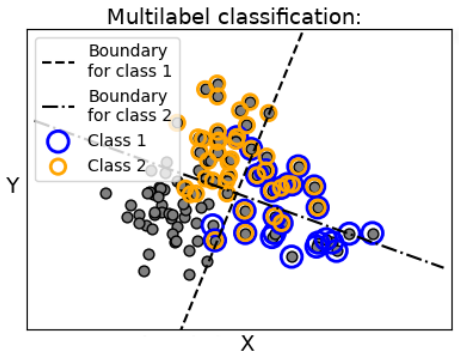
10 Luglio 2018

Il problema della predizione della funzione delle proteine

- Identificare la funzione delle proteine attraverso le analisi di laboratorio è **costosa** e richiede **molto tempo**

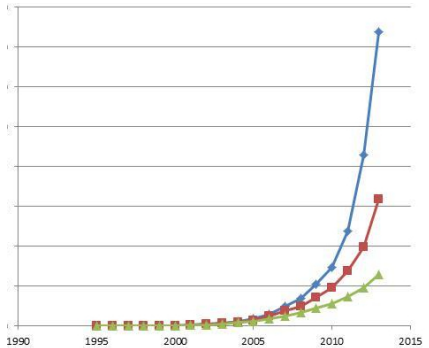
Il problema della predizione della funzione delle proteine

- Identificare la funzione delle proteine attraverso le analisi di laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina, anche contemporaneamente **(problema multiclasse e multietichetta)**



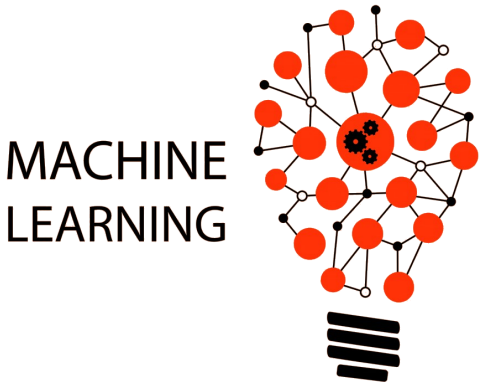
Il problema della predizione della funzione delle proteine

- Identificare la funzione delle proteine attraverso le analisi di laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina, anche contemporaneamente **(problema multiclasse e multietichetta)**
- Il quantitativo di dati genomici cresce molto rapidamente.



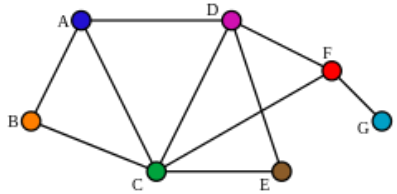
Il problema della predizione della funzione delle proteine

- Identificare la funzione delle proteine attraverso le analisi di laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina, anche contemporaneamente **(problema multiclasse e multietichetta)**
- Il quantitativo di dati genomici cresce molto rapidamente.
- La **classificazione manuale** delle proteine è quindi infattibile. È necessario quindi un approccio **automatico**.



Il problema della predizione della funzione delle proteine

- Identificare la funzione delle proteine attraverso le analisi di laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina, anche contemporaneamente **(problema multiclasse e multietichetta)**
- Il quantitativo di dati genomici cresce molto rapidamente.
- La **classificazione manuale** delle proteine è quindi infattibile. È necessario quindi un approccio **automatico**.
- A complicare ulteriormente il problema è il modo in cui sono *relazionate* tra loro le funzioni delle proteine.



Tassonomie per le funzioni delle proteine

- Esistono infatti due tassonomie principali per l'organizzazione delle funzioni:

Tassonomie per le funzioni delle proteine

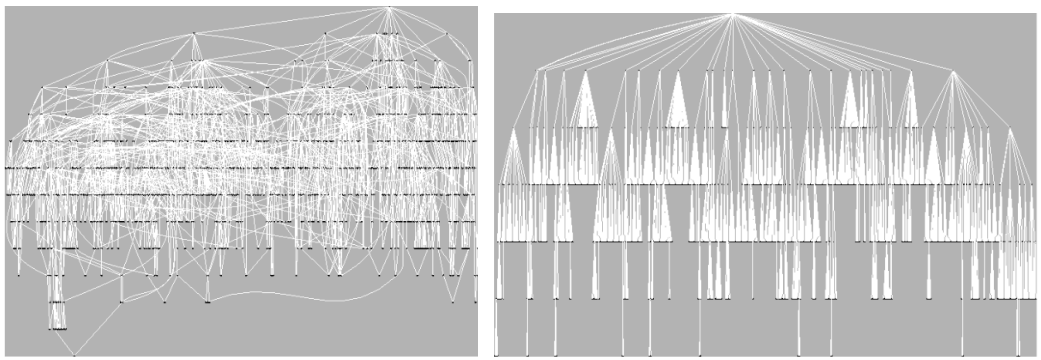
- Esistono infatti due tassonomie principali per l'organizzazione delle funzioni:
  - **Gene Ontology** (GO): che organizza le funzioni come un grafo diretto aciclico (DAG), varia per ogni specie, e possiede tre ontologie differenti (e quindi 3 DAG), e cioè *Biological Process* (BP), *Molecular Function* (MF) e *Cellular Component* (CC).

Tassonomie per le funzioni delle proteine

- Esistono infatti due tassonomie principali per l'organizzazione delle funzioni:
  - **Gene Ontology** (GO): che organizza le funzioni come un grafo diretto aciclico (DAG), varia per ogni specie, e possiede tre ontologie differenti (e quindi 3 DAG), e cioè *Biological Process* (BP), *Molecular Function* (MF) e *Cellular Component* (CC).
  - **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.

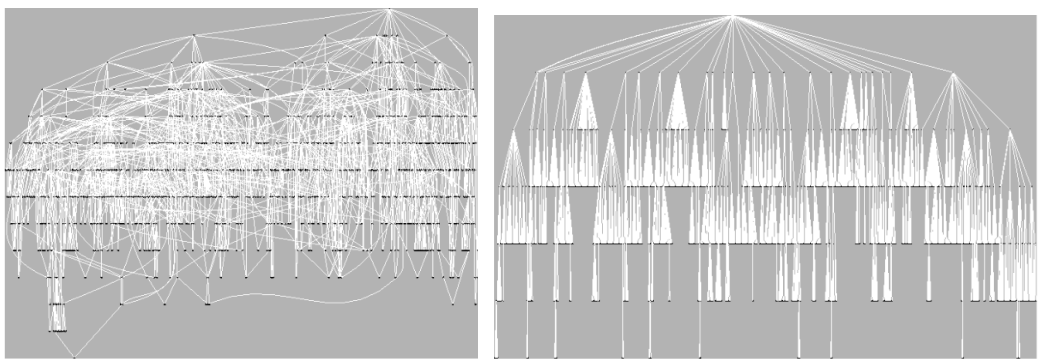
Tassonomie per le funzioni delle proteine

- Esistono infatti due tassonomie principali per l'organizzazione delle funzioni:
  - **Gene Ontology** (GO): che organizza le funzioni come un grafo diretto aciclico (DAG), varia per ogni specie, e possiede tre ontologie differenti (e quindi 3 DAG), e cioè *Biological Process* (BP), *Molecular Function* (MF) e *Cellular Component* (CC).
  - **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.



Tassonomie per le funzioni delle proteine

- Esistono infatti due tassonomie principali per l'organizzazione delle funzioni:
  - **Gene Ontology** (GO): che organizza le funzioni come un grafo diretto aciclico (DAG), varia per ogni specie, e possiede tre ontologie differenti (e quindi 3 DAG), e cioè *Biological Process* (BP), *Molecular Function* (MF) e *Cellular Component* (CC).
  - **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.



- 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.

La predizione della funzione delle proteine tramite metodi automatici

I metodi più noti in letteratura per effettuare predizioni della funzione delle proteine in maniera automatica sono:

La predizione della funzione delle proteine tramite metodi automatici

I metodi più noti in letteratura per effettuare predizioni della funzione delle proteine in maniera automatica 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

I metodi più noti in letteratura per effettuare predizioni della funzione delle proteine in maniera automatica 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

I metodi più noti in letteratura per effettuare predizioni della funzione delle proteine in maniera automatica 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

I metodi più noti in letteratura per effettuare predizioni della funzione delle proteine in maniera automatica 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 1/2

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

Metodi Ensemble Gerarchici 1/2

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 1/2

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 1/2

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 1/2

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*.

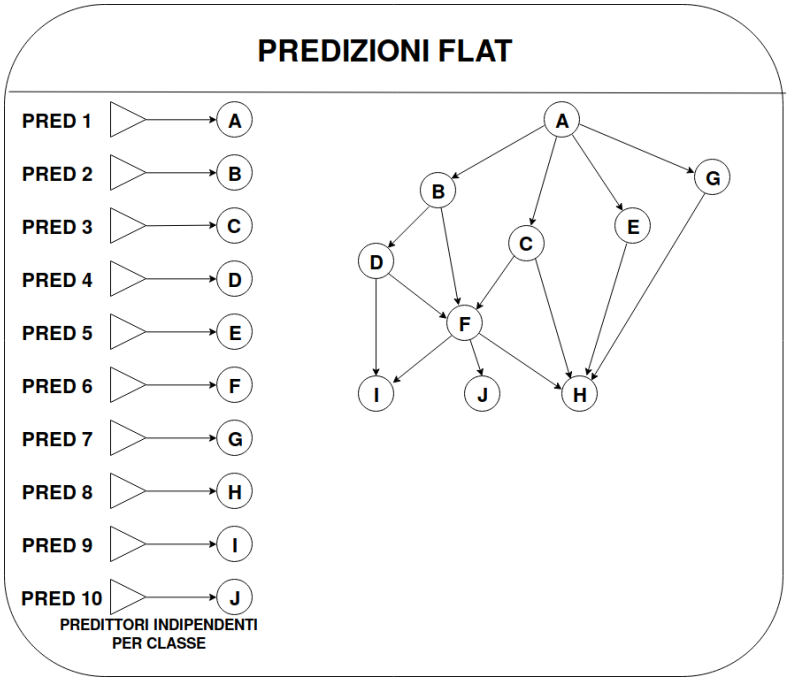
Consistenza & True Path Rule

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

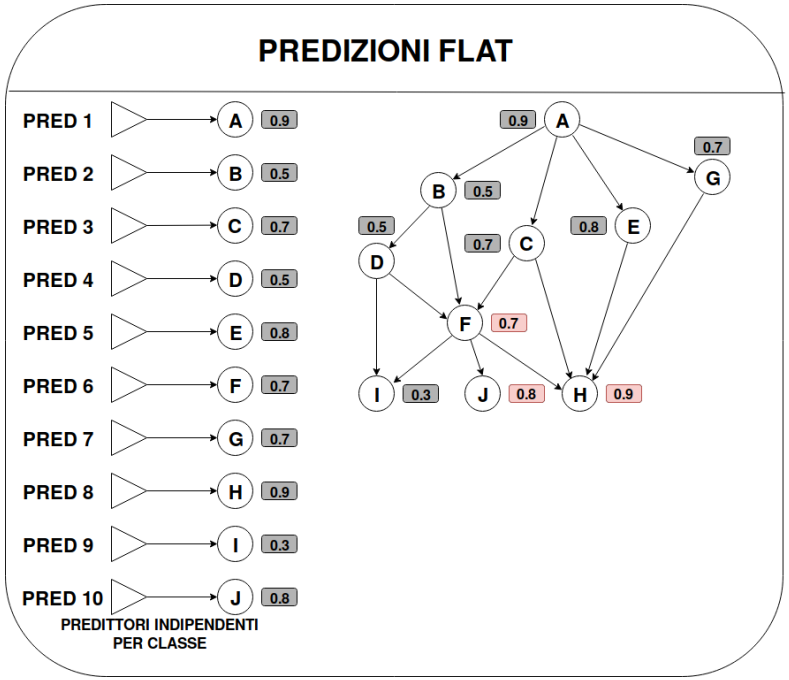
$$y \quad consistente \quad \leftrightarrow \forall i \in N, j \in par(i) \rightarrow y_j \geq y_i$$

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

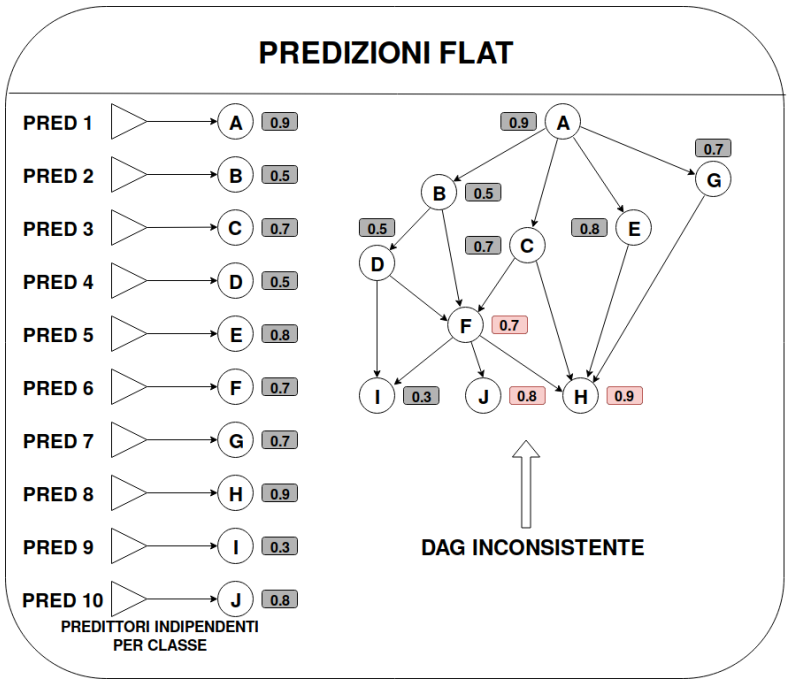
Metodi Ensemble Gerarchici (Esempio) 2/2



Metodi Ensemble Gerarchici (Esempio) 2/2



Metodi Ensemble Gerarchici (Esempio) 2/2



Metodi Ensemble Gerarchici (Esempio) 2/2