*fusion*DB: assessing microbial diversity and environmental preferences via functional similarity networks

Zhu C[1*], Mahlich Y[2*], Bromberg Y[1,2]

[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA

[2]Institute for Advanced Study, Technische Universität München, Lichtenbergstrasse 2 a, D-85748 Garching, Germany.

* Corresponding authors: czhu@bromberglab.org, ymahlich@bromberglab.org
Tel: +1.848.932.5638; Fax +1.848.932.8965

Abstract

Summary: Microbial functional diversification is driven by environmental factors. In some cases, microbes differ more across environments than across taxa. We have recently developed *fusion+,* a network-based method for comparing microbial functionality on the basis of whole proteome sequences. Here we introduce *fusion*DB, a novel database of microbial functional similarities, indexed by available environmental preferences. *fusion*DB entries represent nearly fourteen hundred taxonomically-distinct bacteria annotated with available metadata: habitat/niche, preferred temperature, and oxygen use. Each microbe is encoded as a set of functions represented by its proteome and individual microbes are connected via common functions. Any given database search produces an easily visualizable XML-formatted network file of selected organisms, displaying functions shared by the selected organisms. *fusion*DB thus provides a fast means of associating specific environmental factors with organism functions.

Availability: http://bromberglab.org/databases/fusiondb and as a sql-dump by request.

Contact: czhu@bromberglab.org, ymahlich@bromberglab.org

Supplementary information: Supplementary data are available at *Bioinformatics* online

1. Introduction

Microorganisms carry out many molecular functions relevant to a range of human interests, including health, industrial production, and bioremediation. Experimental study of these functions is expensive and time-consuming; *e.g.* as many as three hundred biochemical/physiological tests only reflect 5-20% of the bacterial functional potential (Garrity GM, 2001). The recent drastic increase in the number of sequenced microbial genomes has facilitated access to microbial molecular functionality from the gene/protein sequence side, *e.g.* via databases like Pfam (Sayers, et al., 2009). In addition, microbial environmental preferences can be found in available databases, *e.g.* GOLD (Pagani, et al., 2012). While it is well known that environmental factors play an important role in microbial functionality (Cohan, 2001), none of the existing resources directly links environment data to microbial function.

We recently mapped bacterial proteins to molecular functions (Zhu, et al., 2015) and studied the functional relationships between bacteria in the light of their chosen habitats. For example, we found that water salinity seems to drive the functional diversification in *Cyanobacteria*. To facilitate further research of this kind, we developed *fusion*DB – a database containing bacterial functional repertoires and similarities from our study and applicable environmental metadata. Our web-interface allows querying for combinations of organism names and environments to produce a network of selected organisms connected via shared functions. Our analyses of the data reveal environmental factors driving microbial functional diversification and illustrate the detection of traces of environmentally driven horizontal gene transfer (HGT).

2. Data Setup and Retrieval

Database setup. *fusion*DB contains 4,284,540 proteins from 1,374 bacterial genomes (Dec. 2011 NCBI GenBank (Benson, et al., 2009). For each bacterium, we store its (1) NCBI taxonomic information (Sayers, et al., 2009) and, where available, (2) environmental metadata (temperature, oxygen requirements, and habitat; GOLD (Pagani, et al., 2012). The environments are generalized, *e.g. thermophiles* include hyper-thermophiles, and "*No data*" indicates missing annotations (Supplementary Online Material, SOM_Table 1). The *fusion* (functional repertoire similarity-based organism network) protocol is described in (Zhu, et al., 2015), but briefly: we used HSSP distances (Rost, 2002) >10 to define two proteins of the same function and clustered all proteins into 1,235,380 function groups (functions) via MCL (Dongen,

2000). Each bacterium was thus mapped to a set of functions, its functional repertoire, consisting of unique functions and those shared with other bacteria.

Web interface. *fusion*DB is web-accessible, allowing queries of organism scientific names (*genus species;* partial matches allowed) and environmental preferences. Users can combine and filter results of multiple searches into a final organism search list (SOM_Figure 1A). From this list, the search produces a downloadable archive containing a network file, a function annotation file, and a ReadMe set of instructions. The network file is a Gephi-formatted (Bastian, et al., 2009) *fusion*+ using ForceAtlas2 (Jacomy, et al., 2014) layout (SOM_Figure 1B). There are two types of vertices (nodes) in *fusion*+: organism nodes and function nodes. Organism nodes are connected to each other only through the function nodes they share. The number of edges (degree) of an organism node represents the total number of functions of the organism. The relative position of each organism node is determined by the pull *towards* other organisms via the common functions and *away* from others via unique functions (SOM_Figure 1B). The network color scheme is pre-set by the user according to genus, temperature, oxygen requirement, or habitat. The functions file contains annotations of the functions in the network, including all non-redundant myRAST (Aziz, et al., 2008) annotations from all proteins. *fusion*DB also allows retrieval of functions (exclusively) shared by a specific subset of organisms (SOM_Figure 1C).

Organism similarity by environment. For each environment, we sampled pairs of organisms from (1) the same condition (SC, *e.g.* both mesophiles) and (2) different conditions (DC, *e.g.* thermophile *vs*. mesophile). Organisms in one pair were always selected from different taxonomic families. The smallest available set of pairs, SC-psychrophile, contained 33 organisms from 17 families (SOM_Table 1; 136 pairs – 48/88 same/different phyla; due to high functional diversity of *Proteobacteria,* its classes were considered separate phyla). For every other environment, we sampled in 100 bootstrap times 136 organism pairs for both SC and DC sets, covering this same minimum taxonomic diversity. We calculated the pairwise functional similarity (Eqn. 1) distributions and discarded organism pairs with less than 5% similarity (SOM_Table 2).

$$similarity = \frac{shared\ \ functions}{the\ larger\ \ functional\ repertoire\ size} \qquad \text{(Eqn. 1)}$$

3. *Environment significantly affects microbial function.*

Members of the mesophile-psychrophile pairs are as similar to each other as psychrophiles are amongst themselves (SC-psychrophile *vs.* DC-mesophile-psychrophile, Figure 1A), while thermophiles are most distinct from all. These results suggest that protein functional adaptation to low temperatures is less drastic than to high temperatures – an interesting finding itself.

Different molecular pathways of aerobic-respiration and anaerobic-respiration/ fermentation explain oxygen requirement-based dissimilarity (Figure 1B). Likewise, high similarity amongst facultative bacteria could be explained by the presence of both (aerobic and anaerobic) pathways and hence a bigger functional overlap.

Habitat-based SC samples also show higher pairwise organism similarity than DC samples (Figure 1C). Since living in a given habitat requires fitting multiple environmental restraints, temperature and oxygen preferences within individual habitats apparently (SOM_Figure 2) play only a minor role in determining organism similarities. SC-soil displays the lowest mean organism similarity of the habitat SC samples. This diversity is likely due to soil's heterogeneity at physical, chemical, and biological levels, from nano- to landscape scale (Lehmann, et al., 2008). In contrast, the more homogenously aquatic and salty (especially in open-ocean) marine environment is most conservative.

SC organisms across all environmental factors are significantly (Kolmogorov-Smirnov test (Massey, 1951), p-val<2.5e-6) more functionally similar than DC organisms (Figure 1; except SC-psychrophile to DC-mesophile-psychrophile, where the difference is not significant). In other words, organisms in the same environment are generally more similar than organisms from different environments. This finding is intuitive and reinforced by extensive HGT within environment-specific microbiomes (Kim, et al., 2012; Liu, et al., 2012; Saye, et al., 1987). Our results, however, for the first time quantify the environmental impact on microorganism function diversification.

4. Case study of a temperature driven HGT event.

We extracted unique thermophilic, mesophilic, and psychrophilic species representatives of the *Bacillus* genus from *fusion*DB. We added two *Clostridia* thermophiles, *D. carboxydivorans* CO-1-SRB and *S. acidophilus* TPY*,* to generate a *fusion*+ (SOM_Table 3, SOM_Figure 3). All *Bacillus* in our set shared 451 functions (pan-function repertoire), including 91 that were not in our *Clostridia* (SOM_Table 4).

Unsurprisingly, the two *Clostridia* were most like the three thermophilic *Bacilli*. Among the shared functions, three were exclusive to all five thermophiles in our set. One is pyruvate, phosphate dikinase (PPDK) that, in extremophiles, works as a primary glycolysis enzyme (Chastain, et al., 2011). Due to our stringent cut-off for defining a function, the versions of PPDK in our organisms must be very similar. However, the *Bacilli* and non-*Bacilli* in our set belong to different classes, within the same phylum. Given a distant evolutionary relationship, the most plausible explanation for this similarity is HGT. The other two shared functions are carried by proteins translated from mobile genetic elements (MGEs) that mediate the movement of DNA within genomes or between bacteria (Frost, et al., 2005). Shared closely related MGEs in distant organisms imply HGT (Krupovic, et al., 2013). We thus suggest that *fusion*DB offers a fast and easy way to trace HGT within niche-specific microbial communities.

5. Conclusions

*fusion*DB links microbial functional similarities and environmental preferences. It offers a novel, fast, and simple way to detect pan-function repertoires, unique functions, as well as traces of HGT. With more microbial genome sequencing and further manual curation of environmental metadata, we expect that *fusion*DB will become an integral part of microbial functional analysis protocols in the near future.

Acknowledgements

We thank Drs. Burkhard Rost (TU Munich), Max Haggblom and Tamar Barkay (both Rutgers), and Tom O. Delmont (U Chicago) for all discussions and to those who deposit their data in public databases.

Funding

This work was supported by the USDA-NIFA (1015:0228906) and the TU Mu□nchen – Institute for Advanced Study Hans Fischer Fellowship, funded by the German Excellence Initiative and the EU Seventh Framework Programme, grant agreement 291763.

Conflict of Interest: none declared

References

Aziz, R.K., *et al.* (2008) The RAST Server: rapid annotations using subsystems technology, *BMC Genomics*, **9**, 75.

Bastian, M., Heymann, S. and Jacomy, M. (2009) *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009.

Benson, D.A., *et al.* (2009) GenBank, *Nucleic Acids Res*, **37**, D26-31.

Chastain, C.J., *et al.* (2011) Functional evolution of C4 pyruvate,orthophosphate dikinase, *Journal of Experimental Botany*.

Cohan, F.M. (2001) Bacterial Species and Speciation, *Systematic Biology*, **50**, 513-524.

Dongen, S.v. (2000) Graph Clustering by Flow Simulation., *PhD thesis, University of Utrecht*.

Frost, L.S., *et al.* (2005) Mobile genetic elements: the agents of open source evolution, *Nat Rev Micro*, **3**, 722-732.

Garrity GM, B.D., Castenholz RW, editors (2001) *Bergey's Manual of Systematic Bacteriology, Volume 1*. Springer, New York (NY).

Jacomy, M., *et al.* (2014) ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, *PLoS ONE*, **9**, e98679.

Kim, S.E., *et al.* (2012) Monitoring of horizontal gene transfer from agricultural microorganisms to soil bacteria and analysis of microbial community in soils, *Journal of microbiology and biotechnology*, **22**, 563-566.

Krupovic, M., *et al.* (2013) Insights into Dynamics of Mobile Genetic Elements in Hyperthermophilic Environments from Five New Thermococcus Plasmids, *PLoS ONE*, **8**, e49044.

Lehmann, J., *et al.* (2008) Spatial complexity of soil organic matter forms at nanometre scales, *Nature Geosci*, **1**, 238-242.

Liu, L., *et al.* (2012) The human microbiome: a hot spot of microbial horizontal gene transfer, *Genomics*, **100**, 265-270.

Massey, F.J. (1951) The Kolmogorov-Smirnov Test for Goodness of Fit, *Journal of the American Statistical Association*, **46**, 68-78.

Pagani, I., *et al.* (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata, *Nucleic Acids Res*, **40**, D571-579.

Rost, B. (2002) Enzyme function less conserved than anticipated, *J Mol Biol*, **318**, 595-608.

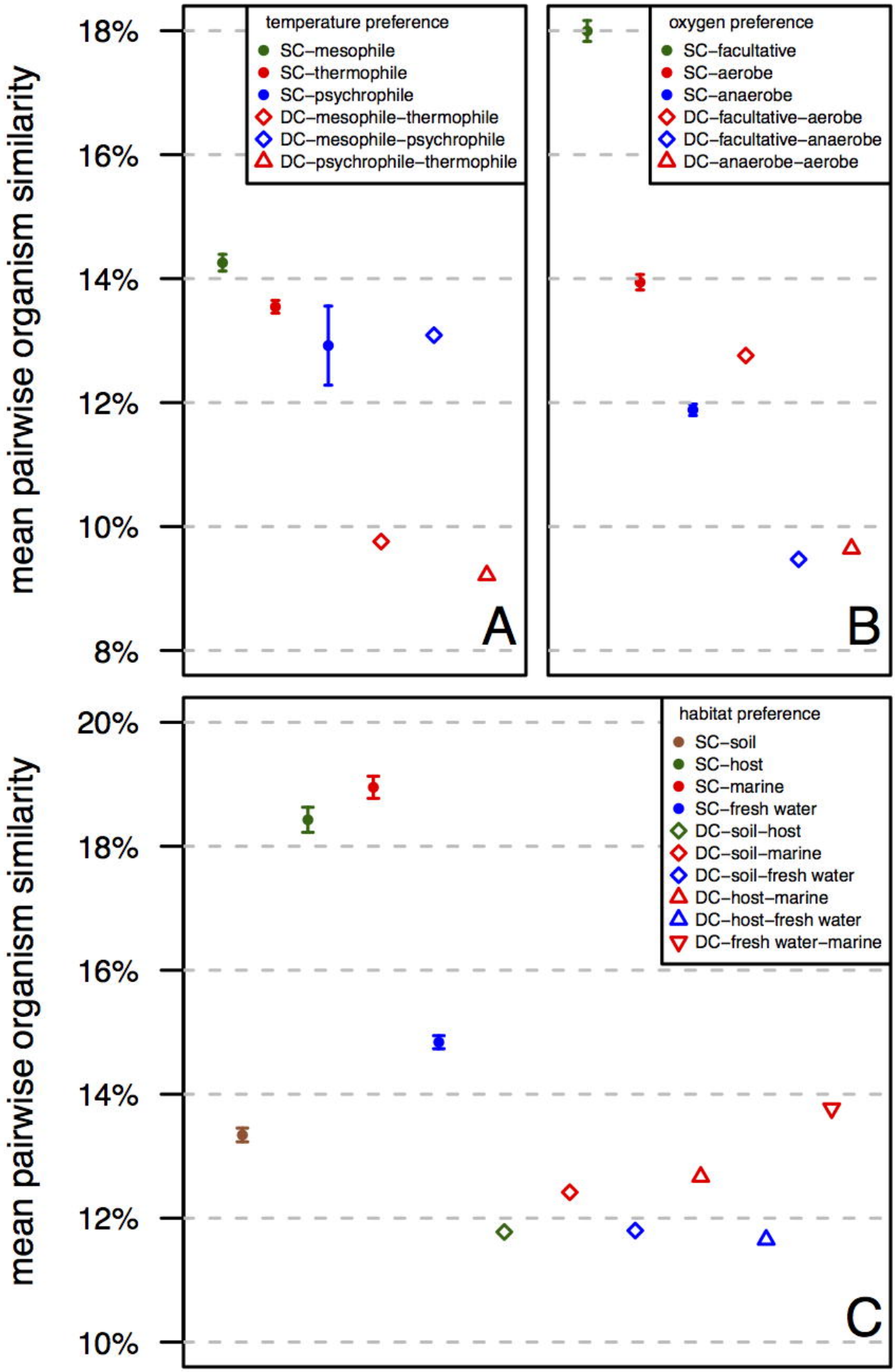
Saye, D.J., *et al.* (1987) Potential for transduction of plasmids in a natural freshwater environment: effect of plasmid donor concentration and a natural microbial community on transduction in Pseudomonas aeruginosa, *Applied and Environmental Microbiology*, **53**, 987-995.

Sayers, E.W., *et al.* (2009) Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res*, **37**, D5-15.

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

Figure captions

Figure 1. Pairwise organism similarity is higher among organisms living in the same environmental conditions. The mean pairwise similarity for same (SC) and different (DC) condition organisms according to their **(A)** temperature preference, **(B)** oxygen requirement, and **(C)** habitat. Note that for all points without error bars the standard errors are vanishingly small.


mean pairwise organism similarity
18%
16%
14%
12%
10%
8%
temperature preference
SC–mesophile
SC–thermophile
SC–psychrophile
DC–mesophile–thermophile
DC–mesophile–psychrophile
DC–psychrophile–thermophile
A
oxygen preference
SC–facultative
SC–aerobe
SC–anaerobe
DC–facultative–aerobe
DC–facultative–anaerobe
DC–anaerobe–aerobe
B
mean pairwise organism similarity
20%
18%
16%
14%
12%
10%
habitat preference
SC–soil
SC–host
SC–marine
SC–fresh water
DC–soil–host
DC–soil–marine
DC–soil–fresh water
DC–host–marine
DC–host–fresh water
DC–fresh water–marine
C