Optimization of minimizer-based k-mer partitioning for genomic sequences

M2 internship - CRSItAL Lille UMR 9189

1 Overview

Let Σ be an alphabet. For any integer $k \geq 1$, a k-mer is a word of size k. The set of all k-mers is denoted by Σ^k. We recall that the lexicographical order over Σ^k is defined as follows. Let $x = a_1 \cdots a_k$ and $y = b_1 \cdots b_k$ be two k-mers; then $x > y$ if and only if either (i) $a_1 > b_1$ or (ii) there exists $1 \leq i \leq k-1$ so that $a_1 \cdots a_i = b_1 \cdots b_i$ and $a_{i+1} > b_{i+1}$. Provided $1 \leq m \leq k$, the *minimizer* of a k-mer x, denoted by $\min(x)$, is the smallest m-mer contained in x – i.e. the smallest substring of size m found in x.

Given a set S of k-mers, we use minimizers as a means of partitioning S: the k-mers sharing the same minimizer are grouped together in the partition; that is, we partition S into the disjoint union $S = S_1 \bigsqcup \cdots \bigsqcup S_p$ so that $\forall x, x' \in S_i$, $\min(x) = \min(x')$. Typically, we can see S as a set of k-mers drawn (with repetitions) from a certain (**non-uniform**) distribution over Σ^k. We are interested in obtaining a partition from S that is as balanced as possible. Unfortunately, even when S is constructed from the uniform distribution on Σ^k, it is folklore that the partitions obtained are highly skewed.

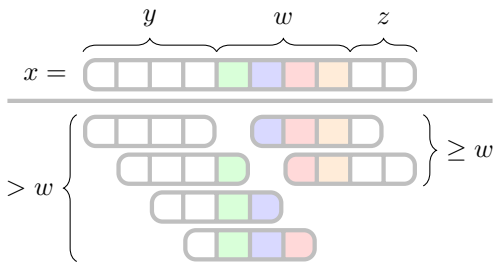

Figure 1: A k-mer x so that w is the minimizer of x.

In the Bonsai team, we've made good progress toward tackling this issue for an alphabet of size 4 corresponding to the DNA bases, that is $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. These bases pair with each other according to specific rules: `A` pairs with `T`, `C` pairs with `G`. This complementary pairing forms the well-known double-helix structure of DNA, where one strand runs in the opposite direction (is reverse) to the other. In this case, m and k are in the 10-50 bases order of magnitude.

k-mers extracted from DNA are commonly used in many downstream bioinformatics analysis. In practice, only a fraction of Σ^k exists in a biological dataset, but this fraction can already top up billions of distinct k-mers in a single dataset. Therefore, specialized datastructures and algorithms have been developed to efficiently handle these sets in RAM or on disk. A typical way of doing this is to partition the k-mers according to their minimizer and create separate indexes, one for each minimizer – hence the interest and the need to be able to construct balanced partitions as mentioned above.

Within the team, we grew interested in quantifying, for a given minimizer w, how many k-mers admit w as a minimizer – thus representing the size of the associated partition in the worst case, if all these k-mers were to be observed. To this particular question, we developed a solution, that works as an oracle of the worst-case size of the partition. **Another question is how can we use this oracle to develop heuristics and methods to obtain an optimal, or close to optimal, minimizer-based partition of the observed k-mers.**

2 Research directions

We aim at creating a set of buckets in which k-mers drawn from a biological dataset are distributed. Our goal is to obtain buckets as balanced as possible, in order to have positive impacts on space footprint and speed through parallelization. The challenge is that we don't know in advance which k-mers will be drawn, and there can be billions of k-mers in a single experiment. The simplest strategy is to assign one bucket for each minimizers, but this strategy is known to produce unbalanced buckets, even when minimizers are no more lexicographical but hashed instead.

Granted we know the expected number of k-mers a minimizer will group, we could redefine how we fill buckets. Partitioning k-mers according to their minimizers could be therefore linked to an optimization problem such as bin packing.

3 Work environment

The Bonsai team in Lille makes significant advancements in the field of DNA sequence data structures. Part of the research in the team focuses on creating space and time-efficient data structures that still support efficient operations such as search or insert. They are located on the Cité Scientifique campus, near the subway station.

The intern will work with several members of the team and will be supervised by Camille Marchet (chargée de recherche, CNRS) and Florian Ingels (postdoc, Université de Lille). Camille has a background in bioinformatics and specialises in data-structures for DNA and RNA, and Florian has a background at the intersection of optimization, combinatorics and graph theory.

The team has 19 members (9 permanent members, 10 PhD students/postdocs) and hosts one artist.

4 Outcomes

We developed one of the early method [3, 14, 8] that leveraged partitioned minimizers to fit the whole human genome in less than 8 GB of RAM [8]. It quickly spread to other highly efficient data-structures [9, 4, 11, 10, 12], genome-level comparison of sequences [13] and for the simultaneous reconstruction of hundreds of bacterial genome [1]. The improvement of the computation of minimizers is actively researched [7, 6, 5]. An improved partition method and its implementation would be very valuable for the computational biological field, and will be submitted to a conference.

An improvement in partitioned minimizers would have direct applications for us, as they are already used in the team:

- to build large scale indexes of biological samples, then deployed e.g. in CHUs to monitor acute myeloid patient samples (for instance at Toulouse's Oncopole)

- to investigate novel candidate genes and biological variants linked to blood, breast and lung cancers [2]
- as inner components of efficient data-structures

Finally, the team frequently offers and funds PhD thesis on algorithms and data-structures motivated by bioinformatics.

5 Candidate profile

We are seeking a candidate with a strong interest in discrete algorithms and their applications. The ideal candidate should be interested in working with efficient implementations in modern programming languages such as Rust, known for its performance and memory safety (prior experience with Rust or C++ is not required). The candidate could have a curiosity for how algorithms can be applied to biological data, especially in tasks related to DNA sequencing.

6 Contacts

Applications should be adressed to

camille.marchet@univ-lille.fr
florian.ingels@univ-lille.fr

References

[1] Gaëtan Benoit et al. "High-quality metagenome assembly from long accurate reads with metaMDBG". In: *Nature Biotechnology* (2024), pp. 1–6.

[2] Chloé Bessière et al. "Exploring a large cancer cell line RNA-sequencing dataset with k-mers". In: *bioRxiv* (2024), pp. 2024–02.

[3] Rayan Chikhi, Antoine Limasset, and Paul Medvedev. "Compacting de Bruijn graphs from sequencing data quickly and in low memory". In: *Bioinformatics* 32.12 (2016), pp. i201–i208.

[4] Jason Fan et al. "Fulgor: a fast and compact k-mer index for large-scale matching and color queries". In: *Algorithms for Molecular Biology* 19.1 (2024), p. 3.

[5] M. Hoang, G. Marçais, and C. Kingsford. "Density and Conservation Optimization of the Generalized Masked-Minimizer Sketching Scheme". In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 31.1 (2024), pp. 2–20. DOI: 10.1089/cmb.2023.0212.

[6] Ragnar Groot Koerkamp and Giulio Ermanno Pibiri. "The mod-minimizer: a simple and efficient sampling algorithm for long k-mers". In: *bioRxiv* (2024). DOI: 10.1101/2024.05.25.595898. eprint: https://www.biorxiv.org/content/early/2024/07/07/2024.05.25.595898.full.pdf. URL: https://www.biorxiv.org/content/early/2024/07/07/2024.05.25.595898.

[7] Patrick Kunzmann. "A fast and simple approach to k-mer decomposition". In: *bioRxiv* (2024), pp. 2024–07.

[8] Camille Marchet, Mael Kerbiriou, and Antoine Limasset. "BLight: efficient exact associative structure for k-mers". In: *Bioinformatics* 37.18 (2021), pp. 2858–2865.

[9] Camille Marchet and Antoine Limasset. “Scalable sequence database search using partitioned aggregated Bloom comb trees”. In: *Bioinformatics* 39.Supplement_1 (2023), pp. i252–i259.

[10] Giulio Ermanno Pibiri. “Sparse and skew hashing of k-mers”. In: *Bioinformatics* 38.Supplement_1 (2022), pp. i185–i194.

[11] Giulio Ermanno Pibiri, Jason Fan, and Rob Patro. “Meta-colored compacted de Bruijn graphs”. In: *International Conference on Research in Computational Molecular Biology*. Springer. 2024, pp. 131–146.

[12] Giulio Ermanno Pibiri, Yoshihiro Shibuya, and Antoine Limasset. “Locality-preserving minimal perfect hashing of k-mers”. In: *Bioinformatics* 39.Supplement_1 (2023), pp. i534–i543.

[13] Timothé Rouzé et al. “Fractional hitting sets for efficient and lightweight genomic data sketching”. In: *bioRxiv* (2023), pp. 2023–06.

[14] Derrick E Wood, Jennifer Lu, and Ben Langmead. “Improved metagenomic analysis with Kraken 2”. In: *Genome biology* 20 (2019), pp. 1–13.