Gene function prediction using Gene Ontology decomposition

Vedrana Vidulin, Sašo Džeroski

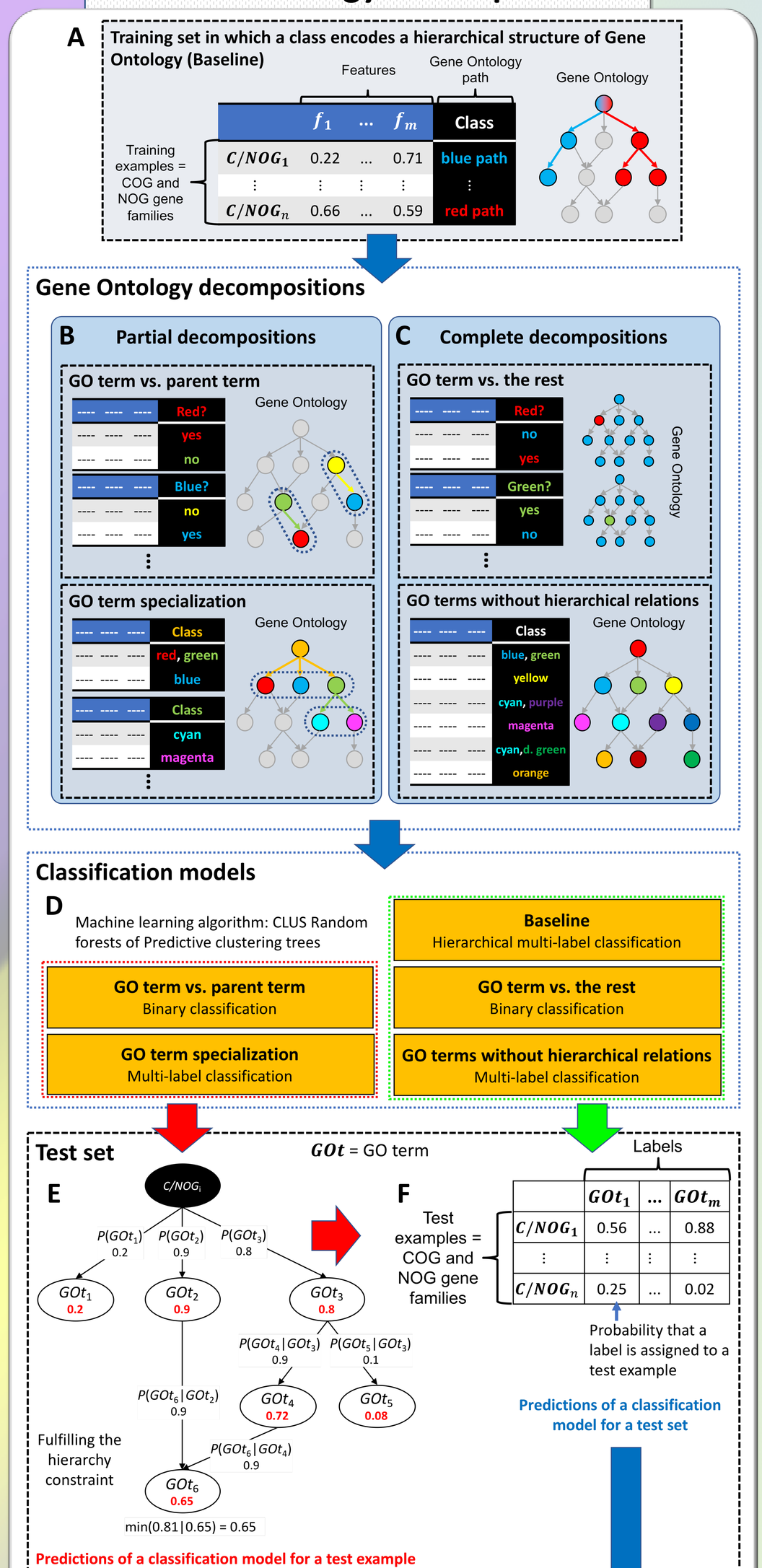
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

Contact: vedrana.vidulin@ijs.si

Introduction

The function of many genes is still not known or it is characterized in rather general terms. The most comprehensive ontology of gene function is Gene Ontology (GO). It interconnects gene functions (GO terms) into a directed acyclic graph. Therefore, we chose to construct a classification model for gene function prediction by applying a hierarchical multi-label classification (HMC) approach [1, 2]. However, the results of a recent study showed that information from the hierarchical organization of the labels/GO terms does not necessarily improve predictive performance in an ensemble setting [3]. Motivated by those results, we pose a question of whether GO decomposition can result in a more accurate model than the HMC approach.

Gene Ontology decomposition



Predictive performance analysis

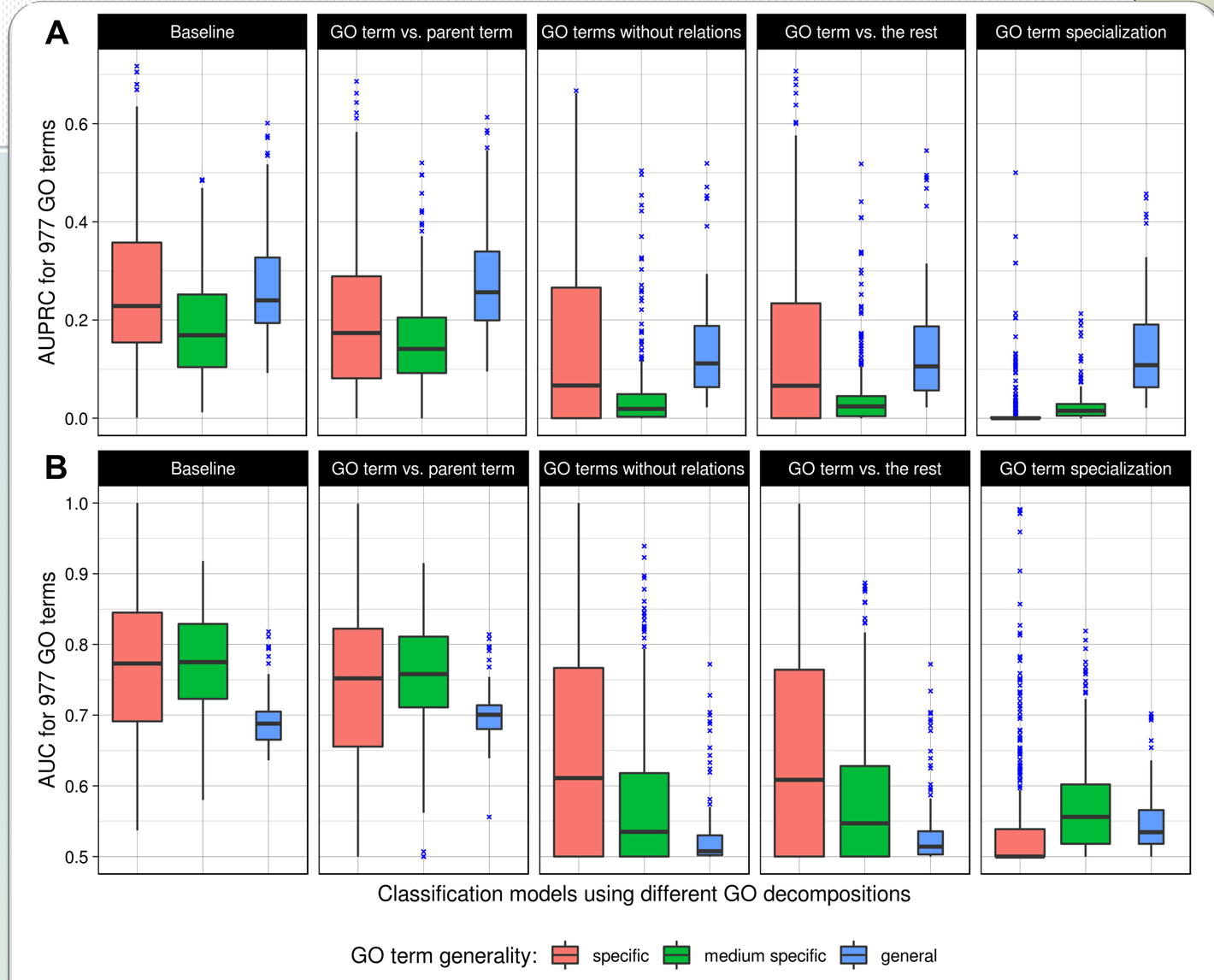


Figure 2. Distributions of GO term(label)-based accuracies of phyletic profile classification models using GO decompositions. The input data set represents phyletic profiles that encode presence/absence patterns of gene families (24,493 examples) over bacterial and archaeal organisms (6,335 features). Gene families are labeled with 1,982 GO terms. Predictive performance is estimated using 5-fold cross-validation and expressed as (**A**) AUPRC and (**B**) AUC. For each model, accuracies are shown for 977 GO terms predicted by at least one model at precision>=50%. GO terms are divided in three groups by their generality, box-plot widths being proportional to the square-roots of the number of GO terms in the bins.

Conclusions and future work

We used the phyletic profiles representation to analyze whether GO decompositions can result in more accurate model than the HMC approach. The results of analysis indicate that hierarchical organization of labels/GO terms plays an important role in constructing well performing gene function predictors. Two out of three models that exploit hierarchy outperform the models based on complete decompositions. Interestingly, the partial decomposition model GO term vs. parent terms performs comparably with the baseline HMC model. These results motivate us to further compare predictive performance of the partial decomposition models against the HMC baseline using additional data sets from the gene function prediction and other domains.

Figure 1. Gene Ontology decomposition pipeline. (**A**) An input into the pipeline is a training set in which a class encodes a hierarchical structure of Gene Ontology (GO). In other words, training examples are annotated with paths from GO. (**B**, **C**) GO decompositions module transforms the input training set into multiple training sets by applying two types of decomposition. (**B**) *Partial decompositions* construct multiple training sets encoding different segments of the GO hierarchy. The first partial decomposition *GO term vs. parent term* constructs a binary training set for each parent-child GO term in GO, composed of the training examples originally labeled with a parent GO term. In a newly created training set, the examples originally labeled with a child GO term are newly labeled as positive, while the rest of the examples are labeled as negative. The second *GO term specialization* constructs a multi-label training set for each parent-children group of GO terms in GO, where examples originally labeled with a parent GO term are added to the training set and labeled with children GO terms. (**C**) *Complete decompositions* construct one or multiple training sets with the same set of examples as the input training set, but labeled only with leaf GO terms from the paths originally assigned to examples. The first complete decomposition, *GO term vs. the rest* constructs a binary training set per leaf GO term, where the examples annotated with the GO term are labeled as positive and the rest of the examples as negative. The second, *GO terms without hierarchical relations* constructs a single multi-label training set that captures GO term cooccurrences by labeling training examples with one or several of the leaf GO terms. (**D**) Classification models module constructs five ensemble classification models from the input training set (baseline) and the training sets outputted by the decomposition module. The models are constructed using CLUS Random forests of Predictive clustering trees. (**E**, **F**) For a test example and a GO term/label, each model outputs a probability that the GO term/label is assigned to the example. (**F**) Probabilities from the baseline and the complete decomposition models are used as is. (**E**, **F**) Probabilities from the partial decomposition models are multiplied to decrease with the depth of hierarchy, fulfilling the hierarchy constraint. (**G**) Predictive performance analysis is performed on predictions/probabilities outputted by the five models for a test set. Performance is measured as area under the precision-recall curve (AUPRC) and area under the ROC curve (AUC).

References

1. Vidulin V, Šmuc T, Supek F. (2016) Extensive complementarity between gene function prediction methods. *Bioinformatics*, 32(23), 3645-3653.
2. Vidulin V, Šmuc T, Džeroski S, Supek F. (2018) The evolutionary signal in metagenome phyletic profiles predicts many gene functions. *Microbiome*, 6(1), 129.
3. Levatić J, Kocev D, Džeroski S (2015) The importance of the label hierarchy in hierarchical multi-label classification. *Journal of intelligent information systems*, 45, 247-271.

LBC 8

Find me at: