PROJEKAT STUDENTSKIH ISTRAŽIVANJA - PSI 2024

Using MinJoin algorithms for detecting large repeats in genomes (Korištenje MinJoin algoritama za pronalaženje velikih duplikacija u genomima)

Ibrahim Numanagić
University of Victoria
Victoria, BC, Canada

Large duplications—also known as segmental duplications (SDs)—are segments of DNA larger than 1kb that are highly similar to other regions within the genome.[1] SDs are among the most important sources of evolution, a common cause of genomic structural variation, and several are associated with various diseases of genomic origin. Despite their importance, SDs—especially those that occurred in the distant past—are hard to detect due to the size of the genome and the computational complexity of the problem. We have recently proposed two methods, SEDEF[2] and BISER,[3] that utilize minimizer-based MinHash sketching[4,5] to quickly identify the potential SD regions in a given genome. However, despite their success in uncovering many novel SDs at scale, both of these methods are probabilistic and rely on heuristics that are not guaranteed to characterize all SDs within a given genome. In this project, we are looking to explore the feasibility of theoretically sounder MinJoin family of sketching algorithms[6,7] for approximating string similarity for this task, and to compare their performance with those of MinHash-based algorithms.

Tasks:

- Read and understand the literature
- Implement and integrate MinJoin++ in Codon programming language
- Integrate MinJoin within the SEDEF/BISER pipelines
- Compare the results between the MinJoin and the MinHash-based implementations

References:

1. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
2. Numanagic, I. *et al.* Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* **34**, i706–i714 (2018).
3. Išerić, H., Alkan, C., Hach, F. & Numanagić, I. Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol. Biol.* **17**, 4 (2022).
4. Broder, A. Z. On the resemblance and containment of documents. in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)* (IEEE Comput. Soc, 2002). doi:10.1109/sequen.1997.666900.
5. Jain, C., Dilthey, A., Koren, S., Aluru, S. & Phillippy, A. M. A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases. *J. Comput. Biol.* **25**, 766–779 (2018).
6. Karpov, N., Zhang, H. & Zhang, Q. MinJoin++: a fast algorithm for string similarity joins under edit distance. *VLDB J.* **33**, 281–299 (2024).
7. Zhang, H. & Zhang, Q. MinJoin: Efficient Edit Similarity Joins via Local Hash Minima. (2018) doi:10.48550/ARXIV.1810.08833.