論文の内容の要旨

Reconsidering representation of complex microbial community structures: Habitat-based analysis based on comparative metagenomics
（比較メタゲノミクスを用いた微生物の生息環境解析
― 微生物群集データの表現方法を再考する）

美世　一守

Introduction

In the era of high-throughput sequencing, huge numbers of prokaryotic community structure datasets are being produced by 16S rRNA gene amplicon and shotgun metagenomic sequencing methods. These abundant datasets from diverse environments have contributed to unveiling the diversities and distributions of environmental microbial communities on earth. However, the methods for interpreting prokaryotic community structure datasets from ecological perspectives are still insufficient, despite the substantial increase in such datasets in microbial ecology and related fields.

As a promising solution to this problem, the trait-based approach aims to enable the ecological interpretation of community structure datasets. The basic idea of this approach is to first classify prokaryotes according to their ecological or physiological traits and then project that trait information to community structure datasets. To date, genomic features, such as genome size and rRNA gene copy number have been adopted for trait-based analyses of prokaryotic community structure datasets. While these approaches performed well, their trait data were limited and biased to cultured prokaryotes with available genomic and physiological data; however, in fact, prokaryotic communities usually contain many uncultured members.

One fundamental prokaryotic trait that has not yet been adopted is habitat information. Species that inhabit seawater are more likely to have the trait of adaptability to saline environments than species that inhabit freshwater only. Likewise, species that inhabit animal gut are more likely to have the trait of adaptability to copiotrophic (eutrophic) environments than species that inhabit soil only. Thus, habitat information is expected to provide mechanistic insights into ecological and physiological characteristics of prokaryotic communities. More importantly, habitat information can be obtained without the reliance on cultivation and isolation experiments because accumulating shotgun metagenomic datasets themselves provide rich information about environmental distribution of both cultured and non-cultured species; such comprehensive prokaryotic habitat data may be referred to as an "environmental atlas".

In this dissertation, I show the effectiveness of habitat-based analysis as a new-trait-based approach in interpreting and investigating prokaryotic community structure datasets. I developed a database, named ProkAtlas that links 16S rRNA gene sequences to prokaryotic habitats. I also developed the ProkAtlas pipeline for the habitat-based analysis of community structure datasets. As proofs of concept, I analyzed datasets from the Earth Microbiome Project (EMP), agricultural soil samples with salinity gradients, lake water samples with different salinity concentrations, human infant gut microbiome samples, developing soil samples along retreating glacier, and potentially polluted river-water samples.

Construction of prokaryotic habitat database: ProkAtlas

ProkAtlas was developed as a comprehensive database of prokaryotic habitat traits based on a meta-analysis of metagenome shotgun sequencing datasets. It comprises 360,474 16S rRNA gene sequences from 5,368 shotgun metagenome projects registered in the INSDC SRA/ERA/DRA databases. Notably, to achieve reliable but efficient prokaryotic habitat estimation, I tried to balance the database comprehensiveness and smallness. It is also notable that the number of 16S rRNA gene sequences in ProkAtlas is comparable to those in Greengenes and SILVA. Each sequence in ProkAtlas is labeled with one of the environmental categories such as *soil*, *marine*, *freshwater*, and *human_gut*, for prokaryotic habitat estimation with 16S rRNA gene sequences.

Furthermore, the ProkAtlas pipeline was developed to estimate the habitats of prokaryotes based on 16S rRNA gene sequences. Basically, the ProkAtlas pipeline uses a sequence similarity search to query every 16S rRNA gene sequence of given data against the ProkAtlas database and obtains a list of environmental categories that are labeled to the hit sequences. Compositions of the retrieved environmental categories are represented by habitat preference scores, which incorporates the number of hits from each environmental category and mitigates the potential biases of ProkAtlas on intensively-investigated environments such as soil and seawater. The ProkAtlas database and pipeline are available at https://msk33.github.io/prokatlas.html.

Evaluating the performance of ProkAtlas

ProkAtlas was applied to 1,021 (nearly) whole-length 16S rRNA gene sequences of pure-isolated bacterial strains from the International Journal of Systematic and Evolutionary Microbiology (IJSEM) phenotypic database. All sequences had significant hits in ProkAtlas, and their estimated habitats were overall consistent with their isolation sources. Bacterial strains isolated from seawater showed the largest MHIs dominated by marine or related environments. Strains isolated from soil, bioreactor, and human feces also showed the largest MHIs consistent with each of the sources. In addition, ProkAtlas was applied to 201 16S rRNA gene sequences retrieved from a large single-cell amplified genome (SAG) sequencing project. Among them, 183 (91.0%) sequences were successfully mapped to ProkAtlas. Again, their estimated habitats were generally in line with their sampling sites. The above results prove that ProkAtlas is an effective tool to estimate habitats of prokaryotes using 16S rRNA gene sequences, regardless of whether the prokaryotes are cultivable or not. ProkAtlas does not rely on data from isolation-based experiments and is free from cultivation biases, which differentiates ProkAtlas from existing prokaryotic trait databases.

Habitat-based analysis of prokaryotic communities

As proofs of concept of the habitat-based analysis of prokaryotic community structures, I applied ProkAtlas to six amplicon-sequencing datasets targeting 16S rRNA genes.

For example, I analyzed 124 agricultural soil samples with salinity gradients.

The dataset contained 12,094 sub-OTUs, most of which were uncharacterized members. When phylum-level taxonomic structures were investigated as many amplicon-sequencing studies do, the estimated compositions were largely similar among all samples, being stably dominated by the phyla *Proteobacteria*, *Bacteroidetes*, and *Actinobacteria*, giving few ecological insights. On the other hand, when the estimated habitats were investigated, we observed a clear trend that the habitat preference scores were affected by soil salinity. While *soil* and *rhizosphere* were major environmental categories, as expected due to the nature of the samples, brine-related categories such as *marine, seawater*, and *salt_lake* showed substantial variation among the samples (2.44–26.4%) and a significant positive correlation with the soil EC (Spearman's correlation test, ρ=0.60, P<0.001). Finally, it should be noted that the habitat-based analysis here gives a more direct and clearer insights into the environment-microbial community relationship than typical taxonomic analyses. Similar trends were observed in lake water prokaryotic communities with different salinities.

I also analyzed the primary succession of infant gut microbiomes, using 654 time-series infant feces samples collected from Finnish infants aged up to 36 months. When phylum-level taxonomic compositions were investigated, the compositions were highly diverse until approximately 400 days after birth, after which the compositions stabilized and were dominated by *Firmicutes* and *Bacteroidetes*. While this process was already well known, the habitat-based analysis gave another view on the process as the convergence to *human gut-related* environmental categories. Similarly, habitat preference scores in the chronosequences of developing soils, sampled along retreating glacier, presented clear traces of primary succession during the maturation of soils: habitat preference scores of soil-related categories were higher in matured soil than in pre-matured ones. These example shows that ProkAtlas can be used to evaluate the "maturity" of prokaryotic ecosystems undergoing temporal successions toward a stable state, without prior knowledge on what "matured" state would be like.