

## Class Descriptions

### **Confronto\_similarità.**

This class computes the Pearson correlation coefficient (PCC) of two similarity vectors in order to compare the GO semantic similarity (obtained by class **Sperimentazioni**) of genes and the sequence similarity (blast) contained in a precomputed vector. These two vectors contains the similarities of one gene and all the other genes in the same corpus and that are annotated with GO terms of the same domain (BP, MF, CC).

**StoreMatrix.** This class check if the similarity matrices computed by the java software are identical to those produced by the original MATLAB code.

**GO.** This class takes the parameters provided by the user in the web interface and load from disk the corresponding files, if they already exist. Moreover, the calss return to the web interface the results of the user search.

### **MachineStore.**

This class correspond to the semantic of the MATLAB INIZIALIZE function, and it generates and stores the data structures of each different semantic measure (Resnik, Lin, Jiang, RWRENIK, RWLIN, RWJIANG). It uses other modules: ParsingOBO, ParsingAnnotationFile, Get\_TREE\_P, Get\_TREE\_LCA, Get\_Resnik, Get\_LIN, Get\_Jiang, Get\_RandomWalk.

### **Get\_Resnik.**

This class provides the method `get_resnik()` to compute the Resnik similarity Matrix . It use previously computed structures previously store by the class MachineStore, like

- TREE\_LCA which contains, for each couple of GO terms i,j, the number of genes annotated with the minimum subsumer of i and j.
- terms\_for\_LCA, list of the GO terms in the TREE\_LCA. These GO terms have at least one gene annotated.
- mapping\_ID\_Indici, hashtable containing couples key-value, storing for each GO term the corresponding index in the matrix TREE\_LCA.
- mapping\_Indici\_ID, hashtable reverse of mapping\_ID\_Indici.
- TREE\_P\_max, for each GO term contains the max number of genes annotated for a GO term weighted by the updegree (number of parents) of the corresponding GO term.

The output is the Resnik similarity matrix of the terms in the selected domains with at least one annotation.

### **Get\_TREE\_LCA.**

Provides the method `get_tree_lca()` to compute the matrix TREE\_LCA. It stores the data structures **terms\_for\_LCA**, **mapping\_ID\_Indici**, **mapping\_Indici\_ID**, **TREE\_LCA**, **termini\_ordinati** (list of increasing ordered ID), **TREE\_P\_max**

### **Get\_TREE\_P.**

Provides the method `get_tree_p(Aspect)` which computes the the following structures according to the chosen Aspect (GO domain) :

-GO\_GENE, corresponding to the TREE\_P object in the original MATLAB code, is an hashtable containing for each GO term the number of annotated genes weighted by the its updegree

-GO\_GENE\_NoBranch, corresponding to TREE\_P\_NoBranch in the original MATLAB code, is an hashtable containing for each GO term the number of annotated genes

### **ParsingOBO.**

Class which by means of the method storeOBO(file) read the obo files containing the GO ontology and provide the following structures:

-go\_ontology: OBOReader object containing the parsed ontology

-ALLGOLIST: table containing all the found GO terms, in which each entry is of kind Tupla\_Tabella\_GO which has the following fields: term ID termine | term name | Aspect | children | parents.

-TreeCC, TreeBP, TreeMF: tables like ALLGOLIST but in which the terms are separated by domain.

-DownCC, DownBP, DownMF, UpCC, UpBP, UpMF: data structures to respectively store the parent and the child terms for a gene in the corresponding domain.

### **ParsingAnnotationFile.**

Provides the method storeAnnotations(file, Aspect, Evidence\_code) which receive the file containing the annotations, the GO domain and the evidence code to be ignored during the parsing.

It produces the following files: 1) Annotation\_table\_reduced, table in which each entry has the format Annotations\_record\_reduced, which is GENE\_ID | TERM\_ID | ASPECT | EVIDENCE\_CODE; 2) classi\_GO, list of GO term IDs in the chosen domain with the chosen evidence; 3) geni, list of the found genes.

### **Get\_LIN, Get\_JIANG.**

Like Get\_Resnik class, these classes compute the Lin e Jiang similarity matrices.

### **Get\_RandomWalk.**

Class providing the method get\_rw(Aspect, host\_measure) to compute downward random walk with the chosen aspect(BP, MF, CC) and host measure (Resnik, Lin o Jiang).

---

### **Supplementary classes :**

**go\_record.** Contains the format of items stored in the object **go\_ontology** (parsed ontology):

ID(term Id) | NAME | TREE(GO domain) | OBSOL(boolean, for obsolete terms).

### **OBOReader**

The class OBOReader(file) provides the method buildHier() which parses the input obo file and build the corresponding ontology. The found GO term are stored as instance of the class TermInfo. Moreover, the class OBOReader provides the methods: 2) get\_Term(id of GO term), which return a TermInfo instance corresponding the chosen term; 3) getList(TermInfo) and 4) getAncestorList( TermInfo), return respectively the list of descendant and ancestor terms of the chosen GO term.

**OBOReaderDriver.**

OBOReaderDriver is a class driver which create an object OBOReader to whom give in input the OBO file

**TermInfo.**

Class containing the info of a GO term, like name, id, synonymous, domain ... and themethods to access these fields.

**Tupla\_Tabella\_GO.** Format of the items in the tables ALLGOLIST, TreeCC, TreeBP, TreeMF.

**Sperimentazioni.**

Class to computer the GO semantic similarity value of a selected gene with all the other genes in the same corpus annotated with terms in the same GO domain.