Inferring exon junction expression from RNA-seq data

Boyko Kakaradov and Brendan Frey
Probabilistic and Statistical Inference group
University of Toronto
Motivation

• Goal: estimate transcript expression
  – absolute counts
  – relative (alternative splicing)

• RNA Sequencing
  – discrete expression counts
  – noisy & biased
Junction Mapping

Input Data

MLCB 2010
Example junction coverage

![Bar graph showing junction coverage across different position with an average line.

MLCB 2010]
Simple Model

\[ X_{ptj} \sim \text{Poisson}(\bar{X}_{tj}) \]

where \( \bar{X}_{tj} = \frac{1}{P} \sum_{p'=1}^{P} X_{p'tj} \)

\[ p(X = k; \bar{X}) = \frac{\bar{X}^k e^{-\bar{X}}}{k!} \]

Assumes uniform read coverage.
T*J shared position parameters. \( \bar{X}_{tj} \)
Junction Coverage is not uniform
Sparse (73% zeros)
Variable (SD = 3 x MEAN)
Tissue / Position Biases
Factored Model

\[ \alpha_{t,j} \quad \text{Junction} \quad j = 1, \ldots, 2553 \]

\[ \beta_{pt} \quad \rightarrow \quad X_{pt,j} \]

\[ \text{Position} \quad p = 1, \ldots, 40 \]

\[ \text{Tissue} \quad t = 1, \ldots, 16 \]
Factored Model

\[ X_{ptj} \sim \text{Poisson}(\beta_{pt} \alpha_{tj}) \]

\( \alpha_{tj} \)  
\( j = 1, \ldots, 2553 \)

\( \beta_{pt} \)

\( X_{ptj} \)

\( p = 1, \ldots, 40 \)

\( t = 1, \ldots, 16 \)
Factored Model: MLE

\[ L = \log p(\alpha, \beta | X) = \log \prod_{pt,j} p(\beta_{pt}, \alpha_{tj} | X_{ptj}) \]

Coordinate Ascent -- not identifiable

\[ \alpha_{tj}^{(i+1)} = \frac{\sum_p X_{ptj}}{\sum_p \beta_{pt}^{(i+1)}} \quad \beta_{pt}^{(i+1)} = \frac{\sum_j X_{ptj}}{\sum_j \alpha_{tj}^{(i+1)}} \]

Closed Form -- if we constrain

\[ \sum_p \beta_{pt} = 1 \]

\[ \alpha_{tj}^* = \frac{\sum_p X_{ptj}}{\sum_p \beta_{pt}} = \sum_p X_{ptj} \]

\[ \beta_{pt}^* = \frac{\sum_j X_{ptj}}{\sum_j \alpha_{tj} - \mu} = \frac{\sum_j X_{ptj}}{\sum_{p'} X_{p'tj}} \]

MLCB 2010
Learned MLE Parameters

$\alpha_{tj}$

$\beta_{pt}$
Factored Model

\[ X_{ptj} \sim \text{Poisson}(\beta_{pt} \alpha_{tj}) \]

\( \alpha_{tj} \)

\( j = 1, \ldots, 2553 \)

\( \beta_{pt} \)

\( X_{ptj} \)

\( p = 1, \ldots, 40 \)

\( t = 1, \ldots, 16 \)

MLCB 2010
Latent Factored Model

\[ X_{ptj} \sim \text{Poisson}(\beta_{pt}\alpha_{tj}) \]

\[ \alpha_{tj} \]

Junction
\[ j = 1, \ldots, 2553 \]

\[ S_{ptj} \]

Position
\[ p = 1, \ldots, 40 \]

\[ X_{ptj} \]

\[ \tilde{X}_{ptj} \]

Tissue
\[ t = 1, \ldots, 16 \]
Latent Factored Model

\[ X_{ptj} \sim \text{Poisson}(\beta_{pt} \alpha_{tj}) \]

\[ S_{ptj} \sim \text{Bern}(\pi_{ptj}) \]

\[ \tilde{X}_{ptj} = X_{ptj} + S \cdot \text{Geometric}(\lambda) \]

Junction
\[ j = 1, \ldots, 2553 \]

Position
\[ p = 1, \ldots, 40 \]

Tissue
\[ t = 1, \ldots, 16 \]
Latent Factored Model: EM

E-step

\[ q(X_{ptj}, S_{ptj}) = \frac{1}{Z_{ptj}} p(X_{ptj} | \alpha_{tj}, \beta_{pt}) p(\tilde{X}_{ptj} | X_{ptj}, S_{ptj}) p(S_{ptj}) \]

\[ Z_{ptj} = \sum_{S_{ptj} \in \{0, 1\}} \sum_{\tilde{X}} p(S_{ptj}) \sum_{X=0} p(X_{ptj} | \alpha_{tj}, \beta_{pt}) p(\tilde{X}_{ptj} | X_{ptj}, S_{ptj}) \]

M-step

\[ \alpha_{tj}^{(i+1)} = \frac{\sum_{p} E_q X_{ptj}}{\sum_{p} \beta_{pt}^{(i)}} \]

\[ \beta_{pt}^{(i+1)} = \frac{\sum_{j} E_q X_{ptj}}{\sum_{p'j} E_q X_{p'jtj}} \]

\[ E_q X = \sum_{s=0}^{\tilde{X}} \sum_{x=0}^{\tilde{X}} x q(x, s) \]

\[ \pi_{ptj} = \sum_{x=0}^{\tilde{X}} q(x, S = 1) \]
Learned MLE Parameters

$\alpha_{tj}$

$\beta_{pt}$
Evaluation

• Model Fit
  – Log-likelihood
  – Reconstruction error

\[
\frac{\| X - \hat{X} \|_1}{\| X \|_1} = \frac{\sum_{ptj} |X_{ptj} - \hat{X}_{ptj}|}{\sum_{ptj} |X_{ptj}|}
\]

• Apply to Alternative Splicing (AS)
  \( j_a = \text{alternative junction} \quad j_c = \text{constitutive junction} \)

\[
\log \left( \frac{\hat{X}_{ptj_a}}{\hat{X}_{ptj_c}} \right) \quad \log \left( \frac{X_{ptj_a}}{X_{ptj_c}} \right)
\]
AS Estimate: Latent Model
Summary

• Junction Coverage
  – not uniform, sparse
  – tissue/position bias

• Poisson Models
  – mean = variance
  – Iterative & closed form

• Latent Model
  – outlier model
  – EM algorithm

• Results
  – model fit
  – alternative splicing