BMC Bioinformatics

Oral presentation

Optimal spliced alignments of short sequence reads

Fabio De Bona*[1], Stephan Ossowski[2], Korbinian Schneeberger[2] and Gunnar Rätsch[1]

Address: [1]Friedrich Miescher Laboratory, Max Planck Society, Spemannstr. 39, 72076 Tübingen, Germany and [2]Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany

Email: Fabio De Bona* - fabio@tuebingen.mpg.de

* Corresponding author

from Fourth International Society for Computational Biology (ISCB) Student Council Symposium
Toronto, Canada. 18 July 2008

Published: 30 October 2008

BMC Bioinformatics 2008, **9**(Suppl 10):O7 doi:10.1186/1471-2105-9-S10-O7

This abstract is available from: http://www.biomedcentral.com/1471-2105/9/S10/O7

Motivation

Next generation sequencing technologies open exciting new possibilities for genome and transcriptome sequencing. While reads produced by these technologies are relatively short and error-prone compared to the Sanger method, their throughput is several magnitudes higher. We present a novel approach, called *QPALMA*, for computing accurate spliced alignments of short sequence reads that take advantage of the read's quality information as well as computational splice site predictions. In computational experiments we illustrate that the quality information as well as the splice site predictions [1] help to considerably improve the alignment quality. Our algorithms were optimized and tested using artificially spliced genomic reads produced with the Illumina Genome Analyzer for the model plant *Arabidopsis thaliana*.

Methods

In this work we aim to develop a method exploiting all available information to accurately align as many as possible spliced reads to the genome. In previous work we already proposed methods taking advantage of splice site predictions and an intron length model (*Palma* [2]). We extend this method to benefit from the read's quality scores. The algorithm is based on extensions of the Smith-Waterman algorithm using more sophisticated parametrized scoring functions. The idea is to tune the parameters of the scoring functions such that the true alignment does not only achieve a large score, but also that all other alignments score lower than the true alignment [3].

In a typical application scenario one needs to align millions of short reads against the genome. In this case the direct application of the extended Smith-Waterman algorithm is not feasible. We therefore propose to combine our method with a fast suffix-array based approach to identify a seed for the alignment. This combined strategy allows us to efficiently align even very large numbers of reads (cf. Figure 1).

Results

We first studied the accuracy of aligning 30, 000 *spliced* sequences using different variants of *QPALMA*: with and without quality information, splice site predictions, and intron length information. From the results given in Table 2 we can conclude that all three components help to reduce the alignment error rate. We also tested the proposed pipeline on about 3 million short reads which contained about 10% spliced reads. The alignment took 12.5 h (on one CPU) and almost all of the reads (98.4%) were aligned correctly. This illustrates that the approach is not only accurate but also fast enough to be used in a next generation mRNA sequencing project.

Conclusion

We have presented a novel approach to solve the difficult task of aligning short reads as generated by NG sequencing techniques over exon boundaries. We were able to successfully exploit all available information sources – the read including its quality score information, splice site predictions, the intron length and, of course, the genome

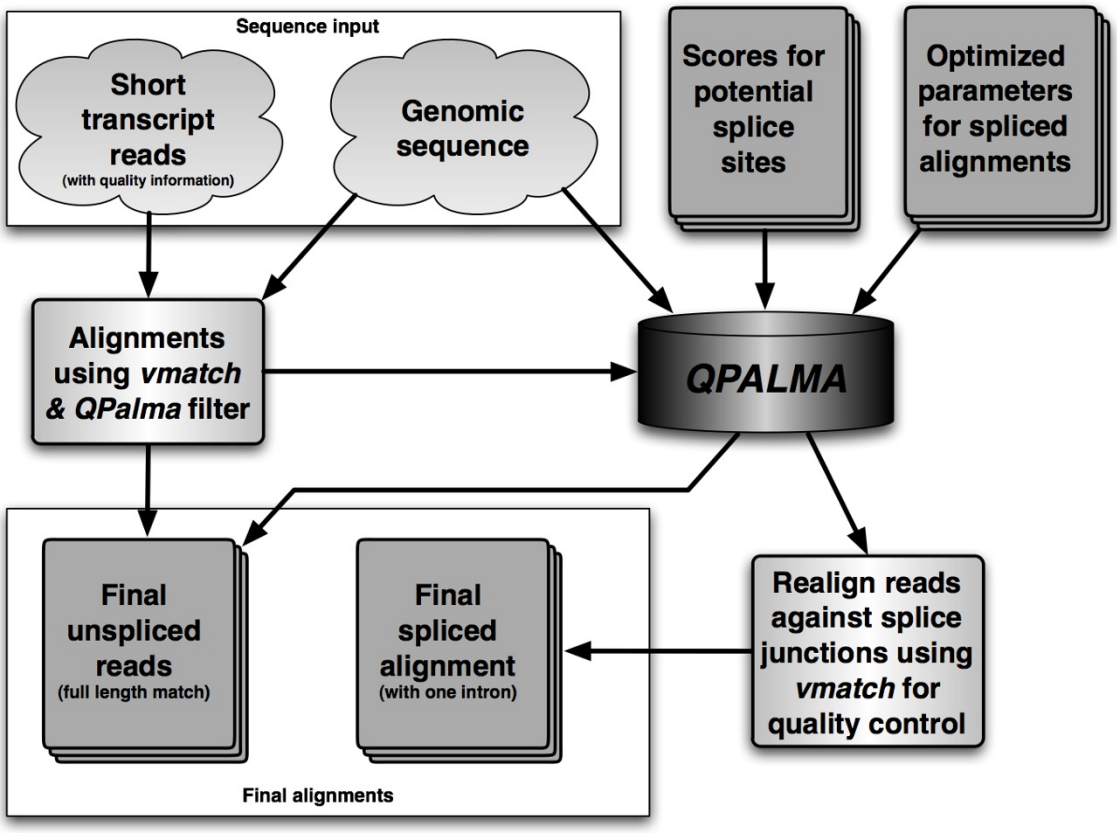

Figure 1
Proposed pipeline: Short reads are first aligned with *vmatch* to identify unspliced reads. Un-mapped or potentially spliced reads are aligned again to identify reads of at least half of the read length to find seeds for *QPALMA*. For each seed position *QPALMA* aligns the read and returns a score. The best scoring alignment is returned as the spliced alignment of the read.

Quality information	Splice site pred.	Intron length	Error rate
-	-	-	14.19 %
+	-	-	13.49 %
-	+	-	3.16 %
+	+	-	2.81 %
-	-	+	9.96 %
+	-	+	9.68 %
-	+	+	1.94 %
+	+	+	1.78%

Figure 2
We compute the fraction of reads that have been accurately aligned at all four boundaries (start and end of first and second exon) with and without using read quality information, splice site predictions and intron length scoring, respectively.

– each significantly contributing to decreasing the alignment error rate.

As future work, it would be interesting to consider the downstream analysis of deriving the gene structure based on these reads and to estimate its error as well. This will be particularly interesting for predicting gene structures with alternative transcripts.

References

1. Sonnenburg S, Schweikert G, Philips P, Behr J, Rätsch G: **Accurate Splice Site Prediction Using Support Vector Machines.** *BMC Bioinformatics* 2007, **8(Suppl 10):**S7.
2. Schulze U, Ong C, Hepp B, Rätsch G: **PALMA: mRNA to genome alignments using large margin algorithms.** *Bioinformatics* 2007, **23(15):**1892-1900.
3. Tsochantaridis I, Hofmann T, Joachims T, Altun Y: **Support Vector Machine Learning for Interdependent and Structured Output Spaces.** *Proceedings of the 16th International Conference on Machine Learning* 2004.