# Predizione della funzione delle proteine con metodi di Machine Learning

Marco Odore
Lorenzo Rossi

18 settembre 2017

Docente: Valentini Giorgio
Corso: Bioinformatica

## 1 Scopo del progetto

L'obiettivo del progetto è di predire la funzione delle proteine di *Drosophila melanogaster*, per determinate ontologie, tramite gli algoritmi di apprendimento Support Vector Machine(SVM) e Multilayer Perceptron(MLP), per poi analizzarne e confrontarne i risultati. Dato che ogni proteina può essere classificata in più di una categoria, il problema trattato è quello della classificazione multi-etichetta.

## 2 Dataset

Il dataset utilizzato per l'apprendimento induttivo è stato generato da un grafo indiretto, i cui nodi sono le proteine e gli archi indicano il grado di similitudine tra due proteine[1]. Tale grafo è rappresentato da una matrice pesata di adiacenza e ogni riga (colonna) si riferisce quindi ad una diversa proteina dell'organismo ed ogni entry al peso dell'arco che connette due proteine.

### 2.1 Istanze degli algoritmi

Le istanze utilizzate per i due algoritmi induttivi sono le righe (colonne) della matrice di adiacenza. Quindi per ogni proteina si avrà un vettore le cui componenti (feature) rappresentano il grado di similitudine che questa ha in relazione alle altre proteine.

[1] Come è stata costruita la matrice:
`https://homes.di.unimi.it/~valentini/SlideCorsi/Bioinformatica1617/Bioinf-Project1617.pdf`

### 2.2 Etichettatura

Per l'etichettatura delle istanze sono state fornite tre ontologie[2]

- BP(Biological Process) con 1951 termini.
- MF (Molecular Function) con 234 termini.
- CC (Cellular Component) con 235 termini.

Rappresentate da matrici di annotazioni, dove sulle righe sono specificate le proteine e sulle colonne i termini delle ontologie. Nell'entry $(i, j)$ della matrice è specificato un 1 se la proteina $i$ appartiene alla categoria\termine $j$, altrimenti 0.

Data la notevole quantità di tempo necessaria per l'addestramento dei classificatori, ci si è soffermati unicamente sull'ontologia CC. Quindi ad ogni istanza del problema è stato associato un sottoinsieme di queste etichette.

## 3 Metodi di apprendimento

I metodi di apprendimento supervisionato utilizzati sono:

- Multilayer Perceptron.
- Support Vector Machine.

### 3.1 MLP

### 3.2 SVM

Si tratta di un algoritmo di apprendimento che performa bene su problemi linearmente separabili nello spazio euclideo delle feature. È inoltre applicabile anche a problemi non linearmente separabili, data la possibilità di introduzione dei *kernel*, che proiettano lo spazio a n-dimensioni (n feature) in uno spazio ad elevata dimensionalità, tramite delle trasformazioni non lineari, dove le probabilità che il problema sia linearmente separabile aumentano notevolmente[3].

Data l'altissima probabilità dei diversi problemi di classificazione di essere non linearmente separabili, nel nostro set-up è stato utilizzato il kernel Gaussiano RBF[4].

[2] Tutti i dataset utilizzati sono scaricabili dal sito: `http://homes.di.unimi.it/valentini/DATA/ProgettoBioinf1617`

[3] Teorema di Cover

[4] `https://en.wikipedia.org/wiki/Radial_basis_function_kernel`

# 4 Set-up sperimentale

Per valutare le performance del metodo si è usato la tecnica sperimentale della 5-fold cross-validation. Si sono poi utilizzate le seguenti metriche:

- Misure per class": Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision Recall Curve (AUPRC);
- Misure per-example" la Precision, Recall ed F-score gerarchica.
- F-score gerarchica.