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

Vedrana Vidulin[*], *Tomislav Šmuc, Fran Supek*

[*]*vedrana.vidulin@irb.hr*

We present a novel pipeline for annotating prokaryotic genes with Gene Ontology functions based on supervised machine learning in a hierarchical multi-label setting. 14,945,154 genes from 5,271 Bacteria/Archaea are used to form a singular learning dataset via mappings of the genes to 60,892 COG/NOG groups, predicting 8,005 different GO terms. Four classifiers are trained on: (i)phyletic profiles indicating presence/absence of COGs in genomes; (ii)signatures of remote homology across genomes; (iii)conserved gene neighborhoods and (iv)a novel method where evolutionary change in codon adaptation is tracked across orthologs. We measured accuracy of classifiers in crossvalidation (*out-of-bag* method in the CLUS-HMC Random Forest), and combined their predictions in a late fusion scheme. This resulted in accuracy higher than the individual components, allowing many genes to receive annotations. For instance, 61.17% *E. coli* COGs received at least one novel and likely correct (precision >50%) function. The four 'genome context' methods were complementary to a large extent: out of 216,694 function predictions made for various *E. coli* genes, 86,285 (40%) were unique to a single method. Moreover, out of 3,041 GO functions that were successfully learned (crossvalidation AUPRC>0.05), 1,316 were learned exclusively by one method. This underscores the need to combine multiple approaches. Finally, we used information accretion to compare the amount of past *vs*. newly predicted knowledge on gene function, and found that model bacteria have ~90–100 bits/gene of known annotations, while our pipeline typically annotates ~30 additional bits/gene. Thus, a comprehensive use of genome context methods allows a sizable increase in our knowledge regarding microbial gene function.