Quality-Aware, Parallel, Multistage Detection and Correction of Sequencing Errors using Storm

Lakshmisha Bhat (lbhat1@jhu.edu)
Department of Computer Science, Johns Hopkins University

Abstract

The sequence data produced by next-generation sequencing technologies is error-prone and has motivated the development of a number of short-read error correctors in recent years. The majority of methods focus on the correction of substitution errors, which are the dominant error source in data produced by Illumina sequencing technology. Our efforts are also aligned towards the same goal. We design a streaming algorithm that takes a stream of sequence reads, builds a distributed abundance histogram assisted by a distributed sketch and use this information to detect which read nucleotides are likely to be sequencing errors, all within the Storm ecosystem. Then, using a maximum likelihood approach, we correct errors by incorporating quality values to achieve the highest accuracy on realistically simulated reads.

1 Introduction

We saw earlier in the course that Sanger reads, typically 700 – 1000 bp in length, are long enough for overlaps to be reliable indicators of genomic co-location, which are used in the overlap-layout-consensus approach for genome assembly. However, the de novo inference of a genome without the aid of a reference genome, is a hard problem to solve. The overlap-layout-consensus approach does poorly with the much shorter reads of second-generation sequencing platforms; for e.g. DNA sequence reads from Illumina sequencers, one of the most successful of the second-generation technologies, range from 35 to 125 bp in length. In this context, de Bruijn graph [5] based formulations that reconstruct the genome as a path in a graph perform better due to their more global focus and ability to naturally accommodate paired read information. As a result, it has become de facto model for building short-read genome assemblers.

In this context it is worth talking about the correctness of these second-generation sequencers. It is well known that the sequence fidelity of these sequencers are high - they have a substitution error rate of 0.5–2.5 (empirically shown in [4]). Errors are attributed to templates getting out of sync, by missing an incorporation or by incorporating 2 or more nucleotides at once. These errors increase in the later sequencing cycles as proportionally more templates fall out of sync and hence their frequency at the 3' ends of reads is higher. This, as we shall see, becomes an important property to be considered for maximum likelihood estimation during error correction.

Error correction has long been recognized as a critical and difficult part of the so called graph-based assemblers (de Bruijn graph). It also has significant impact on alignment and in other next-generation sequencing applications such as re-sequencing. Sequencing errors

complicate analysis, which normally requires that reads be aligned to each other during assemble or to a reference genome for single-nucleotide polymorphism (SNP) detection. Mistakes during the overlap computation in genome assembly may leave gaps in the assembly, while false overlaps may create ambiguous paths [4]. In genome re-sequencing projects, reads are aligned to a reference genome, usually allowing for a fixed number of mismatches due to either SNPs or sequencing errors. In most cases, the reference genome and the genome being newly sequenced will differ, sometimes substantially. Variable regions are more difficult to align because mismatches from both polymorphisms and sequencing errors occur, but if errors can be eliminated, more reads will align and the sensitivity for variant detection will improve.

Fortunately, the low cost of second-generation sequencing makes it possible to obtain highly redundant coverage of a genome, which can be used to correct sequencing errors in the reads before assembly or alignment. Various methods have been proposed to use this redundancy for error correction; for example, the EULER assembler [9] counts the number of appearances of each oligonucleotide of size k (hereafter referred to as k-mers) in the reads. For sufficiently large k, almost all single-base errors alter k-mers overlapping the error to versions that do not exist in the genome. Therefore, k-mers with low coverage, particularly those occurring just once or twice, usually represent sequencing errors. For the purpose of our discussion, we will refer to high coverage k-mers as trusted, because they are highly likely to occur in the genome, and low coverage k-mers as untrusted. Based on this principle, we can identify reads containing untrusted k-mers and either correct them so that all k-mers are trusted or simply discard them. The latest instance of EULER determines a coverage cutoff to separate low and high coverage k-mers using a mixture model of Poisson (low) and Gaussian (high) distributions, and corrects reads with low coverage k-mers by making nucleotide edits to the read that reduce the number of low coverage k-mers until all k-mers in the read have high coverage [10]. A number of related methods have been proposed to perform this error correction step, all guided by the goal of finding the minimum number of single base edits (edit distance) to the read that make all k-mers trusted [11-14].

2 Background

3 Methods and Software

4 Results

5 Conclusions

6 Advisors

- Ben Langmead (langmea@cs.jhu.edu)

References

[1] Andrew McGregor Ely Porat Alexander Andoni, Assaf Goldberger. Homomorphic fingerprints under misalignments: Sketching edit and shift distances. *STOC'13*, 2013.

[2] Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.

[3] Rayan Chikhi and Paul Medvedev. Informed and automated k-mer size selection for genome assembly. *BioInformatics*, pages 1–7, 2013.

[4] Steven L Salzberg David R Kelley, Michael C Schatz. Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, 11(11):13, 2010.

[5] Waterman MS. Idury RM. A new algorithm for dna sequence assembly. *PubMed Central*, 2(2):291–306, 1995.

[6] Liu W Mller-Wittig W. Shi H, Schmidt B. A parallel algorithm for error correction in high-throughput short-read data on cuda-enabled graphics hardware. *Journal of Computational Biology*, 17(4), 2010.

[7] Karin S. Dorman Xiao Yang and Srinivas Aluru. Reptile: representative tiling for short read error correction. *BioInformatics*, 26(20):2526–2533, 2010.

[8] Jan Schrder Yongchao Liu and Bertil Schmidt. Musket: a multistage k-mer spectrum based error corrector for illumina sequence data. *BioInformatics*, 2012.