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

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

RNA-Seq data

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

Regularization approaches

<table>
<thead>
<tr>
<th>Candidate isoforms (up to $2^n - 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$ candidate isoforms</td>
</tr>
</tbody>
</table>

Binary design:

$U = \begin{pmatrix} 1 & \ldots & 0 & 1 & \ldots & 1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 1 & \ldots & 0 & 1 & \ldots & 1 \end{pmatrix}$

Notations

Sparse Regression via Lasso

Estimate $\Phi$ sparse by solving:

$$\min R(U^T \phi) + \lambda \| \phi \|_1,$$

with $R$ a convex loss function. (Computational challenge to enumerate all candidate isoforms for genes with many exons)

Method

Path Selection problem

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:

- Combinations of isoforms are flows

A flow $f$ is a non-negative function on arcs of $(E, \delta)$ that satisfies conservation constraints.

<table>
<thead>
<tr>
<th>Path selection problem</th>
</tr>
</thead>
</table>

Simulations

Speed Comparison

Real Data

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

References

- Z. Xia et al., BMC Bioinformatics, 12:162, 2011.