A Combinatorial Approach to the Inference of Isoforms from Short Sequence Reads

Tao Jiang

Department of Computer Science and Engineering
University of California, Riverside

Joint work with Jianxing Feng and Wei Li
(Tsinghua/Tongji)  (UCR)
1. Background and Existing Work
2. Quadratic Program
3. Valid Isoforms
4. IsoInfer: The Basic Algorithm
5. An Improvement by Lasso Regression
6. Experimental Results and Comparison to Cufflinks and Scripture
7. Concluding remarks
A gene may lead to multiple mRNAs!
Alternative Splicing

5' Pre-mRNA

Isoform 1

Intron retention

Isoform 2

3'
Alternative Splicing

5’ Pre-mRNA

Isoform 1

Isoform 2

Isoform 3

Intron retention

Exon skipping
Alternative Splicing

5’ 3’ Pre-mRNA

Isoform 1

Intron retention

Isoform 2

Exon skipping

Isoform 3

Alternate 3’

Isoform 4

Alternate 5’

Isoform 5
Alternative Splicing

5' Pre-mRNA

 Isoform 1

 Isoform 2

 Isoform 3

 Isoform 4

 Isoform 5

 Isoform 6

Intron retention

Exon skipping

Alternate 3’

Alternate 5’

Mixed
Alternative Splicing

5’ Pre-mRNA

Isoform 1

Intron retention

Isoform 2

Exon skipping

Isoform 3

Alternate 3’

Isoform 4

Alternate 5’

Isoform 5

Mixed

Isoform 6

Widely spread in human genome
More than 92% multi-exon genes
An example: KLF6 gene in human chromosome 10, a tumor suppressor gene, includes 4 alternative splicing variants.
An example: KLF6 gene in human chromosome 10, a tumor suppressor gene, includes 4 alternative splicing variants.

Three of the four variants, if expressed, inactivate the tumor suppressor gene and are associated with increased prostate cancer risk (DiFeo et al. *Cancer Res.* 2005).
To detect mRNA isoforms expressed in a cell and estimate their abundance levels,

- **Traditional methods:**
  - EST (Expressed Sequence Tag)
  - RACE (Rapid Amplification of cDNA Ends)
  - SAGE (Serial Analysis of Gene Expression)
  - CAGE (Cap Analysis Gene Expression)
  - . . .
  - cost ineffective

- **New methods:**
  - RACEArray (Djebali et al, Nat Methods, 2008)
  - PCR+’deep-well’ pooling +sequencing (Salehi-Ashtiani et al, Nat Methods, 2008)
  - . . .
  - large scale? unclear
Assumption: uniformly distributed along expressed isoforms
The Problem

- Single-end short reads (RNA-Seq)
- Paired-end short reads (RNA-Seq)
- Exon-intron boundaries (known or RNA-Seq)
  
  *Trapnell et al, Bioinformatics, 2009*
  
  *Au et al, NAR, 2010*  
  
  **TopHat**
  
  **SpliceMap**
- TSS/PAS pairs (known or GIS-PET)
  
  *Ng et al, Nat Methods, 2005*

  *(Fullwood et al, Genome Res, 2009)*
- PAS profiling (3P-Seq)

  *(Jan et al, Nature 2010)*

So the reference genome is assumed!
The Problem

- Single-end short reads (RNA-Seq)
- Paired-end short reads (RNA-Seq)
- Exon-intron boundaries (known or RNA-Seq)  
  \textit{(Trapnell et al, Bioinformatics, 2009)} \textbf{TopHat}  
  \textit{(Au et al, NAR, 2010)} \textbf{SpliceMap}
- TSS/PAS pairs (known or GIS-PET)  
  \textit{(Ng et al, Nat Methods, 2005)}  
  \textit{(Fullwood et al, Genome Res, 2009)}
- PAS profiling (3P-Seq)  
  \textit{(Jan et al, Nature 2010)}

So the reference genome is assumed!

Problem

\textit{Given the data, find all the isoforms and expression levels of each isoform.}
Existing Work

Expression level estimation

- **RSAT.** (Jiang & Wong, Bioinformatics. 2009)
- **RSEM.** (Li et al, Bioinformatics. 2009)
- ...
Existing Work

Expression level estimation

- RSAT. (Jiang & Wong, Bioinformatics. 2009)
- RSEM. (Li et al, Bioinformatics. 2009)
- ...

Isoform reconstruction

- Scripture. (Guttman et al, Nat Biotechnology. 2010.5)
  - Uses weighting to filter out lowly expressed isoforms. Focuses on full-length transcripts.
- Cufflinks. (Trapnell et al, Nat Biotechnology. 2010.5)
  - Uses a minimal path cover algorithm to find a parsimonious solution to explain the read data.
Outline

1. Background and Existing Work
2. Quadratic Program
3. Valid Isoforms
4. IsoInfer: The Basic Algorithm
5. An Improvement by Lasso Regression
6. Experimental Results and Comparison to Cufflinks and Scripture
7. Concluding remarks
Expressed Segments

$r_i$: # reads mapped to expressed segments $s_i$.

Junction sequences
Quadratic Program

For a gene with isoforms $F$, expressed segments $S$ and junctions $J$: 
For a gene with isoforms $F$, expressed segments $S$ and junctions $J$:

$r_i \sim B(M, p_i), \quad p_i = C \times \sum_{s_i \in f, f \in F} x_f l_i$
For a gene with isoforms $F$, expressed segments $S$ and junctions $J$:

- $r_i \sim B(M, p_i)$, $p_i = C \times \sum_{s_i \in f, f \in F} x_f l_i$

- Approximate it with $N(\mu_i, \sigma_i^2)$
  
  $\mu_i = M p_i$, $\sigma_i^2 = M p_i (1 - p_i) \approx M p_i = \mu_i$
For a gene with isoforms $F$, expressed segments $S$ and junctions $J$:

- \( r_i \sim B(M, p_i) \), \( p_i = C \times \sum_{s_i \in f, f \in F} x_f l_i \)

- Approximate it with \( N(\mu_i, \sigma_i^2) \)
  \[
  \mu_i = M p_i, \sigma_i^2 = M p_i (1 - p_i) \approx M p_i = \mu_i
  \]

- \( \epsilon_i = r_i - \mu_i, \frac{\epsilon_i}{\sigma_i} \) obeys \( N(0, 1) \) approximately.
For a gene with isoforms $F$, expressed segments $S$ and junctions $J$:

- $r_i \sim B(M, p_i), \ p_i = C \times \sum_{s_i \in f, f \in F} x_f l_i$

- Approximate it with $N(\mu_i, \sigma_i^2)$

\[
\mu_i = Mp_i, \sigma_i^2 = Mp_i(1 - p_i) \approx Mp_i = \mu_i
\]

- $\epsilon_i = r_i - \mu_i, \ \frac{\epsilon_i}{\sigma_i}$ obeys $N(0, 1)$ approximately.

\[
\min \quad z = \sum_{s_i \in S \cup J} \left(\frac{\epsilon_i}{\sigma_i}\right)^2
\]

s.t. \[
\sum_{s_i \in f, f \in F} x_f l_i + \epsilon_i = d_i, \quad s_i \in S \cup J
\]

\[
x_f \geq 0, \quad f \in F
\]
min \quad z = \sum_{s_i \in S \cup J} \left( \frac{\epsilon_i}{\sigma_i} \right)^2 \\
\text{s.t.} \quad \sum_{s_i \in f, f \in F} x_f l_i + \epsilon_i = d_i, \quad s_i \in S \cup J \\
\quad \quad x_f \geq 0, \quad f \in F
min \quad z = \sum_{s_i \in S \cup J} \left( \frac{\epsilon_i}{\sigma_i} \right)^2 \\
\text{s.t.} \quad \sum_{s_i \in f, f \in F} x_f l_i + \epsilon_i = d_i, \quad s_i \in S \cup J \\
\quad x_f \geq 0, \quad f \in F

- Convex program
Quadratic Program

\[ \min z = \sum_{s_i \in S \cup J} \left( \frac{\epsilon_i}{\sigma_i} \right)^2 \]

s.t. \[ \sum_{s_i \in f, f \in F} x_f l_i + \epsilon_i = d_i, \quad s_i \in S \cup J \]
\[ x_f \geq 0, \quad f \in F \]

- Convex program
- If \( \sigma_i \) is known \( x_f \) corresponds to the maximum likelihood estimation
Quadratic Program

\[
\begin{align*}
\text{min} & \quad z = \sum_{s_i \in S \cup J} \left( \frac{\epsilon_i}{\sigma_i} \right)^2 \\
\text{s.t.} & \quad \sum_{s_i \in f, f \in F} x_f l_i + \epsilon_i = d_i, \quad s_i \in S \cup J \\
& \quad x_f \geq 0, \quad f \in F
\end{align*}
\]

- Convex program
- If \( \sigma_i \) is known, \( x_f \) corresponds to the maximum likelihood estimation
- Replace \( \sigma_i \) with \( \sqrt{d_i} \)
Quadratic Program

\[ \min \quad z = \sum_{s_i \in S \cup J} \left( \frac{\epsilon_i}{\sigma_i} \right)^2 \]
\[ \text{s.t.} \quad \sum_{s_i \in f, f \in F} x_f l_i + \epsilon_i = d_i, \quad s_i \in S \cup J \]
\[ \quad x_f \geq 0, \quad f \in F \]

- Convex program
- If \( \sigma_i \) is known, \( x_f \) corresponds to the maximum likelihood estimation
- Replace \( \sigma_i \) with \( \sqrt{d_i} \)
- \( z \sim \chi^2(|S| + |J|) \)
Outline

1. Background and Existing Work
2. Quadratic Program
3. Valid Isoforms
4. IsoInfer: The Basic Algorithm
5. An Improvement by Lasso Regression
6. Experimental Results and Comparison to Cufflinks and Scripture
7. Concluding remarks
Theorem

Under the uniform sampling assumption, the probability that an isoform consisting of \( t \) exons, with expression level \( \alpha \) RPKM, has all its junctions covered by single-end reads is at least

\[
P = \left(1 - e^{-\alpha LM/10^9}\right)^{t-1}
\]

RPKM : Reads Per Kilobase of exon model per Million mapped reads.
Theorem

Under the uniform sampling assumption, the probability that an isoform consisting of \( t \) exons, with expression level \( \alpha \) RPKM, has all its junctions covered by single-end reads is at least

\[
P = \left(1 - e^{-\alpha LM/10^9}\right)^{t-1}
\]

RPKM : Reads Per Kilobase of exon model per Million mapped reads.

Example

\( M = 40,000,000, L = 30, \alpha = 6 \) RPKM.

If \( t = 10 \), then \( P = 99.3\% \)

If \( t = 100 \), then \( P = 92.8\% \)
Theorem

The probability that there are no paired-end reads with start positions in the first interval and end positions in the third interval is upper bounded by

\[ P_{M,h,\alpha}(w_1, w_2, w_3) = (1 - P_0)^M \approx e^{-MP_0} \]

where \( P_0 = 10^{-9} \alpha \sum_{0 \leq i < w_1} \int_{l(i)}^{u(i)} h(x) dx, \ l(i) = w_1 - i + w_2, \text{ and } u(i) = w_1 - i + w_2 + w_3. \]
Informative pair

Segment pair \((s_i, s_j), i < j\), on isoform \(f\) is an informative pair if
\[
P_{M,h,\alpha}(l_i + L - 1, g_{i,j}, l_j + L - 1) < 0.05, \text{ where } g_{i,j} = \sum_{i<k<j} l_k f[k]
\]
Informative pair

Segment pair \((s_i, s_j), i < j\), on isoform \(f\) is an informative pair if
\[
P_{M,h,\alpha}(l_i + L - 1, g_{i,j}, l_j + L - 1) < 0.05, \quad \text{where } g_{i,j} = \sum_{i<k<j} l_k f[k]
\]

Valid Isoform

An isoform \(f\) is valid if:

1. All the exon-exon junctions are covered by short reads.
2. All the informative pairs are supported by short reads. \(\alpha\) controls this condition.
3. The start-end expressed segment pair appears in the given start-end pair set. (i.e., the TSS/PAS pairs)
1. Background and Existing Work

2. Quadratic Program

3. Valid Isoforms

4. IsoInfer: The Basic Algorithm

5. An Improvement by Lasso Regression

6. Experimental Results and Comparison to Cufflinks and Scripture

7. Concluding remarks
1. Enumerate all valid isoforms.
Isoform Inference on Each Cluster

1. Enumerate all valid isoforms.
2. Generate subinstances with each one of them focusing on a subset of expressed segments.
Enumerate all valid isoforms.

Generate subinstances with each one of them focusing on a subset of expressed segments.

On each subinstance, enumerate all the subsets of valid isoforms, use $QP$ to find the best one.
1. Enumerate all valid isoforms.
2. Generate subinstances with each one of them focusing on a subset of expressed segments.
3. On each subinstance, enumerate all the subsets of valid isoforms, use $QP$ to find the best one.
4. Combine the results of all the subinstances using set cover.
1. Enumerate all valid isoforms.

2. Generate subinstances with each one of them focusing on a subset of expressed segments.

3. On each subinstance, enumerate all the subsets of valid isoforms, use $QP$ to find the best one.

4. Combine the results of all the subinstances using set cover.
Outline

1 Background and Existing Work
2 Quadratic Program
3 Valid Isoforms
4 IsoInfer: The Basic Algorithm
5 An Improvement by Lasso Regression
6 Experimental Results and Comparison to Cufflinks and Scripture
7 Concluding remarks
LASSO: Least Absolute Shrinkage and Selection Operator


\[
\begin{align*}
\min f(X) &= \sum_i \left( \frac{d_i}{l_i} - \sum_{s_i \in f, f \in F} x_f \right)^2 + \lambda \sum_{f \in F} x_f \\
\text{s.t.} \quad & x_f \geq 0, f \in F
\end{align*}
\]

which is equivalent to the following *constrained form*:

\[
\begin{align*}
\min f(X) &= \sum_i \left( \frac{d_i}{l_i} - \sum_{s_i \in f, f \in F} x_f \right)^2 \\
\text{s.t.} \quad & \sum_{f \in F} x_f \leq \gamma \\
& x_f \geq 0, f \in F
\end{align*}
\]
\[
\min f(X) = \sum_i \left( \frac{d_i}{l_i} - \sum_{s_i \in f, f \in F} x_f \right)^2 \\
\text{s.t.} \quad \sum_{f \in F} x_f \leq \gamma \\
\quad x_f \geq 0, f \in F \\
\quad \sum_{f \in F} x_f \mathbb{1}_{s_i \in f} \geq p, \text{ if } s_i \text{ has mapped reads}
\]
Outline

1 Background and Existing Work
2 Quadratic Program
3 Valid Isoforms
4 IsoInfer: The Basic Algorithm
5 An Improvement by Lasso Regression
6 Experimental Results and Comparison to Cufflinks and Scripture
7 Concluding remarks
Figure: The distribution of simulated isoform expression levels (left), and the accuracy of expression level estimation for different methods (right), using 80M 75 × 2 paired-end reads. Mouse transcriptome is used.
Figure: Sensitivity (left) and precision (right) on single-end reads
Figure: Sensitivity (left) and precision (right) on paired-end reads
Isoform Reconstruction - Running time

- IsoLasso
- Isolnfer without TSS/PAS
- Cufflinks
- Scripture

Running time (seconds)

Number of paired-end reads:
- 20M
- 40M
- 60M
- 80M
- 100M
Figure: The number of known isoforms of mouse (A) and human (B), and the number of predicted isoforms of mouse (C) and human (D), assembled by IsoLasso, Cufflinks and Scripture.
Figure: An alternative 5' start isoform of gene Tmem70 in mouse C2C12 myoblast RNA-Seq data
Mouse transcriptome, junctions predicted by TopHat.
1. Background and Existing Work
2. Quadratic Program
3. Valid Isoforms
4. IsoInfer: The Basic Algorithm
5. An Improvement by Lasso Regression
6. Experimental Results and Comparison to Cufflinks and Scripture
7. Concluding remarks
Concluding remarks

- We presented two combinatorial approaches to infer mRNA isoforms and estimate their expression levels from RNA-Seq data using convex quadratic programming, set cover and Lasso regression. The programs can be downloaded from:

- The programs compete well against Cufflinks and Scripture in terms of accuracy and speed.

- The tools can be further improved by addressing practical issues including sequencing errors, nonuniformly distributed reads, multireads, etc.
