The evolutionary signal in metagenome phyletic profiles predicts many gene functions

Published in: Vidulin, V., Šmuc, T., Džeroski, S., Supek, F. (2018) The evolutionary signal in metagenome phyletic profiles predicts many gene functions. *Microbiome*, 6(1):129.

Introduction

The number of sequenced microbial genomes continuously grows and with them the number of genes with unknown biological function. Many automated function prediction (AFP) methods were proposed to infer gene function from various features derived from whole-genome sequences [1]. Microbes with fully sequenced genomes, however, do not cover the vast phenotypic diversity and the environmental niches spanned by uncultured microbial life. We therefore propose a machine learning approach that uses features derived from metagenomic data in order to improve AFP coverage and accuracy [2].

Methods

Previous research [3-6] has showed that phyletic profiles (PP) [7] are a very highly predictive method compared to others based on genome-derived features. In PP, feature values indicate the presence or absence of a gene family member (example) across genomes (features). Motivated by the predictive accuracy of PP, we propose metagenome phyletic profiles (MPP), where values are relative abundances of gene families in metagenomes. We construct MPP from 5049 metagenomes from the IMG database [8] and compare against PP constructed from 2071 microbial genomes from NCBI [9]. Both data sets cover a common set of 3536 COG gene families, which we annotate with 3358 gene functions from Gene Ontology (GO) [10] by using the UniProt GOA database. Classification models are constructed using the CLUS-HMC Random Forests algorithm [11], which draws on the hierarchical structure of the GO in order to reduce overfitting.

Results

First we compare MPP against PP in terms of the number of 'learnable' gene functions (GO terms) covered by at least one prediction at precision (Pr) thresholds of ≥50%, 70% and 90% (Pr is equivalent to 1 - false discovery rate). We find that MPP is the only method that can provide any predictions for 15-29% of the learnable functions (Fig. 1A; 819, 714 and 664 GO terms were learnable by both MPP and PP, at the 50, 70 and 90% thresholds, respectively). This is a similar proportion to that of the functions learnable by PP but not MPP (29-32%). In other words, metagenomic data can help predict hundreds of gene functions that would not be predicted using only PP constructed from the whole-genome sequences. We then divide the MPP data set into seven parts, according to the environments from which metagenomes were sampled: freshwater, marine, thermal spring, soil, engineered, human-associated and plant-associated (Fig. 1B). Interestingly, at Pr≥50%, 21% out of total 725 GO terms learnable by any of 7 MPPs are learnable by using only one MPP, representing a single environment, but not by others (Fig. 1C). Moreover, 49% functions are learnable from less than half of the environments (up to 3 of 7) but not the remaining half. At the more stringent threshold of Pr≥90%, this proportion of GO terms learnable exclusively from a single environment in our data are even higher, amounting to 30% (out of 601 functions). For example, the GO term 'Alcohol metabolism' is learnable from human-associated metagenomes considerably better than from the rest of the environments (Fig. 1D; cross-validation area under precision recall curve=0.09, next best is "engineered environment" at 0.03). At the same time, distribution of the GO term relative abundances across metagenomes from different environments shows that this function is generally enriched in human-associated microbes, compared to microbes from the other environments (Fig. 1E). These results indicate the importance of sampling metagenomes from diverse environments in order to increase the number of successfully predicted functions by MPP. Our results are additionally confirmed by a simulation in which we gradually add metagenomes to the MPP in order to keep the highest, or the lowest possible diversity of sampled environments. With the high-diversity sampling approach, the maximal accuracy is reached with only 500 metagenomes, since representatives from all seven environments are included (Fig. 1F). With the minimum-diversity approach, the same accuracy is achieved only with the complete set of 5049 metagenomes.

Conclusions

We introduce metagenome phyletic profiling (MPP), a novel genome context method. In combination with machine learning that explicitly accounts for the hierarchy of GO terms, MPP can predict many gene functions that would not be predicted using only the standard PP constructed from whole-genome sequences. The predictive accuracy of MPPs increases with the diversity of the environments from which metagenomes are sampled.

References

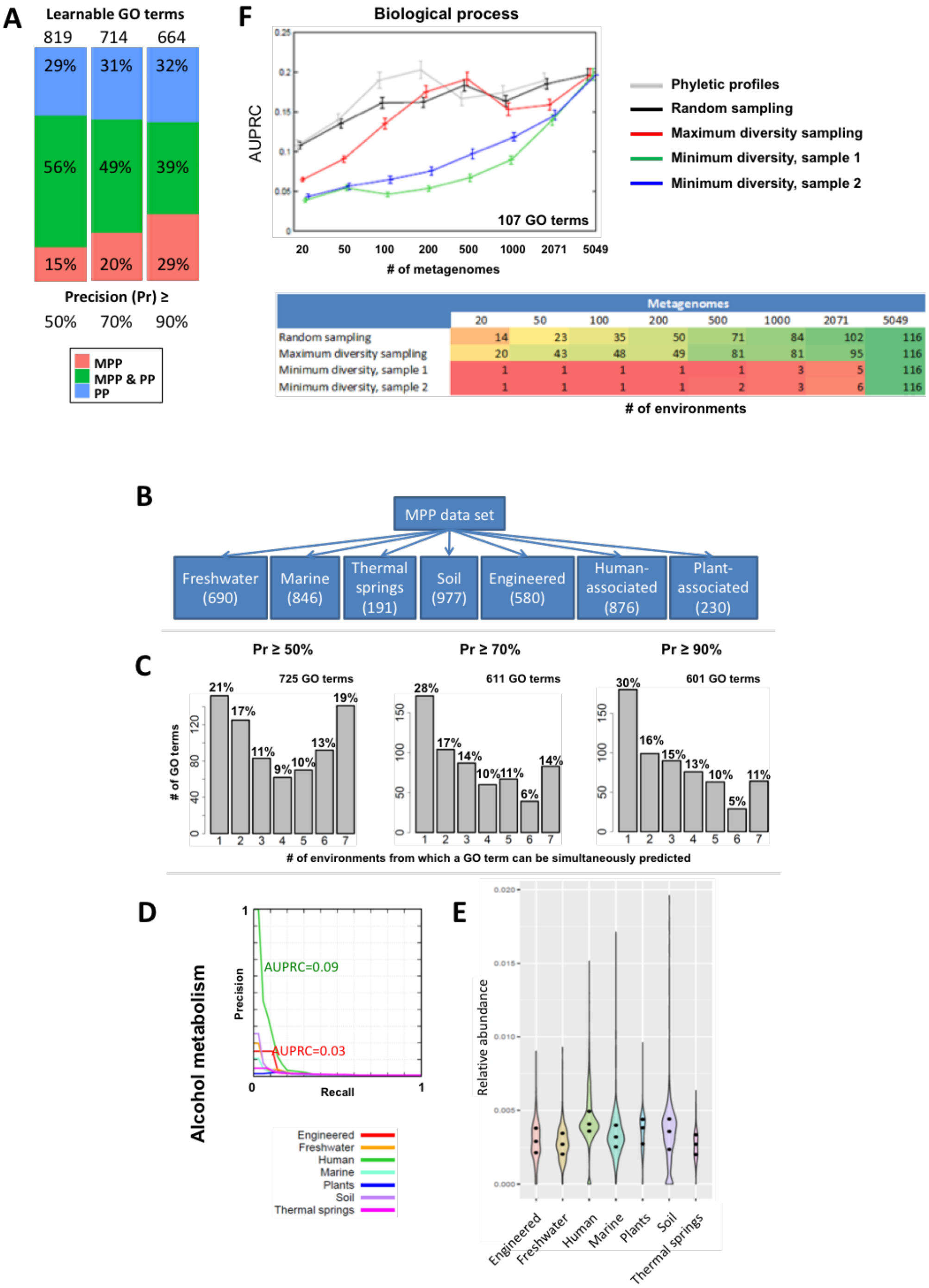
1. Jiang, Y. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(1):184.
2. Vidulin, V. *et al.* (2018) The evolutionary signal in metagenome phyletic profiles predicts many gene functions. *Microbiome*, 6(1):129.
3. Vidulin, V. *et al.* (2016) Extensive complementarity between gene function prediction methods. *Bioinformatics*, 32(23):3645-3653.
4. Enault, F. *et al.* (2005) Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. *BMC Bioinformatics*, 6(1):247.
5. Tian, W. *et al.* (2008) Combining guilt-by-association and guilt-by-profiling to predict *Saccharomyces cerevisiae* gene function. *Genome Biology*, 9(1):1.
6. Huynen, M. A. *et al.* (1998) Measuring genome evolution. *Proceedings of the National Academy of Sciences*, 95(11):5849-5856.
7. Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285-4288.
8. Markowitz, V. M. *et al.* (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*,

42(D1):D568-D573.

9. Sayers, E. W. *et al.* (2012) Database resources of the national center for biotechnology information. *Nucleic acids research*, 40(D1):D13-D25.

10. Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Computational Biology*, 5(7), e1000431.

11. Vens, C. *et al.* (2008) Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185-214.



	Metagenomes							
	20	50	100	200	500	1000	2071	5049
Random sampling	14	23	35	50	71	84	102	116
Maximum diversity sampling	20	43	48	49	81	81	95	116
Minimum diversity, sample 1	1	1	1	1	1	3	5	116
Minimum diversity, sample 2	1	1	1	1	2	3	6	116

# of environments

Fig. 1 A) Overlap between MPP and PP, expressed as percent of GO terms predicted only by MPP, only by PP or by both. **B**) Breakdown of MPP into seven data sets representing environments from the top three levels of the environment-representing tree provided by the IMG database. The numbers in brackets show the number of metagenomes related to each environment. **C**) Proportions of GO terms that can be simultaneously predicted from a certain number of environments, expressed for three different stringencies (Pr thresholds). **D**) Precision-recall curve for a GO term “Alcohol metabolism” associated with a human host data set. **E**) Distribution of GO term relative abundances across metagenomes from different environments. Points in the violin plot represent first quartile, median and third quartile. Width of the violin plots is scaled proportionally to the number of observed metagenomes in the group. **F**) x-axis represents the number of sampled metagenomes. y-axis represents cross-validation AUPRC averaged over 107 specific GO terms (with information content > 8; information content is a negative logarithm of GO term’s frequency in the UniProt GOA database) from the Biological process GO domain. Error bars represent standard error of the mean. Maximum diversity sampling approximately retains the proportions of samples from the environments represented in the full data set. Minimum diversity sampling always begins with the largest environment (e.g., soil); in the second experiment ("sample 2") all samples representing soil were removed from the data and the sampling was started from the second largest environment. The table contains the number of environments contained in each data set. 116 environments are taken from the fourth level of the environment-representing tree provided by the IMG database. Examples of environments are: Environmental -> Terrestrial -> Soil -> Loam, Host-associated -> Plants -> Rhizoplane -> Epiphytes, Engineered -> Wastewater -> Industrial wastewater -> Petrochemical. Abbreviations: MPP, metagenome phyletic profiles; PP, phyletic profiles; AUPRC, area under the precision-recall curve; GO, Gene Ontology; Pr, precision.