

# SoFoCles: Feature Filtering for Microarray Classification Based on Gene Ontology

Sotiris Diplaris<sup>\*</sup>, George Papachristoudis<sup>†</sup> and Pericles A. Mitkas<sup>‡</sup>

Intelligent Systems and Software Engineering Lab  
Dept. of Electrical and Computer Engineering  
Aristotle University of Thessaloniki  
Thessaloniki, Greece 54124  
<http://issel.ee.auth.gr>

<sup>\*</sup>email: [diplaris@issel.ee.auth.gr](mailto:diplaris@issel.ee.auth.gr) (corresponding author)

<sup>†</sup>email: [geopapa@auth.gr](mailto:geopapa@auth.gr)

<sup>‡</sup>email: [mitkas@eng.auth.gr](mailto:mitkas@eng.auth.gr)

## Abstract

A challenging problem in bioinformatics is the analysis of microarray experiments. Microarrays allow for the monitoring of the regulation of thousand of genes or gene products simultaneously, under different conditions. The produced datasets can undergo different means of analysis, such as clustering, classification or density estimation, among others. In microarray analysis problems, the selected features that represent genes are known as marker genes.

Marker gene selection has been an important research topic in the classification analysis of gene expression data. Current methods try to reduce the “curse of dimensionality” by using statistical intra-feature set calculations, or classifiers that lie upon the given dataset. However, statistical methods cannot incorporate the a-priori semantic knowledge that lies behind the feature set. Other bioinformatics tools, though, have already produced such knowledge.

In this paper, we present SoFoCles, an interactive tool that enables semantic feature filtering in microarray classification problems with the use of external, well-defined knowledge retrieved from the Gene Ontology. The quality of microarray classification is enhanced by exploiting the knowledge offered by a structured hierarchy of genetic concepts, the Gene Ontology<sup>1</sup>, instead of the single application of feature selection methods. In this sense, scientifically proofed information with biological meaning is incorporated in the feature filtering procedure, in order to improve the classification accuracy.

To achieve this, within SoFoCles, a semantic similarity methods repository has been developed, by which feature similarities are identified based on the terms of the Gene Ontology that each gene or gene product is tied to. These similarities are used to enrich a feature set that has already been refined using legacy statistical filtering methods.

Information concerning genes or gene products of the refined dataset is preprocessed in order to identify the related GO terms and infer semantically similar genes that are involved in the same biological process. These genes enrich the refined feature set by better describing the biological paths regulated in the conditions tested, thus improving classification accuracy.