FlipFlop: Fast Lasso based Isoform Prediction as a FLOw Problem

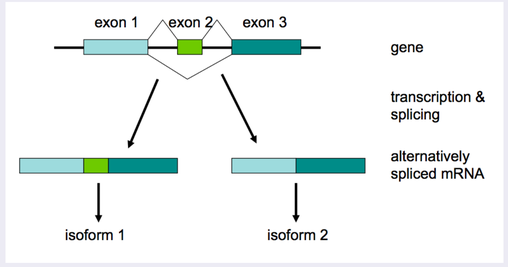
Elsa Bernard[1,2,3], Laurent Jacob[4], Julien Mairal[5,4], Jean-Philippe Vert[1,2,3]

[1]: Center For Computational Biology, Mines ParisTech, Fontainbleau, France;
[2]: INSERM U900, Paris, France;
[3]: Institut Curie, Paris, France;
[4]: Department of Statistics, UC Berkeley, USA;
[5]: LEAR Project-Team, INRIA Grenoble - Rhône Alpes, France

ABSTRACT: Several state-of-the-art methods for isoform identification and quantification are based on sparse probabilistic models, such as Lasso regression. However, explicitly listing the — possibly exponentially — large set of candidate transcripts is intractable for genes with many exons. For this reason, existing approaches using sparse models are either restricted to genes with few exons, or only run the regression algorithm on a small set of pre-selected isoforms. We introduce a new technique called FlipFlop which can efficiently tackle the sparse estimation problem on the full set of candidate isoforms by using network flow optimization. Our technique removes the need of a preselection step, leading to better isoform identification while keeping a low computational cost. Source code is freely available as an R package at http://cbio.mines-paristech.fr/flipflop.

Background

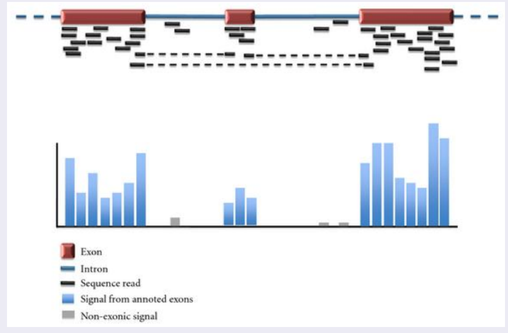
Alternative Splicing



Salzman et al., 2011

- During transcription of eukaryotic genes, exons and introns are alternatively spliced, producing different isoforms.

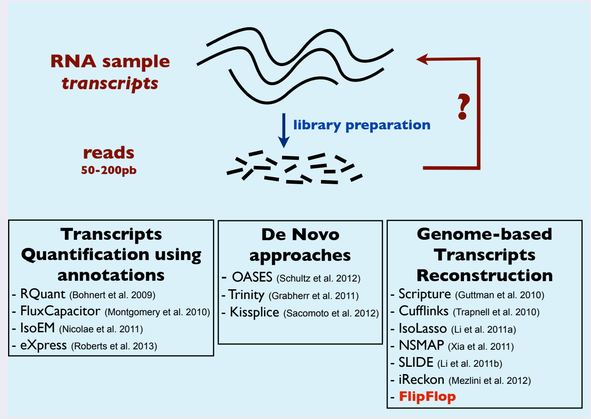
RNA-Seq data



Costa et al., 2011

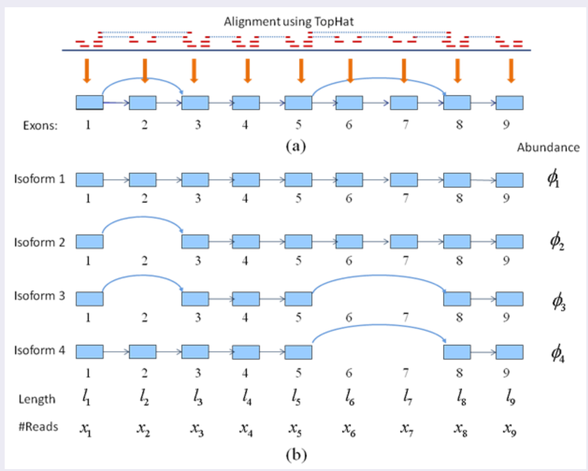
- RNA-Seq measures abundance of each exon and exon-exon junction of a gene.

From RNA-Seq to Isoforms



Regularization approachs

Isoform Deconvolution



Xia et al., 2011

Notations

- n exons
- K candidate isoforms (up to $2^n - 1$)
- Binary design:

$$\mathbf{U} = \begin{pmatrix} & \text{exon}_1 & \dots & \text{exon}_n & \text{junction}_{1,2} & \dots & \text{junction}_{p,n} & \\ & 1 & \dots & 0 & 1 & \dots & 1 & \text{isoform}_1 \\ & & \dots & & & \dots & & \vdots \\ & 1 & \dots & 1 & 0 & \dots & 1 & \text{isoform}_K \end{pmatrix}$$

- $\phi \in \mathbb{R}_+^K$ vector of abundance of isoforms (unknown)
- $\mathbf{U}^T\phi \in \mathbb{R}_+^n$ vector of abundance of exons/junctions (data)

GOAL: estimate isoform abundance ϕ

Sparse Regression via Lasso

Estimate $\mathbf{\Phi}$ sparse by solving:

$$\min_{\phi \in \mathbb{R}_+^K} \mathbf{R}(\mathbf{U}^\top \phi) + \lambda \|\phi\|_1 \ ,$$

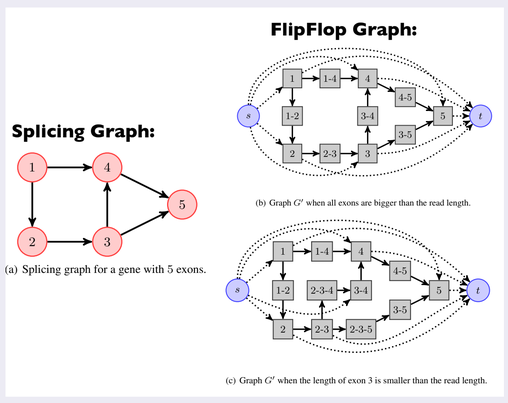
with $\mathbf{R}$ a convex loss function.

- *rQuant (Bohnert et al., 2010)* [1]
- *IsoLasso (Li et al., 2011)* [2]
- *NSMAP (Xia et al., 2011)* [3]
- *SLIDE (Li et al., 2011)* [4]

Computationally challenging to enumerate all candidate isoforms for genes with many exons

Method

Path Selection problem



$G' = (V', E')$
$\mathcal{P}' = \{\text{all paths in } G'\}$

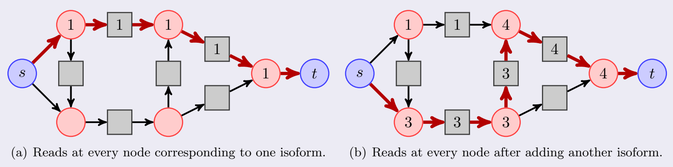
An isoform is a path from source s to sink t

Network Flow Formulation

Isoform detection in sparse regression is equivalent to a **convex cost flow problem** which can be solved in **polynomial time** with the number of exons

Ideas:

- Combinaisons of isoforms are flows



A *flow* f is a non-negative function on arcs on $[f_{uv}]_{(u,v)\in E'}$ that satisfies conservation constraints.

- $f_{uv} = \sum_{p\in\mathcal{P}'} \phi_p \mathbf{1}_{((u,v)\in p)}$ is a flow, and there exists a linear time decomposition algorithm
- Reformulation as Convex Cost Flow problem

$$(\mathbf{U}^T\phi)_v = \sum_{u\in V'} f_{uv} \quad \text{and} \quad \|\phi\|_1 = f_t \ .$$

Then sparse isoform detection is equivalent to

$$\min_{f \text{ flow}} \tilde{R}(f) + \lambda f_t$$

There are efficient algorithms for convex cost flow problem in polynomial time [4,5].

- Related Work: *Traph (Tomescu et al., 2013)* [5] (no sparsity)

Results

Simulations

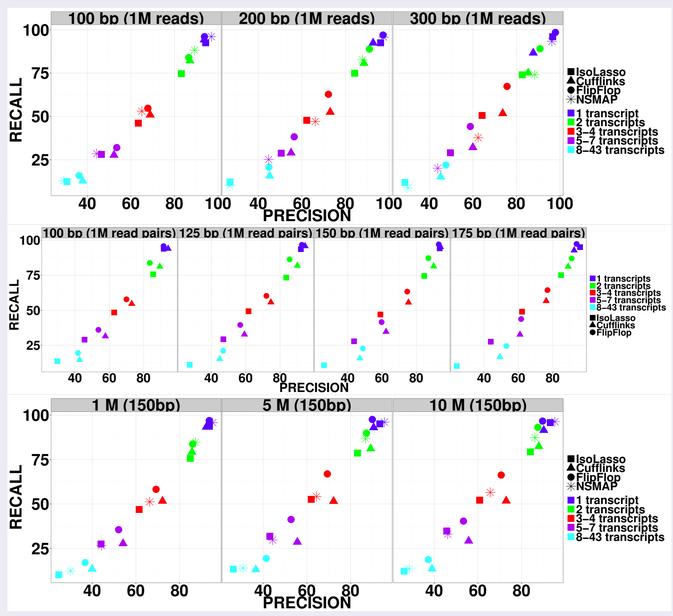
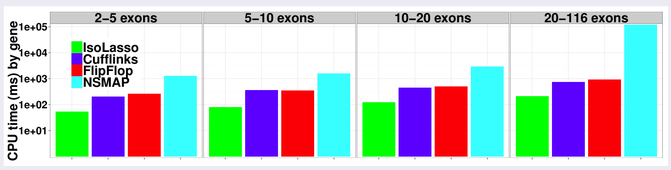
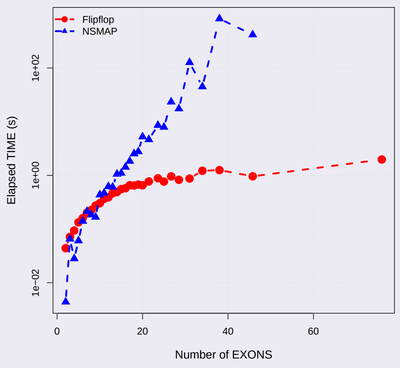


Figure : Precision and Recall and human simulated RNA-Seq single-end and paired-end reads with different read lengths and coverages.

FlipFlop performances increase with read length and coverage

Speed Comparison





FlipFlop: a few seconds regardless the number of exons!

Real Data

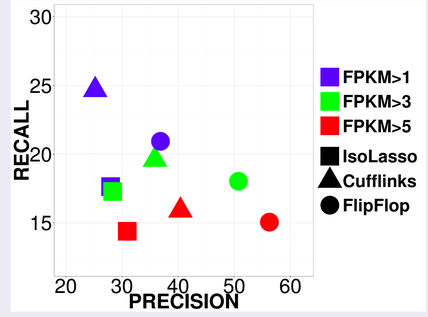


Figure : Precision and Recall on 50 million 75pb paired-end reads of human stem cells

Summary

- Transcript selection over all possible candidates is hard
- We show the problem is equivalent to a simpler one
- The full problem can be solved in polynomial time

References

1 *R. Bohnert et al.*, Nucleic Acids Res, W348âW351. 2009.
2 *W. Li et al.*, J Comput Biol, 18:1693–1707, 2011.
3 *Z. Xia et al.*, BMC Bioinformatics, 12:162, 2011.
4 *J. J. Li et al.*, P Natl Acad Sci USA, 108(50):19867–19872, 2011.
5 *A. Tomescu et al.*, BMC Bioinformatics, 14, S15, 2013.
6 *R. K. Ahuja et al.*, Network Flows, Prentice Hall, 1993.
7 *J. Mairal and B. Yu*, preprint arXiv:1204.4539v1, 2012.
8 *C. Trapnell et al.*, Bioinformatics, 25(9):1105–1111, 2009.