# Alternative Splicing from RNA-Seq Data without the Genome

**S. Beretta, R. Rizzi, G. Della Vedova and P. Bonizzoni**

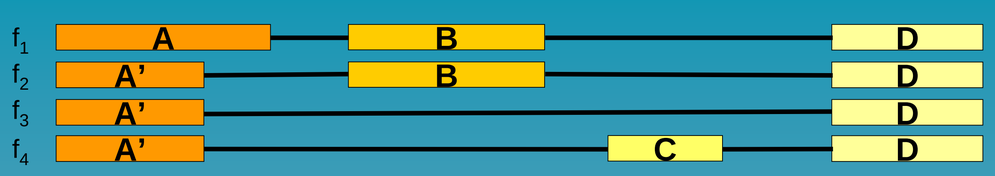
Università degli Studi di Milano-Bicocca

## Abstract

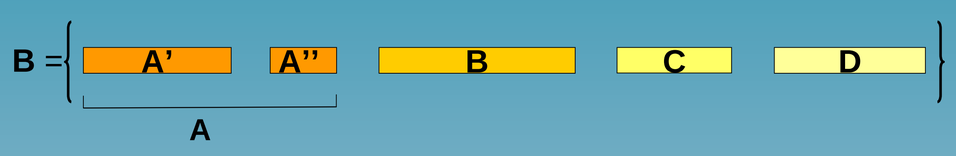
Next Generation Sequencing (NGS) technologies allow massive sequencing of biological molecules and have a strong impact on bioinformatics. In particular, RNA-Seq is a recent technique to sequence expressed transcripts, characterizing both the type and the quantity of transcripts expressed in a cell. Some challenging tasks in RNA-Seq data analysis are to map the reads to a reference genome, to assemble them into contigs, and to predict the exon-intron structure of a gene and its full-length isoforms [1]. Current methods, exploiting RNA-Seq data and used to identify the exon-intron boundaries on the genome, all incorporate a crucial step to compute the spliced alignments of the RNA-Seq reads against the genomic sequence. Few efforts have been devoted to use RNA-Seq data for obtaining a draft structure of a gene. We tackle the problem of predicting, from NGS data, the gene structure induced by the different full-length isoforms, due to Alternative Splicing (AS) [2], without resorting to any kind of alignment of RNA-Seq reads against the genome.

## The expressed gene

A gene **G** with four isoforms $f_1$,$f_2$,$f_3$ and $f_4$

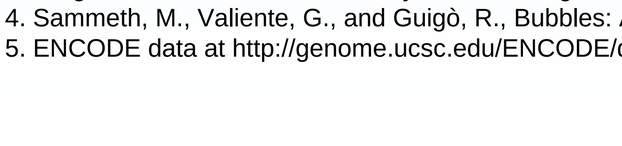



The set **B** of blocks of G




A **block** is a string that appears entirely or not at all in any isoform

## Our goal

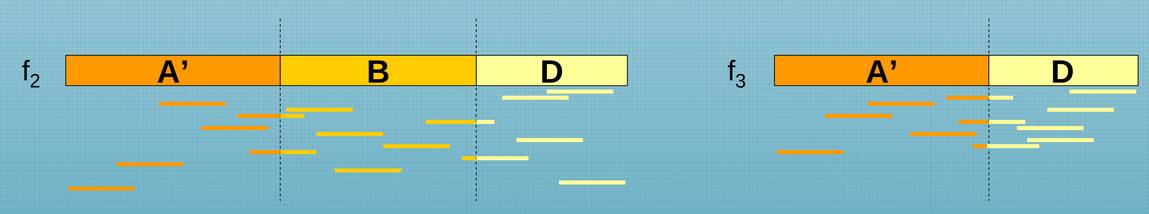



**Isoform Graph $G_F$=(B,E)**

($b_i$,$b_j$) ∈ E iff the two blocks $b_i$ and $b_j$ (∈ B) are consecutive in at least one isoform of G. Each isoform corresponds to a path in $G_F$

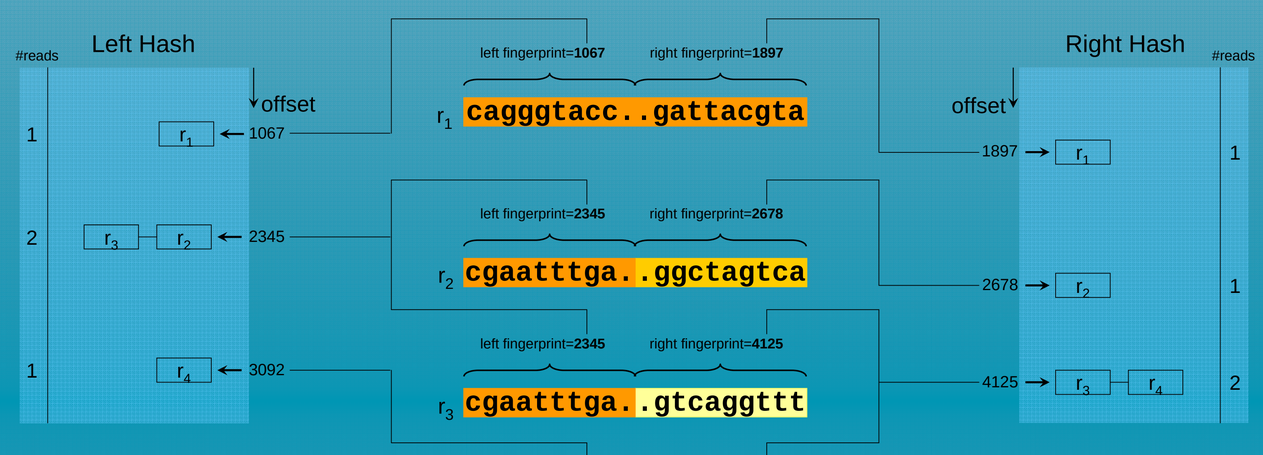
## The method

**INPUT:** a set R of RNA-Seq reads extracted from the unknown transcripts of a gene G




**Reads that are substrings of a block (called *unspliced* in our framework), are represented by a one-colored bar, while reads extracted from a block junction (called *spliced*) are represented by a two-colored bar**

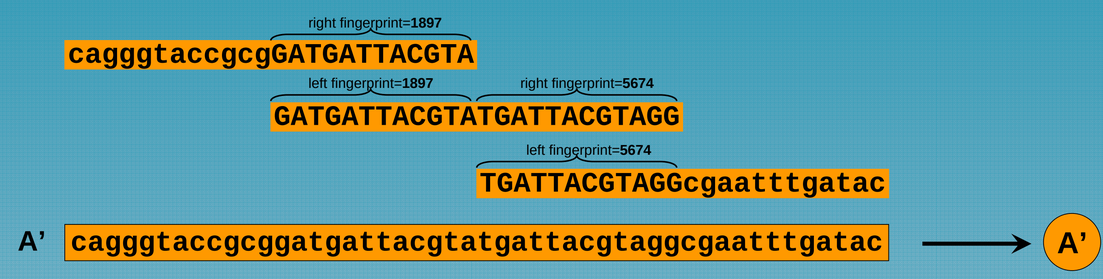
**STEP1**: hashing of the reads in the input set R



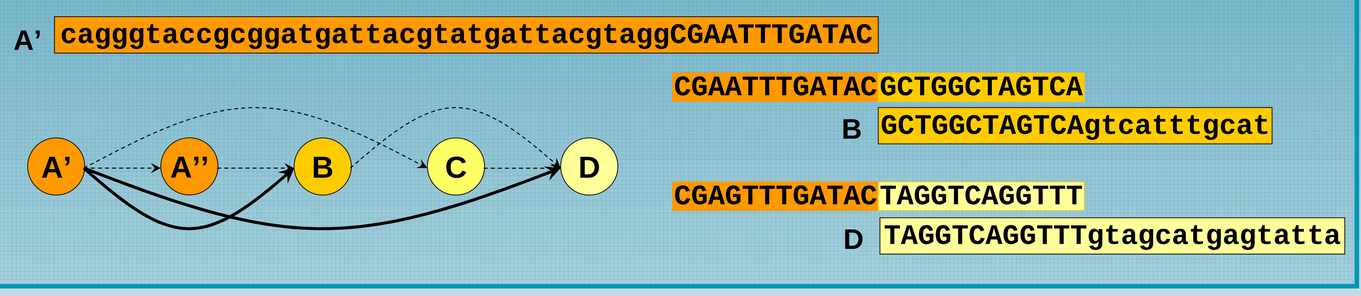

r is *unspliced* when the related #reads is 1 in both tables

r is *spliced* when the related #reads is >1 in at least one table

**STEP2**: assembling *unspliced* reads into blocks (nodes of $G_F$)




**STEP3**: linking blocks (edges of $G_F$) using *spliced* reads




## Results

We implemented our method as C++ programs, and we tested its performance on a standard work-station with a 2.6GHz quad-core processor and 12GB of RAM. Our experiments have been performed on the set of 112 genes used as training set in the EGASP competition [3]. For each gene, we have run our method on simulated and real RNA-Seq data. Successively, we have compared the actual isoform graph with the RNA-Seq graph produced by our method. The results over simulated data (22 million bp of RNA-Seq data have been analyzed in 67 minutes) show that we have computed draft gene structures that fully match those in ENCODE for 40 genes out of 112. Overall, some measures of similarity between graphs that are based on the notions of sensitivity and specificity already used for evaluating AS prediction programs [3] are satisfactory (average Sn=0.868, average Sp=0.765). Due to the novelty of our approach, the quality of the results over real data [5] are not as good (average Sn=0.358, average Sp=0.294), pointing out the need for a refinement phase (currently in the design stage) that considers the graph computed by our program and the genomic sequence (but without any alignment) before outputting the final result. At the same time, we notice that the overall running time has been roughly 3.5 hours for processing 2Gbases of RNA-Seq data.

## References

1. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoforms switching during cell differentiation. Nature Biotechnology 28(5):516-520 (2010).
2. Leipzing, J., Pevzner, P., and Heber, S., The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. Nucleic Acids Research 32(13):3977-3983 (2004).
3. Guigò, R., Flicek, P., Abril, J., Reymond, A., Lagarde, J., at al., EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biology 7(1):S2.1-31 (2006).
4. Sammeth, M., Valiente, G., and Guigò, R., Bubbles: Alternative Splicing events of arbitrary dimension in splicing. Lecture Notes in Computer Science 4955:372-395 (2008).
5. ENCODE data at http://genome.ucsc.edu/ENCODE/downloads.html