

PIntron: a fast method for gene-structure prediction

Paola Bonizzoni

Gianluca Della Vedova

Yuri Pirola

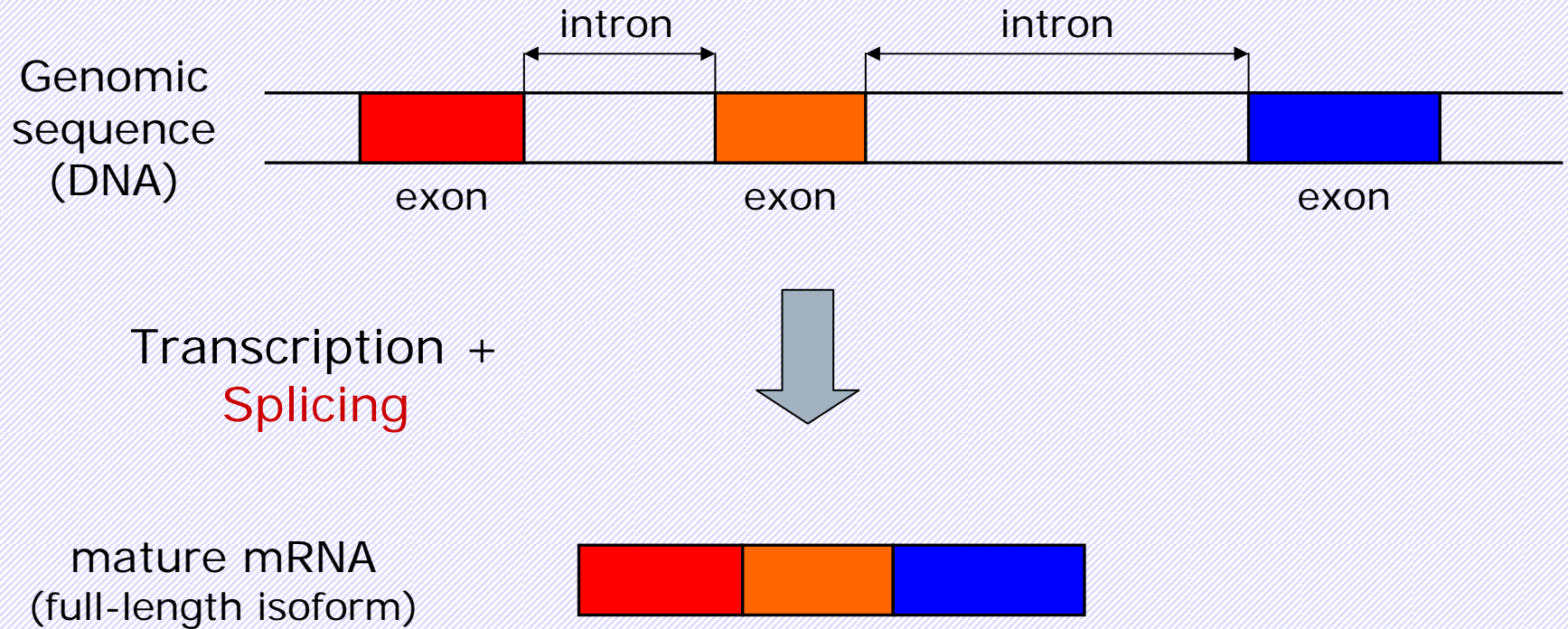
Raffaella Rizzi

Università degli Studi di Milano-Bicocca, Italy

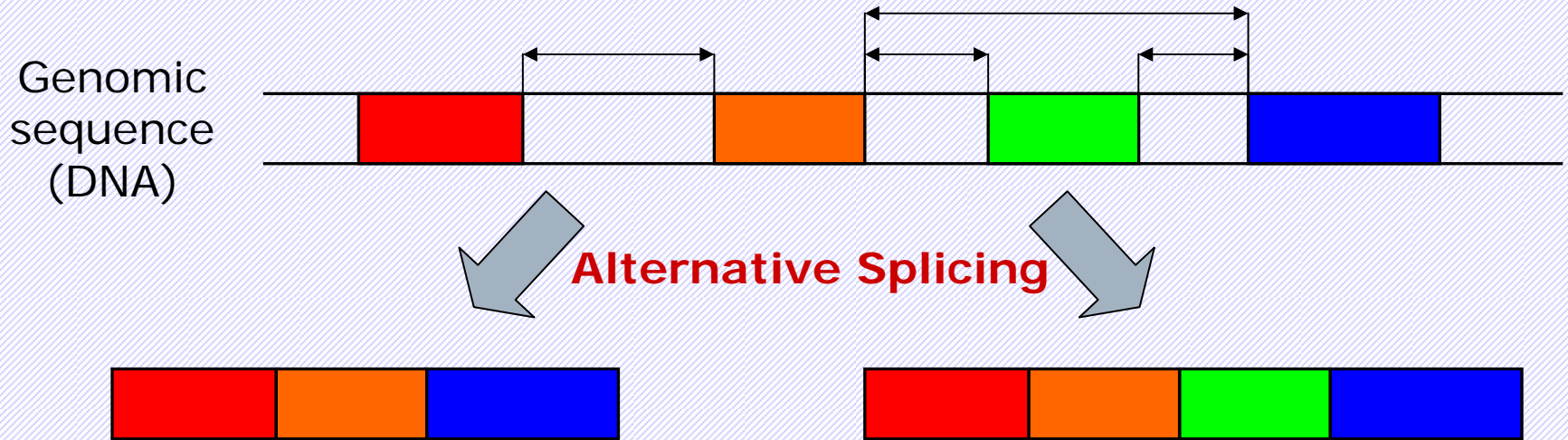
Outline

- The computational problem
- Prior work
- Our pipeline
- Experimental comparison

Eukaryotic Gene Structure



Alternative Splicing

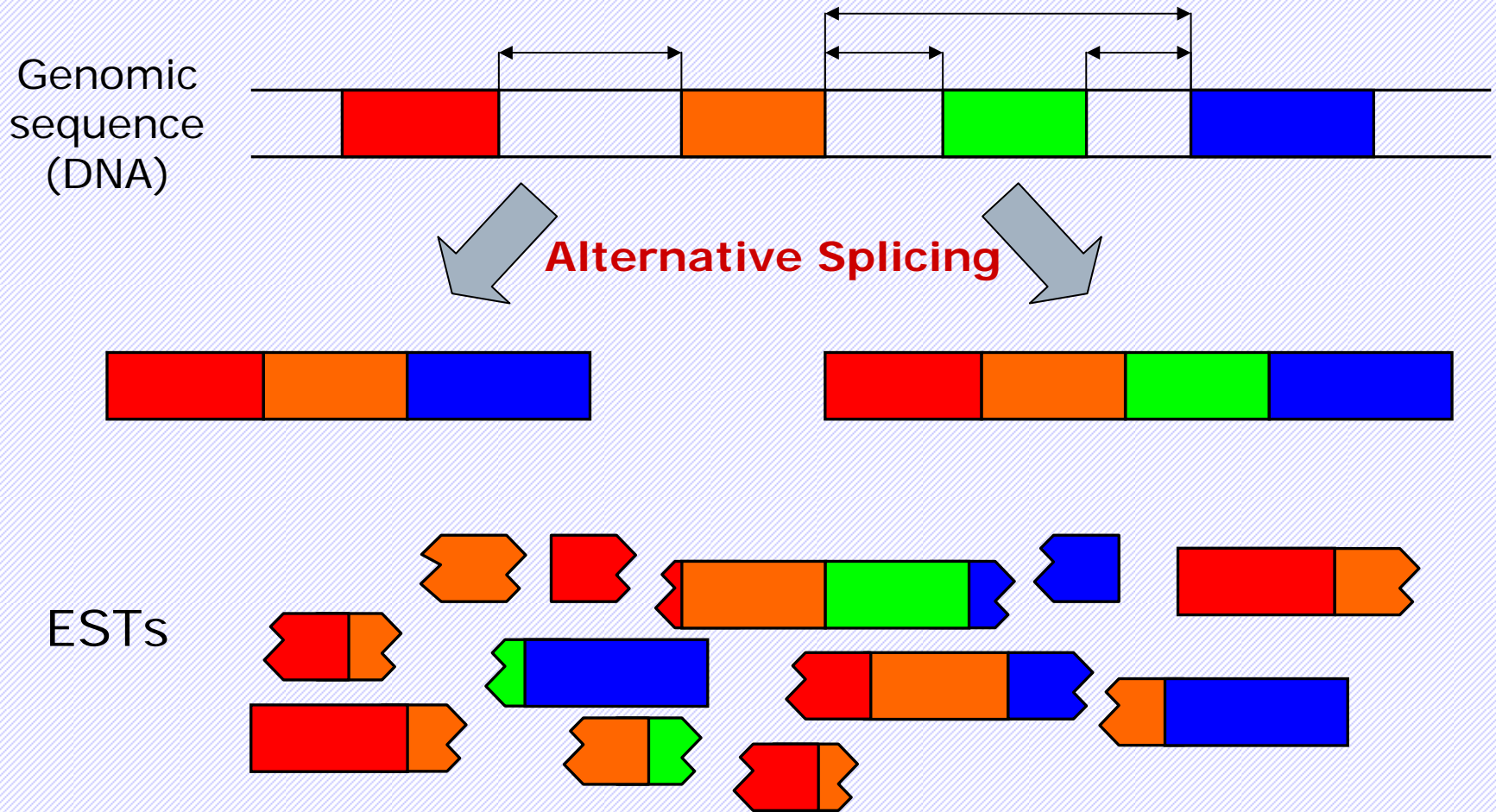


2 distinct mRNAs/full-length isoforms → 2 distinct proteins

- AS is widespread (95% of human genes)
- Aberrant AS events → diseases

(Pan et al., Nat Gen, 2008) and (Matlin et al., Nat Rev, 2005)

Alternative Splicing

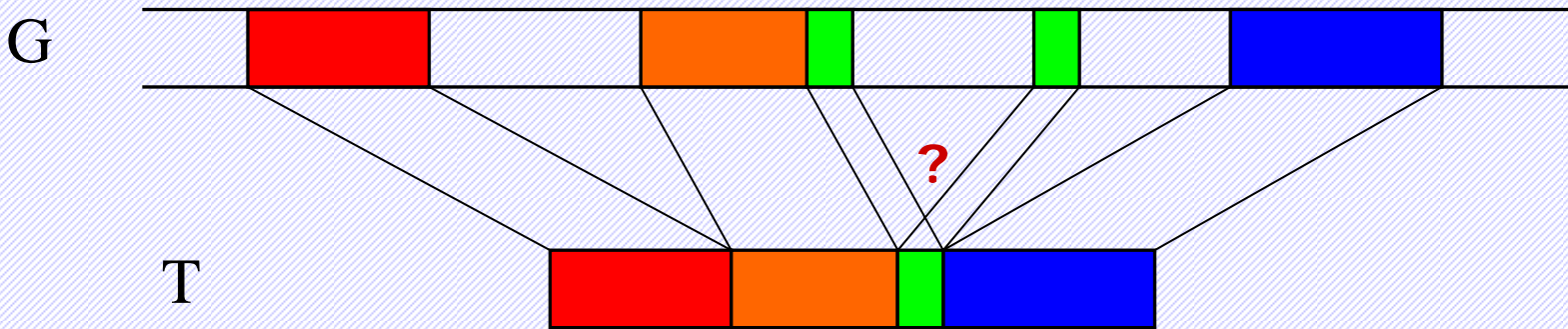


Gene structure prediction: prior work

- Three strategies:
 - ab-initio: “statistical” recognition
 - GENEZILLA, NSCAN
 - expression-based:
alignments of ESTs, mRNAs, or proteins
to the genomic sequence
 - Exogean, ASPic, **PI**ntron
 - (ab-initio + expression-based) methods
 - AUGUSTUS, JIGSAW

Spliced-alignment problem

- **Problem:**
Align a transcript T to the genomic sequence G while admitting introns
- Different alignments exist



PItron: our pipeline

Input: *a cluster of transcripts and a genomic sequence*

1. For each transcript:
 - a. Represent all its possible spliced-alignments
 - b. Extract its biologically-meaningful alignments
2. Compute a raw consensus gene-structure
3. Refine the raw gene-structure

The Pipeline: Step 1.a

Task: represents all the spliced-alignments

How?

Embedding Graph of G and T

- *Vertices* V: maximal common substrings of G and T
- *Edges* E: connect two common substrings that could be consecutive in a spliced-alignment

– Can be built efficiently:

$$O(|G|) + O(|T|+|V|) + O(|V|^2)$$

The Pipeline: Step 1.b

Input: an embedding graph (for each transcript)

Task: extraction of “relevant” spliced-alignments

Output: some biologically-meaningful spliced-alignments (for each transcript)

How?

1. Embedding graph visit +
2. Spliced-alignment reconstruction

The Pipeline: Step 1.b (cont'd)

1. Embedding graph visit

- computes *representative* embeddings (i.e. embeddings which induce distinct spliced-alignments)

2. Spliced-alignment reconstruction

- computes putative exons and raw intron boundaries
- discards low-quality alignments

The Pipeline: Step 2

Input: all the spliced-alignments of all the transcripts

Task: computation of a raw consensus gene-structure

Output: a minimal gene-structure that “explains” a single spliced-alignment of each transcript

Min Factorization Agreement

(Bonizzoni et al., WABI 2009)

Bad news: NP_{hard}

Good news: efficient heuristic

The Pipeline: Step 3

Input: a *raw* gene-structure

Task: refine intron boundaries and discard possible errors

Output: a *final* gene-structure

How?

- classify introns (U2, U12, BSS)
- collapse highly-similar introns unless supported by high-quality alignments
- other heuristic criteria

PItron evaluation

- **Method:** (Guigò et al., Genom Biol, 2006)
 - Sensitivity (Sn) = $TP / (TP + FN)$
 - Specificity (Sp) = $TP / (TP + FP)$
- **Data:** 112 genes on 13 ENCODE regions
- **Gold standard:** GENCODE annotations
- For this prelim. work:
gene structure \equiv set of introns

Evaluation Results

- Overall *Sensitivity* = **0.9780**
- Overall *Specificity* = **0.6846**
- Why *Specificity* is low?
 - our method “over-predicts”, or
 - GENCODE annotation could be “biased”
- If we keep only “canonical” introns
 - *Sensitivity* = **0.9764**
 - *Specificity* = **0.9147**

Compared with... ASPicDB

- ASPicDB (Castrignano et al., Bioinf, 2008)
 - accurate isoform-prediction method (Bonizzoni et al., JCB, 2009)
 - 101 (over 112) also in ASPicDB

	Sn	Sp	Time
ASPic	0.9631	0.7070	> 100 h
PItron	0.9727	0.7363	72 min

Conclusions

- **PIntron:**
fast and accurate pipeline for gene-structure prediction based on transcripts
- **Ongoing/future work:**
 - Comparison with other tools
 - Polishing implementation (open-source)
(www.algolab.eu/PIntron)
 - Extensions to RNA-Seq

PItron: a fast method for gene structure prediction

Paola Bonizzoni

Gianluca Della Vedova

Yuri Pirola

Raffaella Rizzi

Thanks!

Università degli Studi di Milano-Bicocca, Italy