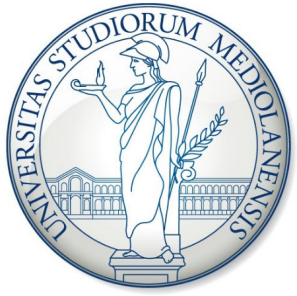
Università Degli Studi Di Milano



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

Relatore

Prof. Giorgio Valentini

Correlatore

Dr. Marco Notaro

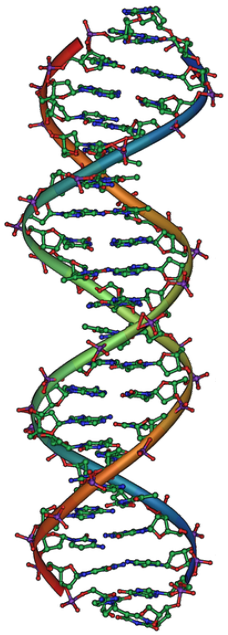
Candidato

Marco Odore

10 Luglio 2018

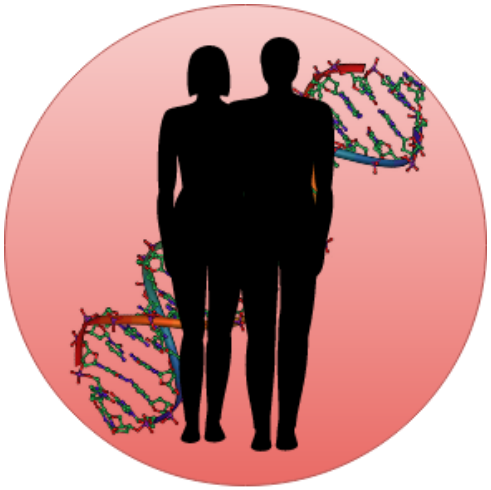
Central Dogma

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



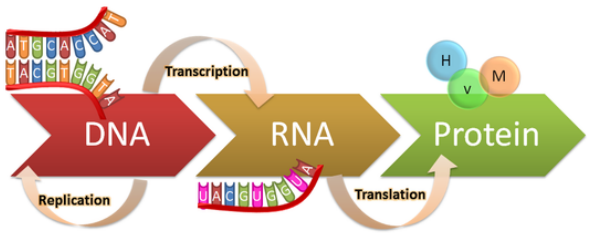
Central Dogma

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



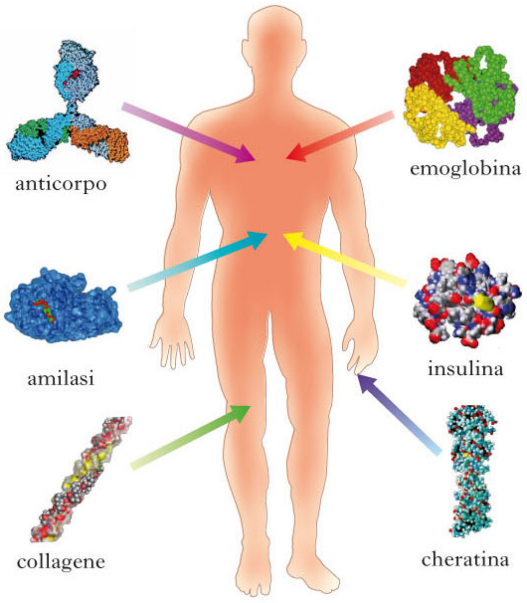
Central Dogma

- All'interno delle molecole di DNA di ogni essere vivente esistono diverse migliaia di geni.
- Si stima che per l'essere umano il DNA possegga 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 possegga 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

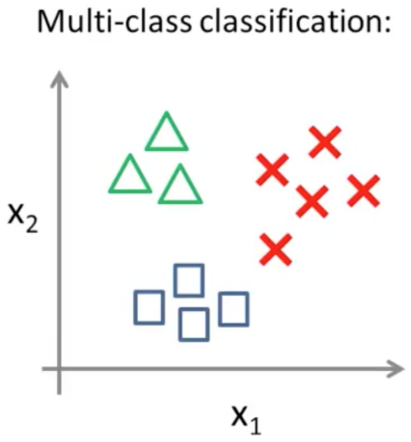
Per molte specie le funzioni di moltissimi geni (e quindi delle corrispettive proteine codificate) è **sconosciuta o parzialmente nota**.

Il problema della predizione della funzione delle proteine 1/3

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

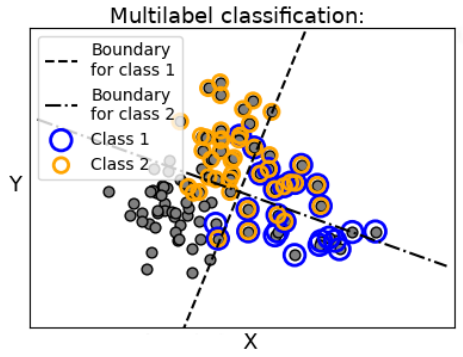
Il problema della predizione della funzione delle proteine 1/3

- 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)**



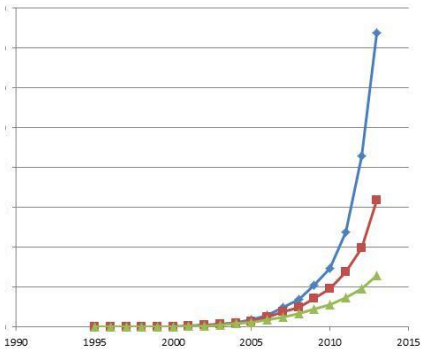
Il problema della predizione della funzione delle proteine 1/3

- 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 1/3

- 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 1/3

- 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 2/3

- 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 2/3

- 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 2/3

- 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 2/3

- 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 (GO) 3/3

- 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) 3/3

- 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) 3/3

- 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) 3/3

- 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) 3/3

- 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**.

Configuration

- The configuration of the standard theme is:
  - `language=italian`
  - `coding=utf8x`
  - `titlepagelogo=name-of-the-logo`
  - `bullet=circle`
  - `pageofpages=of`
  - `titleline=true`
  - `color=blue`
  - `secondcandidate=false`
  - `secondlogo=false`
- Most of them, actually everyone except the *titlepagelogo*, can be omitted if there are no modifications

Behavior of alerts

Each color theme requires different colors to highlight words. To insert alerts by using the *itemize* environment, you can exploit:

```
\begin{itemize}
\item<+-| alert@+> Apple
\item<+-| alert@+> Peach
\end{itemize}
```

For example:

- Apple

Behavior of alerts

Each color theme requires different colors to highlight words. To insert alerts by using the *itemize* environment, you can exploit:

```
\begin{itemize}
\item<+-| alert@+> Apple
\item<+-| alert@+> Peach
\end{itemize}
```

For example:

- Apple
- Peach

Another way to highlight words

If you want to highlight your text out of the enviroment *itemize*, Beamer2Thesis offers you the following possibilities:

- the standard command `\alert{text}`: it simply highlights your text
- the command `\highlight{text}`: it highlights your *text* setting it in italic
- the command `\highlightbf{text}`: it highlights your **text** setting it in bold

Of course, the color used, is set accordingly to your choice in the configuration phase.

Highlighting formulas

- The package hf-tikz allows to highlight formulas and formula parts in Beamer with overlay specifications
- The adaptation of colors to the theme could be done in this way:

```
\usepackage[beamer,customcolors]{hf-tikz}
\hfsetfillcolor{alerted text.fg!10}
\hfsetbordercolor{alerted text.fg}
```

- *Two compilation runs* are required to get the right result!
- Read the package documentation to find more options; an example will be provided in the next frame.

Highlighting formulas (II)

- Example:

$$x + \boxed{y} = 10$$

Highlighting formulas (II)

- Example:

$$\boxed{x + y = 10}$$

- Code:

```
\[\tikzmarkin<2->{a}x+
  \tikzmarkin<1>{b}y\tikzmarkend{b}
  =10\tikzmarkend{a}\]
```

The output

The pdf generated, has automatically, some properties:

- the title
- the name of the author
- the subject:
  - Thesis Presentation by using the english language
  - Presentazione Tesi di Laurea by using the italian language

This is possible thanks to the available options of hyperref. To create references in the text, use:

- `\label{name-reference}` in the starting point
- `\ref{name-reference}` in the point you want to show the reference
- `\href{url}{name-url}` to specify web addresses

Suggestions

- To realize a frame it is possible use the environment *frame* with top (t), center (c) or bottom (b) alignment: I suggest you to use the top alignment; this is the basic code:

  ```
  \begin{frame}[t]{title-of-the-frame}
  ```

  text

  ```
  \end{frame}
  ```

- To make things easier, it has been introduced a new environment which is able to have the top property property intrinsic:

  ```
  \begin{tframe}{title-of-the-frame}
  ```

  text

  ```
  \end{tframe}
  ```

Suggestions (II)

- To realize the titlepage with all options, it has been introduced the command `\titlepageframe`
  - Of course, it is also possible to use the *standard* approach
    ```
    \begin{frame}[plain]
    \titlepage
    \end{frame}
    ```
  - In this case **do not** provide a title for the frame
- If you have to insert some code using *verbatim* or *listings* **do not exploit** *tframe* environment, but:
  ```
  \begin{frame}[t,fragile]{title-of-the-frame}
  \verb!code!
  \end{frame}
  ```

Suggestions (III)

- If the title does not fit in the footer box, it is possible to exploit the so called *shorttitle*; an example:

  ```
  \title[short title]{Long title of the thesis}
  ```

  In this way the long title is just placed in the titlepage.
- In case there are more than two supervisors or assistansupervisors, I suggest you to insert them through commands reported in ?? and separate names thanks to a comma.

On Facebook

The relevance of Facebook is known to everybody: due to this reason, you can find:

- the group Beamer2Thesis
- the page Beamer2Thesis

In this way you can post your comments, hints, suggestion and questions in more familiar way. Morevoer, you can find further examples.

History

Here are shortly reported the main features of the releases:

- basic version (2011-01-17):
  - colors, second logo, second candidate, tframe environment, titleline, bullets, languages, separator string for slide numeration;
- release 2.0:
  - third logo, assistant supervisor, new ways to highlight, new command for the titlepage, new enviroments *adv* and *disadv*, XeTeX and XeLaTeX support, blocks;
- release 2.1:
  - coding option, second supervisor, second assistantsupervisor;
- release 2.2:
  - language, short title, highlighting formulas.

Thanks

I would like to thank people that, with precious hints, help me:

- Alessio Califano
- Alessio Sanna
- Luca De Villa Palù
- Mariano *Dave* Graziano
- Giovanna Turvani
- Mattia Stefano
- Nicola Tuveri
- Giuliana Galati

A special thank to Claudio Beccari for very precise comments on the first version.