MICROBIOME STUDIES: ANALYTICAL TOOLS AND TECHNIQUES[*]

JAMES A. FOSTER

Department of Biological Sciences, University of Idaho, Moscow, ID 83844-3051
USA Initiative for Bioinformatics and Evolutionary STudies (IBEST)
BEACON Center for the Study of Evolution in Action
Email: foster@uidaho.edu

JASON H. MOORE

Institute for Quantitative Biomedical Sciences, Departments of Genetics and Community and Family Medicine, Dartmouth Medical School Lebanon, NH 03756 USA
Email: jason.h.moore@dartmouth.edu

JACK A. GILBERT

Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA
Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA
Email: gilbertjack@anl.gov

JOHN BUNGE

Department of Statistical Science, Cornell University, Ithaca, NY 14853 USA
Email: jab18@cornell.edu

Bacteria and Archaea are major factors in human health and environmental well-being, in ways that we are only now discovering. With new sequencing technologies it is now possible to take deep samples of the microbial species present in a given environment, resulting in very large sequence datasets that require ecologically aware analysis. However, the bioinformatics and statistics this task requires are not fully developed, constituting an active and important research area. This session, the second annual Microbiome Studies session at PSB, presents recent work on the computational and statistical analysis of microbiome data.

[*] This work was funded in part by: NIH COBRE grant P20RR16448, NIH INBRE grant P20RR016454, and NSF STC “BEACON Center for the Study of Evolution in Action” NSF STC DBI-0939454 to JAF; National Science Foundation grant DEB-08-16638 to JB; U.S. Dept. of Energy Contract DE-AC02-06CH11357 to JAG; and NIH grants LM009012, LM010098 and AI59694 to JHM. We also thank the organizers of the Pacific Symposium on Biocomputing.

Introduction to Microbiome Studies

This year's special session on Microbiome Studies at the annual Pacific Symposium on Biocomputing emphasizes practical solutions to the very large data interpretation problems arising from next generation sequencing and microbiome studies.

This is the second year for this special session. Last year's session was well attended, with especially enthusiastic tutorial and discussion sessions. In response to participant feedback, we narrowed the scope of the session to computational and statistical advances in techniques for processing next generation microbial survey data. We also expanded the organizing committee to include more expertise on environmental microbiomes and on statistical approaches. We received several high quality submissions again this year, and anticipate another informative and useful session.

As is only now becoming appreciated, Bacteria and Archaea dominate the Earth's ecology. They are a major factor in global processes such as the carbon cycle and climate. On a more personal level, the human microbiome, including non-pathogenic commensals, affects our health and behavior in surprising ways. Unfortunately, most current tools for analyzing microbial ecosystems were developed first for macro-biology. And most microbial studies have targeted isolates and potential pathogens. This has significantly biased the choice of model systems such as *E. coli*, microbial genome sequencing projects, diversity statistics, and bioinformatics. It is no surprise that many common analytical techniques appear to be ill suited to studying the microbial world.

In the meantime, sequencing technologies continue to improve, with new protocols and methodologies as well as incremental improvements in existing ones. It is no longer wild speculation to talk about sequencing "everything" in a given environment, or to raise the possibility of analyses that reject most of the data. There are now ambitious projects in place to characterize the entire human microbiome and all of the microbial life on earth.

The number of relevant publications continues to increase, along with potential funding. Community resources, both databases and software, are becoming more sophisticated and reliable.

Papers in this session

The papers in this session represent a prospective or leading-indicators view of the direction of the field: away from pure description and toward understanding the functional structure of communities, both internally (i.e., within a given dataset), and externally, with respect to metadata or covariates (the latter being measured directly or derived from database searches).

The paper "MetaDomain: A protein domain classification tool for short sequences" by Zhang and Sun addresses protein homology searching, which supports functional profiling in metagenomic annotation. Since the sensitivity of this process declines with decreasing read length, the authors present a tool specifically designed for short reads (as produced by next-generation sequencing technologies), which compensates for some of the shortcomings and achieves significantly improved sensitivity, even relative to the state-of-the-art profile HMM alignment tool.

The paper "Artificial functional difference between microbial communities caused by length difference of sequencing reads" by Zhang, Doak and Ye also analyzes homology-based approaches to annotation of microbial community data. The authors note that apparent functional differences may be the result of differences in read lengths rather than of differences in, for example, sequencing techniques. They show that certain functional categories are under-annotated and that the accuracy of annotation of short reads varies by function. They present an improved method to address these issues.

The paper "SEPP: SATé-Enabled Phylogenetic Placement" by Mirarab, Nguyen, and Warnow deals with the insertion of short molecular "query" sequences into an existing phylogenetic tree (phylogenetic placement), and subsequent downstream analysis. Using real and simulated data they show that existing methods work on well-behaved inputs (accurate alignments and trees for full-length sequences along with relatively small and not overly diverse sets of sequences), but for more general conditions accuracy declines. They therefore present a "boosting" technique which places metagenomic sequences more accurately when there is a large evolutionary diameter, and also shortens computing time for simpler cases.

The paper "Comparisons of distance methods for combining covariates and abundances in microbiome studies" by Fukuyama, McMurdie, Dethlefsen, Relman and Holmes compares different nonparametric methods for combining abundance and tree data with other metadata or covariates (e.g., clinical information). They look at the use of principal coordinates analysis on UNIFRAC output and consider the power of signal extraction in various settings, based on both real and simulated data. Since this form of analysis is becoming widespread, this is a timely study.

The paper "Estimating population diversity with unreliable low frequency counts" by Bunge, Böhning, Allen and Foster discusses the "singleton problem": when are low-frequency "species" counts derived from high-throughput sequencing real rather than artifactual? The authors present various *ex post facto* statistical adjustments for potential errors in the singletons or low frequency sample counts, while noting that correction at the data source is generally preferable to downstream statistical adjustment.