DNA Sequencing with Longer Reads

Byung G. Kim

(Professor, Computer Science, Department University of Massachusetts Lowell, USA)

■ Time & Location

May 27th 2013 Mon. 4:00PM, Building 500 - ~~L309~~ **(L307호로 변경)**

■ Abstract

A next-generation sequencing (NGS) machine generates numerous reads, from which a complete genome is constructed. An intuitive approach in constructing a genome is to overlap as many reads with common subsequences (OLC: Overlap, layout, and consensus technique). OLC takes each read as a node in a graph and considers all possible overlaps with other reads, creating a connection graph. When millions of reads are involved, however, the OLC technique is known to take a long period of time. Recently, the de Bruijn graph approach breaks up every read into shorter k-mer fragments, and takes each fragment as a link between nodes with (k-1) common sequences. Given million of reads, the algorithm can be completed in a finite period of time.

Newer NGS machines, such as an Illumina HiSeq 2500, for example, generate longer reads (250 bp as opposed to 150 bp in an earlier machine). Breaking up reads into shorter fragments may be acceptable with short reads, but may pose problems when longer reads are generated. For example, one may lose association among fragments when 250-bp reads are broken up into 17-mer fragments. The problem is expected to get worse if even longer reads are generated by future NGS machines.

We consider reverting back to the OLC technique for DNA sequencing with longer reads. Reads are first grouped into a set of buckets. Each bucket contains reads with identical prefix k-mers and postfix k-mers. Namely, the data is partitioned into a number of buckets so that parallel execution of the program is possible. Reads in buckets with identical prefix and postfix k-mers are then combined to examine all possible overlaps between pre- and post-fix reads. As the process is repeated, longer contigs become available, an unlikely contigs are removed from part of the DNA reassembly. Two such algorithms are to be presented as well as their results.

문의: 민상렬 교수 (880-7047)