fusionDB: assessing microbial diversity and environmental preference via functional similarity networks

Zhu C[1], Mahlich Y[1,2,3,4]*, Miller M[1,3], Bromberg Y[1,4]
[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA; [2]Department for Bioinformatics and Computational Biology, Technische Universität München (TUM), Boltzmannstr. 3, 85748 Garching/Munich, Germany; [3]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), 85748 Garching/Munich, Germany; [4]Institute for Advanced Study, TUM, Lichtenbergstrasse 2 a, D-85748 Garching, Germany
*To whom correspondence should be addressed: ymahlich@bromberglab.org

1. INTRODUCTION

Microorganisms carry out much of the molecular functionality relevant to a range of human interests, including health, industrial production, and bioremediation. Microbial functional diversification is driven by environmental factors, *i.e.* microorganisms inhabiting the same environmental niche tend to be more functionally similar than those from different environments. In some cases, even closely phylogenetically related microbes can differ extensively across environments. Nevertheless, current taxonomy assignment is based on evolutionary distance, rather than functional similarity. Additionally, no existing databases directly link microbial functions to the environment. We previously developed *fusion* (1), a method for comparing microbial functional similarities using proteins translated from the sequenced genomes. Here we describe *fusion*DB – a novel web service enabling users to explore bacterial functional repertoires and corresponding environmental niches, as well as to map new microbial genomes to the functional spectrum of 1,374 reference bacteria.

2. MAPPING A MICROBIAL PROTEOME TO A FUNCTIONOME

Theoretically, mapping of a microbial proteome into the function space of *fusion*DB requires extensive computational resources to run a pipeline, which generated the reference *fusion*. That is, for every pair of proteins in the *fusion* reference and in the new proteome, assess shared functionality and cluster the resulting protein network into functional groups. Here we suggest speeding up computation by comparing the new proteome to the reference *fusion* directly – without recomputing the internal *fusion* similarities. The new organism functionome is assessed using our method without compromising the mean per organism precision (99.5% of the functions are correctly assigned to the organism) and recall (98.9% of the organism reference functions can be retrieved by this method).

3. ENVIRONMENT SIGNIFICANTLY AFFECTS MICROBIAL FUNCTION

As an example, we computed the functionome of the *Synechococcus sp.* PCC 7502 (GCA_000317085.1) bacterium, currently not present in *fusion*DB (Figure 1); its 3,318 proteins where mapped to 2,206 functions in under three and a half hours. This fresh water Cyanobacterium was found to share the highest functional similarity with *Synechocystis* PCC 6803, a fresh water organism closely related to *Synechococcus.* Interestingly, marine *Synechococcus* share far lower similarity to the mapped microbe.

We also evaluated the effect of the environment on bacterial functionality. We found that

organisms sharing the same environmental niche (e.g. marine bacteria), unsurprisingly, show higher functional similarity amongst themselves, than when comparing organisms of different environmental niches (e.g. aerobic microbes vs. anaerobic microbes). Notably, we observed a significantly higher functional similarity between thermophilic than between psychrophilic microbes and between marine than freshwater organisms, indicating more room for diversity and/or less environmental pressure in cold (vs. hot) and freshwater (vs. salty) environments.

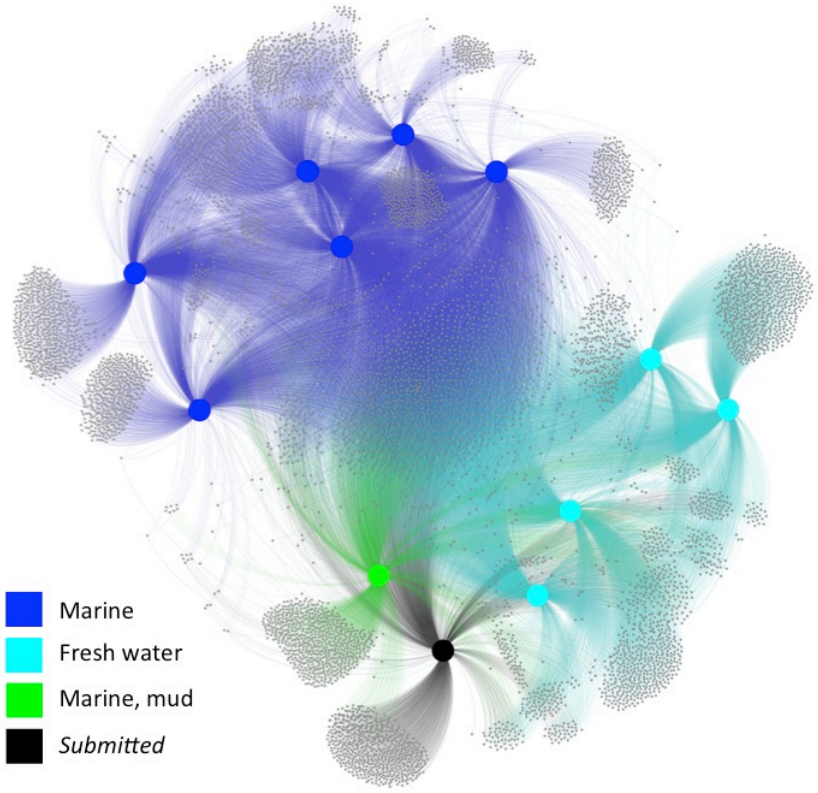

Figure 1: The fusion+ view of all Synechoccocus genomes. The new *Synechococcus sp.* PCC 7502 (black) clusters with the fresh water *Synechococcus* organisms (light blue).

4. REFERENCES

1. Zhu, C., *et al.* (2015) Functional Basis of Microorganism Classification, *PLoS Comput Biol*, **11**, e1004472