MICROBIOME STUDIES: PSB 2011 SPECIAL SESSION INTRODUCTION*

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 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

Recent advances in sequencing technologies have made is possible, for the first time, to take a thorough census of the microbial species present in a given environment. This presents a particularly exciting opportunity since bacteria and archea comprise the dominant forms of life on earth, and since they are vital to human health and to the wellbeing of our environment. However, the bioinformatics for interpreting these very large sequence datasets are not fully developed. This session presents recent work supporting the computational analysis of microbiome data.

1. Introduction to Microbiome Studies

1.1. *Producing Hard Copy Using MS-Word*

Microbes, including both eubacteria and archaea, are the dominant forms of life on earth, in absolute numbers, biomass, and diversity of ecosystems. During more than half of the Earth's 3.5 billion year biological history, only microbes were present. Microbial physiology is a dominant factor in carbon cycling, greenhouse gas emission, and oxygen production. In the human body there are ten times more microbial cells than human cells, and there are two orders of magnitude more microbial gene products than human gene products. Unfortunately, over 97% of microbes cannot be cultivated with current techniques, which has significantly biased the choice of model systems, as well as microbial genome sequencing and bioinformatics. Even our evolutionary and ecological theories and software were developed with macro-biology in mind, and often appear to be ill suited to studying the microbial world. Consequently, we have until recently been unable to fully understand and appreciate some of the most important ecological systems on earth.

Fortunately, new sequencing and bioinformatics technologies, such as tagged barcoding, community genomics, pyrosequencing, and metagenomics have made it possible to study the structure and dynamics of microbial communities. Some natural microbiomes that have recently been characterized include surveys of human microbiomes and their relationship to human health

* This work is partially supported by NIH P20RR16448 from the NCRR IDEA/COBRE program and by NSF "Evolution in Action" STC DBI 0939454

† Work partially supported by NIH R01 LM009012 and P20 RR018787.

in the human gut, skin, mouth, and reproductive tracts and ecological surveys of soil, air, and water to understand the effects of climate change and pollution.

A crude estimate of the number of publications in microbiome studies (using a Pubmed search for “microbiome OR metagenome OR community genomics OR microbial ecology”) has exploded in recent years, growing from 248 in 2000 to 1102 in 2009, with a total of 7117 hits to date. Publications to date are on track to double or triple in the coming year.

During this time, funding for microbiome studies has significantly increased, including signature areas such as the Human Microbiome Project from NIH. Community resources have become widely used (such as RDP, Greengenes/Silva, CAMERA, VAMPS, HMP DACC) and comprehensive bioinformatics tool suites are (such as mg-rast, mothur, catchall, and unifrac) being developed and widely used. Moreover, “the personal microbiome” may prove to be as important as “personal genomics”, the theme of PSB 2010.

These were the considerations that led to this PSB special session on “microbiome studies”. Our intention is to bring the expertise of the PSB community to bear on this increasingly important new field.

2. Papers in this session

One major challenge posed by microbial sequence data is to infer function and ecology of complex microbial communities from very large sequence datasets. This is the challenge addressed in this year’s session.

Most of the papers in this session present tools or frameworks to facilitate interpreting community function or composition from 16S fragments extracted from the environment and analyzed directly. The 16S gene codes for the small subunit of the ribosome, which is essential to DNA replication. Therefore this molecule is strongly conserved even in very ancient lineages. Woes introduced the use of the 16S gene as a phylogenetic marker for microbes, thereby showing that the archea comprise a distinct third kingdom of life. It has since become standard practice to use the similarity of several hypervariable regions in the 16S genes of microbes to identify and distinguish populations of microbes.

Holmes et al., “Visualization and Statistical Comparisons of Microbial Communities using R packages on Phylochip Data” introduces packages for the very popular R statistical analysis package which interpret data from the PhyloChip. This is a microarray with 16S targets for 8743 distinct bacteria and archea. Thus this paper represents a technology for identifying microbial communities using microarray technologies. The utility of this tool is limited by the set of targets on the Phylochip, of course. However, the R packages will be useful for any data from similar microarrays, including potentially customized arrays for specific complex screenings.

So called “metagenomic” or “barcode pyrosequencing” techniques have been developed to avoid the bias inherent in microarray design, and the even stronger bias that comes from culturing sequences and building clone libraries. (Note, “metagenomics” has multiple meanings, and we use it here to refer to 16S fragment analysis.). One approach, popularized by Roche, is to attach DNA “barcodes” to primers and then to perform very large scale PCR in micro-droplets that contain nano-beads with complementary sequences attached. This makes it possible to generate over million reads between 200bp and 500bp long, which is ideal for the hypervariable regions of 16S. Other technologies exist and are emerging that can generate far more reads, and these technologies are progressing rapidly.

This much data creates a major bioinformatics challenge. Several software packages or web services are emerging to provide bioinformatics support for this type of data. Two are presented in this session. Eran et al., "A FRAMEWORK FOR ANALYSIS OF METAGENOMIC SEQUENCING DATA" presents a software framework that allows scientists to build custom workflows for their "next generation" data analysis—with a particular emphasis on 16S microbial community sequence data.

The paper by Moore et al., "Human microbiome visualization using 3D technology", addresses the problem of making sense of microbiome data analysis visually. Scientists often need to "play with" possible interpretations of very large datasets, searching for more precise hypotheses to be verified or just getting a handle on what the data are like. This paper presents a possible framework for this challenge. This differs from the many existing "pipelines" in that it enables the user to directly customize their software to support individual workflows.

The paper by Bunge, "Estimating the Number of Species With CatchAll," presents the newly expanded CatchAll package for microbial community analysis. Most statistical techniques for inferring species richness and other ecological measure of diversity are nonparametric and based on relatively small samples from relatively small populations, having been derived to interpret mega-biome data such as that from forests or reefs. But techniques that work for hundreds of species often fail to work robustly for thousands, especially when there are typically dozens to hundreds of "rare" species in the sample. CatchAll provides the classical measures of diversity, but also adds some novel parametric estimates that seem to work very well for microbiome analyses. (Note: be sure to see Dr. Bunge's tutorial on ecological diversity estimations as well.)

In summary, we are pleased with the orientation of this new PSB special session toward practical solutions to the very large data interpretation problems arising from next generation sequencing and microbiome studies.