FAST: Fast Alignments for Short Tags

Del Fabbro C[1,2], Morgante M[1,3], Policriti P[1,2]

[1] Istituto di Genomica Applicata, Udine
[2] Department of Mathematics and Computer Science, Faculty of Science, University of Udine
[3] Department of Crop Science and Agricultural Engineering, Faculty of Agriculture, University of Udine

Motivation

The recently emerged new sequencing technologies are completely redesigning the algorithmic alignment problem in Bioinformatics. The principal character of these technologies (Illumina/Solexa) is that they can produce a huge amount of short (30-75bp) tags in a few days, whose alignment to a reference sequence is crucial when judged on many different parameters. In particular, the problem of performing a fast AND accurate alignment of every single tag, must now be tackled under a new perspective. While standard alignment programs (BLAST[1], BLAT[2]) are not suitable to align short tags, many new alignment softwares (Eland, MAQ[3], Soap[4]) have been proposed to do the job in an efficiently way. However, often, such tools lack precision: essentially, different solutions use heuristics having the goal of a fast alignment process and the resulting output is not optimal. The most common sub-optimal character of the resulting alignments, being the missing of acceptable tag's positions and misclassification of tag's occurrence profiles. Often these errors have a small size in absolute terms but a significant impact in analysis. We propose a software tool for re-sequencing applications (genomes comparison, structural variations and transcriptome profiles) that can handle short tags in a fast and completely accurate way.

Methods

Our alignment tool (called "FAST": Fast Alignments for Short Tags) is implemented to align millions of tags in a very accurate way. Using an entirely deterministic algorithmic approach and implementing only combinatorially justifiable heuristics, we are able to align (allowing mismatches) the full amount of short tags of a Illumina/Solexa run in 5-7 hours on a twin quad-cores system with 8G RAM, with no placement or classification errors. The tool is based on suffix arrays[5] and the core system uses an efficient dichotomy algorithm that allow fast alignment with few (1-2) mismatches. Using a BLAST-like strategy we allow up to 5-6 mismatches (for 75bp tags). The algorithms are precise and return the best alignment (if any) for each tag. FAST is able to align single or pair ends short tags (either for DNA-seq and mRNA-seq) and determine exons junction and alternative splicing sites (for mRNA-seq).

Results

The comparison of FAST with other alignment tools showed that using an exact tool we can obtain a significantly more precise analysis. FAST was successfully used for massive re-sequencing projects with Illumina/Solexa instruments as well as for transcriptome profile determination, exons detection, and alternative splicing sites analysis in a pre-mRNA-seq project of four Vitis vinifera tissues[6,7]. Always in the context of grape genome sequencing initiative, FAST was used for aligning single and pair-ends DNA tags for structural variation analysis of 14 varieties/clones.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. "Basic local alignment search tool". J Mol Biol, 215(3) :403–410, 1990.
[2] W. J. Kent. "BLAT - The BLAST-Like Alignment Tool". Genome Res., 12(4):656–664, 2002.
[3] H. Li, J. Ruan, R. Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores". Genome Res. (25 September 2008)
[4] R. Li, Y. Li, K. Kristiansen, and J. Wang. "SOAP: short oligonucleotide alignment program". Bioinformatics, page btn025, 2008.
[5] Udi Manber and Gene Myers. "Suffix arrays: A new method for on-line string searches". SIAM Journal on Computing, 22(5):935–948, 1993.
[6] E. Mica, V. Piccolo, M. Delledonne, A. Ferrarini, M. Pezzotti, C. Casati, C. Del Fabbro, G. Valle, A. Policriti, M. Morgante, G. Pesole, M.E. Pè, D. S. Horner. "High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera" (submitted to Genome Biology)
[7] C. Del Fabbro, A. Casagrande, D.S. Horner, A. Policriti, M. Morgante et al. "Vitis Vinifera Transcriptome Analysis Using Next Generation Sequencers" (article in preparation)

Contact : delfabbro@dimi.uniud.it