Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli Studi di Milano-Bicocca
Milan - Italy

Alternative Splicing from RNA-seq Data without the Genome

8th Special Interest Group meeting on Alternative Splicing
AS-SIG, ISMB/ECCB 2011
Vienna — July 15th/16th, 2011

Stefano Beretta Raffaella Rizzi
Gianluca Della Vedova Paola Bonizzoni

Motivations and Challenges

Detecting Alternative Splicing (AS) variations from RNA-seq data

- No specific tools for large-scale inference of AS variations among gene transcripts
- **Our goal**: identification of AS variations **without a reference genome**

Motivations and Challenges

- Reference genome is not always available
- RNA-seq data alignment against the genome is too expensive

Our Solution

fast* construction of a graph representation of AS variations from RNA-seq data **without a reference genome**

*linear time w.r.t. the number of reads

RNA-seq Data

- Basic Features:
  - Short sequences ($30 - 400$bp)
  - Depth sequencing $\rightarrow$ Millions / Billions of sequences
  - Quality
  - Unknown error
- High-throughput sequencing platforms:
  - SOLiD, Illumina, Roche's 454, HeliScope

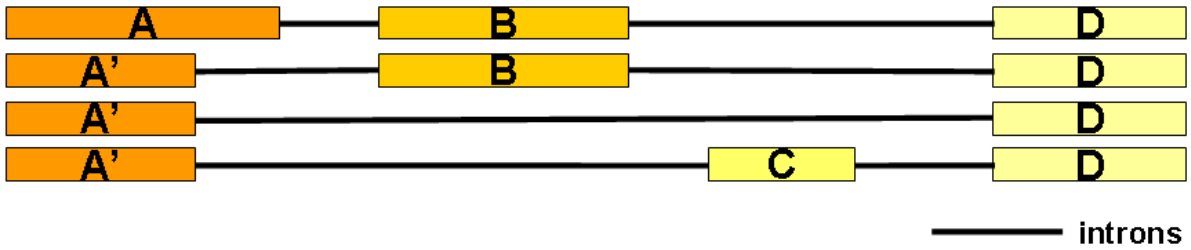
RNA-seq Analysis: State of Art

- Read Mapping (Spliced Aligners)
  - Exon-first methods (MapSplice, SpliceMap, Tophat)
  - Seed-extend methods (GSNAP, QPALMA)
- Expression Quantification
  - Gene quantification (Alexa-seq ,ERANGE, NEUMA)
  - Isoform quantification (Cufflinks, MISO, RSEM)
- Transcriptome Reconstruction
  - Genome-guided assembly (Scripture, Cufflinks)
  - Genome-independent assembly (Velvet, TransABySS, Trinity)

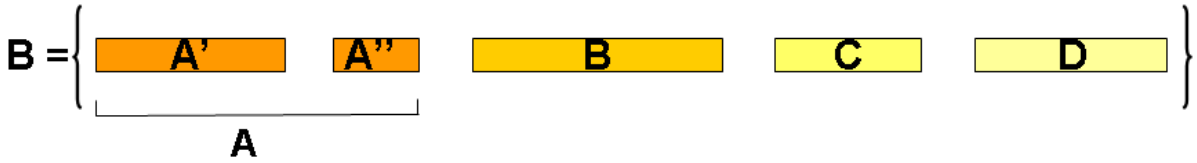
NATURE METHODS, June 2011

Our Goal: Isoform Graph

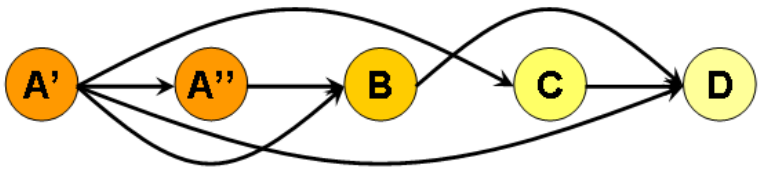
- Gene isoforms



- Set of blocks



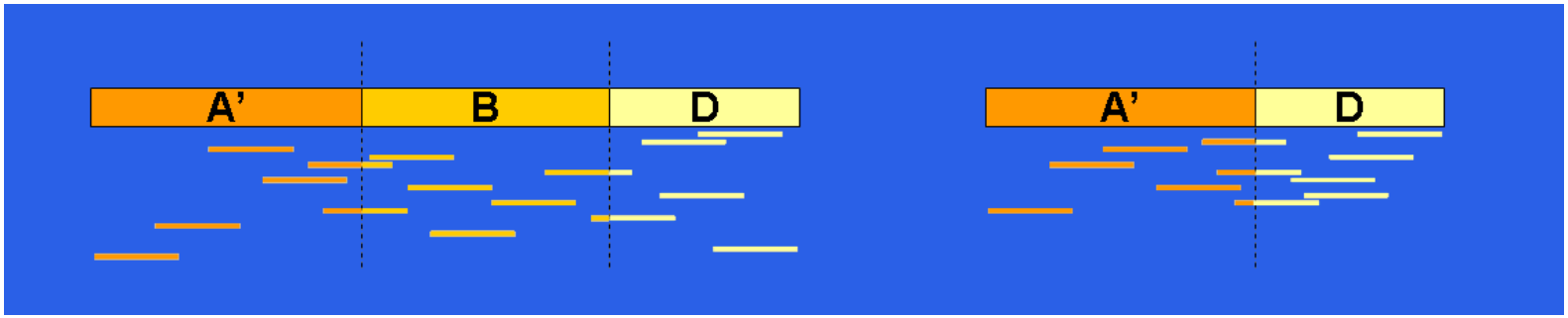
- **Isoform graph**



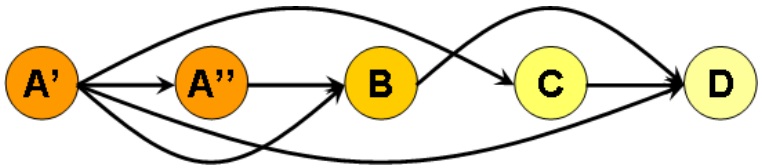
Our Goal: Isoform Graph

Isoform Graph Reconstruction

Input: a set of RNA-seq reads from unknown gene transcripts



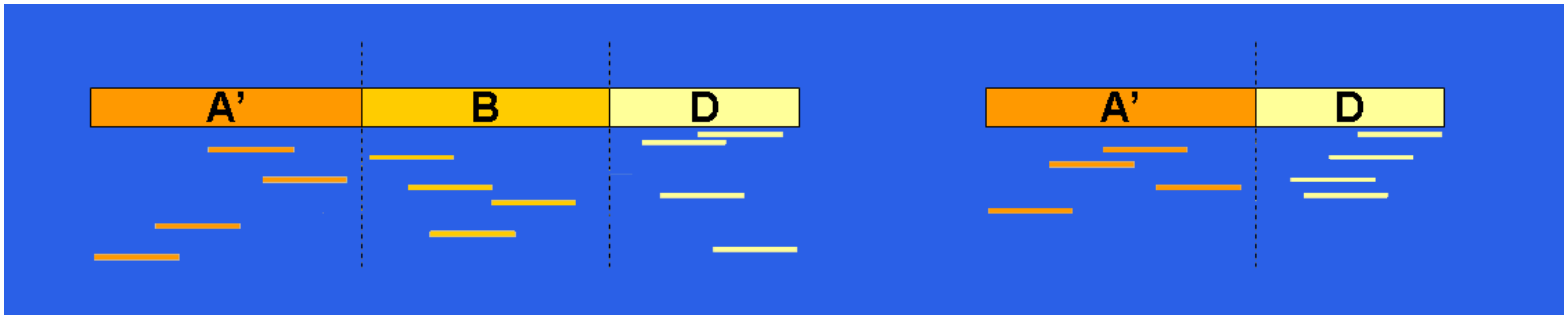
- *Isoform Graph*:



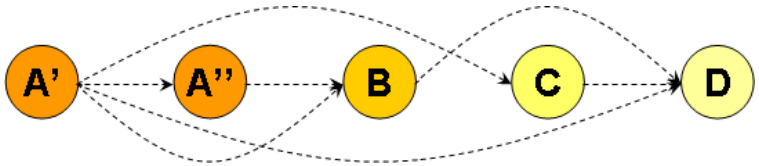
Our Approach

Isoform Graph Reconstruction

Input: a set of RNA-seq reads from unknown gene transcripts



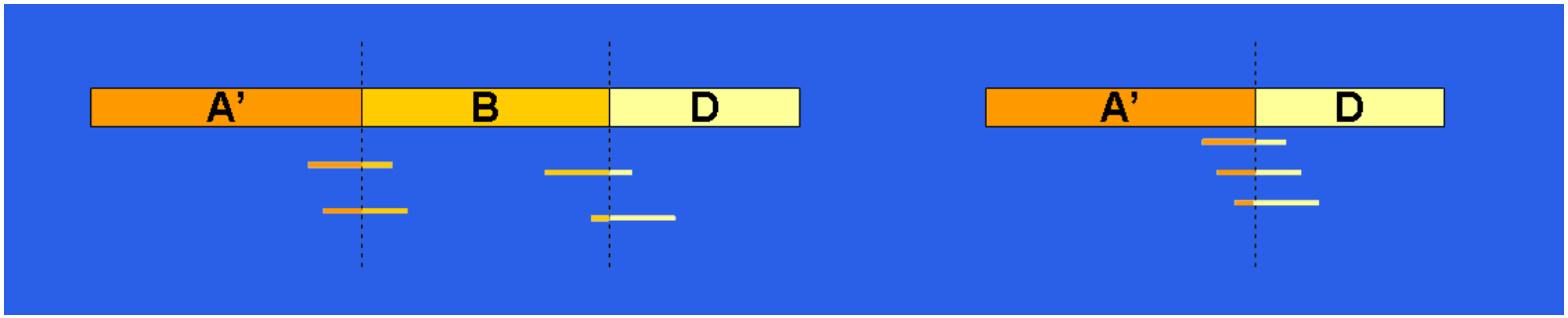
- *Unspliced reads*:



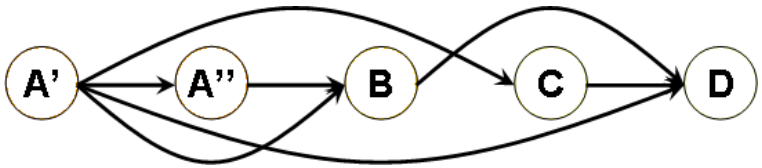
Our Approach

Isoform Graph Reconstruction

Input: a set of RNA-seq reads from unknown gene transcripts

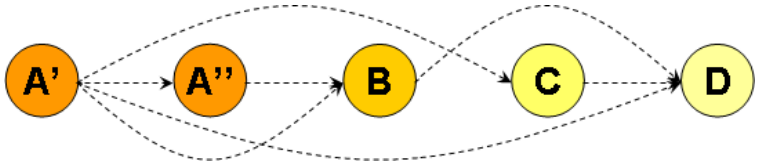


- *Spliced reads*:

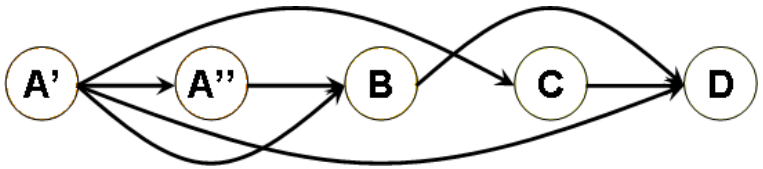


Method Outline

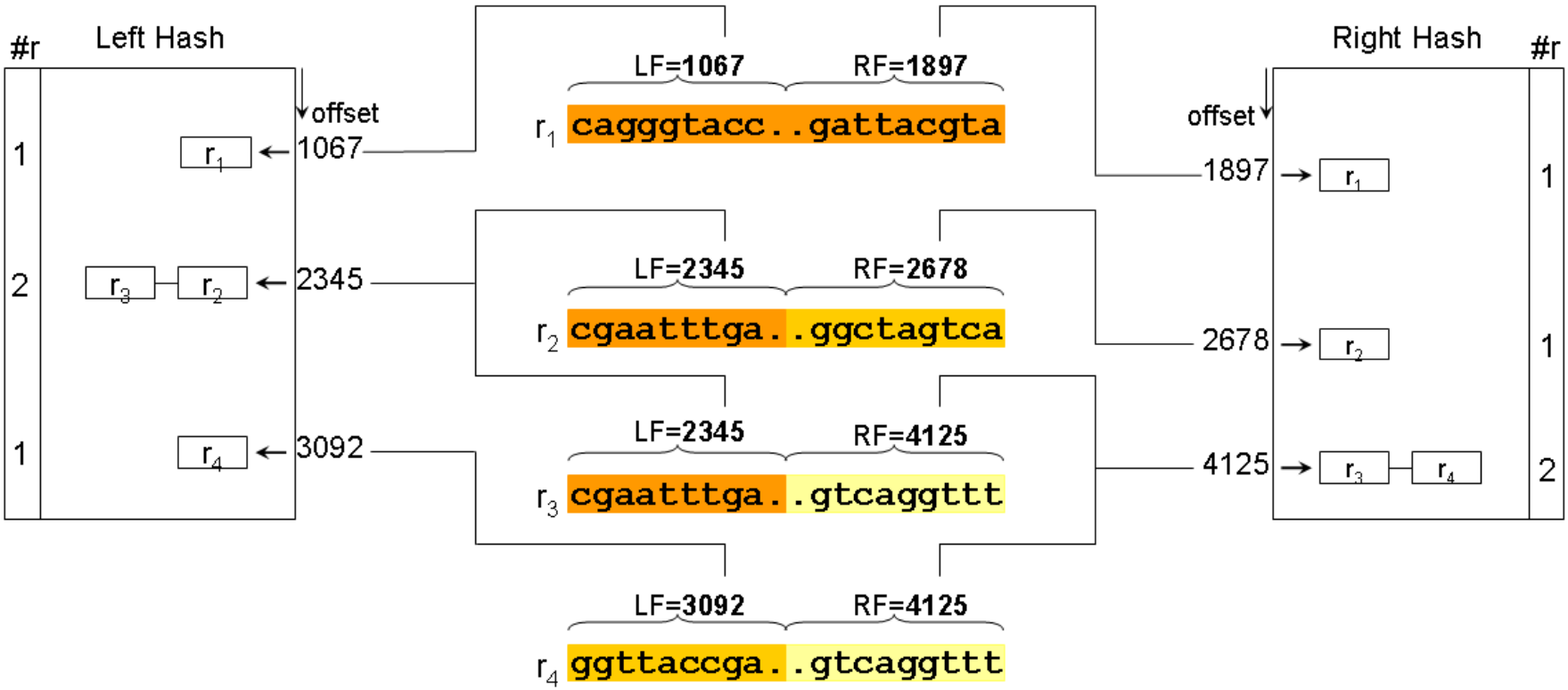
- Hashing input reads
  - Input set partitioning $\rightarrow$ Unspliced/Spliced
  - Constant time access to RNA-seq reads
- Assembling *unspliced* reads into blocks (graph nodes)



- Linking blocks with *spliced* reads (graph edges)

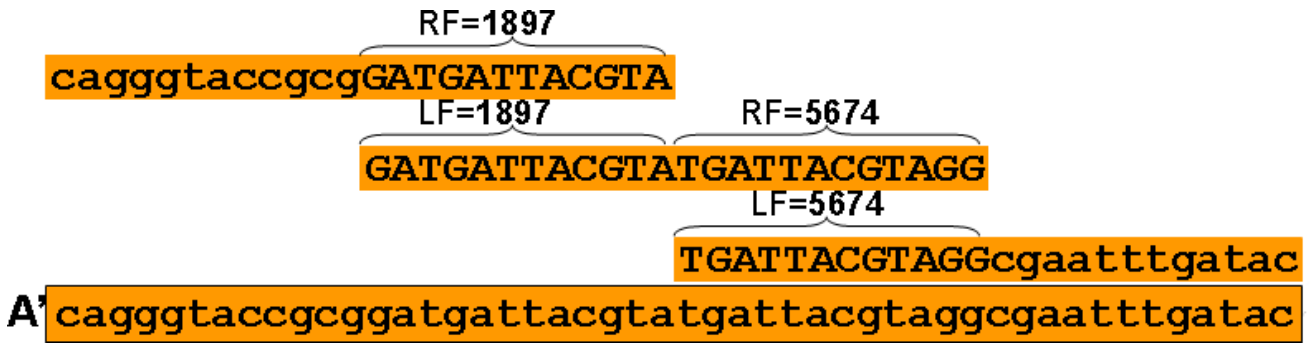


Hashing of the input reads

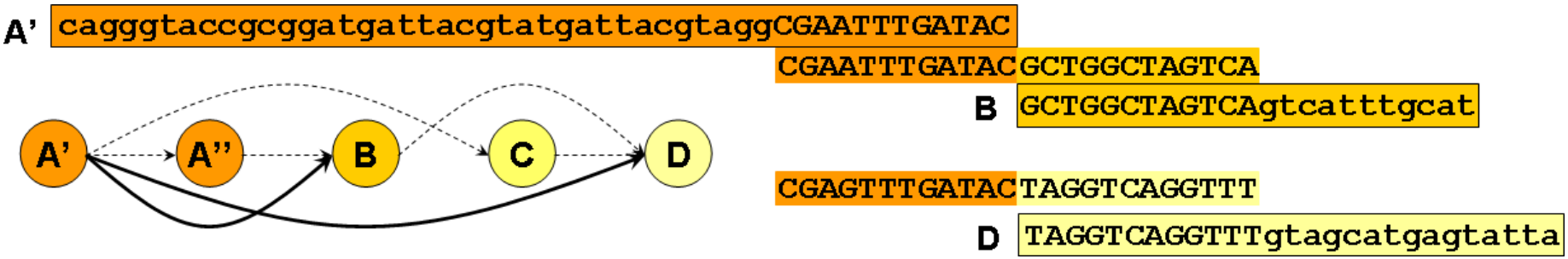


Assembling and Linking

- Assembly of unspliced reads



- Linking with spliced reads



Experiment on Simulated Data

Data from: 112 genes used as training set in EGASP*

- 22.8×10^6 simulated reads
- read length: 64bp
- % of mutated reads: $0, 2, 4, 8, 16$

Results

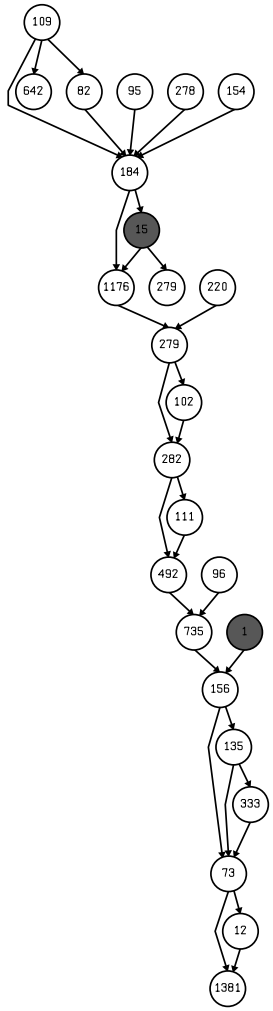
- $\sim$ 40 genes "correctly reconstructed"
- 67 minutes
- Average $S_n = 0.868$
- Average $S_p = 0.765$

*Guigò et al., *Genome Biology*, 2006

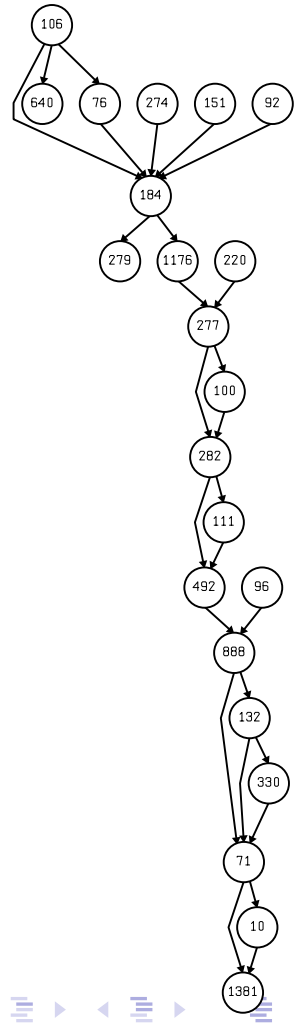
An Example: gene L1CAM

- Prediction summary
  - Predicted nodes: 22
  - Predicted arcs: 27
  - S_n (nodes): 0.84
  - S_p (nodes): 0.95
  - S_n (arcs): 0.71
  - S_p (arcs): 0.82

- Annotation



- Prediction



Experiment on Real Data

RNA-seq data from ENCODE/Caltech

- 2×10^9 reads
- read length: 75bp (Illumina)
- unknown error

Results

- 210 minutes
- Average $S_n = 0.358$
- Average $S_p = 0.294$

Issues and Future Work

- Issues
  - SNP
  - Read error
  - Splice junctions not uniquely identified
  - Some AS variations are hard to characterize
- Future Works
  - Extract AS events (exon skipping, mutually exclusive exons, etc.) from isoform graph
  - Use a reference genome to predict AS variants in a donor genome (also represented with RNA-seq reads)
  - Genome-wide experiment on real data from different sequencing technologies

Conclusions

- New method for AS variants inference from NGS data
- Efficient in theory and practice
- 2 k-mers/read
- No error $\rightarrow$ Good performance
- Extremely scalable approach
- Ongoing implementation development
  - Improving performances on real data
  - SNP
  - Error correction
  - Intron/Exon refinement (involving the genome)