45. Giles Miclotte, Pieter Audenaert & Jan Fostier

Iterative Seeding for Sequence to Graph Alignments

Giles.Miclotte@ugent.be

Several tools have been developed for aligning long third generation DNA sequencing reads to a genome assembly graph, with the aim of correcting errors in long reads or producing better assemblies. These methods rely on the seed-and-extend paradigm using only exact matches as seeds. Due to the high error rate in long reads, only short exact seeds can be identified. In large, repeat-rich genomes, this leads to an overabundance of uninformative seeds. We propose the use of longer, inexact seeds in combination with a reseeding scheme to find additional smaller, exact seeds in sparsely seeded regions. We found that this approach can be as fast as traditional exact seeding, while resulting in fewer uninformative seeds and covering a significantly larger portion of the reads. We applied these concepts to a proof-of-concept error correction tool, which resulted in high quality assemblies, with a significantly lower run time than current state-of-the-art software.

FACULTY OF ENGINEERING AND ARCHITECTURE

IDLab

Giles Miclotte, Pieter Audenaert, Jan Fostier

ITERATIVE SEEDING FOR SEQUENCE TO GRAPH ALIGNMENTS

Abstract

Several tools have been developed for aligning long third generation DNA sequencing reads to a genome assembly graph, with the aim of correcting errors in long reads or producing better assemblies. These methods rely on the seed-and-extend paradigm using only exact matches as seeds. Due to the high error rate in long reads, only short exact seeds can be identified. In large, repeat-rich genomes, this leads to an overabundance of uninformative seeds.

We propose the use of longer, inexact seeds in combination with a reseeding scheme to find additional smaller, exact seeds in sparsely seeded regions. We found that this approach can be as fast as traditional exact seeding, while resulting in fewer uninformative seeds and covering a significantly larger portion of the reads. We applied these concepts to a proof-of-concept error correction tool, which resulted in high quality assemblies, with a significantly lower run time than current state-of-the-art software.

Seeds between reads and nodes

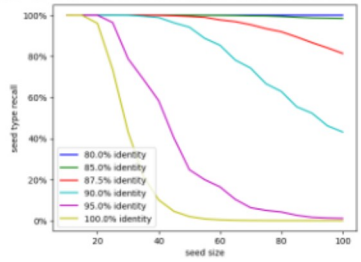

Figure 1: Seed type recall for inexact seeds for various identity levels. A seed is a tuple of start and end positions. The alignment identity of the seed is the percentage of nucleotide matches in the alignment of the implied subsequences. Typically, one is interested in seeds with high identity. This is reflected in the widespread use of exact *k*-mer matches in literature. However, longer inexact seeds contain more information than shorter exact seeds. Additionally, inexact seeds are more likely to occur than exact seeds of the same size, especially in the case of a high error rate in the reads.

Iterative seeding

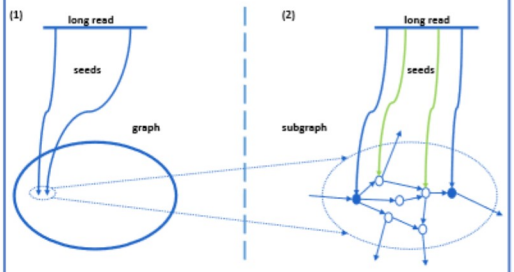

Figure 2: (1) Long inexact seeds between the long read and the graph are found. A flow graph between these seeds on the graph is identified. **(2)** Long seeds can not be found in short nodes. A new seed finding procedure is performed with a shorter seed size, and restricted to the subgraph. Performing this search on the entire graph would result in too many false positives, but by restricting the search to the subgraph, this is not an issue. These new seeds between the read and the nodes of the subgraph, guide the alignment between the initial long inexact seeds.

Pipeline overview

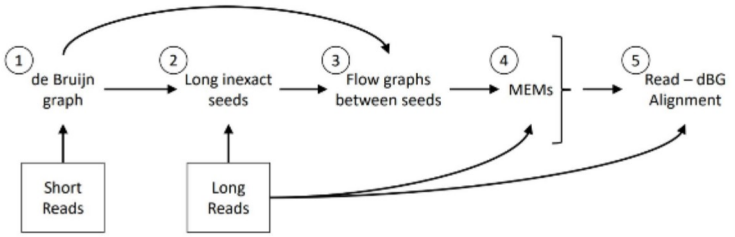

Figure 3: (1 - 4) pipeline from short and long reads to seeds. (1) Short reads are used to construct a de Bruijn graph; (2) long reads are aligned to the nodes of the de Bruijn graph producing long, inexact seeds; (3) between consecutive seeds a flow graph is constructed; (4) for each flow graph additional short, exact seeds (MEMs) are detected between the flow graph's nodes and the corresponding subsequence of the read; (3-4) this process can be repeated using flow graphs between the newly found seeds, (5) seeds are used to align the read to the de Bruijn graph

Results

Initial assembly results obtained with our iterative seeding approach on small genomes had higher contiguity than the state of the art assemblers (Canu, Unicycler). Additionally, our pipeline is an order of magnitude faster on these data. We are currently improving the implementation of these ideas to obtain similar results on larger genomes.

Contact

Giles.Miclotte@ugent.be
www.IDLab.ugent.be
www.IDLab.technology