Gene function prediction using Gene Ontology decomposition

Vedrana Vidulin, Sašo Džeroski
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
vedrana.vidulin@ijs.si

Introduction

The function of many genes is still not known or it is characterized in rather general terms. This is the case even with well-studied model organisms, where a quarter or more of the genes are poorly characterized [1]. The most comprehensive ontology of gene function is Gene Ontology (GO) [2], which describes molecular functions of genes, biological processes in which the genes participate and cellular components in which the processes are active. 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 [3, 4, 5]. However, the results of a recent study show that information from the hierarchical organization of the labels/functions does not necessarily improve predictive performance in an ensemble setting [6]. Motivated by those results, we pose a question of whether GO decomposition can result in a more accurate model than the HMC approach.

Method

We begin the analysis with a data set that encodes a hierarchical structure of GO. In this data set, an example labeled with a GO term is automatically labeled with all parent terms from GO. A classification model is constructed with the HMC approach. We name this approach a *baseline* approach. Then, we apply two complete decompositions that label examples with the most specific GO terms associated with the examples. The first *GO term vs. the rest* constructs a binary data set per a GO term, where the GO term represents a positive label and all the other GO terms a negative label. It decomposes an HMC problem into multiple binary classification problems. The second *GO terms without hierarchical relations* constructs a single multi-label data set that captures GO term cooccurrences. It decomposes an HMC problem into a multi-label classification problem by treating all GO terms as being of equal weight. Finally, we apply a partial decomposition *GO term vs. parent terms* where each model training step exploits a segment of a hierarchical structure. For each parent-child pair in GO, a binary training set is constructed, where examples annotated with a child GO term are labeled as positive and the rest of examples annotated with a parent GO term as negative.

We constructed the four ensemble models using CLUS Random forests [3], which handles HMC, multi-label and binary classification tasks using the same framework of predictive clustering trees. For a test example and a GO term, each model outputs a confidence that the term is assigned to the example. Confidence is a value between zero and one, where a higher value indicates a higher level of certainty. For the baseline approach and the complete decompositions we extracted confidences from CLUS. For GO term *vs.* parent terms approach we computed a confidence as suggested in [3], that is, by multiplying a confidence from a parent-child model P(GO term|parent(GO term)) with a confidence P(parent(GO term)) from the GO term *vs.* the rest model for parent(GO term). Since GO is a directed acyclic graph, a GO term can have multiple parents. The final confidence is a minimum of confidences computed for a GO term *vs.* all of its parents.

Results and discussion

We constructed the four gene function predictors from the phyletic profiles feature set [4, 7], which represents the presence and absence patterns of gene families across genomes. More specifically, the data set is composed of 24493 eggNOG gene families [8] as examples and 6335 bacterial and archaeal

genomes as features. The examples are labeled with 1982 GO terms (from the UniProt GOA database [9] of November, 2016) covered with at least five examples.

Predictive accuracy is evaluated using 5-fold cross-validation and measured for each GO term as an area under the precision-recall curve (AUPRC). Using a precision-recall curve for a GO term, we compute a precision at which a classification is made and consider as learnable those GO terms that receive at least one prediction at precision ≥ 0.5 by any of the four classification models.

Comparisons among the four models (for 949 learnable GO terms) show that the baseline model performs the best, while the partial decomposition model performs the worst (Fig. 1). The latter performs the worst on specific GO terms, which provide the most detailed characterization of gene function. The specific GO terms are represented with low number of training examples and often yield low confidences. When low confidences are multiplied and minimum applied, the resulting confidences are too low, which leads to low AUPRC.

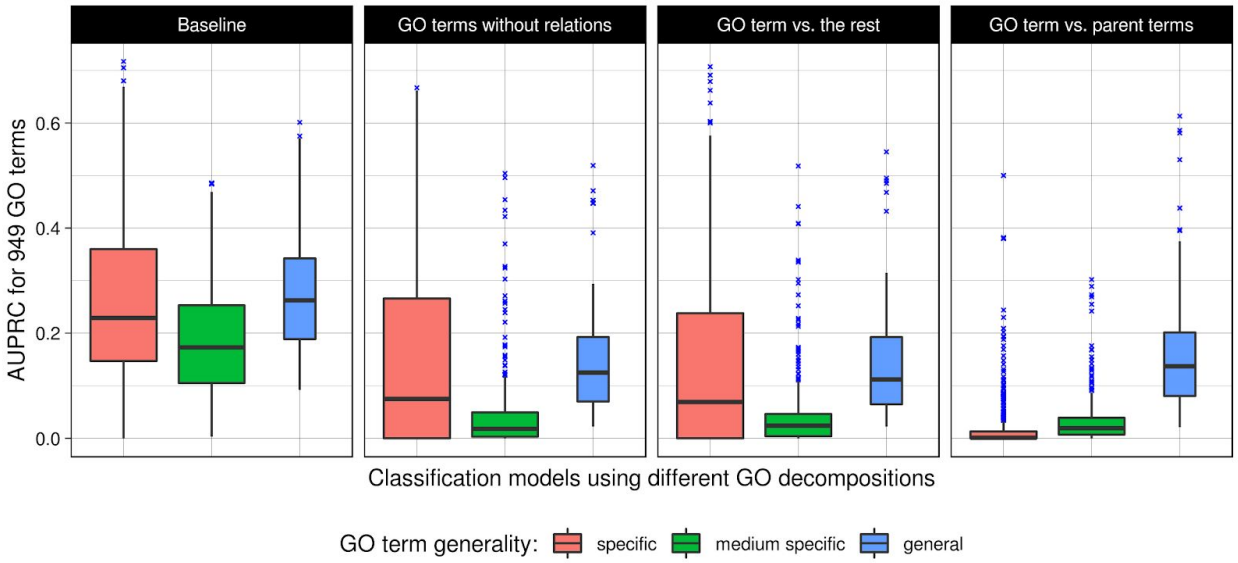

Fig. 1. Distributions of GO term-based accuracies of phyletic profile classification models using different GO decompositions. For each model (panel), accuracy (expressed as AUPRC) is shown for 949 learnable GO terms, stratified by the GO term generalty. Box-plot widths are proportional to the square-roots of the number of GO terms in the bins.

Conclusions and future work

We examined whether an accuracy of gene function predictor can be improved by decomposing GO. The results show that decomposition does not help. However, the role of GO hierarchy in constructing accurate models remains unclear since the best and the worst performing models both exploit information from the hierarchy, leaving the approaches that do not in the middle. This leads to the question of whether we can improve accuracy using different variations of partial decomposition. For example, by constructing training sets from examples representing all parent nodes in a directed acyclic graph of GO, by computing maximum of confidences obtained for different parents, or by performing a multi-label classification on children GO terms. An alternative approach would be to analyze the role of hierarchy decompositions on domains other than gene function prediction, which is a rather challenging domain. Finally, different feature sets for gene function prediction could potentially lead to different conclusions.

References

1. Zhou N *et al.* (in press) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*.
2. Ashburner M *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1), 25.
3. Vens C, Struyf J, Schietgat L, Džeroski S, Blockeel H (2008) Decision trees for hierarchical multi-label classification. *Machine learning*, 73, 185-214.
4. Vidulin V, Šmuc T, Supek F. (2016) Extensive complementarity between gene function prediction methods. *Bioinformatics*, 32(23), 3645-3653.
5. Vidulin V, Šmuc T, Džeroski S, Supek F. (2018) The evolutionary signal in metagenome phyletic profiles predicts many gene functions. *Microbiome*, 6(1), 129.
6. 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.
7. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *PNAS*. 96, 4285–8.
8. Powell S *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic acids research*, 42(D1), D231-D239.
9. Camon E. *et al.* (2004) The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic acids research*, 32(suppl_1), D262-D266.