The landscape of microbial phenotypic traits and associated genes

Maria Brbić, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, Fran Supek

Bacteria and Archaea inhabit a wide spectrum of ecological niches, including growth in extreme environments and association to plant or animal hosts. This is made possible by a plethora of physiological adaptations observed in prokaryotes, such as the use of different carbon sources and electron acceptors, resistance to stressors and molecular interactions with host cells. Broadly construed, the notion of a microbial phenotypic trait encompasses all of the above facilities – the ability to colonize ecological niches and the underlying physiological features. The amount of prokaryotic genomes is increasing rapidly, but efforts to obtain systematic, high-quality phenotype annotation of microbes are not keeping pace, meaning that the potential for comprehensive gene-trait association studies cannot be realized. In contrast, the scientific literature abounds with trait descriptions stored as unstructured text, which are therefore not directly accessible to automated analyses. Relying on manual curation to organize these data does not scale with the increasing volume of scientific publications.

Motivated by the above, we have developed ProTraits, a resource containing ∼545,000 novel phenotype inferences, spanning 424 traits assigned to 3,046 bacterial and archaeal species; available online at http://protraits.irb.hr/. These annotations are assigned by a computational pipeline that associates microbes with phenotypes by text-mining the scientific literature and the broader World Wide Web, while also being able to define novel concepts from unstructured text. In particular, we applied non-negative matrix factorization algorithm (NMF) to model phenotypic concepts across the texts resulting in 113 non-redundant traits discovered from biological literature *de novo*.

Furthermore, we complement text-mining inferences by using an orthogonal approach, which relies on genome data. These include three data types well-known to be informative of microbial phenotypes, in particular (i) the amino acid content of the proteome (ii) the gene repertoire of the genome, and (iii) the patterns of co-occurrence of prokaryotic taxa. Moreover, we employ two methods able to infer phenotypes (iv) from conserved gene neighborhoods and (v) from evolution of synonymous codon biases. Notably, we find that gene synteny is highly predictive of many phenotypes, and highlight examples of gene neighborhoods associated with spore-forming ability. A global analysis of trait interrelatedness outlined clusters in the microbial phenotype network, suggesting common genetic underpinnings.

Given that the resulting machine learning models may be complex and thus challenging to interpret, we systematically gauged the accuracy of the FDR estimates provided for the inferences in ProTraits by an expert evaluation of a large sample of the predictions *via* a literature review. The results suggest that our FDR estimates are trustworthy. Since the supplied confidence scores estimates have a probabilistic interpretation, they enable the users of the ProTraits resource to make informed decisions about how best to use this massive data set in their work.

The broad coverage with phenotype annotations allows us to systematically discover 57,088 high confidence statistical associations that link genes to traits, recovering many known associations involving sporulation, flagella, catalase activity, aerobicity, photosynthesis and other traits. These significant gene family-trait links represent a 6.6-fold increase over the genes that could be implicated by using only the previously available databases. Finally, our analyses suggest that over 99% of the commonly occurring gene families are involved in genetic interactions conditional on at least one phenotype, suggesting that epistasis has a major role in shaping microbial gene content. This work was recently published in *Nucleic Acids Research* (Brbić *et al.*, 2016, doi:10.1093/nar/gkw964).