Flip-Flop: Fast lasso-based isoform prediction from RNA-seq data

Jean-Philippe Vert
(joint work with Elsa Bernard, Laurent Jacob, Julien Mairal)

Kyoto University, Bioinformatics Center, July 12th, 2013
(old) Central dogma

DNA

transcription

RNA

translation

Protein

mRNA Transcription

Mature mRNA

Transport to cytoplasm for protein synthesis (translation)
Alternative splicing: 1 gene = many proteins

In human, 28k genes give 120k known transcripts (Pal et al., 2012)
Importance of alternative splicing

(Pal et al., 2012)
Opportunities for drug developments...

(Pal et al., 2012)
The isoform identification and quantification problem

Given a biological sample (e.g., cancer tissue), can we:

1. identify the isoform(s) of each gene present in the sample?
2. quantify their abundance?
RNA-seq measures mRNA abundance by sequencing short fragments

(Wang et al., 2009)
RNA-seq and alternative splicing

(Costa et al., 2011)
From RNA-seq to isoforms

RNA sample
transcripts

reads
50-200pb

library preparation

Transcripts Quantification using annotations
- RQuant (Bohnert et al. 2009)
- FluxCapacitor (Montgomery et al. 2010)
- IsoEM (Nicolae et al. 2011)
- eXpress (Roberts et al. 2013)

De Novo approaches
- OASES (Schultz et al. 2012)
- Trinity (Grabherr et al. 2011)
- Kissplice (Sacomoto et al. 2012)

Genome-based Transcripts Reconstruction
- Scripture (Guttman et al. 2010)
- Cufflinks (Trapnell et al. 2010)
- IsoLasso (Li et al. 2011a)
- NSMAP (Xia et al. 2011)
- SLIDE (Li et al. 2011b)
- iReckon (Mezlini et al. 2012)
- FlipFlop
De Novo methods

- OASES (Schultz et al. 2012)
- Trinity (Grabherr et al. 2011)
- Kissplice (Sacomoto et al. 2012)
Genome-based methods

- Scripture (Guttman et al. 2010)
- Cufflinks (Trapnell et al. 2010)
- IsoLasso (Li et al. 2011a)
- NSMAP (Xia et al. 2011)
- SLIDE (Li et al. 2011b)
- iReckon (Mezlini et al. 2012)
- FlipFlop
Genome-based

Input:
spliced alignment of reads
against reference genome

Job:
reconstruct transcripts
multi-assembly problem
Isoforms are Paths in a Graph

(a) Splicing graph for a gene with 5 exons.
(b) Graph $G'$ with junctions, source $s$ and sink $t$ nodes.

- **Cufflinks** → *overlap graph*
- **Scripture, IsoLasso** → *connectivity graph*
- **SLIDE, NSMAP, iReckon** → *splicing graph*
- **FlipFlop** → ‘customized’ *splicing graph*
How to select a small number of paths?

\[ n \text{ exons} \rightarrow \sim 2^n \text{ paths/candidate isoforms} \]

\sim 1000 \text{ candidates paths for 10 exons and } \sim 1000000 \text{ for 20 exons}

Minimum Path Cover
- Cufflinks, IsoLasso.

Regularization approaches
- NSMAP, SLIDE, iReckon, FlipFlop.
Cufflink strategy

A two-step approach:

1. Find a set of *minimal paths* in the graph (independently from the read abundance value) to identify a good set of isoforms

2. Estimate isoform abundance using read abundance

(Trapnell et al., 2010)
Regularization approach

- Suppose there are \( c \) candidate isoform (\( c \) large)
- Let \( \phi \) the unknown \( c \)-dimensional vector of abundance
- Let \( L(\phi) \) quantify whether \( \phi \) explains well the observed read counts (e.g., minus log-likelihood)
- Regularization approach solve a problem:

\[
\min_{\phi} L(\phi) \quad \text{such that} \quad \phi \text{ is sparse.}
\]
Pros and cons of both paradigms

**Separate** identification and abundance estimation

- Find a small set of transcripts which covers all reads, *then* estimate $\phi$.
- Cufflinks, Isolasso.

**Simultaneous** identification and abundance estimation

- Estimate sparse $\phi$ over set of all possible transcripts.
- NSMAP, SLIDE, iReckon, Flip-Flop
Pros and cons of both paradigms

**Separate** identification and abundance estimation
- Find a small set of transcripts which covers all reads, *then* estimate $\phi$.
- Cufflinks, Isolasso.
- Pros: fast.
- Cons: loss of power.

**Simultaneous** identification and abundance estimation
- Estimate sparse $\phi$ over set of all possible transcripts.
- NSMAP, SLIDE, iReckon, Flip-Flop
- Pros: More powerful.
- Cons: Exponential complexity (up to $2^n - 1$ candidates).
Simultaneous identification and abundance estimation: more power

(Li et al., 2011)
The isoform deconvolution problem

(Xia et al., 2011)
More formally

e exons, \( n \) "bins" (exons+junctions)
c candidate isoforms (up to \( 2^e - 1 \))
\( \phi \in \mathbb{R}_+^c \) the vector of abundance of isoforms (unknown!)
\( U \) binary matrix:

\[
\begin{bmatrix}
\text{exon}_1 & \cdots & \text{exon}_e & \text{junction}_{1,2} & \cdots & \text{junction}_{e_1,e}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\text{isoform}_1 & 1 & \cdots & 1 & 1 & \cdots & 1 \\
\text{isoform}_2 & 1 & \cdots & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\text{isoform}_c & 0 & \cdots & 1 & 0 & \cdots & 0
\end{bmatrix}
\]

\( U^T \phi \) the abundance of each exon/junction.

Goal: estimate \( \phi \) from the observed reads on each exon/junction
The log likelihood of $\phi \in \mathbb{R}^c$ only depends on the abundance of each exon/junction in $U^T \phi \in \mathbb{R}^n$

Example: Gaussian (IsoLasos, SLIDE) or Poisson (NSMAP, FlipFlop) negative log-likelihood

Regularization-based approaches try to solve:

$$\min_{\phi \in \mathbb{R}^c} R(U^T \phi) \quad \text{such that} \quad \phi \text{ is sparse},$$

where $R : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex

This is generally a NP-hard problem, so we use a convex relaxation akin to Lasso regression
The Lasso idea

The $\ell_1$ penalty (Tibshirani, 1996; Chen et al., 1998)

If $R(\beta)$ is convex and "smooth", the solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

is usually sparse.

Geometric interpretation with $p = 2$
Lasso example

Typically solved in $O(n^3)$
Isoform deconvolution with the Lasso

Estimate $\phi$ sparse by solving (IsoLasso, NSMAP, SLIDE):

$$\min_{\phi \in \mathbb{R}^c_+} R(U^T \phi) + \lambda \| \phi \|_1$$

Complexity $O(c^3) = O(2^{3e})$...

Works well BUT computationally challenging to work with all candidate isoforms for large genes!
The isoform deconvolution problem

\[ \min_{\phi \in \mathbb{R}_{+}^c} R(U^T \phi) + \lambda \| \phi \|_1 \]

can be solved in \textit{polynomial time} in the number of exon.

Key ideas

1. \( U^T \phi \) corresponds to a \textit{flow} on the graph
2. Reformulation as a \textit{convex cost flow problem} (Mairal and Yu, 2012)
3. Recover isoforms by flow decomposition algorithm

"Feature selection on an exponential number of features in polynomial time"
A **flow** $f$ is a nonnegative function on arcs that respects conservation constraints (Kirchhoff’s law).
Combinations of isoforms are flows

(a) Reads at every node corresponding to one isoform.
(b) Reads at every node after adding another isoform.

- **Linear combinations of isoforms** ⇒ **Flow value on every nodes**
- **Flow value on every nodes** ⇒ **Paths with given value/abundance**
  - **Flow Decomposition (linear time algorithm)**
From isoforms to flow (key trick!)

(a) Reads at every node corresponding to one isoform.

(b) Reads at every node after adding another isoform.

- $U^\top \phi \in \mathbb{R}^n$ when $\phi \in \mathbb{R}^c$ is the set of flows
- Moreover, $\| \phi \|_1 = f_t$

Therefore,

$$\min_{\phi \in \mathbb{R}^c_+} R(U^\top \phi) + \lambda \| \phi \|_1$$

is equivalent to

$$\min_{f \text{ flow}} R(f) + \lambda f_t$$
\[
\min_{\phi \in \mathbb{R}_+^c} R(U^T \phi) + \lambda \| \phi \|_1
\]

- Cufflink: *a priori* selection of isoforms (minimum graph cover)
- IsoLasso: pre-filtering of candidate isoforms using various heuristics
- NSMAP, SLIDE: limit the maximum number of exons
- FlipFlop: exact optimization without pre-filtering in polynomial time, by solving a convex problem in the space of flows (dimension \( n \)) and recovering path with the flow decomposition algorithm.
Human Simulation: Precision/Recall

hg19, 1137 genes on chr1, 1 million 75 bp single-end reads by transcript levels.

Performance increases with read length

PRECISION
RECALL
IsoLasso
Cufflinks
FlipFlop
NSMAP
1 transcripts
2 transcripts
3–4 transcripts
5–7 transcripts
8–43 transcripts
Performance increases with coverage

- IsoLasso
- Cufflinks
- FlipFlop
- NNSMAP

1 transcripts
2 transcripts
3−4 transcripts
5−7 transcripts
8−43 transcripts
Extension to paired-end reads OK.

- 100 bp (400 bp fragments, 1M reads)
- 125 bp (400 bp fragments, 1M reads)
- 150 bp (400 bp fragments, 1M reads)
- 175 bp (400 bp fragments, 1M reads)

Graph showing recall and precision for different transcript counts and tools:
- IsoLasso
- Cufflinks
- FlipFlop

Legend:
- 1 transcript
- 2 transcripts
- 3–4 transcripts
- 5–7 transcripts
- 8–43 transcripts
- IsoLasso
- Cufflinks
- FlipFlop

JP Vert (ParisTech)

Kyodai 33 / 36
Speed trial

![Graph showing elapsed time (s) vs. number of exons for Flipflop and NSMAP.]

- **Number of EXONS**: 2−5 exons, 5−10 exons, 10−20 exons, 20−116 exons
- **CPU time (ms) by gene**: IsoLasso, Cufflinks, FlipFlop, NSMAP, SLIDE

Table summarizing CPU time for different gene sets:

<table>
<thead>
<tr>
<th>Number of EXONS</th>
<th>IsoLasso</th>
<th>Cufflinks</th>
<th>FlipFlop</th>
<th>NSMAP</th>
<th>SLIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2−5 exons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5−10 exons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10−20 exons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20−116 exons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

http://cbio.mines-paristech.fr/flipflop

Summary

- Transcript selection over all possible candidates is hard.
- We show the problem is equivalent to a simpler one.
- With our approach, the full problem is solved as quickly as the more heuristic one (Cufflinks approach).

Future work

- Some loose ends: GC content, decomposition, post-processing...
- Ongoing: abundance estimation comparison.
- Adapt framework to paired-end reads,
- Applications: differential expression, classification, clustering.
Elsa Bernard (Mines ParisTech / Institut Curie), Laurent Jacob (UC Berkeley / CNRS), Julien Mairal (INRIA)