Benchmarking pangenome graph mapping with strobemer-based seeding

Marouane Boumlik, Benjamin Linard, Matthias Zytnicki

Benchmarking pangenome graph mapping with strobemer-based seeding

Marouane BOUMLIK[1], Benjamin LINARD[1] and Matthias ZYTNICKI[1]

1 Unité de Mathématiques et Informatique Appliquées, INRAE,Castanet-Tolosan, France
Corresponding Author: benjamin.linard @inrae.fr

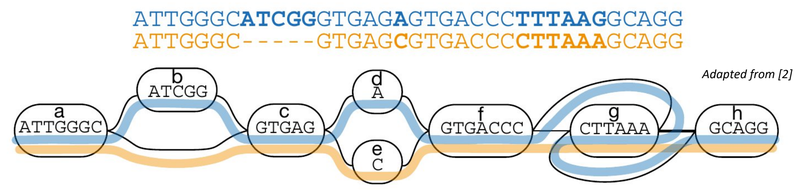
CONTEXT

A pangenome represents the total genetic diversity of a species or a species complex. A recent model, the variation graph, aims to integrate full length genomes [1]. Contrary to typical mapping approaches which rely on a single reference genome and are intrinsically biased, mapping query sequences to the full genetic variability of a species leads to more accurate mappings and improved genotype/phenotype analyses.

Mapping to a graph is similar to classic genome mapping and the first step involves a seed & extend approach, to target regions in the graph involving similar sequences before the steps of path selection and alignment. When divergent query sequences are analyzed, this step can be problematic : not enough k-mer seeds are detected, and the following steps do not happen.

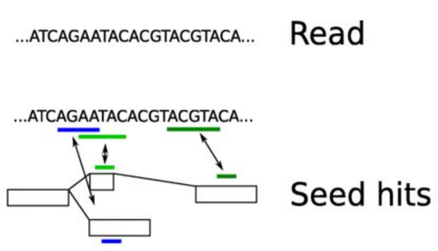
This work aims to test the potential of strobemers, an alternative to classic k-mer+minimizer seeding that allows more sequence divergence [4], in the context of mapping to pangenome graphs.

Input : high-quality assemblies of ful genomes, aligned.
Output : a graph modeling the full variation, e.g. a non-biased reference.

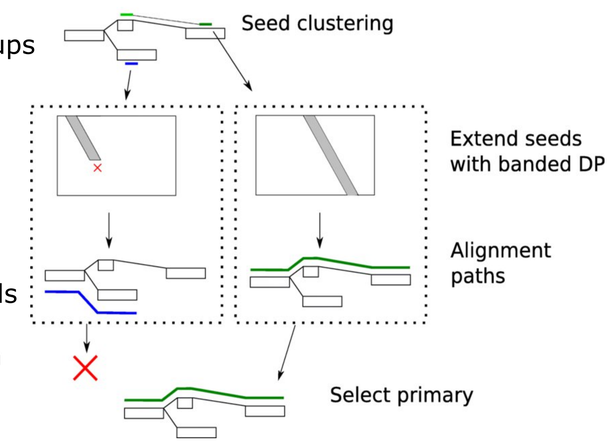


Seed & extend process in GraphAligner *[3]*

- Seeds are short sub-sequences that show a high degree of similarity and serve as anchors for following steps in the mapping process



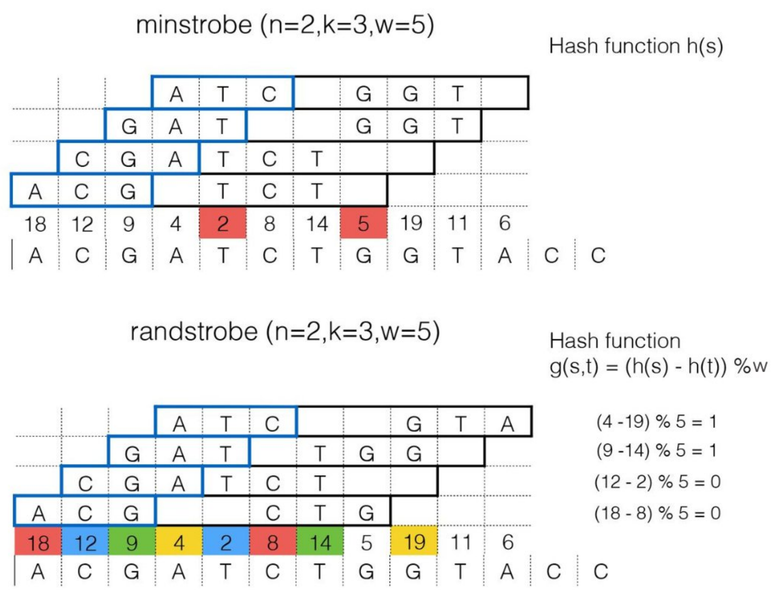
- Clustering groups closely related seeds
- Aligning is the process of extending seeds by alignement and choosing a path.



STROBEMERS

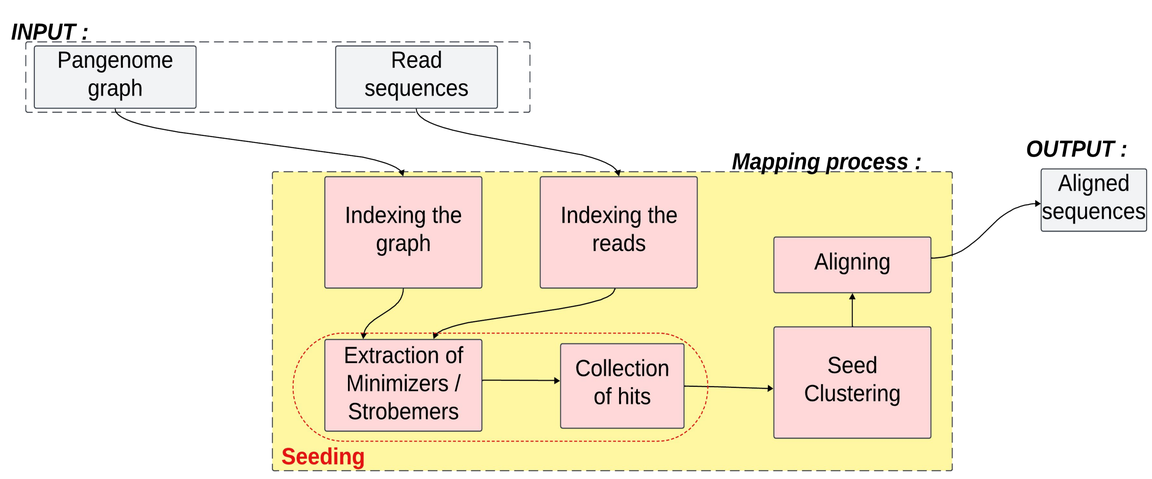
- Strobemers [4] is a combination of multiple discontinuous subsequences, known as "strobes".
- Strobemers have two sub-schemes: the minstrobes and the randstrobes [4].
- A C++ library prototype was available to exploit them in another software as an alternative seeding technique.

Minstrobe vs Randstrobe with parameters n, k and w (number of strobes, strobe length and window length)



METHOD

- Strobemers are injected as an alternative seeding scheme into GraphAligner.
- Other steps remain untouched.
- A basic strobemer filter "1 out of 5" is applied (for scalability).
- Long reads are simulated from the graph and mapped back to the graph.
- We measure mapping quality differences between classic k-mer + minimizers and strobemers approaches.



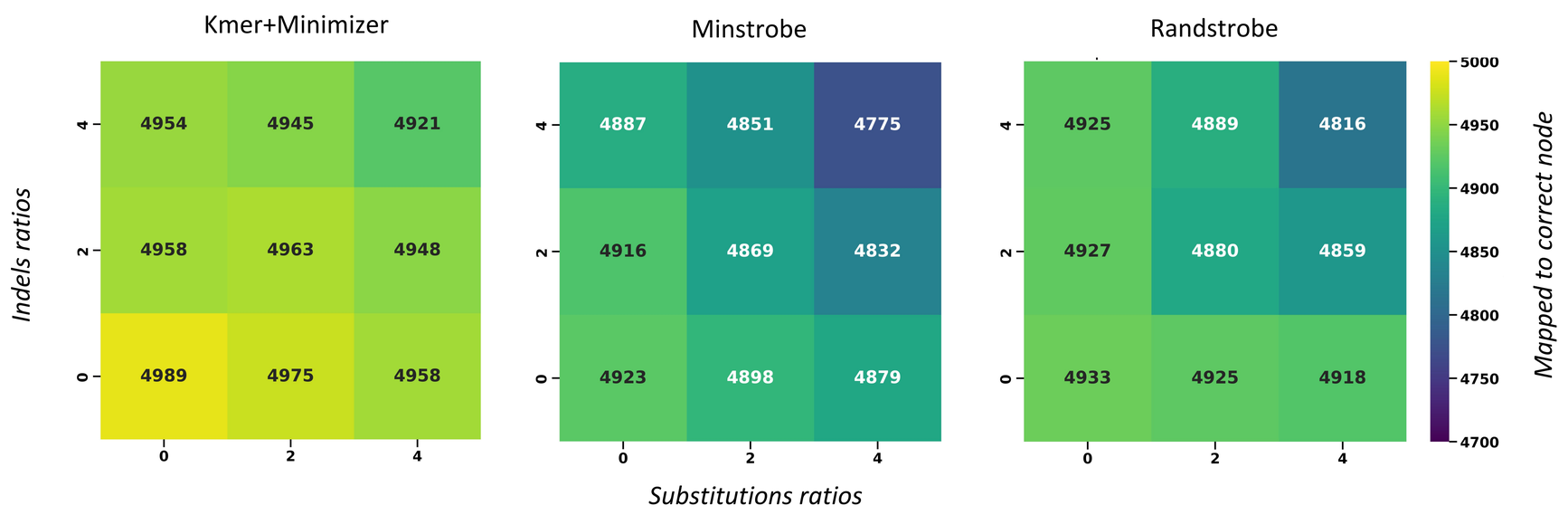
PRELIMINARY RESULTS

Software engineering :

- Improvement the C++ strobemer library: bug-fixes, cmake compilation, functionality ... (accepted in author's GIT repository).
- Development of benchmaring prototype "GraphAligner-strobe" : A hybrid software allowing to choose between k-mer + minimizers and strobemer seeding.
- Allows to set different parameters : rand|minstobe approaches, and strobemer parameters k, n, w_min, w_max.

Benchmark :

Simulated reads (vg sim)
Variable SNP/indel ratio
5000 reads of 5kbp



DISCUSSION

- Randstrobe performs better than Minstrobe, but in our preliminary tests, both perform worse than kmer+minimizer.
- We used a basic strobemer filter (filter 1 out of 5), what about a strobemer+minimizers ?
- Critical choice: we index k-mers or strobemers from node's sequences, which excludes nodes with length < k (GraphAligner's approach). What about path indexing ?
- Strobemer parameters chosen for this early benchmark may be unadapted, and strobemer seeding may be improved with different parameters.
- The benchmark should be extended to more parameter testing before further conclusions.

[1] Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017 May;27(5):665-676. [2] Liao, WW., Asri, M., Ebler, J. et al. A draft human pangenome reference. Nature 617, 312–324 (2023). [3] Rautiainen, M., & Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. Genome biology, 21(1), 253. [4] Sahlin, K. (2021). Effective sequence similarity detection with strobemers. Genome research, 31(11), 2080-2094.