EcoFun-MAP: an Ecological Function Oriented Metagenomic Analysis Pipeline

Zhou Shi (jason.shi@ou.edu)[1], Zhili He[1], Erin Nuccio[3], **Jennifer Pett-Ridge[3], Mary Firestone[2]** and **Jizhong Zhou**[1]

[1]University of Oklahoma, Norman, OK; [2]University of California, Berkeley CA; [3]Lawrence Livermore National Laboratory, Livermore, California;

Project Goals: The EcoFun-MAP has been developed as an accurate, efficient and highly accessible strategy to fish out reads of important environmental functional genes from shotgun metagenome sequence data. EcoFun-MAP will facilitate functional analysis of large complex data by 1) constructing reliable and comprehensive reference databases specializing on the functional genes that are important to the ecological functions and geochemical processes and 2) providing efficient tools, graphical interface and automated workflow to easily access the databases for reducing repetitive effort by metagenome researchers.

Increasingly large amount of next generation sequencing (NGS) data has on one hand resulted in unprecedented insights into microbial ecology studies, and on the other hand, created a burden onto the computational analysis due to a lack of quick, accurate and dedicated tools. Here we present EcoFun-MAP, an Ecological Function oriented Metagenomic Analysis Pipeline, which was developed for automatic analyses of metagenome sequencing data from an ecological function perspective. EcoFun-MAP was built upon two functional gene databases, including a protein sequence based Hidden Markov Model database, and nucleotide sequence based functional OTU database, and both of them were manually curated and specifically tailored for fitting the applicable scope. EcoFun-MAP allows to profile a large amount of raw reads down to the functional OTU level, and annotating them into hierarchical ecological functional categories.

We refine the applicable scope of EcoFun-MAP to the functional genes encoding proteins/enzymes that play crucial roles in the major geochemical processes and ecological functions, including carbon (C), nitrogen (N), sulfur (S), and phosphorus (P) cycling, electron transferring, metal homeostasis, organic remediation, stress responses, secondary metabolism, and virus and virulence activity. Within those categories, a total of 1399 functional genes (e.g., *nifH*, *nirS*, *nirK*, *amoA*) were selected, and for each of them, a keyword-based query was manually crafted and submitted to the National Center for Biotechnology Information (NCBI) online databases for the retrieval of both protein and nucleotide based candidate reference sequences. The number of sequences retrieved for each gene could be ranged from a few to tens of thousands.

Two reference databases were constructed for the EcoFun-MAP to fully function: a Hidden Markov Models (HMMs) based database (HMMDB), and functional OTUs (FOTU) based database (FOTUDB). To construct HMMDB, we manually selected a minimum of five to a few hundred distinguished representative sequences as seed sequences (SS's) from protein based candidate sequences for each gene. Then, the selected seed sequences were aligned in ClustalW,

and the produced alignments were manually verified and later used as inputs for another program HMMBUILD to build function gene HMMs. The process has been done repeatedly for all functional genes and resulted in HMMDB finally. To construct FOTUDB, the candidate reference sequences of each gene were searched back against corresponding HMM from HMMDB using HMMSEARCH with applying an e-value cutoff to ensure that irrelevant sequences were excluded from further procedures. Due to the heterogeneity among the sequence sets of different genes, the most appropriate cutoff value for each gene could differ from others, therefore needs tremendous human interference by make repeated adjustments. After that, the output sequences were considered to be highly reliable reference functional gene sequences. Next, FOTUs were generated by clustering the confirmed sequences into a number of OTUs using CD-HIT99 with group similarity threshold of 95%, and corresponding BLAST databases were also constructed using MAKEBLASTDB. To this end, both of two reference databases have been established.

In the annotating workflow, HiSeq sequence results were resampled in each sample based on the minimal reads number in samples. The resampled sequences were input as raw unknown nucleotide sequences and were trimmed by Btrim with setting window size to 5 and average quality to 20, so as to remove unreliable sequences indicated by poor quality score. Next, all trimmed nucleotide sequences were translated into protein sequences using FragGeneScan with an error ratio 10%, which is widely accepted as estimated normal Illumina sequencing error ratio. Then HMMSEARCH was used for annotating the predicted protein sequences with the HMMDB database and an e-value cutoff can be customized by the users, and both global and local model hits were counted as valid results. In an additional filtering step, all confirmed sequences were compared together back against the FOTUDB with BLASTN. Only the best hits (Rank No. 1 in BLAST results) were kept as final fish-out results. All processing steps, statistical analysis methods, and bioinformatics tools are organized into a pipeline and will be integrated into a DOE KBase.

This research is based upon work supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research Genomic Science program under Award Numbers DE-SC0004730 and DE-SC0010570 at UC Berkeley and by UC-subcontract number 00008322 at the University of Oklahoma.