TAPYR: An efficient high-throughput sequence aligner for re-sequencing applications

Francisco Fernandes[1,2], Paulo G.S. da Fonseca[2], Luis M.S. Russo[1,2], Arlindo L. Oliveira[1,2], Ana T. Freitas[1,2]
[1]Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento, Lisboa, Portugal
[2]Instituto Superior Tecnico, Universidade Tecnica de Lisboa (IST/UTL), Lisboa, Portugal

During the last two decades most laboratories used Sanger's "shotgun" method in many significant large-scale sequencing projects, being this method considered the 'gold standard' in terms of both read length and sequencing accuracy. Recently, several next generation sequencing (NGS) technologies have emerged, including the GS FLX (454) Genome Analyzer, the Illumina's Solexa 1G Sequencer, the SOLiDTM and the Ion Torrent Systems, which are able to generate three to four orders of magnitude more sequences and are considerably less expensive than the Sanger method. However, the read lengths of NGS technologies create important algorithmic challenges. While the 454 platform (using Titanium technology) is able to obtain reads in the 400-600 base pairs (bp), the Illumina's Solexa 1G Sequencer and the Ion Torrent Systems present reads with an average length of 100 bp and the SOLiD platform is currently limited to 25-50 bp.

Several assembly tools have recently been developed for generating assemblies from short, unpaired sequencing reads. However, the sheer volume of data generated by these technologies (0.4 Gbp/run for the 454 and 16 Gbp/run for the SOLiD), and the need to align reads to increasing large reference genomes limits the applicability of standard methods.

One way to speed up the read alignment task is to resort to software based on approximate indexing technologies. This means that the whole reference genome is scanned while applying a dynamic programming algorithm. Indexed alignment algorithms, which preprocess the reference genome into an index data structure that can then be searched, correspond to more efficient approaches. On one hand it can discard irrelevant portions of the reference genome much more efficiently. On the other hand the computation on relevant regions can be factored out. However, building indexes is time and space consuming. State of the art algorithms are using techniques from a new class of indexes, compressed indexes, which have smaller space requirements by using data compression techniques to eliminate regularities in the indexes.

In this work we present TAPyR (http://www.tapyr.net) a new method for the alignment of NGS reads that uses compressed indexing build an index of the reference genome sequence to accelerate the alignment. Being firstly proposed to handle the 454 GS FLX data, it can also be used with Illumina and Ion Torrent data. Like other algorithms, TAPyR uses in a second stage a multiple seed heuristic to anchor the best candidate alignments. This heuristics has the advantage that it dispenses the need of determining the number and length of the seeds beforehand, relying on the assumption that the optimal alignments are mostly composed of relatively large chunks of exact matches interspersed by small, possibly gapped, divergent regions. At the ultimate stage banded dynamic programming is used to finish up the candidate multiple seed alignments considering user-specified error constraints.

TAPyR was evaluated against other mainstream mapping tools namely BWA-SW, SSAHA2, Segemehl, GASSST, and Newbler. The analyses were performed with real and simulated data sets, with the objective of assessing the efficiency and accuracy of the aforementioned tools in the context of re-sequencing projects. As the results show the new method manages to achieve convincing performance in terms of speed and in terms of the number and precision of aligned reads. In fact, TAPyR has displayed class-leading CPU-time performance and excellent use of input reads in comparison to other mainstream tools.