ASGAL: Aligning RNA-Seq Data to a Splicing Graph to Detect Novel Alternative Splicing Events

Luca Denti*, Raffaella Rizzi, Stefano Beretta, Gianluca Della Vedova, Marco Previtali, Paola Bonizzoni

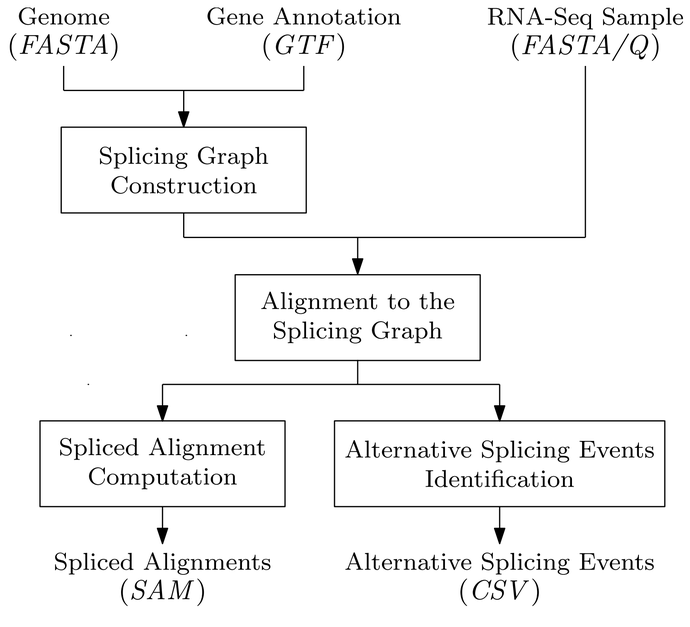
Dept. of Informatics, Systems, and Communication (DISCo), University of Milano–Bicocca, Milan (Italy)
*luca.denti@disco.unimib.it

Background

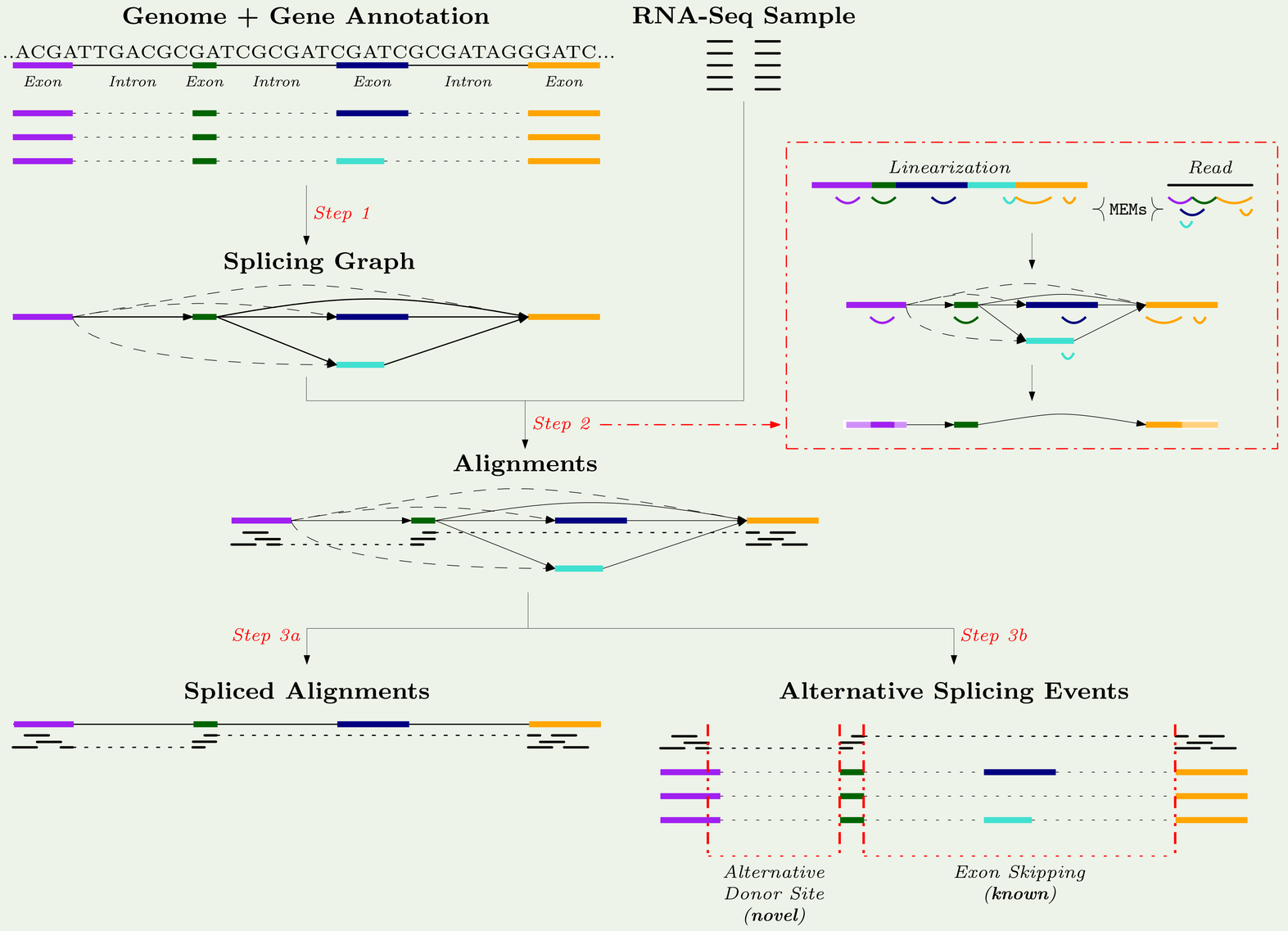
While the reconstruction of transcripts from a sample of RNA-Seq data is a computationally expensive and complicated task, the **detection of alternative splicing events** from RNA-Seq data and a gene annotation is computationally feasible. This task, which is adequate for many transcriptome analyses, is usually achieved by **aligning the reads** to a reference genome and by **comparing the alignments** with a gene annotation.

Main Contribution

`ASGAL` (Alternative Splicing Graph ALigner) is the first tool which aligns a RNA-Seq sample directly to the **splicing graph** of a gene with the main goal of detecting alternative splicing events. Differently from other tools which search for alternative splicing events differentially expressed in a sample, `ASGAL` detects the alternative splicing events expressed in a sample which are novel with respect to the input gene annotation.



Method



Step 1 Construction of the *(augmented)* splicing graph from the genome and the annotation

Step 2 Alignment of the RNA-Seq read sample to the splicing graph

a) *Linearization* of the splicing graph into a single string (one-time step)

b) Computation of the Maximal Exact Matches (`MEMs`) between this *linearization* and each read

c) Computation of the splicing graph alignments by linking `MEMs` together (allowing gaps both on the read and the splicing graph)

Step 3a Mapping of the alignments to the input reference genome using the exon boundaries

Step 3b Identification of the events expressed by the input sample

a) Computation of the splice sites used by the alignments

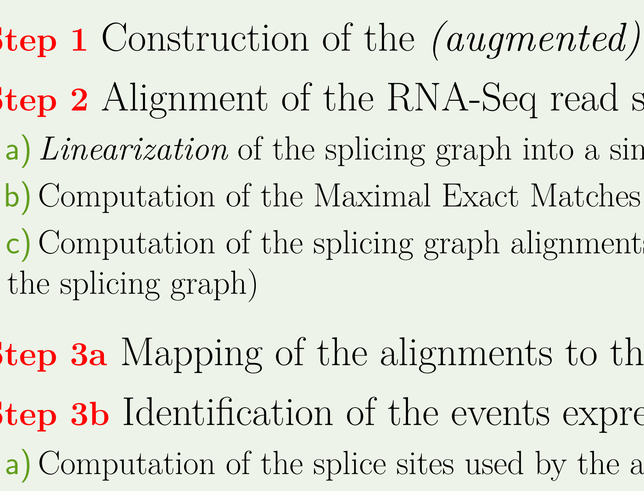
b) Reconciliation of the introns identified from the splice sites by using canonical intronic patterns

c) Comparison of the identified introns with the annotated ones

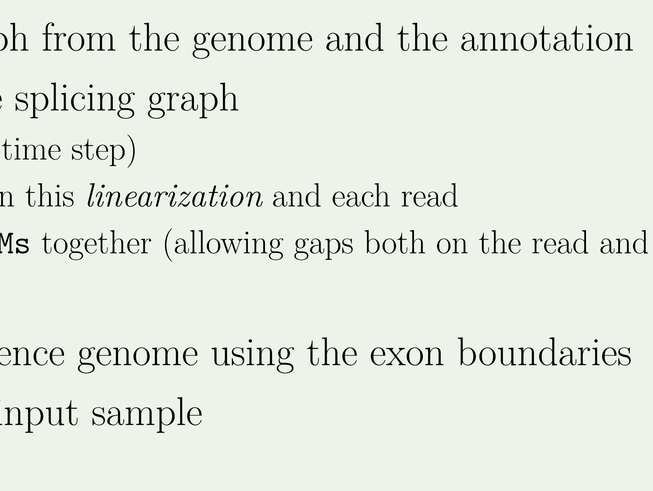
Experimental Results

Simulated data: 1000 genes and reads from [4]

- Ground truth built using `AStalavista` [6] from GENCODE annotation (v19)
- Comparison with `SplAdder` [4]
- Alignment phase *(mean time and max RAM per gene)*
  - `ASGAL` required 45 secs and 267MB
  - `STAR` [7] required 564 secs and 275MB
  - `Hisat2` [8] required 3 secs and 14MB
- Events identification phase *(as above)*
  - `ASGAL` required 2.7 secs and 443MB
  - `SplAdder` required 3.25 secs and 47MB



Real data: 630 genes and reads from [5]



Discussion

- Aligning reads directly to a splicing graph allows for better alternative splicing event prediction with respect to tools requiring spliced alignments of the RNA-Seq data to a reference genome. However, using a splicing graph is more resource-consuming than using a reference genome
- `ASGAL` shows better accuracy in managing real samples that are likely to be the result of the expression of a single transcript for each gene
- `ASGAL` is freely available at `http://asgal.algolab.eu`

References

[1] Beretta et al. (2017). AlCoB
[2] Denti et al. (2018). *bioRxiv*
[3] Ohlebusch et al. (2010). *SPIRE*
[4] Kahles et al. (2016). *Bioinformatics, 32.12*
[5] Flockhart et al. (2012). *Genome research 22.6*
[6] Foissac et al. (2007). *Nucleic acids research 35.suppl_2*
[7] Dobin et al. (2013). *Bioinformatics, 29.1*
[8] Kim et al. (2015). *Nature methods 12.4*