LoRDEC: a tool for correcting errors in long sequencing reads

Eric Rivals

LoRDEC: a tool for correcting errors in long sequencing reads

Montpellier Omics Days 2016

Eric Rivals, LIRMM & Institut de Biologie Computationnelle (IBC)

CNRS et Université de Montpellier, France

High-throughput DNA and RNA sequencing has become a routine experiment in molecular biology and life sciences in general. It is increasingly used in the hospital as a key procedure of personalized medicine.

Compared to the second generation, third generation sequencing technologies produce longer reads with comparatively lower throughput and higher error rate. Those errors include substitutions, indels, and they hinder or at least complicate downstream analysis like mapping or de novo assembly. However, these long read data are often used in conjunction with short reads of the 2nd generation.

I will present a hybrid strategy for correcting the long reads using the short reads that we introduced last year. Unlike existing error correction tool, ours, called LoRDEC, avoids the alignment of short reads on long reads, which is computationally very intensive. Instead, it takes advantage of a succinct graph to represent the short reads, and compares the long reads to paths in the graph. Experiments show that LoRDEC outperforms existing methods in running time and memory while achieving a comparable correction performance. In conclusion, i will comment on the impact of read error correction.

Work in collaboration with L. Salmela from University of Helsinki. LoRDEC is available at: http://www.lirmm.fr/~rivals/lordec/