A Thesis for the Degree of Ph.D. in Science

Development of \textit{de novo} assemblers for metagenomic sequencing data

January 2015

Graduate School of Science and Technology
Keio University

Afiahayati
## Thesis Abstract

### Thesis Title

Development of *de novo* assemblers for metagenomic sequencing data

### Thesis Summary

Next-generation sequencing (NGS) technologies allow the exploration of complex metagenomes in an effective manner, at lower cost and higher throughput. The throughput of metagenomic sequencing data generated by NGS technologies raise the bottlenecks in the computational analysis. Metagenomes has presented a number of additional assembly challenges, how to assemble multiple genomes from mixed sequence reads of multiple species. Assemblers that can generate longer and more accurate scaffolds help to extract more information for analysis tasks of metagenomes. A *de novo* assembler for single-genome is not capable of resolving metagenomic sequence data, therefore *de novo* assemblers designed specifically for metagenomes are indispensable.

In this dissertation, we present two *de novo* assemblers for metagenomic sequencing data generated by the NGS technologies. First, we propose a *de novo* metagenomic assembler specifically designed for the huge numbers of short reads (36-150 bp) generated by Illumina-type next-generation sequencers. This NGS technology enables deep sequencing of the inhomogeneous and divergent species in a microbial community. Second, we propose a *de novo* metagenomic assembler specifically designed for rather long sequence reads generated by 454 sequencing technology (200-500 bp). The 454 sequencing technology has been used for metagenomic analysis.

In Chapter 1, the importance of *de novo* metagenomic assemblers was introduced.

In Chapter 2, a *de novo* metagenomic assembler for short sequence reads was described. We extended a *de novo* assembler for single genome, Velvet, to a *de novo* metagenomic assembler utilizing supervised learning, named MetaVelvet-SL. MetaVelvet-SL improved a metagenomic assembler, MetaVelvet, in classifying chimeric nodes. MetaVelvet detects chimeric nodes in the assembly (de Bruijn) graph using simple heuristics which results in low accuracy and low sensitivity preventing the generation of longer contigs and scaffolds. MetaVelvet-SL significantly improved the original MetaVelvet in classifying chimeric nodes and generating accurate longer assemblies. MetaVelvet-SL also outperformed other state-of-the-art metagenomic assemblers, IDBA-UD, Ray Meta and Omega to reconstruct accurate longer assemblies.

In Chapter 3, a *de novo* metagenomic assembler for sequence reads generated by
454 sequencing technology was described. We extended a *de novo* metagenomic assembler, Genovo, by incorporating paired-end information, named Xgenovo, so that it generates accurate longer assemblies with paired-end reads. Genovo, is a de novo assembler for metagenomes under a generative probabilistic model. Paired-end sequencing is currently widely-used yet Genovo was designed for 454 single end reads. Unlike other assemblers which used paired-end information to generate scaffold, we attempted to increase the assembly performance without the aim of generating scaffold but attempted to map reads to the contigs in correct location. Xgenovo successfully generated longer N50 than the original Genovo and another metagenomic assembler for 454 sequencing technology, MAP. Xgenovo also demonstrated the potential to decrease the computational cost.

In Chapter 4, this study was summarized and future works were discussed.