The Environmental Niche Atlas: Global Mapping of Microbial Functions (#3300)
Katherine S. Pollard (lead) and Jonathan A. Eisen (co-lead)

PROGRESS REPORT – June 1, 2013

Research Progress

Aim 1. Build automated tools to characterize protein families from metagenomic sequencing data.

SFams Protein Family Database

- Using the design approach described in our January 2013 progress report, we developed software for auto-updating SFams to include new genomes. The code is freely available here: https://github.com/gjospin/Sfam_updater.
- The first SFams auto-update was initiated in May 2013 and will incorporate >10 million proteins from >2,000 new genomes released by IMG since the first build of SFams.
- We implemented a quality control pipeline that quantifies the performance of each SFam for metagenome and metatranscriptome annotation. Most SFams have high precision and high recall. Poor quality SFams will be annotated in the database and used to improve the SFam updating algorithm in future iterations.
- We developed bioinformatics tools for functionally annotating SFams, including running InterProScan and mapping SFams to KEGG pathways. These annotations will facilitate metabolic pathway level analyses of metagenomes and metatranscriptomes.
- Additional new SFams features include a re-designed repository structure that is better for version control and modifications to the database organization to facilitate disambiguation and bug fixes.
- We identified 40 protein families that are phylogenetically informative markers for bacteria and archaea (highly universal, low copy number variation, congruent phylogenies). Using these genes, we developed a taxonomic classifier that produces OTUs that are comparable to those from classifiers based on SSU-rRNA sequence similarity.
- We identified SFams that are ecologically and phylogenetically informative markers for subgroups within the bacterial and archaeal lineages. A paper has been submitted on this analysis.
- We have been testing a new approach to identify clusters of protein families with similar distribution patterns using an "Extreme Sets" approach. Our goal is to predict functions of SFAMs based on similarity of distribution patterns SFams with known functions (akin to phylogenetic profiling). We have compared extreme sets with MCL clusters of different inflation parameters. We will be testing this on metagenomic distribution patterns next.

Metagenomic Read Classifier (MRC)

- We implemented algorithmic improvements to MRC that significantly improved run times for analyses of large metagenomic and metatranscriptomic datasets. We also experimented with different database engines to try to improve computational performance.
- We expanded our simulation analysis of MRC performance. We found that pair-wise search methods (e.g., BLAST) perform better for reads <200 bases long, whereas profile search methods (e.g., HMMER) perform better on longer reads (Figure 1). The two approaches also perform differently across protein families of different sizes and sequence diversity). We improved MRC performance by

combining the two methods in a two-step search procedure. We are collaborating with Sean Eddy (Janelia Farms) to further optimize MRC.

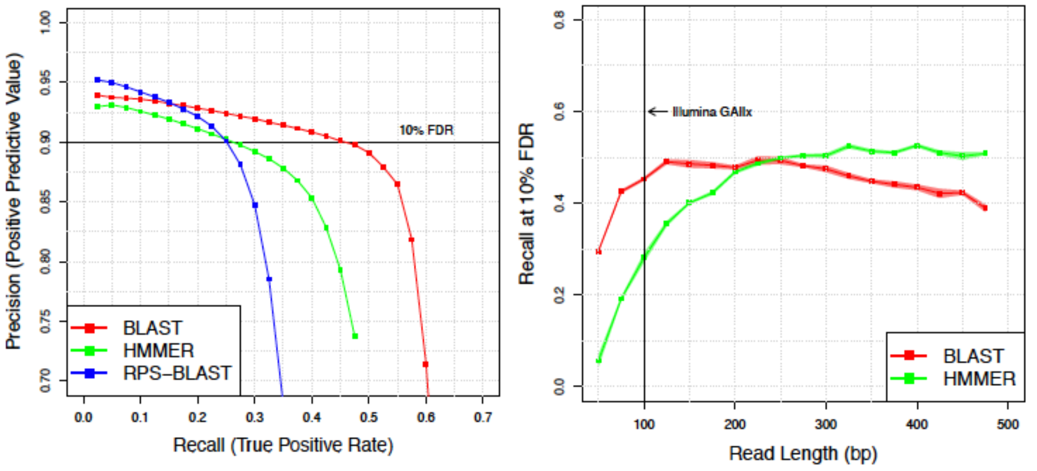

Figure 1: Profile vs. Pairwise Alignment Methods for Metagnomic Read Classification. (Left) Precision-Recall curves for an Illumina GAIIx metagenomic simulation from 250 genomes annotated with Pfam domains. At a conservative 10% false discovery rate (FDR) cutoff, BLAST is able to accurately classify 45% of reads overlapping Pfam domains, while the two profile-based methods (RPS-BLAST and HMMER) detect only 25%. (Right) Recall at a 10% false discovery rate across read length. At short read lengths (< 200 bp) BLAST is more sensitive, while for reads longer than 250 bp, HMMER becomes more sensitive.

Aim 2. Predict niches of protein families from environmental data.

- We made algorithmic improvements to our software for niche modeling to allow for exhaustive model selection using larger sets of environmental variables and to enable future extensions of the approach through a more modular Java code base. All niche modeling software is freely available by request. We are also exploring the possibility of integrating our niche modeling tools into QIIME (http://qiime.org).
- We continued to develop and validate a clustering approach to locate ecoregions (i.e., similar communities) based on diversity maps. This approach models differences between the taxonomic or functional composition communities as a function of environmental differences, uses this model to predict community similarity globally, applies clustering techniques to identify groups of similar communities, and maps these globally. We found that ecoregions are primarily stratified by latitude, but that the arctic and Antarctic harbor distinct ecoregions (Figure 2).
- We implemented methods for survey-gap analysis, which will enable us to map locations where marine microbial communities are likely to be most dissimilar to communities that have been sampled to date. These maps could inform future sampling efforts and may also identify specific environmental niches that are least studied to date and may harbor many novel, potentially endemic taxa and protein families.
- We are working on a novel approach to combine environmental data and a few deeply sampled locations (e.g., HOT/ALOHA, English Channel) to estimate the

total diversity of taxa and protein families at regional, ocean-wide, and potentially global scales.

Aim 3. Analyze niche maps to characterize global functional diversity.

- We continued to network and search for publicly available marine metagenomic datasets. We downloaded new data from the English Channel and a Pacific transect to the project server.
- We applied our methods to available data:
    - **SFams/MRC:** We processed English Channel metagenomes and metatranscriptomes (L4 sampling location, day/night, 3 seasons; Gilbert et al. PLoS ONE 2011). We found that metatranscriptomes at this location encode fewer known SFams and may encode a greater diversity of novel protein families than do the metagenomes from the same samples.
    - **Niche Mapping:** There is insufficient global metagenomics data to perform protein family niche mapping. We continued to test our methods using OTU data (from 16S sequences) and soil metagenomes from North America and Tibet.

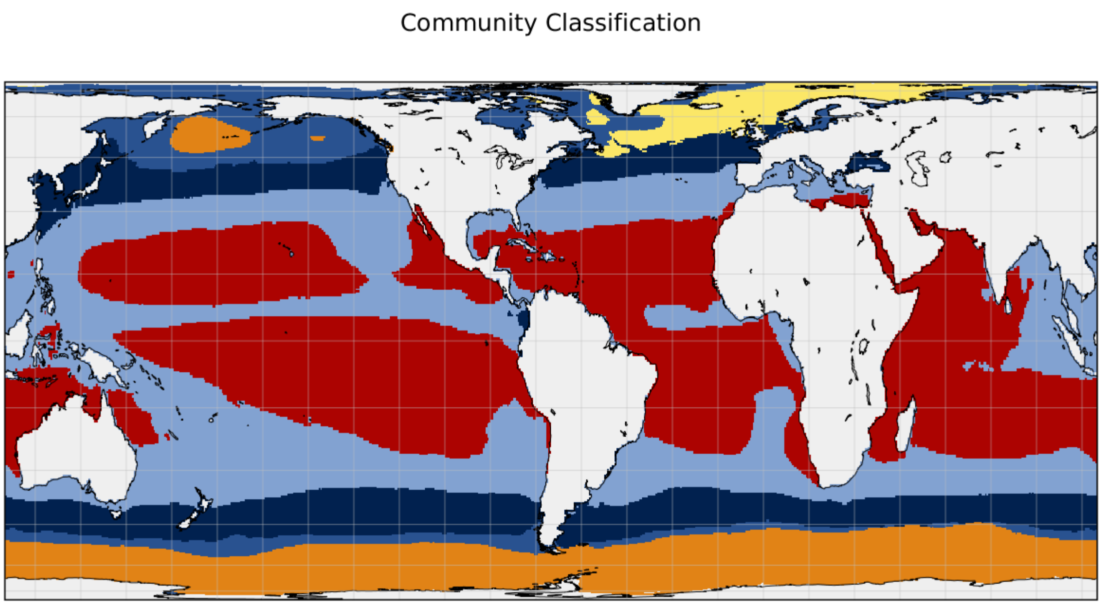

Figure 2: Ecoregion analysis of marine surface water bacterial diversity maps. Groups of similar communities were mapped using niche-modeling techniques coupled with cluster analysis to identify regions with similar diversity patterns. Each cluster is plotted in a different color. This analysis is based on community taxonomic composition (16S data). In future work, we plan to apply this approach to community functional composition based on protein families detected in marine metagenomes using SFams and MRC. This analysis will require a random sample of shotgun metagenomic studies.

Publications

- Ladau, J., T.J. Sharpton, M.M. Finucane, G. Jospin, S.W. Kembel, J. O'Dwyer, A.F. Koeppel, J.L. Green, K.S. Pollard. 2013. *Global marine bacterial diversity peaks at high latitudes in winter.* **The ISME Journal**, advance online publication.
- Cindy J. Castelle, Laura A. Hug, Kelly C. Wrighton, Brian C. Thomas, Kenneth H. Williams, Dongying Wu, Susannah G. Tringe, Steven Singer, Jonathan A. Eisen, Jillian F. Banfield. *Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment.* **Nature Communications**, In press.
- Christian Rinke, Patrick Schwientek , Alexander Sczyrba , Natalia Ivanova , Iain Anderson , Jan-Fang Cheng , Stephanie Malfatti , Aaron Darling , Brandon Swan , Esther Gies , Jeremy Dodsworth , Brian Hedlund , George Tsiamis , Stefan Sievert , Wen-Tso Liu , Jonathan Eisen , Steven Hallam , Nikos Kyrpides , Ramunas Stepanauskas , Edward Rubin , Philip Hugenholtz, Tanja Woyke. *Insights into the Phylogeny and Coding Potential of Microbial Dark Matter.* **Nature**, In press.

Talks

- *May 2013:* Josh Ladau - Predicting Microbial Distributions on a Global Scale using Niche Models. 113th General Meeting American Society for Microbiology. Denver, CO.
- *May 2013:* Jonathan Eisen - Phylogeny-Driven Approaches to Genomics and Metagenomics. Fresno State.
- *April 2013:* Katie Pollard – Quantifying the taxonomic and functional diversity of metagenomes from shotgun sequencing data. Society for Molecular Biology & Evolution Eukaryotic ‘Omics Meeting. Davis, CA.
- *April 2013.* Jonathan Eisen – The Need for a Phylogeny Driven Genomic Encyclopedia of Eukaryotes. Society for Molecular Biology & Evolution Eukaryotic ‘Omics Meeting. Davis, CA.
- *January 2013:* Katie Pollard – Decoding our genomes: what can our DNA tell us about our close relatives and who is traveling with us? Castro Valley Educational Foundation. Castro Valley, CA.