Maximizing the number of aligned reads from RNA-Seq data

Thomas Bonfert, Gergely Csaba, Ralf Zimmer and Caroline C. Friedel
Institute for Informatics, Ludwig-Maximilians-University Munich
bonfert@bio.ifi.lmu.de

RNA sequencing by next generation sequencing technologies (RNA-Seq) provides a novel way to characterize the transcriptome of different cell types. One important application is the detection of different transcripts from the same genetic locus.
For this purpose, the position of reads on the transcriptome has to be determined first. If an organism with a sequenced and well-annotated genome is investigated, a common approach is to align the generated reads against a reference transcriptome. Based on the alignment, the reads are mapped onto exons and exon-exon junctions of annotated genes. This mapping then provides the basis for quantification and analysis of the transcriptome. Unfortunately, depending on the quality of the sequenced reads, this approach generally fails to identify the origin for a significant large amount of reads.
In order to address this problem, we have developed a pipeline which maximizes the number of reads whose origin can be explained by investigating several alternative possibilities. This pipeline is suitable for RNA-Seq data for any organism with available transcriptome and genome annotation. In a first step, a transcriptome mapping is performed as described above. Second, alignment to the genome identifies both incompletely spliced transcripts as well as novel splicing events. Finally, novel transcripts are predicted and potential sources of contaminations are analyzed. Our evaluation results show that using this advanced read processing pipeline, the number of aligned reads can be increased considerably compared to alternative approaches.