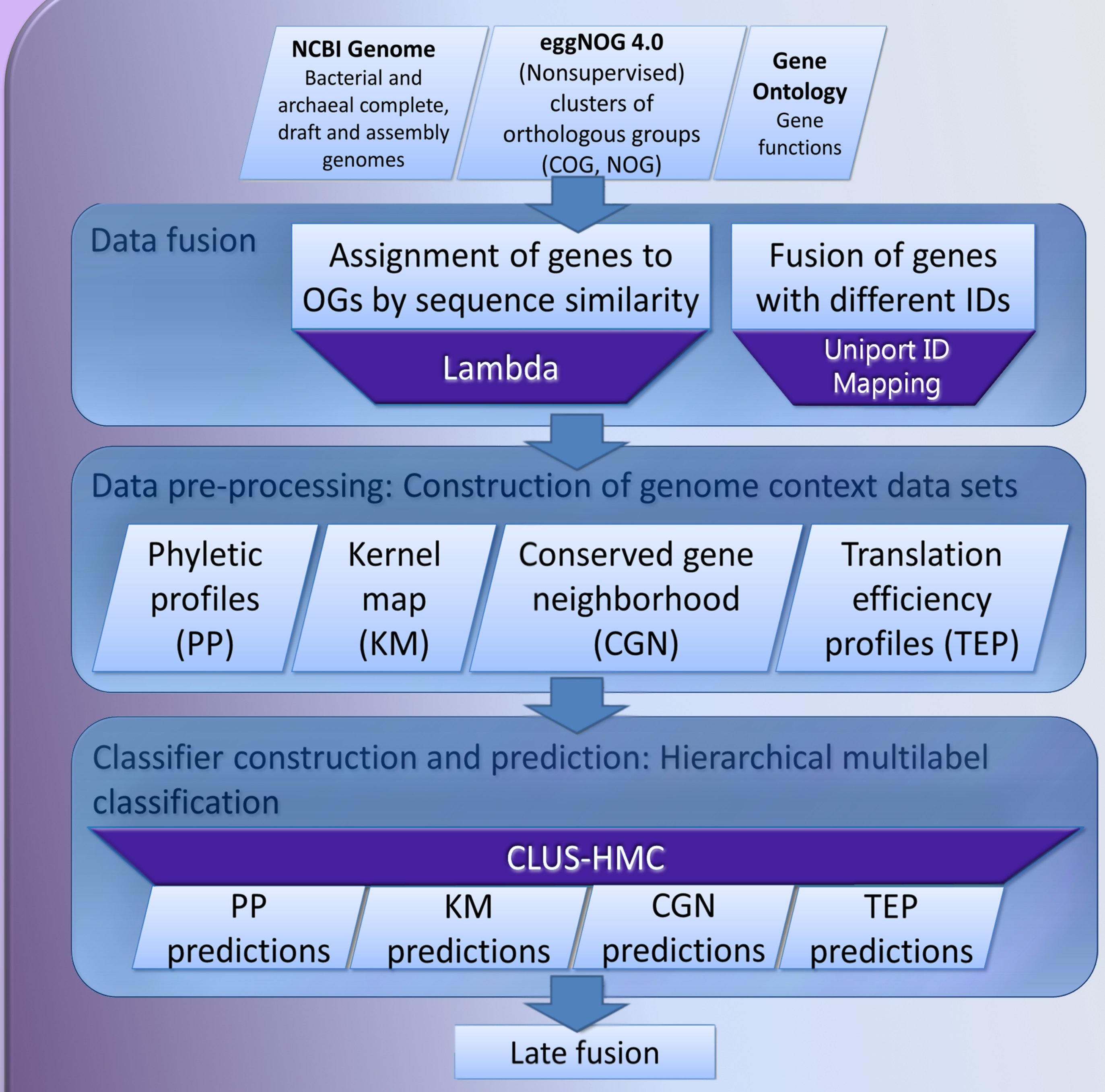
Predicting Microbial Gene Function on a Massive Scale Reveals Extensive Complementarity between Genome Context Methods

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

Abstract

We present a novel pipeline for annotating prokaryotic genes with Gene Ontology functions based on supervised machine learning in a hierarchical multi-label setting. The pipeline includes four genome context methods, which are complementary to a large extent. The four methods are combined and information accretion measured to compare the amount of past vs. newly predicted knowledge on gene function. Results indicate that a comprehensive use of genome context methods allows a sizable increase in knowledge regarding gene function.

Methods and Data



NCBI Genome
Bacterial and archaeal complete, draft and assembly genomes

eggNOG 4.0
(Nonsupervised) clusters of orthologous groups (COG, NOG)

Gene Ontology
Gene functions

Data fusion

Assignment of genes to OGs by sequence similarity
Lambda

Fusion of genes with different IDs
Uniport ID Mapping

Data pre-processing: Construction of genome context data sets

Phyletic profiles (PP)

Kernel map (KM)

Conserved gene neighborhood (CGN)

Translation efficiency profiles (TEP)

Classifier construction and prediction: Hierarchical multilabel classification

CLUS-HMC

PP predictions

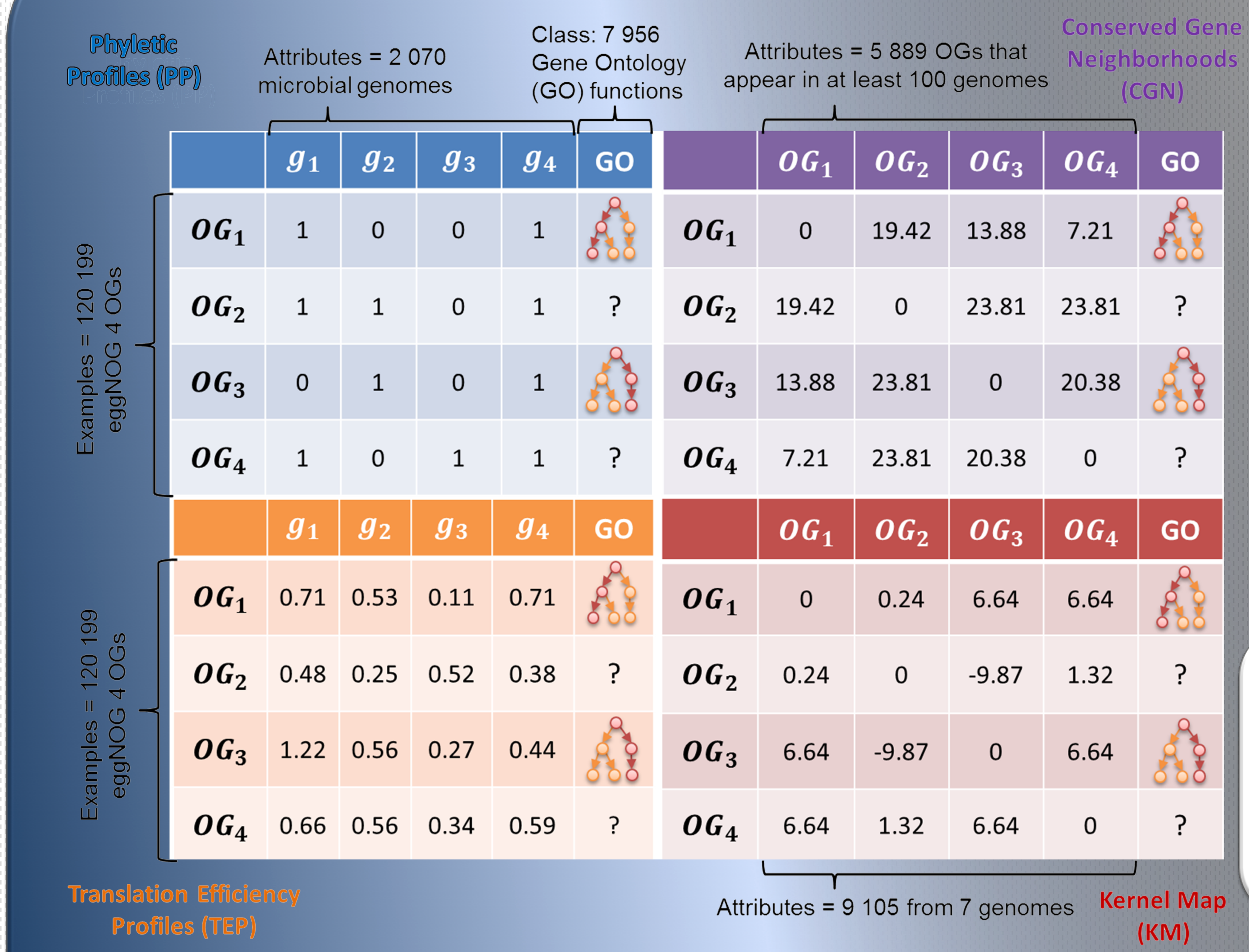
KM predictions

CGN predictions

TEP predictions

Late fusion

Figure 1. *Gene function prediction pipeline.* Accuracy of classifiers is measured on cross-validation (out-of-bag error in CLUS-HMC Random Forest) and their predictions are combined in a late fusion scheme.



Phyletic Profiles (PP)

Attributes = 2 070 microbial genomes

Class: 7 956 Gene Ontology (GO) functions

Examples = 120 199 eggNOG 4 OGs

	g_1	g_2	g_3	g_4	GO
OG_1	1	0	0	1	
OG_2	1	1	0	1	?
OG_3	0	1	0	1	
OG_4	1	0	1	1	?

Conserved Gene Neighborhoods (CGN)

Attributes = 5 889 OGs that appear in at least 100 genomes

	OG_1	OG_2	OG_3	OG_4	GO
OG_1	0	19.42	13.88	7.21	
OG_2	19.42	0	23.81	23.81	?
OG_3	13.88	23.81	0	20.38	
OG_4	7.21	23.81	20.38	0	?

Translation Efficiency Profiles (TEP)

Examples = 120 199 eggNOG 4 OGs

	g_1	g_2	g_3	g_4	GO
OG_1	0.71	0.53	0.11	0.71	
OG_2	0.48	0.25	0.52	0.38	?
OG_3	1.22	0.56	0.27	0.44	
OG_4	0.66	0.56	0.34	0.59	?

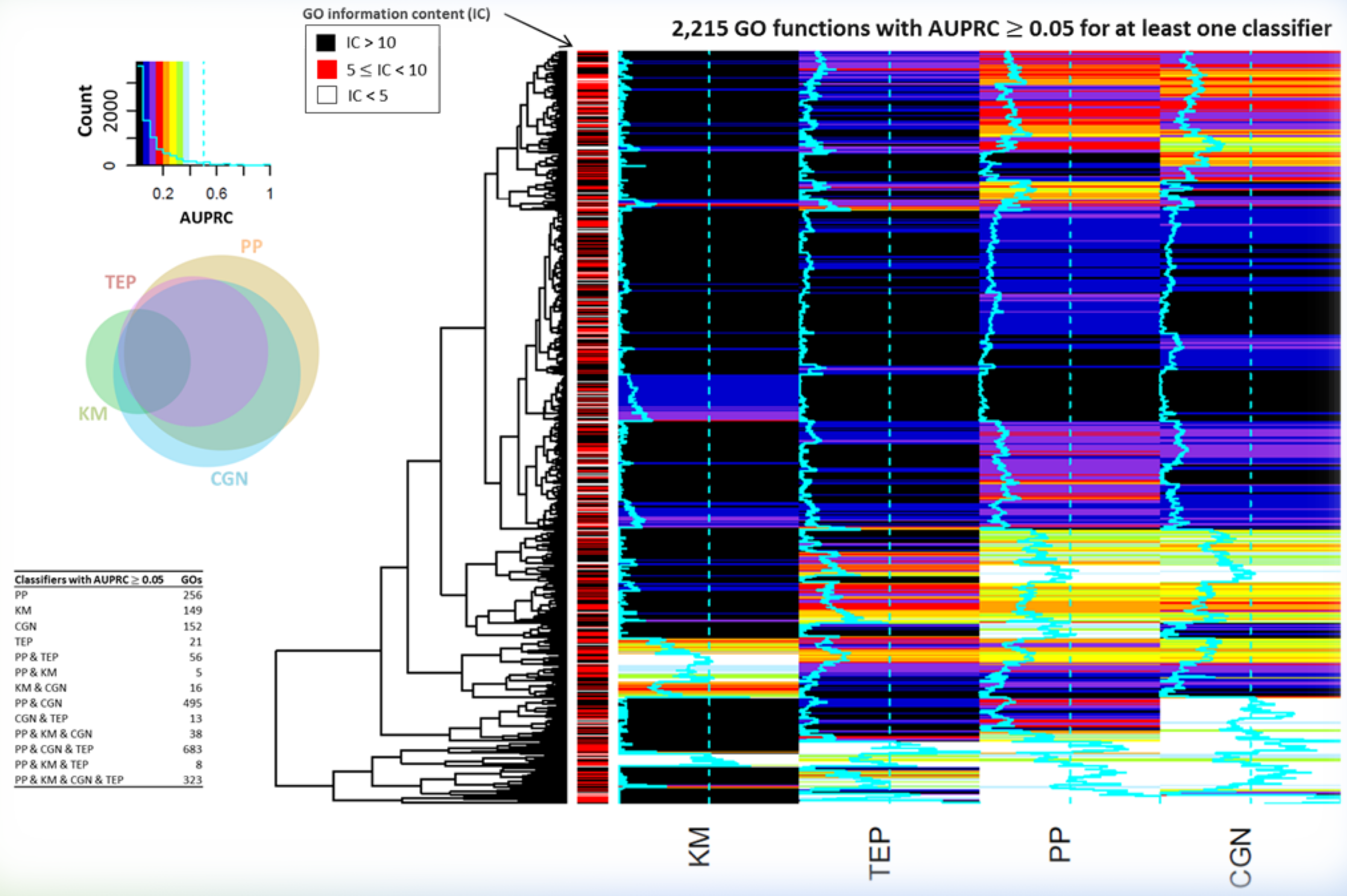
Kernel Map (KM)

Attributes = 9 105 from 7 genomes

	OG_1	OG_2	OG_3	OG_4	GO
OG_1	0	0.24	6.64	6.64	
OG_2	0.24	0	-9.87	1.32	?
OG_3	6.64	-9.87	0	6.64	
OG_4	6.64	1.32	6.64	0	?

Figure 2. *Genome context representations.* Attribute values represent presence of genes in genomes in PP, predicted gene expression levels in TEP, logarithm of average chromosomal pairwise distances between genes across all genomes in CGN, and logarithm of e-values in KM. 43% of OGs are used for training classifiers.

Results



Classifiers with AUPRC ≥ 0.05	GOs
PP	256
KM	149
CGN	152
TEP	21
PP & TEP	56
PP & KM	5
KM & CGN	16
PP & CGN	495
CGN & TEP	13
PP & KM & CGN	38
PP & CGN & TEP	683
PP & KM & TEP	8
PP & KM & CGN & TEP	323

Figure 3. *Complementarity between genome context methods*. From 2,215 functions that were successfully learned (cross-validation AUPRC>0.05) 578 were learned exclusively by one method.

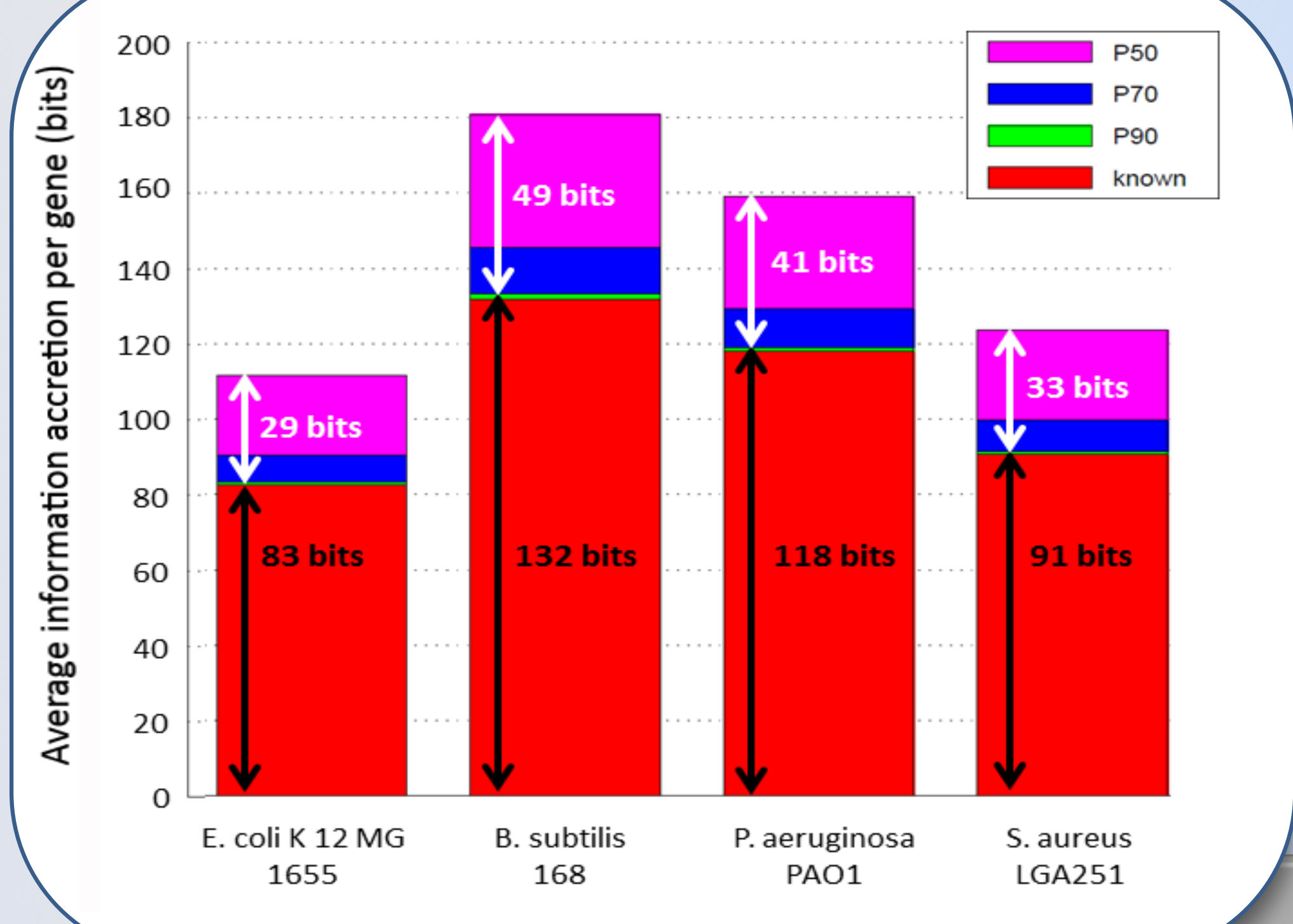
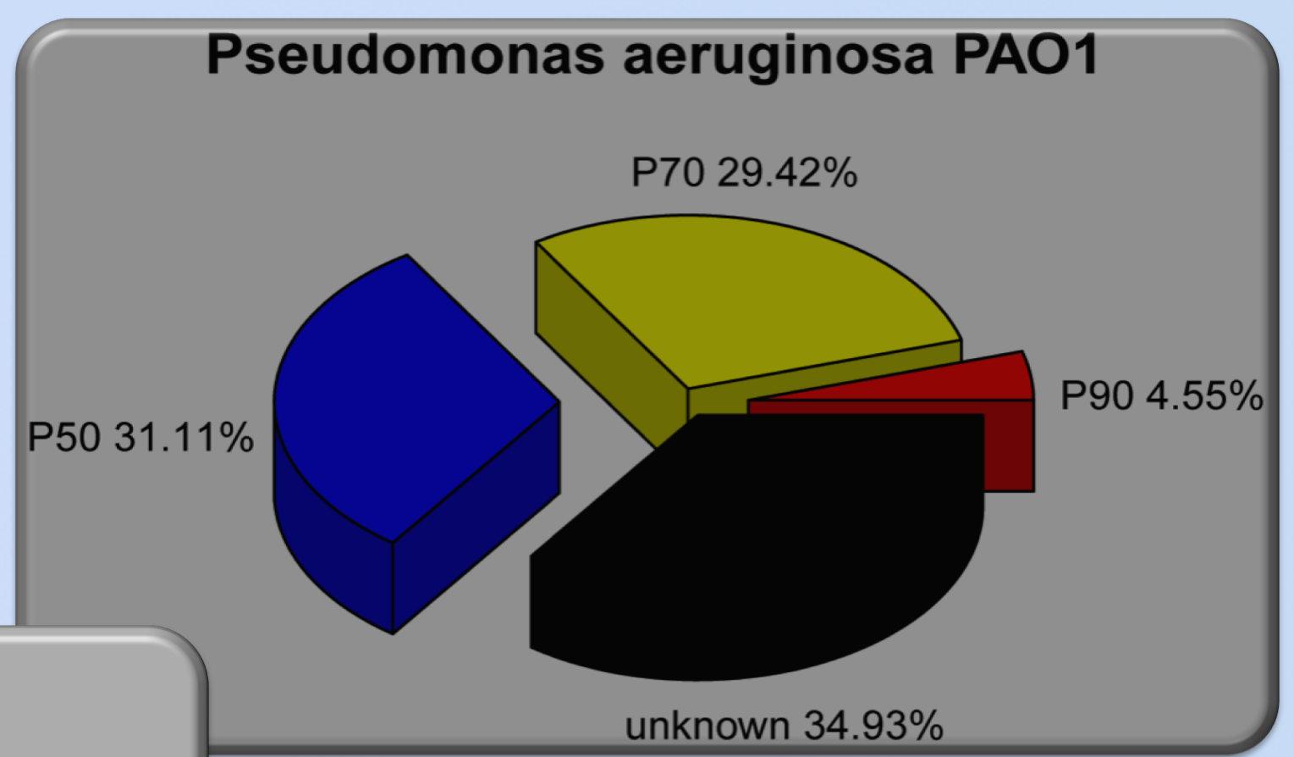


Figure 4. *Information accretion indicating the amounts of past vs. newly predicted knowledge on gene functions.*



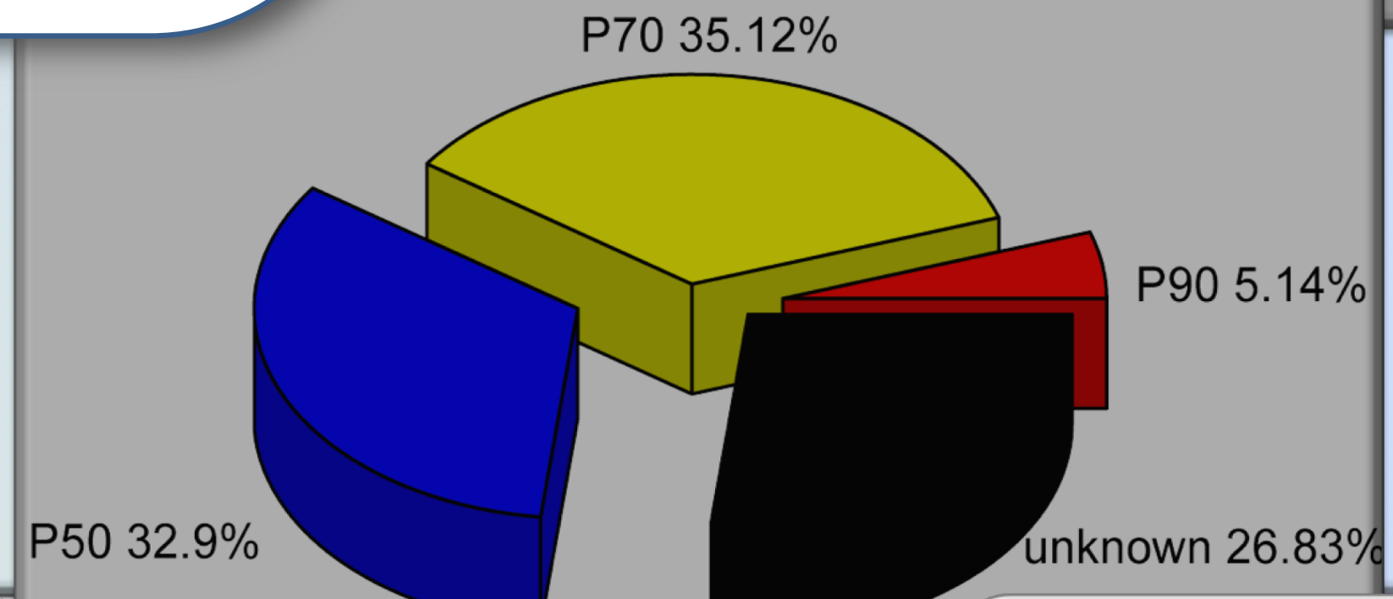
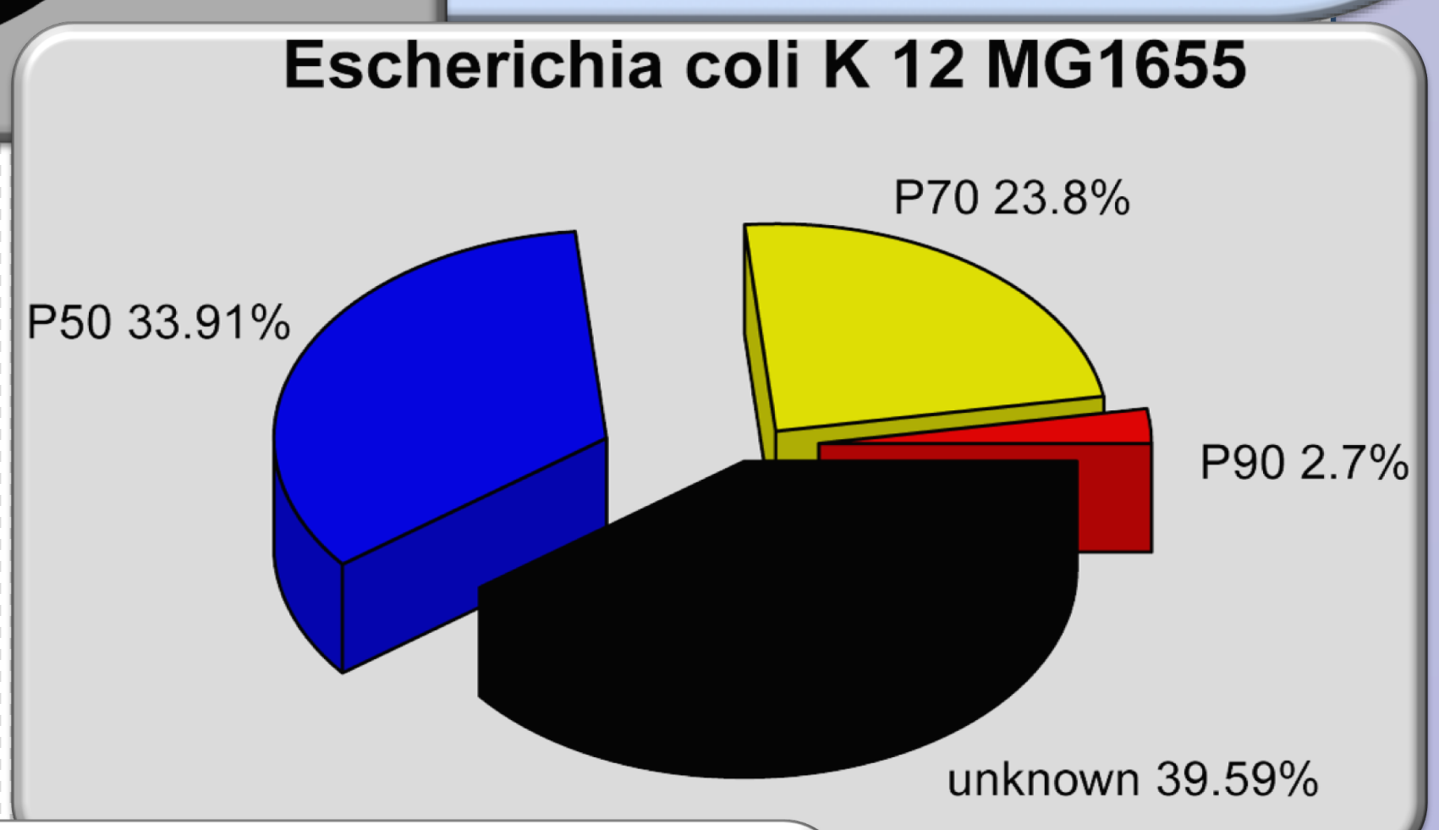
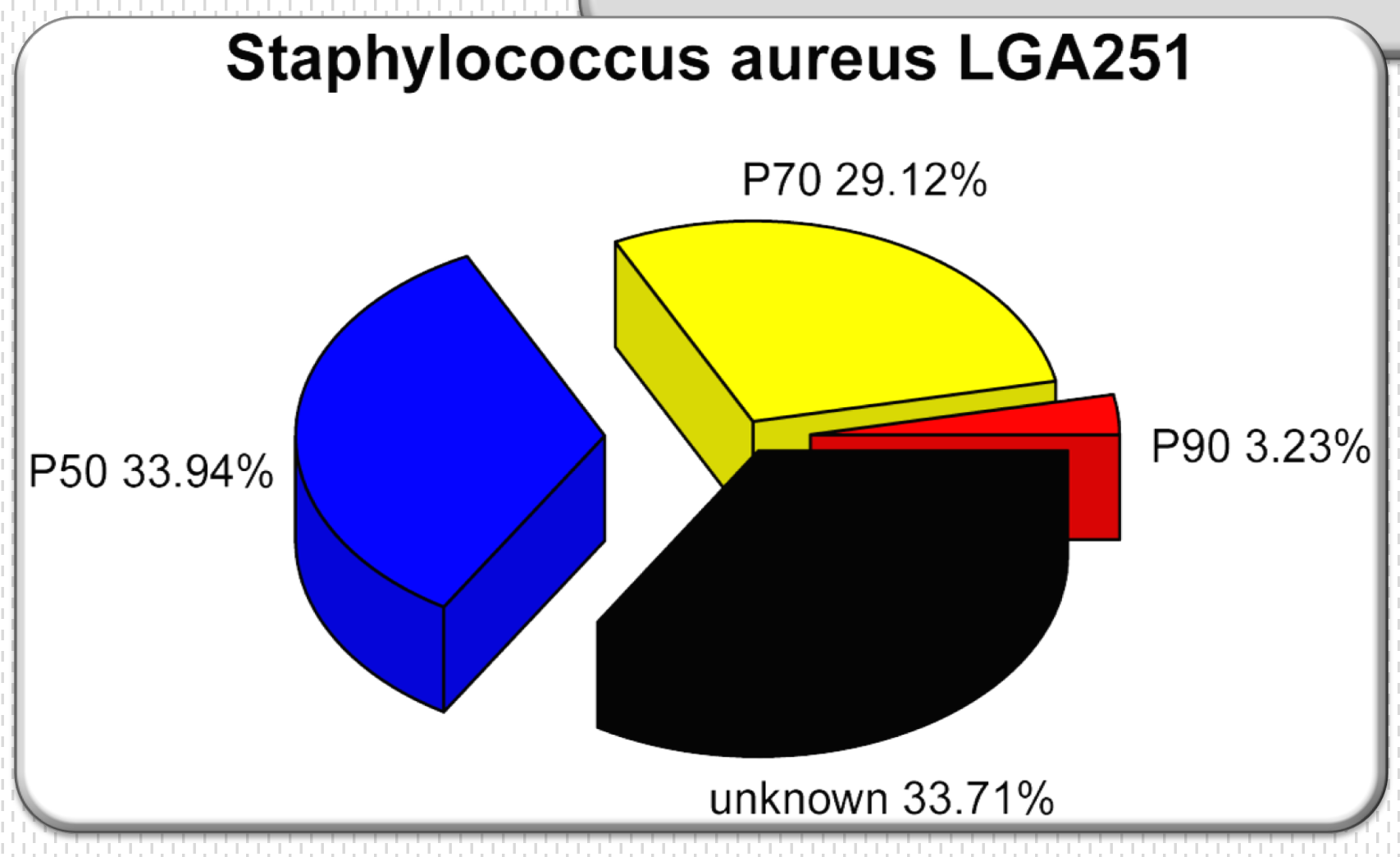


Figure 5. *Genes from model organisms that received new annotations.*





Conclusions

Evaluation of proposed pipeline's performance shows a sizable increase in knowledge regarding gene function. 60.41% E. coli OGs received at least one novel and likely correct (precision >50%) function. Model bacteria have 83–132 bits/gene of known annotations, while our pipeline annotates on average 38 additional bits per gene.

Acknowledgement

We would like to acknowledge the support of the European Commission through the projects MAESTRA (grant number ICT-2013-612944) and InnoMol (FP7-REGPOT-2012-2013-1, grant agreement number 316289).