Variant Tolerant Read Mapping with Locality Sensitive Hashing

Marcel Bargull, Kada Benadjemia, Benjamin Kramer, David Losch, Jens Quedenfeld, Sven Schrinner, Jan Stricker, Dominik Köppl, Dominik Kopczynski, Henning Timm, Johannes Fischer and Sven Rahmann
Department for Computer Science, TU Dortmund
pg583.cs@lists.tu-dortmund.de

The rapid development of genomic sequencing technologies in the past decades has outgrown the advances of computing power and therefore requires efficient read mapping algorithms [LBB12]. Read mappers align sequenced reads to a reference genome, where a set of reads aligned to the same position can hint at possible mutations. Even though many fast read mappers have been published in the recent years, most of them do not consider common variants of the reference genome. Variant tolerance highly increases accuracy of read mappers when aligning reads against a species' pangenome.

We have developed a new read mapper for variant tolerant alignment by usage of hash based filtering in combination with an alignment algorithm based on dynamic programming. In the first step, we use locality sensitive hashing (LSH), initially designed for finding similarities in documents, for candidate filtering [AI06]. We treat reads and windows of the reference genome as documents, which are compared by LSH.

As a result, we obtain an approximative mapping to the reference regions. This leads to a dramatical reduction of the reference length and therefore semi-global alignment becomes reasonable. The aligner handles variants like SNPs, insertions and deletions and decides which variants lead to the best alignment. New genetic variants and gene mutations can be found by observing the mismatches from the alignments.

References

[AI06] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 459–468. IEEE, 2006.

[LBB12] Po-Ru Loh, Michael Baym, and Bonnie Berger. Compressive genomics. *Nature biotechnology*, 30(7):627–630, 2012.