Progetto Metodi Bioinformatici

SVMLab: apprendimento supervisionato di signature tumorali.

Il progetto si base su un tutorial condotto da S. Lianoglou, X. Li e C. Leslie al dipartimento di Biologia Computazionale del Memorial Sloan-Kettering Cancer Center di New York. Basandosi su un lavoro precedentemente pubblicato [1], il tutorial si propone di addestrare dei classificatori binari supervisionati per predire colture cellulari umane che esprimono specifiche attivita' tumorali rispetto a linee cellulari di controllo e di individuare signature di geni per pathway tumorali associati con l'espressione di oncogeni specifici. Inoltre si vuole valutare se le medesime signature individuate nelle linee celluari umane possono permettere di effettuare predizioni non solo per le linee cellulari umane ma anche per corrispondenti tessuti tumorali di topo.

Obiettivo del progetto: individuare "gene signature" per 3 differenti pathway tumorali associati con l'espressione dei seguenti oncogeni: c-Myc, H-Ras ed E2F3.

Il documento "SVM Lab: Supervised learning of oncogenic pathway signatures" scaricabile da:
homes.di.unimi.it/valentini/MB201314/SVMLab/svmLab-nopackage.pdf
spiega passo per passo come individuare gene signature per pathway tumorali associati all'espressione del gene c-Myc., utilizzando metodi e funzioni disponibili nei seguenti package R:
- e1071: libreria per la classificazinoe con SVM
- gplots: libreria per produrre heatmap
- scatterplot3d: libreria per scatter plot tridimensionali
- caret: libreria di machine learning
- affy: libreria per la gestione di dati di espressione con piattaforme Affymetrix

Schematicamente i passi sperimentali necessari per individuare le gene signature per i pathway associati a c-Myc sono i seguenti (vedi il precedentemente citato documento SVM Lab":

0. Dati di espressione di linee cellulari umane (piattaforma Affymetrix hgu133plus2) relative a campioni: (i) GFP (controlli); (ii) MYC; (iii) Ras e (iv) E2F3.
1. Analisi esplorativa tramite clustering gerarchico e visulizzazione con heatmap
2. Applicazione di Support Vector Machine (SVM) e metodi di filtering basati sull'analisi della varianza per: (a) selezionare i geni della signature (MYC-VAR Gene signature); (b) per generare un modello (classificatore) in grado di classificare i campioni MYC con tecniche di valutazione dell'errore di held-out e cross-validation.
3. Applicazione di un metodo di gene selection basato sul valore dei pesi della SVM lineare per individuare subset di geni associati al pathway Myc (MYC-SVM Gene signature).
4 Visualizzazione con heatmap e comparazione dei geni delle signature MYC-VAR e MYC-SVM.
5. Analisi PCA Prinicipal Component Analysis) delle signature MYC per la visualizzazione tridimensionale dei campioni MYC e non-MYC (parte opzionale)
6. Applicazione del modello appreso dai dati di espressione delle linee cellulari umane al topo. La signature appresa dalle linee cellulari umane e' in grado di predire correttamente anche i campioni murini?
7. Stima delle predizioni Myc tramite SVM probabilistiche.
8. (Opzionale) Usare il package GOstats per individuare i termini GO sovrarappresentati nelle gene signature (metodo hyperGTest).

Il tutorial "SVM Lab: Supervised learning of oncogenic pathway signatures" spiega in dettaglio quali package R applicare e come usarli per effetture gli esperimenti.
I dati di espressione necessari per gli esperimenti sono scaricabili da:
homes.di.unimi.it/valentini/MB201314/SVMLab/common.data.clean.rda (.RData file 3.9 M)
homes.di.unimi.it/valentini/MB201314/SVMLab/human.base.clean.rda (.RData file 15 M)

Per il progetto d'esame ripetere gli esperimenti per individure una pathway signature per l'oncogene umano H-Ras, o E2F3, usando la medesima procedura sperimentale utilizzata per la MYC signature. Oltre agli script (debitamente commentati) necessari per realizzare i punti da 1 a 7 (ed eventualmente il punto 8 opzionale), si prepari un report sintetico che illustri (anche con opportuni grafici e/o tabelle) la procedura sperimentale ed i risultati ottenuti.

Parte opzionale: ripetere i medesimi esperimenti utilizzando uno o più differenti classificatori supervisionati (ad es: un classificatore Naive-Bayes o una rete neurale artificiale), utilizzando il medesimo package e1071, o altri package disponibili in R.

Bibliografia:

[1] Andrea H. Bild, Guang Yao, Jeffrey T. Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Johnathan M. Lancaster, Andrew Berchuck, Jeffrey R. Olson, Holly K. Dressman, Mike West, and Joseph R. Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature, 439, 2006.