Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2015
Colin Dewey
cdewey@biostat.wisc.edu
Overview

- Part I - Alternative splicing and the challenges it poses
- Part II - A solution: *Probabilistic Splice Graphs (PSGs)*
- Part III - Evaluating PSG methodology
Alternative splicing

pre-mRNA

splicing

alternatively spliced mRNAs

translation

protein isoforms
Classes of alternative splicing events

- Exon skipping
- Mutually exclusive exons
- Alternative 5' donor sites
- Alternative 3' acceptor sites
- Intron retention
Complication 1: De novo transcriptome assembly

- RNA-Seq reads/fragments are relatively short

- Often insufficient to reconstruct full-length isoforms in the presence of alternative splicing

- Transcriptome assemblies perhaps best left in “graph” form
  
  - De Bruijn graph
  
  - String graphs

Graph constructed by the “Butterfly” module of Trinity (Grabherr et al. 2011)
Complication 2: Non-identifiability of full-length isoform models

Lacroix et al. 2008; Hiller et al. 2009
Complication 3: Combinatorial explosion of distinct isoforms

- Combinatorial explosion of the number of possible isoforms for each gene

- Insufficient data to accurately estimate abundances of thousands of isoforms

Drosophila *Dscam*: more than 38,000 possible isoforms
(Schmucker et al., 2000)
Overview

• Part I - Alternative splicing and the challenges it poses

• Part II - A solution: *Probabilistic Splice Graphs (PSGs)*

• Part III - Evaluating PSG methodology
Splice Graphs

- Heber et al. 2002

- Compact **data structure** for representing the possible isoforms of a gene
Splice Graphs with EST and RNA-Seq data

• Xing et al. 2006

  • EM algorithm for estimating abundances of all possible isoforms given splice graph and EST data

• Montgomery et al. 2010, Singh et al. 2011

  • Graph flow-based methods for quantification/differential splicing given RNA-Seq data

• Rogers et al. 2012

  • SpliceGrapher: construct splice graph structure given RNA-Seq data
**Probabilistic Splice Graphs**

- Jenkins et al. 2006

- Compact **probabilistic model** representing isoform frequencies in terms of frequencies of individual splice events

- Originally used by Jenkins et al. for EST analysis

![Diagram of probabilistic splice graph with frequencies 0.8, 0.4, 0.2, 0.6, 0.32, 0.48, 0.08, 0.12]
Probabilistic Splice Graph Complexity

- Known isoforms
- “Line graph”
- “Exon graph”
- “Higher-order exon graph”
- “Unfactorized graph”
Advantages of PSGs

• Compact description of the possible isoforms of a gene

  • Models the frequencies of potentially exponentially many isoforms with a polynomial number of parameters

  • Models dependence or independence of splice events

• The parameters of a PSG are more often identifiable than a model that has a parameter for every possible isoform

• Splice graphs are naturally-produced structures from transcriptome assemblers
PSGs are alternative “parsimonious” models

- Other methods find smallest set of isoform structures that explain the data
  - Cufflinks (Trapnell et al., 2010)
  - IsoLasso (Li et al., 2011)
  - NSMAP (Xia et al., 2011)
  - SLIDE (Li et al., 2011)
- PSG models are another form of parsimonious model
  - Minimize the number of splice event parameters
  - Assumption of independence between splice events
Our contributions

• Application of PSGs to RNA-Seq data
  
  • Combined model of PSG with RNA-Seq generative model
  
  • Efficient PSG parameter estimation with EM and dynamic programming
  
  • Identifiability proofs for PSG with RNA-Seq data
  
  • Differential processing (splicing) tests

The PSG parameter inference task

• Given: RNA-Seq reads and a PSG structure

• Do: Estimate the (ML or MAP) parameters for the model
A model of RNA-Seq from PSGs

- RSEM model extended to probabilistic splice graphs
  - fragment length distribution, quality scores, read mapping ambiguity
- Dynamic programming algorithms → polynomial time inference for genes with an exponential number of isoforms

Probability of including vertex $j$ given that vertex $i$ was in transcript

$$ f(i,j) = \sum_{a:a_i = t, a_j = j} w(s) = \begin{cases} 1 & i = j \\ \frac{1}{\sum_k \alpha_{kj} f(i,k)} & i \neq j \end{cases} $$

Expected prefix length

$$ d_p(i) = \ell_i + \frac{1}{f(0,i)} \sum_j f(0,j) \alpha_{ij} d_p(j) $$

Expected suffix length

$$ d_q(i) = \ell_i + \sum_j \alpha_{ij} d_q(j) $$
EM for PSG parameter estimation

- E-step: compute the expectation of the number of times edge \((i,j)\) is used

\[
E[Z_{n(i,j)}] = \frac{\sum_{(b,s) \in \pi(r)} g(s, i, j)}{\sum_{(b,s) \in \pi(r)} g(s)}
\]

\[
g(s) = f(0, s_1)w(s)
\]

\[
g(s, i, j) = \begin{cases} 
  f(0, s_1)w(s) & (i, j) \in s \\
  f(0, i)\alpha_{i1} f(j, s_1)w(s) & \text{if } \exists \text{ path from } v_j \text{ to } s_1 \\
  f(0, s_1)w(s)f(s_{|s|}, i)\alpha_{i1} & \text{if } \exists \text{ path from } s_{|s|} \text{ to } v_i \\
  0 & \text{otherwise}
\end{cases}
\]

- M-step: maximize the completely-observed likelihood given the edge counts

\[
\alpha_{i,j} = \frac{\frac{z_{i,j}}{|z_{x,i} + z_{y,j}|}}{\sum_{k} \frac{z_{i,k}}{|z_{x,i} + z_{y,k}|}}
\]
Identifiability of PSGs with RNA-Seq data

- Identifiability: \[ P(D|M, \theta) = P(D|M, \theta'), \forall D \iff \theta = \theta' \]

- Proposition: If for all edges \((u, v)\), there exists a read that is uniquely derived from that edge, or \(v\) has indegree 1 and there exists a read that is uniquely derived from \(v\), then the PSG is identifiable.
The differential processing (DP) task

- **Given:** RNA-Seq reads from two conditions and a PSG structure

- **Do:** Determine if the processing frequencies are different

\[
\alpha_1 = \alpha'_1 \text{ and } \alpha_2 = \alpha'_2 \quad ?
\]

\[
\alpha_1 = \alpha'_1 \text{ or } \alpha_2 = \alpha'_2 \quad ?
\]
Our approach to the differential processing (DP) task

• Simple likelihood ratio tests with PSG model

• Test for null hypothesis that all frequencies are the same

\[
LR = \frac{P(R^1|\hat{\alpha}^1)P(R^2|\hat{\alpha}^2)}{P(R^1 \cup R^2|\hat{\alpha}^{12})}
\]

• Test for null hypothesis that frequencies of edges out of one vertex \((i)\) are the same

\[
LR = \frac{P(R^1|\hat{\alpha}^1_i)P(R^2|\hat{\alpha}^2_i)}{P(R^1, R^2|\hat{\alpha}^1_i, \hat{\alpha}^2_i, \hat{\alpha}^{12}_i)}
\]
Overview

• Part I - The problem

• Part II - A solution: Probabilistic Splice Graphs (PSGs)

• Part III - Evaluating PSG methodology
Efficient inference for highly-spliced genes

- DSCAM running time test
  - 23,976 isoforms
  - 184 read pairs from a modENCODE sample

<table>
<thead>
<tr>
<th>Method</th>
<th>RSEM</th>
<th>Cufflinks</th>
<th>PSG EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running time</td>
<td>Not possible</td>
<td>&gt; 6 hours</td>
<td>&lt; 3 seconds</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(&gt; 90 GB RAM)</td>
<td></td>
</tr>
</tbody>
</table>
A simple method for comparison

• The **Junction-Read** (JR) method

• Keep only reads that align to the splice junctions (edges in the PSG)

• Throws away data, but is very robust to model assumption violations
Convergence with simulated data
Comparisons on real data

• Require notion of “distance” between estimates from different methods

• Our distance measure:
  - per vertex
  - maximum difference between probability estimates on out-edges of vertex (L-∞ norm)

```
\text{method A}

\begin{tikzpicture}
  \node (a) at (0,0) [draw, fill=gray!30] {0.3};
  \node (b) at (1,0) [draw, fill=gray!30] {0.5};
  \node (c) at (0,1) [draw, fill=gray!30] {0.2};
  \draw (a) -- (b);
  \draw (c) -- (a) -- (b);
\end{tikzpicture}

\text{method B}

\begin{tikzpicture}
  \node (a) at (0,0) [draw, fill=gray!30] {0.1};
  \node (b) at (1,0) [draw, fill=gray!30] {0.3};
  \node (c) at (0,1) [draw, fill=gray!30] {0.6};
  \draw (a) -- (b);
  \draw (c) -- (a) -- (b);
\end{tikzpicture}
```

\[\text{distance}_v(A, B) = \max(0.6 - 0.2, 0.5 - 0.3, 0.3 - 0.1) = 0.4\]
How close are the estimates from JR and EM on real data?

Vertices from 88 most abundant (> 5000 reads) alternatively-spliced genes in a modENCODE fly data set
Convergence of estimates on real data
Comparing PSGs of different complexity

- Same set of fly data
- Estimated with three classes of PSG: line, exon, full-length
- Compared estimates to those from JR (gold-standard)
- No statistically-significant difference between exon and full-length graph estimates
Summary of Junction-Read comparison results

• Estimates using PSG models are generally close to those from the simplistic JR-method
  • ⇒PSG model assumptions appear to be reasonable

• PSG estimates converge more quickly as the data set increases in size
  • ⇒Our EM estimation procedure uses information from all reads, not just those that span splice junctions

• Exon-graph estimates as good as those using traditional full-length isoform models
  • ⇒Independence assumptions of exon graphs appear to be reasonable
### Differential processing detection

**DP Accuracy on real data**

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
<th>PSG</th>
<th>FDM</th>
<th>Cuffdiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEU Rep 1</td>
<td>CEU Rep 2</td>
<td>0</td>
<td>0</td>
<td>1187</td>
</tr>
<tr>
<td>CEU Rep 1</td>
<td>Yoruban Rep 1</td>
<td>39</td>
<td>24</td>
<td>269</td>
</tr>
<tr>
<td>CEU Rep 1</td>
<td>Yoruban Rep 2</td>
<td>46</td>
<td>24</td>
<td>282</td>
</tr>
<tr>
<td>CEU Rep 2</td>
<td>Yoruban Rep 1</td>
<td>45</td>
<td>22</td>
<td>253</td>
</tr>
<tr>
<td>CEU Rep 2</td>
<td>Yoruban Rep 2</td>
<td>38</td>
<td>29</td>
<td>260</td>
</tr>
<tr>
<td>Yoruban Rep 1</td>
<td>Yoruban Rep 2</td>
<td>0</td>
<td>0</td>
<td>1253</td>
</tr>
<tr>
<td>CME_W1.CL8+ Rep 1</td>
<td>CME_W1.CL8+ Rep 2</td>
<td>16</td>
<td>32</td>
<td>204</td>
</tr>
<tr>
<td>CME_W1.CL8+ Rep 1</td>
<td>Kc167</td>
<td>365</td>
<td>207</td>
<td>7</td>
</tr>
<tr>
<td>CME_W1.CL8+ Rep 1</td>
<td>ML-DmBG3-c2</td>
<td>232</td>
<td>164</td>
<td>6</td>
</tr>
<tr>
<td>CME_W1.CL8+ Rep 1</td>
<td>S2-DRSC</td>
<td>406</td>
<td>228</td>
<td>12</td>
</tr>
<tr>
<td>CME_W1.CL8+ Rep 2</td>
<td>Kc167</td>
<td>319</td>
<td>211</td>
<td>16</td>
</tr>
<tr>
<td>CME_W1.CL8+ Rep 2</td>
<td>ML-DmBG3-c2</td>
<td>260</td>
<td>126</td>
<td>16</td>
</tr>
<tr>
<td>CME_W1.CL8+ Rep 2</td>
<td>S2-DRSC</td>
<td>353</td>
<td>220</td>
<td>17</td>
</tr>
<tr>
<td>Kc167</td>
<td>ML-DmBG3-c2</td>
<td>384</td>
<td>321</td>
<td>12</td>
</tr>
<tr>
<td>Kc167</td>
<td>S2-DRSC</td>
<td>419</td>
<td>209</td>
<td>12</td>
</tr>
<tr>
<td>ML-DmBG3-c2</td>
<td>S2-DRSC</td>
<td>431</td>
<td>287</td>
<td>4</td>
</tr>
<tr>
<td>HUVEC Rep 1</td>
<td>HUVEC Rep 2</td>
<td>35</td>
<td>43</td>
<td>440</td>
</tr>
<tr>
<td>HUVEC Rep 1</td>
<td>K562 Rep 1</td>
<td>376</td>
<td>344</td>
<td>8</td>
</tr>
<tr>
<td>HUVEC Rep 1</td>
<td>K562 Rep 2</td>
<td>379</td>
<td>302</td>
<td>12</td>
</tr>
<tr>
<td>HUVEC Rep 2</td>
<td>K562 Rep 1</td>
<td>443</td>
<td>382</td>
<td>8</td>
</tr>
<tr>
<td>HUVEC Rep 2</td>
<td>K562 Rep 2</td>
<td>355</td>
<td>285</td>
<td>10</td>
</tr>
<tr>
<td>K562 Rep 1</td>
<td>K562 Rep 2</td>
<td>224</td>
<td>308</td>
<td>168</td>
</tr>
</tbody>
</table>
### Differential processing detection

#### DP accuracy on simulated data

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Predicted DP</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSG</td>
<td>A Rep 1</td>
<td>A Rep 2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A Rep 1</td>
<td>B Rep 1</td>
<td>257</td>
<td>0.60</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>A Rep 1</td>
<td>B Rep 2</td>
<td>230</td>
<td>0.54</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>A Rep 2</td>
<td>B Rep 1</td>
<td>251</td>
<td>0.59</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>A Rep 2</td>
<td>B Rep 2</td>
<td>235</td>
<td>0.54</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>B Rep 1</td>
<td>B Rep 2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Cuffdiff

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Predicted DP</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Rep 1</td>
<td>A Rep 2</td>
<td>379</td>
<td>0.11</td>
<td>0.92</td>
</tr>
<tr>
<td>A Rep 1</td>
<td>B Rep 1</td>
<td>49</td>
<td>0.13</td>
<td>0.88</td>
</tr>
<tr>
<td>A Rep 1</td>
<td>B Rep 2</td>
<td>58</td>
<td>0.12</td>
<td>0.98</td>
</tr>
<tr>
<td>A Rep 2</td>
<td>B Rep 1</td>
<td>48</td>
<td>0.11</td>
<td>0.88</td>
</tr>
<tr>
<td>A Rep 2</td>
<td>B Rep 2</td>
<td>51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B Rep 1</td>
<td>B Rep 2</td>
<td>148</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### FDM

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Predicted DP</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Rep 1</td>
<td>A Rep 2</td>
<td>11</td>
<td>0.39</td>
<td>0.51</td>
</tr>
<tr>
<td>A Rep 1</td>
<td>B Rep 1</td>
<td>311</td>
<td>0.28</td>
<td>0.44</td>
</tr>
<tr>
<td>A Rep 1</td>
<td>B Rep 2</td>
<td>255</td>
<td>0.37</td>
<td>0.47</td>
</tr>
<tr>
<td>A Rep 2</td>
<td>B Rep 1</td>
<td>320</td>
<td>0.37</td>
<td>0.47</td>
</tr>
<tr>
<td>A Rep 2</td>
<td>B Rep 2</td>
<td>242</td>
<td>0.24</td>
<td>0.40</td>
</tr>
<tr>
<td>B Rep 1</td>
<td>B Rep 2</td>
<td>148</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Simulations based on two ENCODE cell lines, 10% of genes selected to be DP
Next steps for modeling RNA-Seq with PSGs

- Graph construction
- Exon discovery
- Splice junction discovery
- Model selection
- Learning dependencies between splice events
Summary

• Alternative splicing is a significant complication in RNA-Seq analysis.

• Probabilistic Splice Graphs enable identifiable models for alternatively spliced genes with efficient inference algorithms.

• Differential processing (splicing) tests with PSG models look promising.