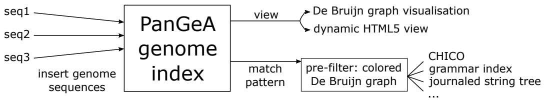
PanGeA: Pan-Genome Annotation
Indexing Annotated Human Genome Collections

Andre Brehme, Sven Brümmer, Jonas Charfreitag, Jonas Ellert, Jannik Junghänel, Dominik Köppl,
Christopher Osthues, Sven Rahmann, Dennis Rohde, Julian Sauer, Jonas Schmidt, Lars Schäpers,
Uriel Elias Wiebelitz, Jens Zentgraf
Chair for Algorithm Engineering, Technical University of Dortmund
pgpangea.cs@lists.tu-dortmund.de

Novel high-throughput sequencing methods make it possible to create huge sets of genomes. A genome of these sets is nowadays often found to have annotations of their gene sequences. By the study and analysis of the similarities and differences of an annotated area among the genomes of individuals, opportunities for personalized genome-based medicine arise.

The genome of a human individual consists of roughly 3 billion base pairs that can be stored plainly in 1 GB of space, which is possible because each base pair takes 2 bits. To provide a tool for genome-based medicine, the genomes of all patients have to be stored and maintained in a sophisticated index that supports answering queries efficiently, while compressing the input genome sequences. We highlight the latter point as crucial, for instance because just storing the population of a middle-sized city like Dortmund already costs 500,000 GB of space. Our idea to compress the input is to exploit the circumstance that two individuals of the same species share around 99% of the genetic information [FCS06].



This high inter-similarity of the genomes motivated us to devise a novel index data structure that explicitly takes advantage of the facts that (a) the input consists of a collection of DNA sequences that are highly similarly to each other, and that (b) the annotations can help to cluster together common gene sequences. To make this all possible, we have implemented different indexes and compression techniques so far, like a colored De Bruijn [BBB^{+}17] graph as a pre-filter, the JST [RWR14] and a tailored-version of CHICO in [Val16] and relative Lempel-Ziv [KPZ11].

As an outlook, we currently work on combining one of the aforementioned indexes with compression and on implementing other approaches like grammar-based self-indexes [CN12]. In addition, we plan to use the additional information of the annotations for annotation based queries. For example it should be possible to request in what area of the sequences of the indexed pangenome a specific variation of a gene is contained.

References

[BBB^{+}17] Keith Belk, Christina Boucher, Alexander Bowe, Travis Gagie, Paul Morley, Martin D Muggli, Noelle R Noyes, Simon J Puglisi, and Rober Raymond. Succinct Colored de Bruijn Graphs. *Bioinformatics*, to appear, 2017.

[CN12] Francisco Claude and Gonzalo Navarro. Improved Grammar-Based Compressed Indexes. In *Proc. SPIRE*, volume 7608 of *LNCS*, pages 180–192. Springer, 2012.

[FCS06] Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. Structural variation in the human genome. *Nature Reviews. Genetics*, 7(2):85–97, 2006.

[KPZ11] Shanika Kuruppu, Simon J. Puglisi, and Justin Zobel. Relative Lempel-Ziv Compression of Genomes for Large-scale Storage and Retrieval. In *Proc. SPIRE*, volume 7024 of *LNCS*, pages 201–206. Springer, 2011.

[RWR14] René Rahn, David Weese, and Knut Reinert. Journaled string tree – a scalable data structure for analyzing thousands of similar genomes on your laptop. *Bioinformatics*, 30(24):3499–3505, 2014.

[Val16] Daniel Valenzuela. CHICO: A Compressed Hybrid Index for Repetitive Collections. In *Proc. SEA*, volume 9685 of *LNCS*, pages 326–338. Springer, 2016.