Strategies to update a compacted de Bruijn graph

Khodor Hannoush, Camille Marchet, Pierre Peterlongo

▶ **To cite this version:**

Khodor Hannoush, Camille Marchet, Pierre Peterlongo. Strategies to update a compacted de Bruijn graph. SeqBIM 2024, Nov 2024, Rennes, France. hal-04861344

HAL Id: hal-04861344
https://inria.hal.science/hal-04861344v1

Submitted on 2 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Abstract

SeqBIM

Strategies to update a compacted de Bruijn graph

Khodor Hannoush[1]*, Camille Marchet[2], Pierre Peterlongo[1]

[1]*Univ. Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes, F-35000 France*
[2]*Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France*
***Corresponding author**: khodor.hannoush@inria.fr*

Abstract

With the rapid increase in genomic data due to advances in DNA sequencing, there's a growing need for efficient methods to store and handle the massive amount of sequence information. De Bruijn graphs have become a fondamental data structure in computational biology. These graphs compactly represent relationships between k-mers (subsequences of length k) in sequence data, minimizing redundancy.
Several construction methods for de Bruijn graph were proposed in the literature such as BCALM2 [1], TwoPaCo [2] and GGCAT [3]. However, these methods require full graph reconstruction when new sequences need to be added, which introduces significant computational overhead.
Dynamic solutions such as BufBOSS [4], DynamicBOSS [5] and Bifrost [6] have attempted to address this limitation by allowing updates to the graph. However, these approaches still involve redundant computations, limiting their scalability and efficiency, particularly as datasets grow. Efficient updates to existing graphs are necessary to handle the increasing volume of genomic data while avoiding the time and memory costs of complete graph reconstruction.
In this context, we introduce Cdbgtricks, a novel method designed to efficiently update an indexed compacted de Bruijn graph when adding new sequences, including entire genomes. Cdbgtricks uses a new indexing approach that helps identify the regions in the graph that need to be updated, thus accelerating the update process. When adding new sequences, Cdbgtricks dynamically updates both the graph and its index.
Experimental evaluations on a dataset composed of a hundred human genomes and a larger dataset composed of thousands of *E. coli* genomes show use-cases where Cdbgtricks offers better scalability than the state of the art tools. Cdbgtricks is an open source software available at `https://github.com/khodor14/Cdbgtricks`.

References

[1] Rayan Chikhi, Antoine Limasset, and Paul Medvedev. Compacting de bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, 32:i201–i208, 06 2016.

[2] Ilia Minkin, Son Pham, and Paul Medvedev. TwoPaCo: an efficient algorithm to build the compacted de Bruijn graph from many complete genomes. *Bioinformatics*, 33(24):4024–4032, 09 2016.

[3] Cracco A. and Tomescu A. Extremely-fast construction and querying of compacted and colored de Bruijn graphs with GGCAT. *bioRxiv*, 2022.

[4] Jarno Alanko, Bahar Alipanahi, Jonathen Settle, Christina Boucher, and Travis Gagie. Buffering updates enables efficient dynamic de bruijn graphs. *Computational and Structural Biotechnology Journal*, 19:4067–4078, 2021.

[5] Bahar Alipanahi, Alan Kuhnle, Simon J Puglisi, Leena Salmela, and Christina Boucher. Succinct dynamic de Bruijn graphs. *Bioinformatics*, 37(14):1946–1952, 07 2021.

[6] Guillaume Holley and Páll Melsted. Bifrost - Highly parallel construction and indexing of colored and compacted de Bruijn graphs. *bioRxiv*, 2019.