Speed and accuracy benchmarks of large-scale microbial gene function prediction

Vedrana Vidulin[1], Tomislav Šmuc[1], Fran Supek[1,2]
[1]Division of Electronics, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
[2]EMBL/CRG Systems Biology Unit, Centre for Genomic Regulation, Dr. Aiguader 88, 08003 Barcelona, Spain

Introduction

State-of-the-art machine learning approaches for microbial gene function prediction (MGFP) from genome context data can be divided into:

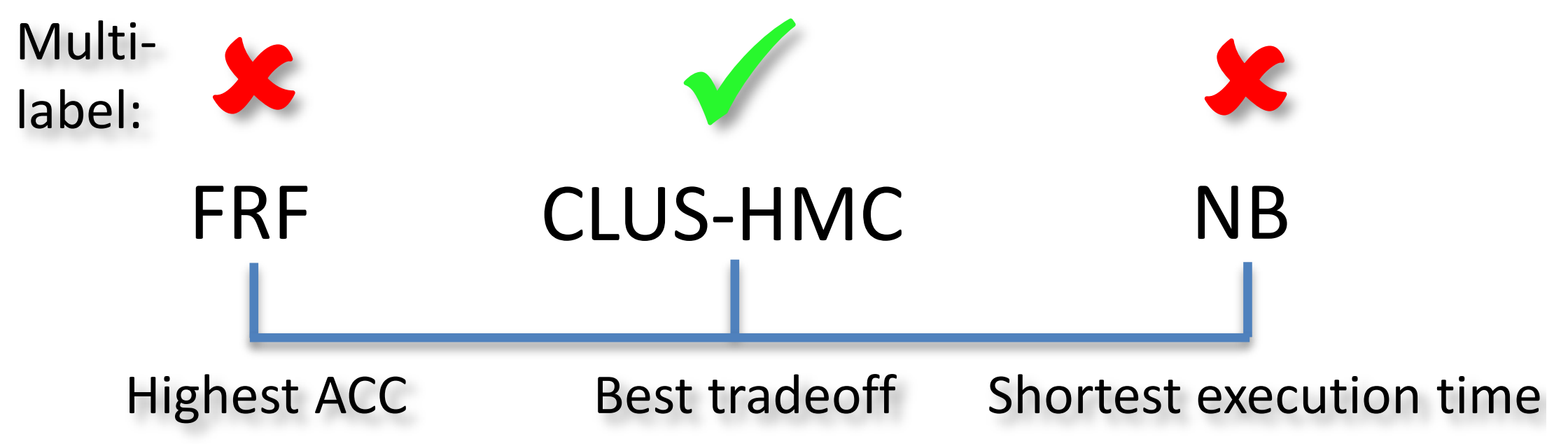
- *Unsupervised*: based on pairwise distances between individual examples arranged into "functional interaction networks";
- *Supervised*: typically predicting a limited set of functions and/or using a single-label approach to classification, constructing a separate classifier for each function and ignoring the relationships between the functions.

Multi-label MGFP approaches may perform better, especially those that can exploit the relationships between functions readily available in gene function ontologies.

We compared single *vs.* multi-label approaches on MGFP based on:

- *Predictive accuracy*: high accuracy is a prerequisite for applying the classifier in real-life tasks;
- *Computational time*: a lower demand for computational time is of importance when the number of considered functions is high.

Conclusions



Multi-label:

FRF ✗ — Highest ACC

CLUS-HMC ✓ — Best tradeoff

NB ✗ — Shortest execution time

Predictive accuracy:

- *Single-label FRF outperforms multi-label CLUS*: CLUS does not fully benefit from the hierarchical relationships between the functions in an ensemble random forest setup as in the case in a non-ensemble single tree applications to other data sets.
- Ensembles (FRF and CLUS) outperform single model classifiers (NB and kNN).
- While kNN outperforms NB when the data does not contain missing values (PP), NB outperforms kNN when values are missing in the data set (TEP, CGN).

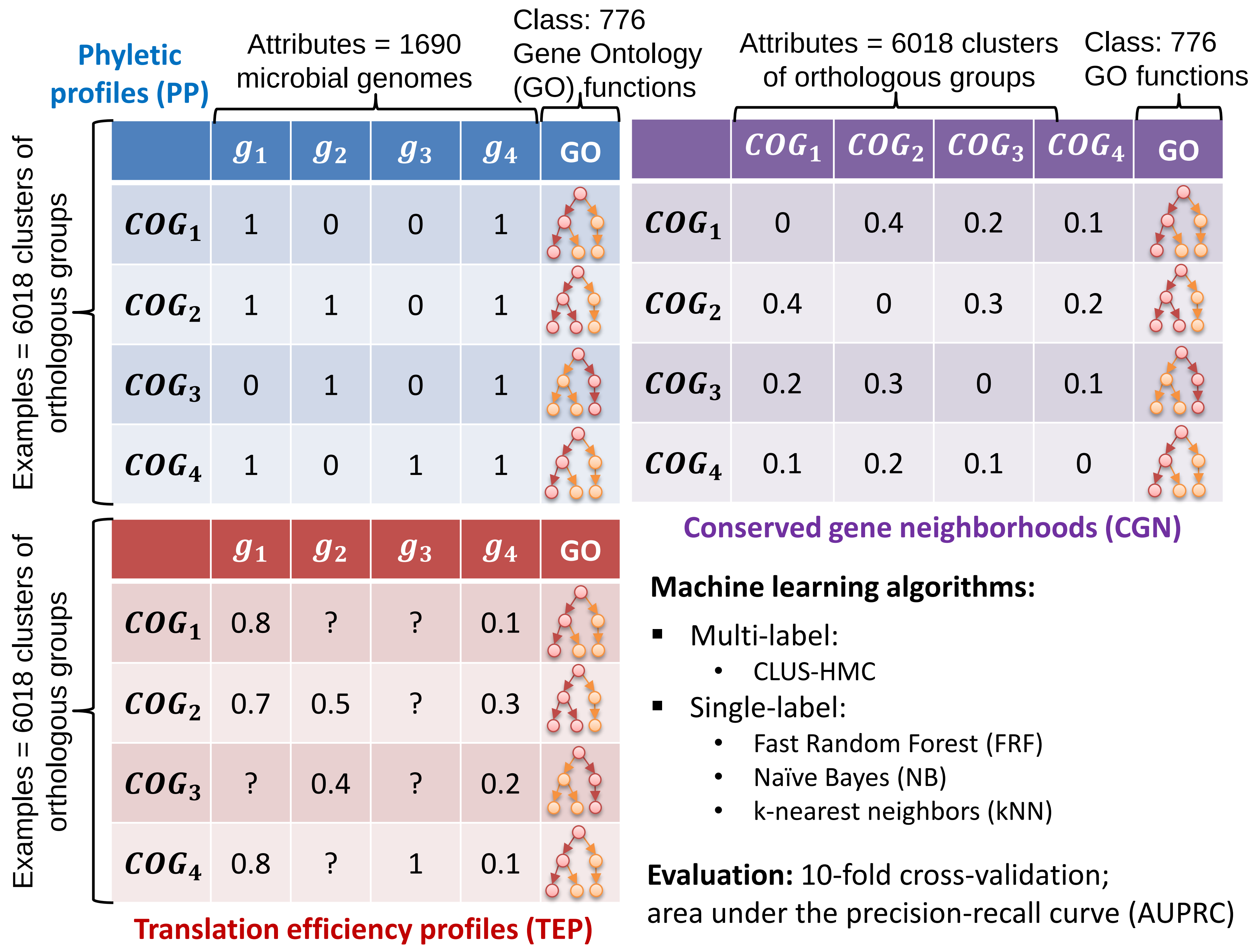
Computational time:

- *Construction of random forest ensembles is faster in the multi-label setup*: CLUS outperforms FRF.
- NB has the shortest execution time when values are missing from the data.

Acknowledgement

We would like to acknowledge the support of the European Commission through the project MAESTRA (Grant number ICT-2013-612944).

Data and Experimental Setup



Phyletic profiles (PP)

Attributes = 1690 microbial genomes

Class: 776 Gene Ontology (GO) functions

Examples = 6018 clusters of orthologous groups

	g_1	g_2	g_3	g_4	GO
COG_1	1	0	0	1	
COG_2	1	1	0	1	
COG_3	0	1	0	1	
COG_4	1	0	1	1	

Attributes = 6018 clusters of orthologous groups

Class: 776 GO functions

	COG_1	COG_2	COG_3	COG_4	GO
COG_1	0	0.4	0.2	0.1	
COG_2	0.4	0	0.3	0.2	
COG_3	0.2	0.3	0	0.1	
COG_4	0.1	0.2	0.1	0	

Conserved gene neighborhoods (CGN)

Examples = 6018 clusters of orthologous groups

	g_1	g_2	g_3	g_4	GO
COG_1	0.8	?	?	0.1	
COG_2	0.7	0.5	?	0.3	
COG_3	?	0.4	?	0.2	
COG_4	0.8	?	1	0.1	

Translation efficiency profiles (TEP)

Machine learning algorithms:

- Multi-label:
  - CLUS-HMC
- Single-label:
  - Fast Random Forest (FRF)
  - Naïve Bayes (NB)
  - k-nearest neighbors (kNN)

Evaluation: 10-fold cross-validation;
area under the precision-recall curve (AUPRC)

Figure 1: *Three data representations and machine learning setup used in experiments.* Attribute values represent presence/absence of genes in genomes in PP representation, predicted gene expression levels in TEP, and average chromosomal pairwise distances between genes across all genomes in CGN.

Results

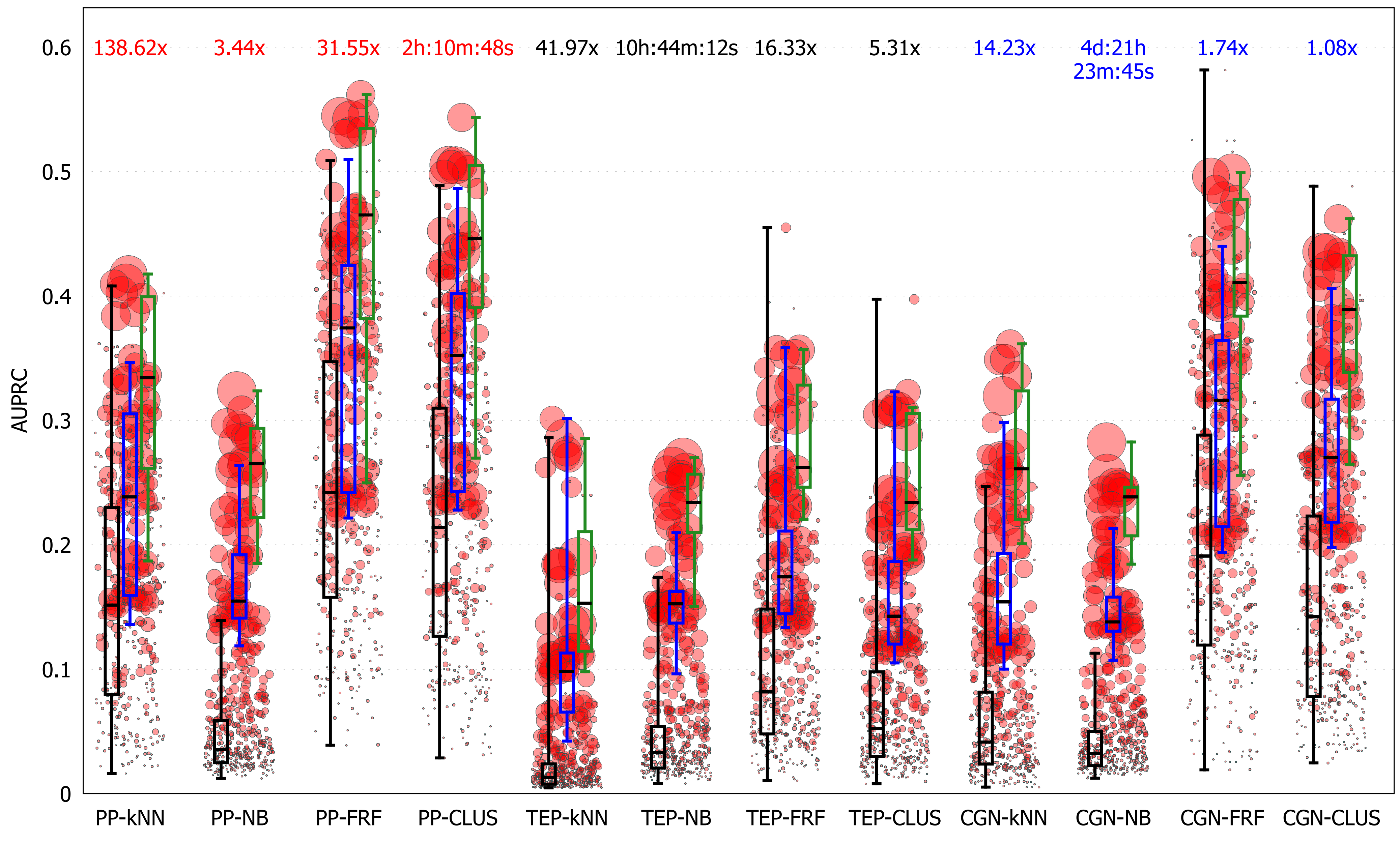


Figure 2: *Results of comparisons.* The names of classifiers on x-axis are composed of representation and algorithm name abbreviations. Circles represent GO functions and their size denote GO categories' frequencies. Very general functions (frequency > 30%) are omitted. The three box-plots for each classifier represent classifier's performance on small (frequency <= 10% – black), medium (10% < frequency <= 20% – blue) and large (20% < frequency <= 30% – green) categories. Box-plots represent minimum, first quartile, median, third quartile and maximum accuracy. Execution times are given above the box-plots, where comparisons for each representation are marked in different color. The fastest execution times are expressed in absolute numbers, while the other executions times are expressed relative to the fastest times.

Future Work

Since new microbial genomes are sequenced at a high pace, we plan to extend used data sets by including new genomes and to repeat the analysis. A high proportion of missing values makes high prediction accuracy challenging to attain for the CGN and particularly for the TEP representation. We plan to group genomes by environments and phylogeny in order to reduce the number of missing values. We also plan to make the late fusion of predictions made by the models constructed from the three representations since we expect that the three models will make complementary predictions.