OT-03

STATISTICAL METHODS FOR FUNCTIONAL METAGENOMICANALYSIS BASED ON NEXT GENERATION SEQUENCING DATA

*Lingling An[1]and Naruekamol Pookhao[1]
[1]*Department of Agricultural and Biosystems Engineering, University of Arizona,*
Tucson, AZ, USA 85721-0038
Corresponding author: Lingling AN. Email: anling@email.arizona.edu

Keywords: metagenomes;

ABSTRACT

Metagenomics is a relatively new but fast growing field within environmental biology and medical science. It enables researchers to understand the diversity of microbes, their functions, cooperation, and evolution in a particular ecosystem. Traditional methods in genomics and microbiology cannot capture the structure of the broad microbial community within the environmental sample (e.g, soil, seawater, or human gut). Nowadays, high-throughput next generation sequencing technologies provide a powerful way in metagenomic studies. However, due to the massive short DNA sequences produced by the new sequencing technologies, there is an urgent need to develop efficient statistical methods to rapidly analyze the massive sequencing data generated from microbial communities and to accurately detect the features/functions present in a metagenomic sample/ community. Although several issues about functions of metagenomes at pathways or subsystems level have been investigated, it is lack of investigation on functional analysis of metagenoimics at a low level, i.e., more specific level.

This study is focusing on identifying all possible functional roles that are at the low level and present in a metagenomic sample/community. In this research we propose a statistical mixture model at the codon level of the genes to globally assign short reads to the candidate functional roles based on the SEED classification, with sequencing error considered. Comparing with other available algorithms and tools designated for metagenomic analysis through comprehensive simulation studies, our proposed approach is able to more specifically detect functional roles and more accurately estimate their relative abundance. The methods are also employed to analyze a real metagenomic data set.

METHODS

Method overview

The overview of the workflow is shown in Figure 1. DNA sequencing data of microbial communities is aligned against the NCBI-NR database, which is the reference database containing non-redundant protein sequences using BLASTX which is a Basic Local Alignment Search Tool (BLAST) that measures of local similarity to score the sequence alignments in such a way as to identify regions of good local alignments. On the output from BLASTX, mixture model proposed by Jiang et al. (2012) was employed to perform functional metagenomic analysis based on the SEED classification (Overbeek et al. 2005).

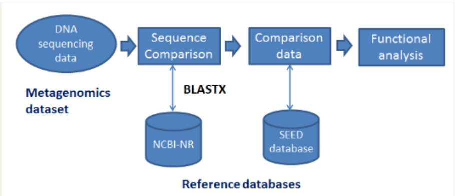


Figure 1. The overview of the workflow for functional metagenomic analysis

Mixture model

Mixture model identified multiple functional roles on the hits obtained by aligning sequence reads against the NCBI-NR database. The numbers of identical matches together with the query sequence length are used to evaluate the likelihood of alignment of a sequence with a given functional roles among the BLAST-generated candidate alignment list. A scoring matrix *M* below, where rows represent reads and columns represent functional roles, is an input for mixture model.

$$M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1K} \\ M_{21} & M_{22} & \cdots & M_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nK} \end{bmatrix} \qquad (1)$$

To identify which of the K candidate functional roles in the scoring matrix are truly present in the metagenomic sample, we propose a statistical framework to model the matches between the reads and reference sequences. Let R_i denote the proportion of reads contributed to functional role i (i = 1,2,…,K) in the sample, where $R_i \geq 0$ and ΣR_i =1. Let p denote the probability of observing a mismatched base pair, then 1-p is the probability of observing a matched base pair. The probability that a read x_j is contributed to functional role i with M_{ii} matched base pairs and L_j - M_{ii} mismatched base pairs is $R_i p^{Lj\text{-}Mji}(1\text{-}p)^{Mji}$, where L_j = max{L_{ji},i = 1,…K} is the maximum alignment length. Then the probability of observing a read x_j in the dataset is

$$\Pr(x_j) = \sum_{i=1}^{K} \left[R_i p^{L_j - M_{ji}} (1-p)^{M_{ji}} \right].$$

Assuming that the reads are independent of each other, then the likelihood function of the data is:

$$\ell(p, R_1, \cdots, R_K) = \prod_{j=1}^{n} \Pr(x_j) = \prod_{j=1}^{n} \left\{ \sum_{i=1}^{K} \left[R_i p^{L_j - M_{ji}} (1-p)^{M_{ji}} \right] \right\}, \qquad (2)$$

where the values of L_j and M_{ji} are observable, and the parameters p and R_i ($i = 1,\ldots K$) are to be estimated, with constraints $R_i \geq 0$ and $\Sigma R_i = 1$. For this mixture model, the expectation maximization (EM) algorithm (Dempster et al. 1977) is used to calculate the maximum likelihood estimation for the parameters.

RESULTS

Results for simulation study

A comparative performance study based on sophisticated simulation settings was conducted to compare the performance of mixture model with the performances of other existing methods including MEGAN4 (Huson et al. 2007) and MG-RAST (Meyer et al. 2008). Simulated data sets were generated from the sequence database downloaded from the SEED website (http://pseed.theseed.org/). We conducted 6 simulations and only the result from one simulation is shown here since all studies draw consistent conclusion. In this simulation 3,550 reads with average length of 100bp and 2% of errors were generated for 10 functional roles. The distributions of assigned reads to different functional roles by using mixture model, MEGAN, and MG-RAST are compared with the true values (Figure 2). It is evident that the number of reads assigned to different functional roles by the mixture model is closer to the true value.

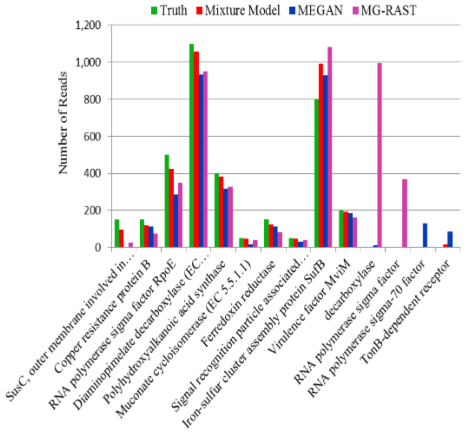


Figure 2. Numbers of reads assigned to different functional roles using mixture model, MEGAN, and MG-RAST server are compared with the true values (Truth) for simulation study.

Results for real data analysis

A real data set, DNA-Time1-Bag1, which comprises 209,073 reads with an average 250 bp read length from a marine sample (Gilbert et al. 2008) was used to compare the performance of mixture model, MEGAN, and MG-RAST server. The resulting functional analysis by using DNA-Time1-Bag1 is shown in Figure 3. In general, such a depiction shows the comparison of the number of reads assigned to the top 25 significant functional roles by using mixture model, MEGAN, and MG-RAST server. The numbers of reads assigned by the three methods are similar for some functional roles such as *Ribonucleotide reductase of class Ia (aerobic), alpha subunit (EC 1.17.4.1), Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)*, and *Aldehyde dehydrogenase (EC 1.2.1.3)*. However, the numbers of reads assigned by the three methods are very different for some functional roles such as *TonB-dependent receptor, Flagellar hook-length control protein FliK*, and *decarboxylase*. For these functional roles, the mixture model result may be closer to the true distribution of the functional roles in the marine sample.

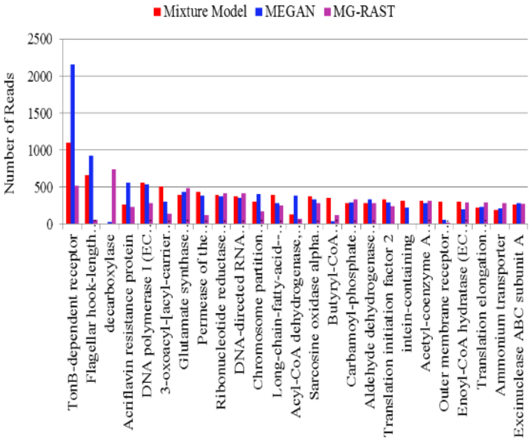


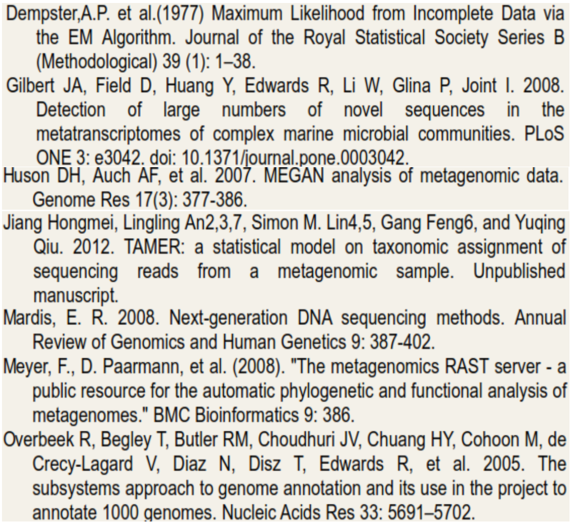
Figure 3. The resulting functional analysis by using DNA-Time1-Bag1. The comparison of the number of reads assigned to the top 25 significant functional roles by using the mixture model, MEGAN, and MG-RAST server.

DISCUSSION

Metagenomics is currently getting more interest as an efficient tool to investigate environmental samples. One of the main challenges about using metagenomics to investigate an environmental sample that scientists would like to answer is how to identify all possible functional roles present in an environmental sample. Due to the complexity of metagenomics and the huge volume of sequencing reads of short lengths obtained from the next generation sequencing technologies, the need of efficient statistical tools to accomplish this challenge is increasing.

We performed functional metagenomic analysis by using mixture model to assign short sequence reads to functional roles. From the results of the simulation study, compared to MEGAN and MG-RAST, mixture model is more accurate in assigning reads to relating functional roles. In the mixture model approach a sequence read is assigned to the functional role that has the largest posterior possibility while MEGAN and MG-RAST use the best match (or hit) score for assignment. However, a short sequence or gene could be involved in multiple functional roles simultaneously. How to define an appropriate criteria to allow multiple assignment for a single read is very challenging, and it will be our next research focus.

REFERENCES



Dempster,A.P. et al.(1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39 (1): 1–38.

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Glina P, Joint I. 2008. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS ONE 3: e3042. doi: 10.1371/journal.pone.0003042.

Huson DH, Auch AF, et al. 2007. MEGAN analysis of metagenomic data. Genome Res 17(3): 377-386.

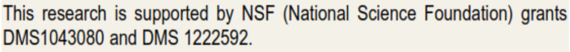
Jiang Hongmei, Lingling An2,3,7, Simon M. Lin4,5, Gang Feng6, and Yuqing Qiu. 2012. TAMER: a statistical model on taxonomic assignment of sequencing reads from a metagenomic sample. Unpublished manuscript.

Mardis, E. R. 2008. Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9: 387-402.

Meyer, F., D. Paarmann, et al. (2008). "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes." BMC Bioinformatics 9: 386.

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, et al. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33: 5691–5702.

ACKNOWLEDGEMENT



This research is supported by NSF (National Science Foundation) grants DMS1043080 and DMS 1222592.