Long Reads Alignment

Giulia Guidi[1,2], Aydın Buluç[2]
{gguidi, abuluc}@lbl.gov
[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
[2]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, USA

June 30, 2017

1 Problem Statement and General Overview

High-throughput sequencing technologies produce a large number of short and low-quality DNA sequences, called *reads*. These data are used as starting point to reconstruct the whole DNA sequence in a process called *de novo* genome assembly.

Recent advances in this field, such as the Pacific Biosciences Single Molecule Real-Time (SMRT) Sequencing technology, led to higher consensus accuracy, unbiased error distribution and longer reads. These improvements are crucial to assemble high-quality DNA sequences. On the other hand, the SMRT technology is prone to an higher error rate with respect to previous sequencing technology, about 15%. In general, high error rates lead to a significant computational effort in throwing away useless data. The presented work proposes a novel approach to select significant data from the input and recognize overlapping reads pairs to reconstruct the whole DNA sequence.

Overview of the idea, slide 6-7.

2 Reliable k-mers

The first computed analysis regards the identification of a reliable set of k-mers (RKS). Our algorithm relies on detecting overlaps between long reads efficiently. We treat the k-mer occurrences in each long read as the feature vector of that read. However, due to high error rates, the number of distinct k-mers in a dataset can be orders of magnitude larger the actual correct k-mers. Keeping all the k-mers in our feature set would not only increase the computational costs and memory requirements, it would also lower our precision.

Ideally, we want k-mers that occur only once in the genome. Multiple occurences of the same k-mer in the genome correspond to repeat regions. If we kept non-unique k-mers in our feature set, they would increase the number of spurious alignments and hence increase the computational costs. Our rationale for ignoring non-unique k-mers comes from the observation that either (a) the repeated region is small enough compared to the length of the read that the unique part of the read can still be used to find overlaps of this read with other reads, or (b) that the repeated region is almost as long as the read itself, in which case there is no benefit in aligning this read to other reads because it does not increase our information about the final genome.

We see by looking at the k-mer histogram that the majority of k-mers in the right tail either occur multiple times or do not occur in the genome at all (Figure 1). The probability of a k-mer being sequenced correctly is approximately $(1-e)^k$ where e is the error rate. If the sequencing depth is d, then observing this k-mer in the input data d times is a very slim $(1-e)^{dk}$.

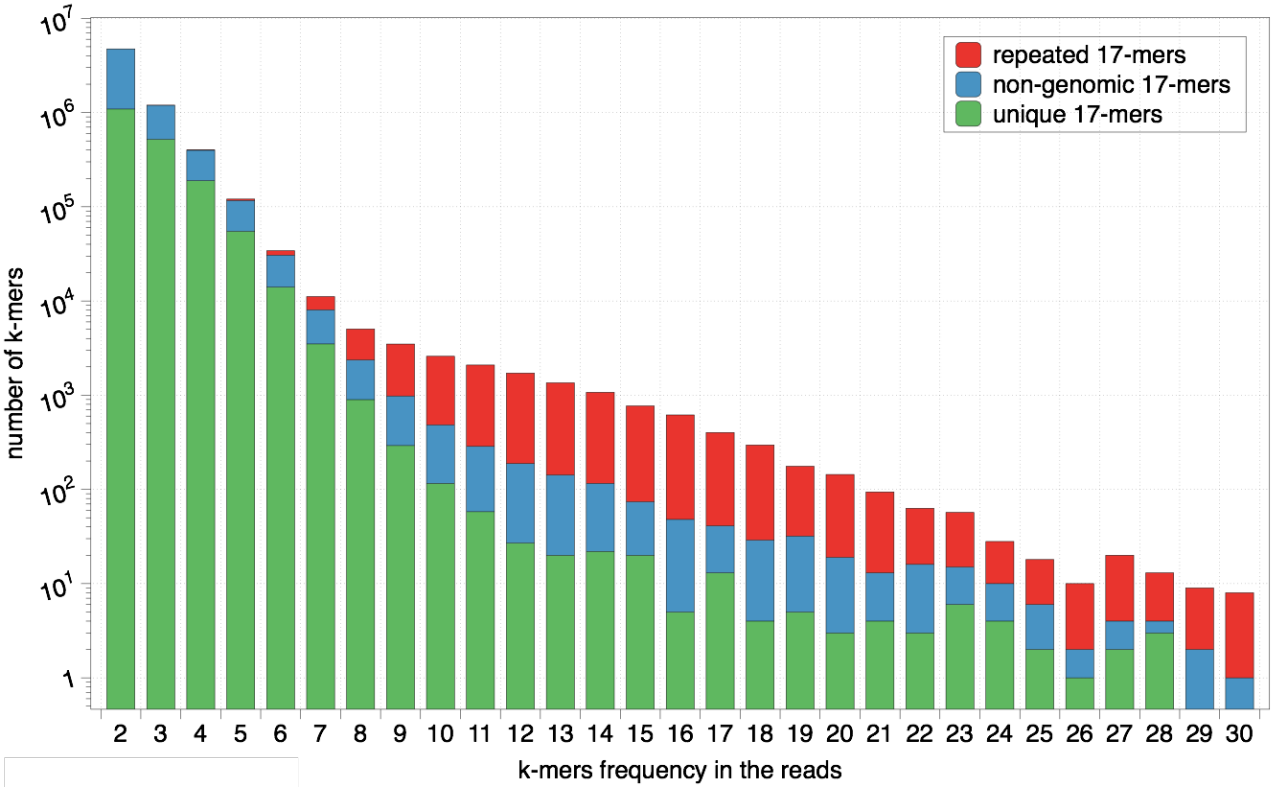

Figure 1: Percentage of unique, non-genomic and repeated k-mers over the total amount of k-mers belonging to a certain frequency.

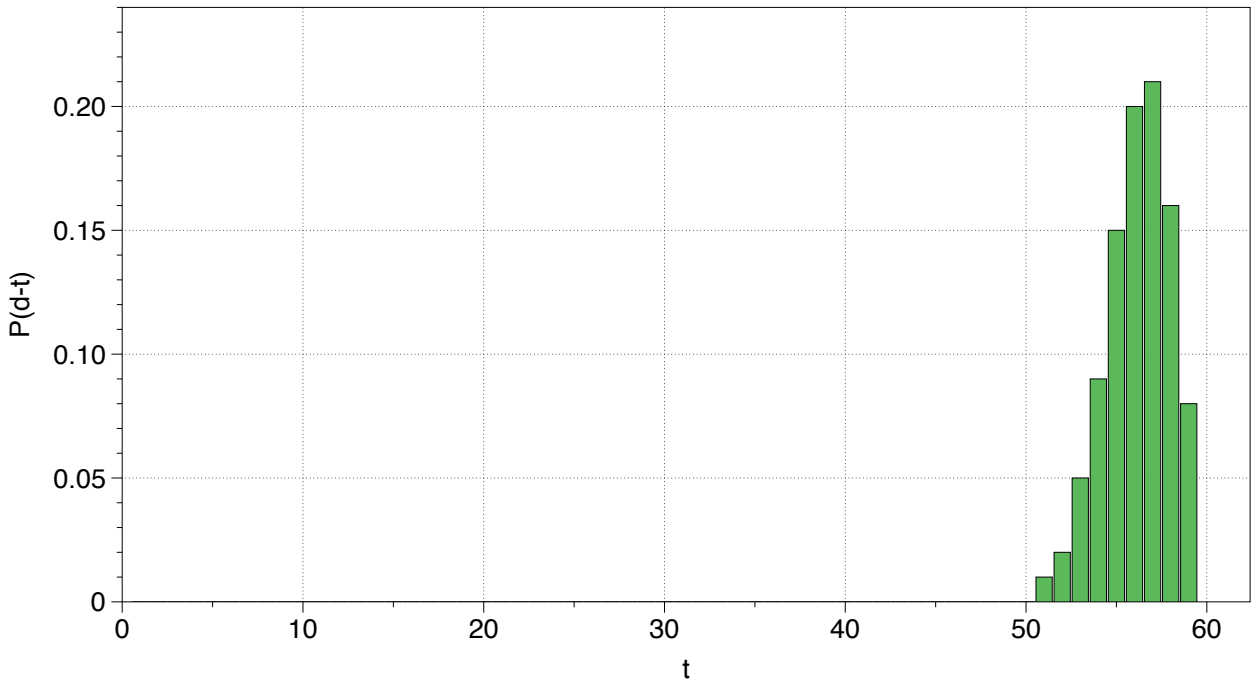

Figure 2: Probability that observing a k-mer that corresponds to a non-repetitive region $d-1$ times in the input.

Our analysis here is only correct if no other distinct section of the genome has been morphed into this k-mer by error. We acknowledge that such morphing occurs in practice but the majority of the high-frequency k-mers in the input set are due to correct sequencing if the value of k is chosen appropriately. Take the Human genome that is approximatyely 3 Gbp and for the sake of argument, only consider substitution errors. For $k = 17$, It encodes $4^{17} = 16$ billion different k-mers. Assuming independence, every possible k-mer exists in the genome with probability 3/16. The probability that a 17-mer we have seen being the result of off-by-1 error in sequencing is $(3/16) \cdot 17 \cdot e \cdot (1-e)^{17-1} \approx 3e(1-e)^{16}$, whereas the probability of sequencing a 17-mer correctly is $(1-e)^{17}$.

The probability that observing a k-mer that corresponds to a unique (non-repetitive) region $d-1$ times in the input would be approximately $P(d-1) = d(1-e)^{(d-1)k}(1-(1-e)^k)$. To generalize, the probability to observe $d-t$ times is:

$$P(d-t) = \binom{d}{t}(1-e)^{(d-t)k}(1-(1-e)^k)^t$$

Figure 2 shows the probability that a k-mer with input occurrence $d-t$ corresponds to a unique region in the genome given $d = 60$, $k = 17$ and $e = 0.15$.

3 Proposed algorithm

To be fixed, slide 11-16.

The ultimate goal of the previous analysis consists in obtaining a [k-mer -by- read] matrix (where A_{ij} is the occurrence/absence of k-mer i in read j). This matrix is used as starting point for the construction of a feature vector in finding alignments among the reads.
From the previous statistics we decided to take into account the k value equal to 15 and consider the

k-mers that occur in the range [4,8] as this range provides the highest percentages of unique existence in the *Escherichia Coli* genome.
Firsly, our approach consists in the creation of a dictionary containing all the k-mers belonging to the defined range. Then, we construct a [kmer -by- read] sparse matrix, where the $\{i, j\}$ cell is a *pair* data structure. The first value of the pair correspond to the identification number of the k-mer i contained in the read j, while the second value is a *vector* data structure where all the positions of that k-mer in the considered read are saved.
Once creating the matrix, we compute its transpose [read -by- k-mer] in order to multiply them and obtain a [read -by- read] matrix. We implement the calculation to obtain as final cell $\{i, j\}$ a *map* data structure organized as follow. The *keys* correspond to the identification numbers of the shared k-mers between the two reads, while the *values* are pairs of vectors containing the k-mers positions on the two reads.
To do: filter on the matrix to identify reads pairs sampled from the same region of the genome, implement an Apply() function to compute the delta-pos among k-mers belonging to the same read and compare k-mers delta-pos between reads sharing the same k-mers. Filter to identify *true overlap* reads pair:

- Compute the delta-pos among k-mers belonging to the same read.
- Comparing k-mers delta-pos between reads sharing the same k-mers.
- If at least one of the comparing is rejected by our filter, we discharge the considered pair.
- The filter consist in a parametric analysis that takes into account the probability of indention and deletion (from the PacBio simulator) to calculate the minimum and maximum length of the two given delta length.
- If there is not overlap between the L_{max} of the shortest read and the L_{min} of the longest one, the pair is rejected as considered a *fake overlap*.

3.1 Light version

4 Evaluation