In silico processing of Metagenomic data

Rohita Sinha
Dept of Food Science & Technology,

Metagenomics

A giant leap

Culture based analysis
Massively parallel DNA sequencing

Sequencing metagenomic samples

DNA extracted from environmental samples

Amplification of 16S rRNA genes

Complete sequencing of genetic materials

1. Data size in Gb/sample
2. Well established data processing protocols
3. Quick profiling of bacterial diversity
4. No information about the proven content of bacterial population

1. Data size in Gb/sample
2. Data processing protocols are still evolving
3. Relatively slower than 16S rRNA processing (quick profiles like MetaPhlAn also exist)
4. Detailed analysis proven content of the bacterial population is still possible

High resolution of information
Computational resources & skills are essential to analyze metagenomic data.

Data pre-processing
Nextera library prep + Illumina Sequencing

1. Shorter insert length may lead to sequencing of "sequencing primers", hence removal of those fragments is necessary.
2. Illumina sequencer tends to produce low quality base call towards 3' end of the reads. These low quality bases should be removed before the analysis.

Data pre-processing (Fastqc report)
Data analysis (metagenome assembly)

- Assembly terms:
  - Contigs: Long contiguous chain of nucleotides (ATGC) generated through an assembly operation are called contigs.
  - Mosaics: Two contigs are concatenated by multiple NC, when paired-end reads strongly suggest the proximity of contigs.
- Majority of metagenomes assemble via 'de Bruijn' graph to generate contigs.
- Output of a 'de Bruijn' graph-based assembly mostly depends on the single input parameter (kmer size).

Metagenome assembly (de Bruijn graph)

Taxonomic annotation of contigs

1. Simple "BLAST" based analysis of a good quality contigs may be classified easily if:
   - All the BLAST hits go to single reference sequence (multiple CQs having same genus)
2. Feature based methods like Hafaya, Mytax can be used to classify these contigs which could not be classified by "BLAST".
3. Pre-processed reads can be aligned to taxonomically annotated contigs to calculate the microbial abundance profiles.
Biological importance of the functional profiling of a microbiome

- Metabolite system is in perfect symbiosis with our animal systems.
- It is governed by the biochemical potential of the microbiome (protein content).
- Theories: correct functional profiling of the microbiome is essential to understand the delicate balance between microbes and the environment.

Functional annotation of metagenomic data

- Metagenome approach: leads to millions of DNA fragments
- Majority are originating from microbial genes
- In protein space: millions of short peptides

For functional annotation:

1. Prediction of ORFs in reads.
2. Blast: Match those ORFs to known proteins.
3. Apply sequence homology to assign function to ORFs.