USING SEMANTIC SIMILARITY FOR GENE SET ANALYSIS OF MICROARRAY DATA

A. Sanchez-Pla[†], J.Ll. Mosquera

Universitat de Barcelona. Barcelona, Spain

[†] E-mail: *asanchez@ub.edu*

New technologies such as microarrays or proteomics allow to perform experiments resulting in long lists of genes. Most methods developed to help interpret these data perform some kind of analysis based on the annotations in databases such as the Gene Ontology (GO). The GO is a complex structure formed by 3 huge directed acyclic graphs. Most methods perform some kind of *enrichment analysis* aimed at establishing if a given GO category changes under different experimental conditions. A main drawback of this approach is that it ignores the underlying graph structure. An intermediate approach between the previous one and the use of the whole graph is to weight its nodes in order to select those which are most informative and base the analysis on them. This does not restrict the analysis to arbitrarily fixed levels but it does not require using the whole graph either. We introduce an index of cumulative semantic similarity which allows to weight the nodes based in the information they contain. This may be computed on a graph yielding a categorization of its nodes. Based on this index a distance measure between graphs can also be defined in order to compare experimental results. Simulation studies and application to real data sets show that the index defined has good properties and is advantageous to alternative approaches in some situations. They also show that useful biological insight can be obtained using this method.