

# Riconoscimento e Recupero dell'Informazione per Bioinformatica

LAB. 11 – Ripasso

Pietro Lovato

Corso di Laurea in Bioinformatica  
Dip. di Informatica – Università di Verona  
A.A. 2016/2017



# Problema: classificazione e clustering di dati derivanti da DNA Microarray

Task da risolvere:

- **Classificazione:** lungo la dimensione delle condizioni, per addestrare un sistema che sia in grado di predire lo stato di un nuovo campione (es. capire se è sano / malato)
- **Clustering:** lungo la dimensione dei geni, per cercare geni che mostrino livelli di espressione coerenti fra loro.

# Esercizio 1

```
>> load( 'dataset_prostate.mat' )
```

- Dataset contenente livelli di espressione di pazienti in 3 classi (sani, malati, malati con tumore in metastasi)
  - Quanti geni ci sono nel dataset?
  - Quanti pazienti?
- Testare due diversi classificatori che discriminino le tre classi, utilizzando le specifiche seguenti

# Esercizio 1

- Si consiglia di utilizzare PRTools (non è obbligatorio)
- Classificatori: Support Vector Machine, K-NN
- Protocollo di validazione:
  - 1. Leave-one-out
  - 2. 10-fold cross-validation (con 2 ripetizioni)
- Quale dei due classificatori funziona meglio? Con quale protocollo di validazione?

# Esercizio 2

- Effettuare un clustering k-means dei geni nel dataset
- Numero di cluster = 5
- Si può utilizzare il codice scritto a lezione o la funzione Matlab

```
>> [cidx, ctrs] = kmeans(data, K);
```

## Esercizio 2

- Plottare il profilo di espressione medio di ogni cluster e sceglierne uno che si ritiene particolarmente “interessante” (dove per esempio i geni sembrano variare tra sani e malati)
- Salvare i nomi di tali geni in un file di testo:

```
>> cluster = gene_names(cidx==1);  
>> dlmwrite('cluster.txt', char(cluster), '');
```

# Esercizio 2

- Caricare questo file nel server GOstat (<http://gostat.wehi.edu.au/cgi-bin/goStat.pl>). Nel campo “Gene-association database” mettere goa\_human (uomo), e nel campo “subset of GO hierarchy” indicare biological\_process (siamo interessati al processo biologico cui i geni nel cluster prendono parte). Lasciare gli altri valori a default.
- Che informazioni si riescono a dedurre? È possibile ricavare i geni maggiormente responsabili dello sviluppo / morte cellulare?