Abstract

SeqBio 2016

Rconnector: a resource-frugal probabilistic dictionary and applications in (meta)genomics and transcriptomics

Camille Marchet[1]*, Antoine Limasset[1], Pierre Peterlongo[2]

[1] *Genscale team, University of Rennes 1, Rennes, France*
[2] *Genscale team, INRIA, Rennes, France*
***Corresponding author**: camille.marchet@irisa.fr*

Abstract

Genome and transcriptome sequencing fields generate huge sets of sequences [1], that are often chopped in voluminous sets of k-mers for their further analysis, which brings its own share of high performance problems. To extract relevant pieces of information from the large data sets generated by current sequencing techniques, one must rely on extremely scalable methods and solutions. In this work we present a straightforward indexing structure that scales to billions of elements and we propose three direct applications based on k-mer diversity to explore (meta-)genomics and transcriptomics data sets.

Indexing has often been proven extremely expensive for large scale problems, while being a fundamental need in this field. For addressing the problem of scalability, we rely on an in-house library (BBhash) to construct Minimal Perfect Hash. This implementation distinguishes itself in its ability to construct a Minimal Perfect Hash Function for up to 100 billions elements in hours [2, 3], with limited memory fingerprint (< 4 bits per element). We combine this Minimal Perfect Hash Function with a quasi-dictionary to associate information to the elements indexed.

We present two applications for short reads data: Short Read Connector Counter and Short Read Connector Linker. We use the quasi-dictionary for indexing k-mers and keeping track of their related pieces of information, enabling to scale up large (meta-)genomic instances. Short Read Connector Counter links any read to its estimated abundance in a collection of samples. Short Read Connector Linker second connects any read from a given sample to similar reads in the data set [4].

Furthermore we present advances in the latest application, Long Read Connector RNA. Contrary to the previous applications, the input reads are long, erroneous reads from last generation of transcriptome sequencing (PacBio's IsoSeq, Nanopore [5, 6]). These sequences give a novel insight in transcriptomic analysis because a read often corresponds to a full length transcript. In this context, Long Read Connector RNA is designed to enable recruiting reads partly sharing their exon contents. In particular, alternative transcripts coming from a same gene will be connected.

In addition to the applications we propose, we believe that many tools could benefit from the fundamental data structure (Minimal Perfect Hash Function + quasi-dictionary) we propose.

MPHF : BBhash is available at `https://github.com/rizkg/BBHash`

Applications: `https://github.com/GATB/short_read_connector`

References

[1] Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nature*, 200(8):16–18, 2007.

[2] Paolo Ferragina and Giovanni Manzini. Indexing Compressed Text. *Journal of the ACM*, 52(4):552–581, 2000.

[3] Denis Charles and Kumar Chellapilla. Bloomier filters: A second look. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5193 LNCS, pages 259–270, 2008.

[4] Veronika B Dubinkina, Dmitry S Ischenko, Vladimir I Ulyantsev, Alexander V Tyakht, and Dmitry G Alexeev. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*, 17(1):38, dec 2016.

[5] Donald Sharon, Hagen Tilgner, Fabian Grubert, and Michael Snyder. A single-molecule long-read survey of the human transcriptome. *Nature biotechnology*, 31(11):1009–1014, 2013.

[6] Hagen Tilgner, Fabian Grubert, Donald Sharon, and Michael P Snyder. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proceedings of the National Academy of Sciences*, 111(27):9869–9874, 2014.