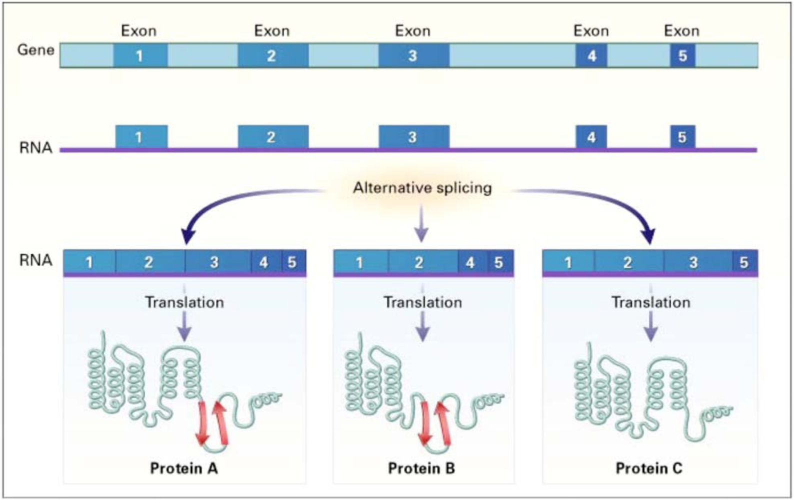
Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli Studi di Milano-Bicocca
Milan - Italy

Reconstructing Isoform Graphs from RNA-Seq data

IEEE International Conference on
Bioinformatics and Biomedicine (BIBM2012)
Philadelphia – October 4th/7th, 2012

Stefano Beretta Raffaella Rizzi
Gianluca Della Vedova Paola Bonizzoni

Alternative Splicing



Motivations and Challenges

Detecting Alternative Splicing (AS) variations
from RNA-Seq data without a reference genome

- No specific tools for large-scale inference of AS variations among gene transcripts
- Reference genome is not always available

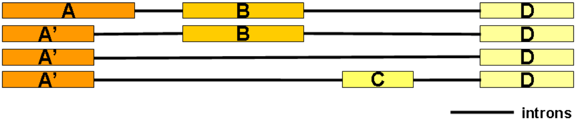
Motivations and Challenges

Characterization of AS variations from RNA-Seq data

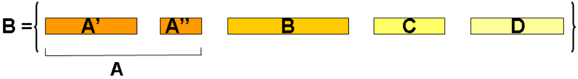
- Background
  - Tools for transcript reconstruction/quantification from NGS data (and genomic sequence) are known
- Goal
  - Building a isoform graph that explains all AS events derived from several transcripts
    - without explicit transcript assembly
    - without a reference sequence

Isoform Graph

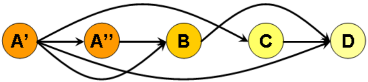
- Gene isoforms



- Set of blocks



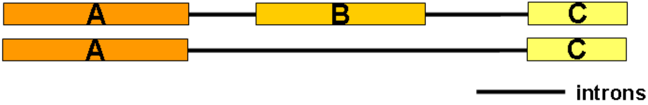
- **Isoform graph**



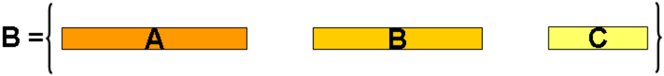
Block Definition

A *block* is a string that appears entirely (not partially) or not at all, in each isoform

- Isoforms



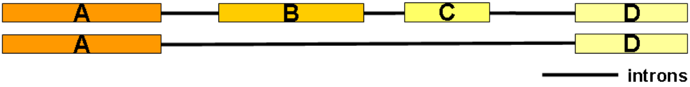
- Set of Blocks



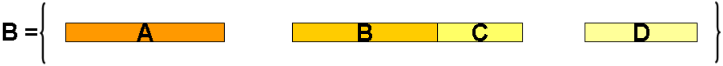
Block Definition

A *block* is a string that appears entirely (not partially) or not at all, in each isoform

- Isoforms



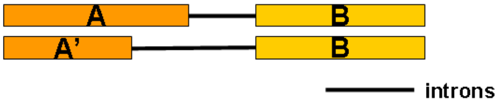
- Set of Blocks



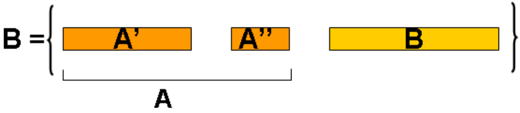
Block Definition

A *block* is a string that appears entirely (not partially) or not at all, in each isoform

- Isoforms

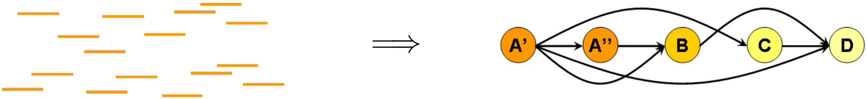


- Set of Blocks



Computational Problem

Input: set of RNA-Seq reads from unknown gene transcripts
Output: Isoform Graph

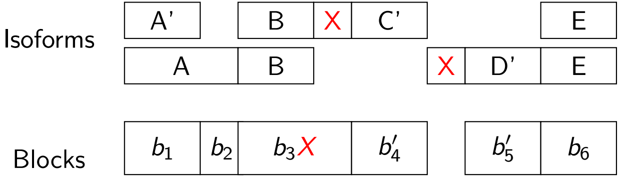


- Question 1: what are the conditions under which the isoform graph of a gene can be reconstructed from a sample of RNA-Seqs?
- Question 2: can we build efficiently such a graph or an approximation of it?

Solution - Question 1

Conditions

- Starts and ends of blocks on branches derived from the same block must be different characters

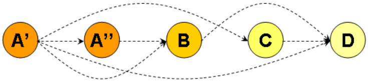


- There are no repeated substrings in the block sequences

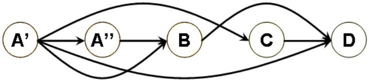
Solution - Question 2

Method Outline

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



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



Experimental Validation

- Data
  - 112 genes used as training set in EGASP
    - Separated and Mixed reads
    - Low and High coverage
- Analysis
  - Graph mapping (nodes and arcs)
  - Accuracy (S_n and PPV)

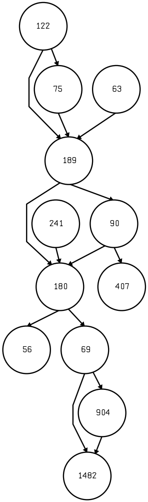
Experimental Validation

Dataset	Perfectly Predicted Genes	S_n (nodes)	PPV (nodes)	S_n (arcs)	PPV (arcs)
High cov. (separated)	43	0.86	0.92	0.72	0.82
Low cov. (separated)	39	0.87	0.91	0.75	0.81
High cov. (mixed)	-	0.84	0.78	0.71	0.68

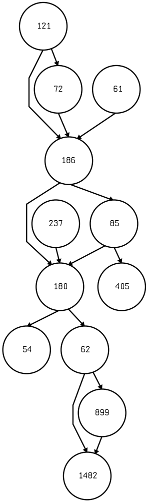
Example: gene TUFT1

- Prediction summary
  - Predicted nodes: 12
  - Predicted arcs: 14
  - S_n (nodes): 1
  - S_p (nodes): 1
  - S_n (arcs): 1
  - S_p (arcs): 1

- Annotation



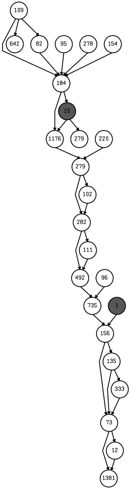
- Prediction



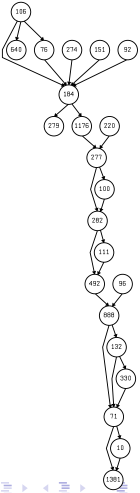
Example: gene L1CAM

- 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



Conclusions and Developments

- Conclusions
  - Computational method for building isoform graph (from millions of sequences)
  - Efficient in theory and practice
  - Characterization of conditions for reconstructing splicing graph without a reference sequence
  - Extremely scalable approach
- Developments
  - 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)

References

- Acknowledgements
    - Paola Bonizzoni
    - Gianluca Della Vedova
    - Raffaella Rizzi
- Software
    - `http://algolab.github.com/RNA-seq-Graph/`

Thanks!