

Internship subject: Compressed text indexing data structures for genomic sequences.

Eric Rivals

November 2011

Keywords : text algorithms, indexing data structures, genomic applications

Tutor : Eric RIVALS <http://www.lirmm.fr/rivals/>; rivals@lirmm.fr

Location : LIRMM (<http://www.lirmm.fr>), Montpellier, F; Tel. : 04 67 41 86 64

Description

The current revolution of sequencing technologies for DNA and RNA molecules enable the acquisition of millions, and soon billions, of short sequencing reads (or simply reads) in a few days from a biological sample. Novel and scalable algorithms are needed to search, compare, classify and index such amount of sequence data. Clearly, traditional indexing data structures like the well-known suffix arrays, suffix tree, Directed Acyclic Word Graph cannot scale up to such data size [Crochemore, Rytter, 02]. Compressed indexes, also called self indexes, have been designed to store large sequence in main memory, see for instance the Burrows-Wheeler Transform and its variants [Ferragina et al. 08]. However, further and more complex developments are required to match current needs of applications in genomic, medicine, and other domains.

We proposed recently an uncompressed data structure, named Gk-arrays, devised to index large read collections [Philippe et al 11]. Both the complexity analysis and experiments confirm that the construction and querying of the Gk-arrays are competitive with alternative solutions like the sparse hash tables implemented by Google (). This research topics of this internship aims at survey the literature and propose ideas of algorithms for one of the following goals :

1. compressing or sampling the Gk-arrays to limit their memory consumption
2. allowing easy updates of the Gk-arrays and avoiding reconstruction from scratch when the underlying read collection changes over time [Salson et al 10]
3. constructing them without building the intermediary generalised suffix array.

Literature

- M. Crochemore and W. Rytter, *Jewels of Stringology*, World Scientific, 2002, 310 pages.
- P. Ferragina, R. González, G. Navarro, R. Venturini : Compressed text indexes : From theory to practice. *ACM J Experimental Algorithmics* 13 : (2008)
- N. Philippe, M. Salson, T. Lecroq, M. Leonard, T. Commes and E. Rivals, Querying large read collections in main memory : a versatile data structure. *BMC Bioinformatics*, 12, p. 42, doi :10.1186/1471-2105-12-242, 2011.
- M. Salson, T. Lecroq, M. Léonard and L. Mouchard. Dynamic extended suffix arrays. *J Discrete Algorithms* p. 241-257, 8(2) 2010.