



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

Relatore

Prof. Giorgio Valentini

Correlatore

Dr. Marco Notaro

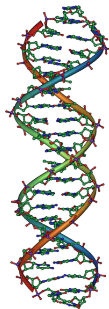
Candidato

Marco Odore

10 Luglio 2018

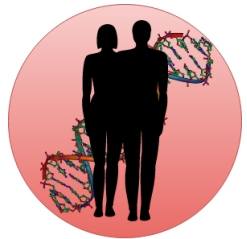
Central Dogma

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



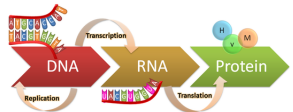
Central Dogma

- 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.



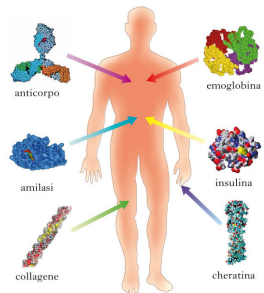
Central Dogma

- 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 1/5

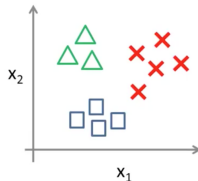
- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**



Il problema della predizione della funzione delle proteine 1/5

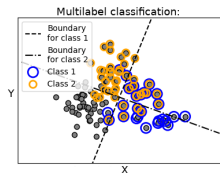
- L'individuazione della funzione delle proteine attraverso le analisi con sperimentazione diretta in laboratorio è **costosa** e richiede **molto tempo**
- Esistono centinaia di funzioni a cui poter associare un gene/proteina (**problema multiclasse**)

Multi-class classification:



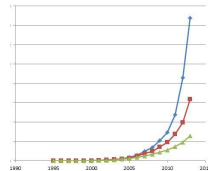
Il problema della predizione della funzione delle proteine 1/5

- 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 1/5

- 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 1/5

- 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 2/5

- 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 2/5

- 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 2/5

- 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 2/5

- 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 3/5

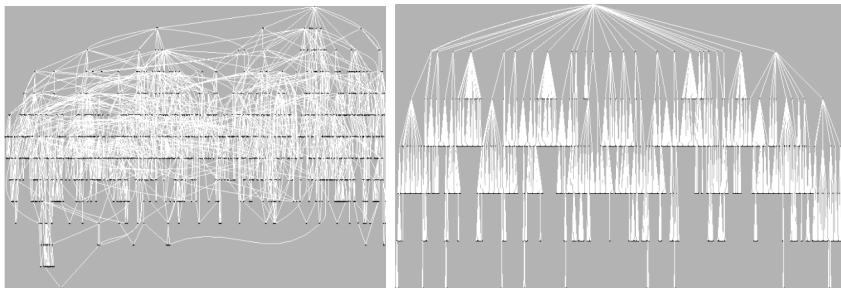


Figura: A sinistra un DAG della GO per la specie *S. cerevisiae*. A destra FunCat.

Il problema della predizione della funzione delle proteine (GO) 4/5

- 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) 4/5

- 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) 4/5

- 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) 4/5

- 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) 4/5

- 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 automatica

Per gestire il problema della predizione della funzione delle proteine si rende quindi necessario un approccio **automatico**.