

Proposition de stage de recherche
*Comparing large genomic datasets with Invertible Bloom
Lookup Tables*

Supervisor: Gregory Kucherov (Gregory.Kucherov@univ-eiffel.fr)
Directeurs de recherche CNRS

Lab: Laboratoire d'Informatique Gaspard-Monge
Université Gustave Eiffel, Cité Descartes, Marne-la-Vallée (RER A Champs-sur-Marne)

Context: Invertible Bloom Lookup Table (abbreviated IBLT, also called Invertible Bloom filter) [2, 4] is a probabilistic data structure based on random hash functions. It can be seen as a “magic hash table” that supports insertions and deletions of keys, can store an *unlimited number of keys*, and is capable of listing the stored keys as long as their number is below a given threshold. The threshold determines the size of the data structure.

One of the applications of this data structure is so-called *set reconciliation* when *arbitrarily large but similar* sets (such as different versions of a file or a database, individual genomes of a species, etc.) are each represented by a small IBLT (*sketches*) so that the difference between the sets can be recovered by comparing their sketches [3, 1]. This property has many interesting potential applications that have not yet been much explored. In [5], we applied this idea to bioinformatics, namely to comparing genomic datasets, through their sets of short fixed-length strings, called *k*-mers.

Work description: The work will essentially consist of improving and extending the method proposed in [5], in light of many new ideas appeared since that paper was published. The work will include both a theoretical and an experimental component, in particular it is planned that the student will implement new improvements and will perform experiments on genomic data. In case of successful work, a journal publication is planned.

Prerequisites: The student is expected to have a good background in discrete mathematics and algorithms as well as necessary programming skills (Python, C++) for performing computer experiments. If successful, the internship can be continued to PhD studies.

References

- [1] Djamal Belazzougui, Gregory Kucherov, and Stefan Walzer. Better Space-Time-Robustness Trade-Offs for Set Reconciliation. In Karl Bringmann, Martin Grohe, Gabriele Puppis, and Ola Svensson, editors, *51st International Colloquium on Automata, Languages, and Programming*

- (ICALP 2024), volume 297 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 20:1–20:19, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.
- [2] David Eppstein and Michael T Goodrich. Straggler identification in round-trip data streams via Newton’s identities and invertible Bloom filters. *IEEE Transactions on Knowledge and Data Engineering*, 23(2):297–306, 2011.
 - [3] David Eppstein, Michael T. Goodrich, Frank Uyeda, and George Varghese. What’s the difference? Efficient set reconciliation without prior context. In *Proceedings of the ACM SIGCOMM 2011 Conference*, SIGCOMM ’11, page 218–229, New York, NY, USA, 2011. Association for Computing Machinery.
 - [4] Michael T Goodrich and Michael Mitzenmacher. Invertible Bloom lookup tables. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 792–799. IEEE, 2011.
 - [5] Yoshihiro Shibuya, Djamal Belazzougui, and Gregory Kucherov. Efficient Reconciliation of Genomic Datasets of High Similarity. In Christina Boucher and Sven Rahmann, editors, *22nd International Workshop on Algorithms in Bioinformatics (WABI 2022)*, volume 242 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 14:1–14:14, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.