Progetto d'esame per il corso di Informatica Avanzata 2014-15

Ricerca di associazioni geni associati a patologie specifiche tramite metodi metodi di Random Walk applicati a reti funzionali.

Il progetto ha l'obiettivo di applicare metodi basati su reti per il problema della *gene-disease prioritization*.
Dati una o più reti che rappresentano interazioni funzionali fra geni, si vogliono individuare quei geni che potenzialmente potrebbero essere correlati ad una patologia specifica, sfuttando la topologia della rete (cioe' l'insieme delle relazioni funzionali fra i geni) e la conoscenza a priori relativa ai geni la cui associazione con la patologia è già nota.
A tal fine si implementi e si utilizzi l'algoritmo di *Random Walk* studiato durante il corso.

Il progetto si articola in 2 parti + 1 parte opzionale

1. Implementazione in R dell'algoritmo Random Walk a k passi.
2. Ordinamento dei geni umani rispetto a 33 malattie definite tramite il MeSH thesaurus (Medical Subject headings – http://www.nlm.nih.gov/mesh/)
3. (opzionale) Implementazione in R dell'algoritmo Random Walk con restart e sua applicazione all' ordinamento dei geni umani rispetto a 33 malattie definite tramite il MeSH thesaurus

1. *Implementazione in R dell'algoritmo Random Walk.*

Si implementi la funzione RW.
Tale funzione riceve in ingresso una matrice di adiacenza pesata di un grafo, l'insieme dei nodi positivi e ritorna lo score (probabilita') predetta dall'algoritmo per tutti i nodi del grafo.

Argomenti di ingresso:
- W : matrice di adiacenza (pesata) del grafo
- ind.positives: indici del "core" dei vertici positivi del grafo (cioe' dei geni la cui associazione con la malattia è nota a priori). Gli indici si riferiscono alle righe della matrice W.
- tmax : numero massimo di iterazioni (passi di random walk) consentite.

Output:
- un vettore p delle probabilità predette per ciascun vertice/gene.

Si possono anche aggiungere altri parametri purche' debitamente documentati e giustificati.

2. *Ranking dei geni umani rispetto a 33 malattie MeSH*.

La funzione RW dovrà essere applicata all'ordinamento dei geni umani rispetto a 33 malattie (MeSH descriptor).

a. Dati
Dalla directory http://homes.di.unimi.it/valentini/IA1415/esame/ sono scaricabili sia i file relativi alle reti funzionali dei geni (3 diversi file corrispondenti a 3 diverse reti funzionali (finet,hnnet,cmnet) relative a circa 8000 geni umani), sia i dati relativi alle annotazioni dei geni per ognuna delle malattie MeSH.
Si veda il file README per i dettagli e la letteratura relativi ad ognuno dei file.
Si noti che il file MESH.labels.rda e' un file binario (con valori 0 e 1) per le associazioni note a priori fra i geni (righe della matrice) e classi MeSH

(colonne). Nel file sono presenti le etichette per 100 MeSH descriptor, ma ogni gruppo dovra' analizzare solo 33 malattie, secondo questo ordine:
ll gruppo A studierà le prime 33 MeSH disease (prime 33 colonne della matrice MESH.labels), il gruppo B le seconde 33 (colonne 34-66), il gruppo C le MeSH disease corrispondenti alle colonne (67-100).

b. Ranking dei geni e valutazione delle prestazioni.
Per la stima dell'errore di generalizzazione, si applichi la tecnica di 5-fold cross-validation.
Applicare RW e calcolare l'AUC (Area Under the ROC Curve) per ogni MeSH diseae.
L' AUC va calcolata:
- separatamente per ognuno dei 3 data set finet, hnnet e cmnet.
- per ogni classe
- come valor medio fra tutte le classi.
Per calcolare i valori dell'AUC si usi la funzione AUC.single.over.classes del package PerfMeas (scaricabile da CRAN: http://cran.r-project.org/).

Si ripeta l'esperimento con Random walk ad 1, 2, e 3 step e si confrontino i risultati.

3. *Implementazione in R dell'algoritmo Random Walk con restart*

Si implementi la funzione RWR
Tale funzione riceve in ingresso una matrice di adiacenza pesata di un grafo, l'insieme dei nodi positivi e ritorna lo score (probabilita') predetta dall'algoritmo per tutti i nodi del grafo.

Argomenti di ingresso:
- W : matrice di adiacenza (pesata) del grafo
- ind.positives: indici del "core" dei vertici positivi del grafo (cioe' dei geni la cui associazione con la malattia è nota a priori). Gli indici si riferiscono alle righe della matrice W.
- theta : parametro di restart
- tmax : numero massimo di iterazioni (passi di random walk) consentite.

Output:
- un vettore p delle probabilità predette per ciascun vertice/gene.

Si possono anche aggiungere altri parametri purche' debitamente documentati e giustificati.

Si ripeta l'analisi effettuata al punto 2, usando questa volta la funzione RWR. Si sperimenti con diversi valori di theta.

Scrivere un report sintetico che illustri l'algoritmo Random Walk (e opzionalmente l'algoritmo Random Walk con restart) ed i risultati ottenuti ai punti 2 e 3, anche con opportune tabelle e/o grafici.
Riportare il codice R per la funzione RW e lo script degli esperimenti per il punto 2 (e opzionalmente per il punto 3). Il codice va commentato: il codice di ogni funzione deve essere preceduto da commenti che spieghino sinteticamente a cosa serve la funzione stessa, il tipo ed il significato di ogni argomento di ingresso, il tipo ed il significato dell'output della funzione.

Bibliografia:
[1] G. Wu, X. Feng, L. Stein, A human functional protein interaction network and its application to cancer data analysis, Genome Biology 11 (2010) R53.
[2] M. Re, G. Valentini, Cancer module genes ranking using kernelized score functions, BMC Bioinformatics 13 (2012) S3.
[3] I. Lee, U. Blom, P. Wang, J. Shim, E. Marcotte, Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Genome Res 21 (2011) 1109-1121.
[4] E. Segal, N. Friedman, D. Koller, A. Regev, A module map showing conditional activity of expression modules in cancer, Nat Genet 36 (2004) 1090-1098.