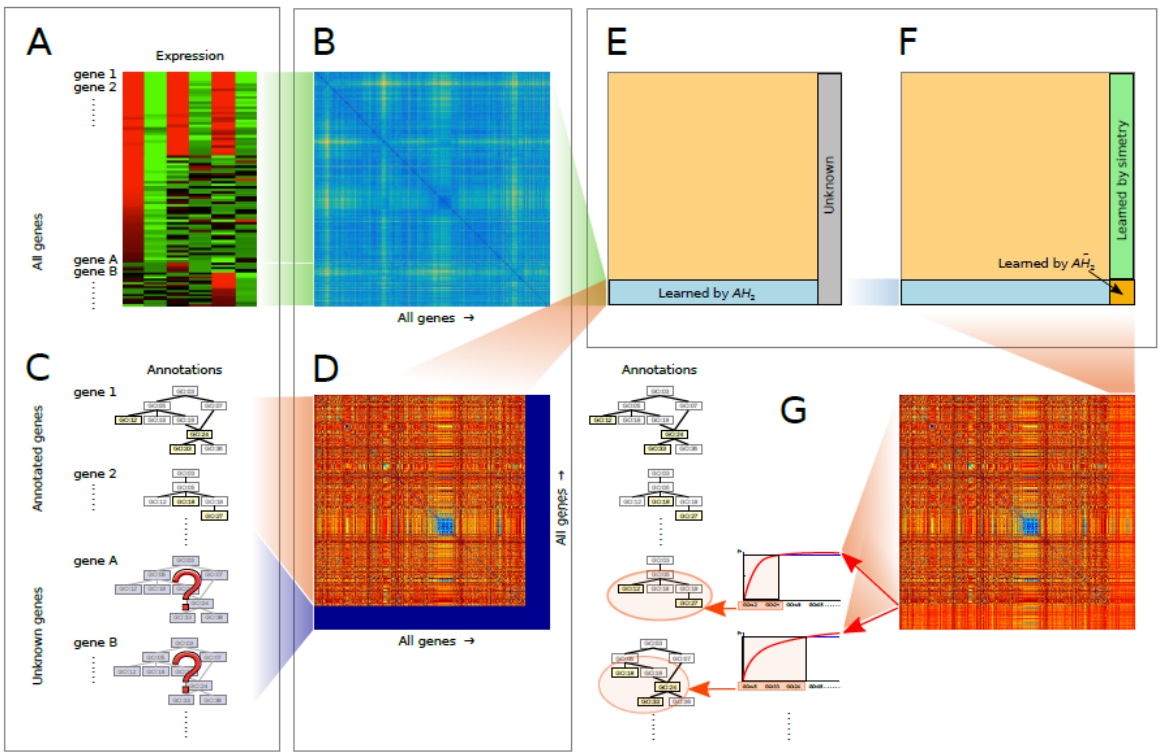
Annotation pipeline for inferring gene functions integrating GO annotations and expression data

L. Di Persia, D.H. Milone and G. Stegmayer
Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH/UNL-CONICET, (3000) SF, ARG

Background: Computational methods for the prediction of gene function refers to automatically finding associations between a gene and a set of Gene Ontology (GO) terms. Since the hand-made curation process of novel annotations are very time-consuming, computational tools that can reliably predict likely annotations and boost the discovery of new gene annotations are urgently needed. **Results:** This work proposes a novel pipeline (see Figure) for inferring gene annotations based on the automatic reconstruction of the semantic similarity between genes. The semantic similarity is a metric defined over a set of terms, where the distance between them is based on the likeness of their meaning or semantic content. We benchmarked the proposal against state-of-the-art methods on three published data sets (*Arabidopsis thaliana*, *Saccharomyces cerevisiae* and *Dictyostelium discoideum*). Independent experiments have shown that the proportion between annotated and unannotated genes does not influences the model accuracy. We have used a leave-one-out cross-validation technique. Being the state-of-the-art an average F_1 = 15% for related methods, we have achieved a F_1 = 30% in average, for all 3 species. It can be stated that our proposal has shown the most balanced results, not missing true GO labels and not assigning, either, a large number of false GO terms to un-annotated genes.



Pipeline for inferring GO labels. A) Expression data. **B)** Expression pairwise Euclidean distance matrix d_E among all genes in the study: full matrix, no missing values in any row/column. **C)** Semantic data: GO annotations for well-known genes (gene 1, gene 2,...); some genes in the study (gene A, gene B, ...) are completely unknown and do not have GO terms associations. **D)** Semantic distance matrix d_{GO} among all genes: with missing rows and columns because many genes are not semantically annotated, thus a semantic distance among them cannot be calculated. **E)** and **F)** The missing distances in d_{GO} in D) are completed using the information available in B) and D). **G)** Once the d_{GO} matrix is completed, GO annotations are assigned according to the "closest genes" in the reconstructed semantic space. From the closest set of genes with known GO labels, the potential labels (according to a Bayesian model) are sorted in descending order by their posterior probability. Finally, the candidate GO labels (in yellow) with an accumulated probability of 0.95 are assigned to each un-annotated gene (gene A, gene B, ...).

Conclusions: We presented a novel method for gene annotation, which is capable of annotating genes according to gene expression distance and genes annotations distance in the GO. Our approach can be very useful in the case of: i) a model genome, when there is still a minority of un-annotated genes; ii) a partially studied genome where annotation mostly based on homology with similar genomes has been done; iii) when there is a de-novo sequenced genome, partially annotated, whose initial annotations can be then reviewed when transcriptomics data becomes available. We have found that our proposal could effectively enhance the prediction of GO annotations.