Using Ontologies and Semantic Similarity Measures for Prioritization of Gene Regulatory Networks

Marianna Milano, Pietro Hiram Guzzi, Mario Cannataro

Motivation
Omics sciences are widely used to analyze diseases at a molecular level. Usually, results of omics experiments are sets of candidate genes potentially involved in different diseases [1]. The interpretation of results and the filtering of candidate genes or proteins selected in an experiment is a challenge in some scenarios. This problem is particularly evident in clinical environments in which researchers are interested in the behavior of few molecules related to some specific disease while results may contain thousands of data and have very relevant dimensions. The filtering requires the use of domain-specific knowledge that is usually encoded into ontologies. Consequently, to filter out false positive genes, different approaches for selecting genes have been introduced. Such approaches are often referred to as Gene prioritization methods. They aim to identify the most related genes to a disease among a larger set of candidates genes, through the use of computational methods.

Methods
We implemented GoD (Gene ranking based On Diseases), an algorithm that ranks a given set of genes based on ontology annotations. The algorithm orders genes by the semantic similarity computed with respect to a disease among the annotations of each gene and those describing the selected disease. The current version of GoD enables the prioritization of a list of input genes for a selected disease. It uses HPO (Human Phenotype Ontology), GO (Gene Ontology), and DO (Disease Ontology) ontologies for the calculation of the ranking. It takes as input a list of genes or gene products annotated with GO Terms, HPO Terms, DO Terms and a selected disease described regarding annotation of GO, HPO or DO (user may also provide novel annotations). It produces as output the ranking of those genes with respect to the input disease. Package consists of three main functions: hpoGoD (for HPO based prioritization), goGoD (for GO based prioritization), and doGoD (for DO based prioritization).

Results
We tested GoD on Gene Regulatory Networks (GRNs). Biological network inference aims to reconstruct network of interactions (or associations) among biological genes starting from experimental observations. The reconstruction of GRNs is an important area since it enables, for instance, the comparison of GRNs of different states (e.g. health or disease) or the investigation of the progression of illnesses. We selected three expression datasets: Dataset 1 (GDS3285) [2], related to breast cancer disease; Dataset 2 (GDS5072) [3], related to prostate cancer disease; and Dataset 3 (GDS5093) [4], related to Dengue virus (DENV) infection. Initially, experimental data are given as input to five GRN inference algorithms, i.e. ARACNE, CLR, MRNET, GENIE3 and GGM, to produce 5 inferred GRN networks. For each inferred GRN, GoD receives as input the list of top genes and produces for each gene a semantic similarity value on a selected disease considering one of the previous ontologies (e.g. Disease Ontology). For each GRN, the genes are ranked and reordered on the basis of the computed semantic similarity and are compared allowing to rank each GRN inference method with respect to the initially selected disease.
Figure 1 shows the average gene-disease semantic similarity computed by GoD for the top 5 genes extracted from the GRNs inferred using respectively ARACNE, CLR, MRNET, GENIE3 and GGM on Dataset 1, Dataset 2 and Dataset 3.

![Performance of Inference Methods](image)

Figure 1: Average gene-disease similarity for top 5 genes extracted from inferred networks

