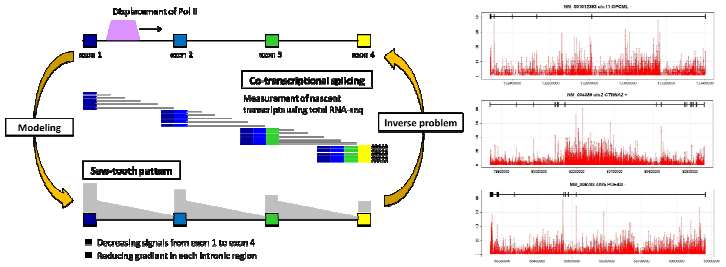
全トランスクリプトーム計測に基づく転写伸長速度およびスプライスパターンの予測

河村　優美　　総合研究大学院大学 複合科学研究科 統計科学専攻 博士後期課程2年　（指導教員：吉田亮准教授）

1. 概要

Total RNA-seq（Total RNA-sequencing without poly(A) selection）という手法を用いることで，細胞中のRNA分子（新生RNAを含む）の量を網羅的に計測することができる．本研究では，Total RNA-seqの解析から，RNA ポリメラーゼ II（以下，Pol II）の転写伸長プロセスを再構成できることを示す．Pol II は，転写（RNAの合成）を触媒する酵素で，DNA上の遺伝子領域を5'から3'方向に移動しながら，段階的にmRNAの合成を進める．この生体プロセスのことを転写伸長という．本研究は，Total RNA-seqを用いて転写伸長の速度やスプライス・パターンを同定するデータ解析手法を整備し，生物学における有用性を実証することである．



2. 背景と目的

本研究では，Pol IIの存在確率とリード分布の関係を状態空間モデルで表現し，ベイズ推定によりPol IIの存在確率（転写伸長の相対速度）とスプライス部位を同定することを試みる．

Total RNA-seqという方法によって，細胞中に存在するRNA分子の総量を計測することが可能となる．データは，転写伸長途中のRNA分子の発現量のスナップショットになっている．データは，リードカウントという形式で表現される．これは，RNAの核酸塩基の位置を横軸に，個数（リード数）を縦軸にとったものである．リードの分布は，左図に示されるような鋸歯状の分布(saw-tooth pattern)が現れることが知られている．このpol IIの勾配が転写伸長速度を反映しており，このパターンをモデル化して，逆問題を解けば，転写伸長のプロセスを再構成できると考えられる．推定された速度分布に基づき，転写伸長速度とヒストン修飾，クロマチンの状態（エピジェネティクス），スプライシング異常との関係を調べ，また細胞種による転写伸長過程の違い等を明らかにしたい．

Pol IIの存在確率とリード分布の関係について

リード分布の形状を決定する主要因子は，塩基配列の位置上に与えられるPol IIの存在確率である．観測データである，リード分布からpol IIの存在確率を推定する．位置tにおけるpol IIの存在確率をρ(t)とする．Pol IIの移動速度（転写伸長速度）は，存在確率の逆数に比例する．また，各位置の期待リード数はρ(t) の積分値で与えられる．

Transformation formula

$$\lambda(x)=\begin{cases}\int_x^{t_{\text{intron }k}}\rho(s)\,\mathrm{d}s & x\in I_k\\ \int_x^{t_{\text{last exon}}}\rho(s)\,\mathrm{d}s & x\in E_k\end{cases}$$

3. 状態空間表現

粒子フィルタ(pf)でPol IIの存在確率を推定

■ 観測モデル

リード数:

(1) $y_n=\lambda_n(x_{1:n},s_n)w_n,\ \ w_n\sim\text{lognormal}(\mu,\sigma)$

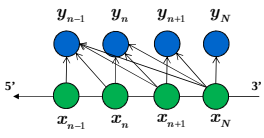
期待リード数:　$\lambda_n(x_{1:n},s_n)=\sum_{i=n}^{s_n}x_i\ \ (s_n\le n)$

$n\ (=1,\cdots,N)$：位置（RNAの各核酸塩基）

x_n：Pol IIの存在確率

s_n：スプライス部位（位置 n の塩基が除去される位置）

グラフィカルモデル



■ システムモデル

Pol IIの存在確率は滑らかに変化

(2) $\log x_n=\log x_{n+1}+v_n,\ \ v_n\sim N(0,\gamma)$

4. スプライスサイトの推定

観測モデル

$y_n=\lambda_n(x_{n:N},s_n)w_n,\ \ w_n\sim\text{lognormal}(\mu,\sigma)$

期待リード数:　$\lambda_n(x_{1:n},s_n)=\sum_{i=n}^{s_n}x_i\ \ (s_n\le n)$　状態変数の和 (n から s_n)

s_n は n 番目の塩基が切除（スプライス）される位置
未知パラメータであり，転写伸長速度と同時に推定

多様なスプライシングパターンであるオルタナティブスプライシングや，複数回のスプライシング反応を示すリカーシブスプライシング(RS)などがあるため，スプライス部位は未知である．
期待リード数λ_nは，s_nからnまでの状態変数の総和によって表現される．s_nは位置nの塩基が除去される位置を表す．位置nがエキソンの場合，途中でエキソンの除去がなければ，s_nは3'末端の位置（$s_n=N$）になる．nがイントロンの場合，RSがないと仮定すれば，s_nはイントロンの終末点となる．

5. 粒子フィルタによる状態推定（Pol II density, splice pattern）

粒子フィルタは状態空間モデルの状態推定を実行するシミュレーションベースのアルゴリズムである．ここで行った粒子フィルターは以下のようなアルゴリズムである．N(3'末端の位置)から1(5'末端の位置)へ逆向きに粒子フィルタリングを行う．状態変数（pol II）であるx_{n+1}について各粒子ごとにシステムノイズv_{n+1}を発生させ，$n+1$からnの予測粒子群を出す．期待リード数λ_nは，状態変数（splice site）s_nからnまでの状態変数x_nの総和によって表現される．s_nはセクション4で示したように，3'末端の位置（$s_n=N$），RSの位置，またはイントロンの終末点となる．RSの発生頻度はチューニングを行い決定した．nの観測データy_nから尤度計算を行い，各粒子iの重みw_iが計算される．重みに基づいてn:Nのリサンプリングを行い，適合度の高いものが復元抽出されて，粒子フィルタによる状態推定が行われる．

パラメータ推定について

式(1)の観測モデルと式(2)のシステムモデルを使ってパラメータ推定を行った．

■ 式(1)から　$w_n=\dfrac{y_n}{\sum_{i=n}^{s_n}x_i}$

■ 式(1)から　$f(n)=\log\sum_{i=n}^{s_n}x_i$

$\exp(f(n))=\sum_{i=n}^{s_n}x_i$

$\hat{x}_i=\max\{0.0001,\exp(f(n))-\exp(f(n+1))\}$

■ 式(2)から　$v_n=\log x_n-\log x_{n+1}$

6. Total RNA-seqリードの選定

Total RNA-seqリードにイントロン勾配が見られるもののみを解析対象の遺伝子として選んだ．

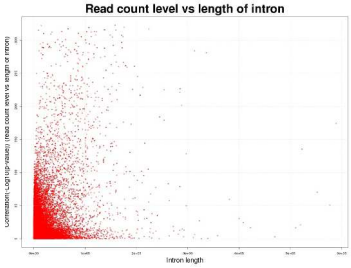
Read count at position x　　Distance from the intron terminal

$$\mathbf{R}_{\text{intron}}(x)\approx\mathbf{G}_{\text{gene}}\left(t^{\text{intron}}-x\right)\ \Rightarrow\ \rho\left[\mathbf{R}_{\text{intron}}(x),\ \left(t^{\text{intron}}-x\right)\right]$$

Correlation

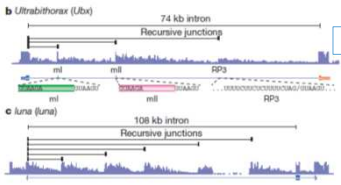
Constant
(intrinsic gene expression level)

イントロンの終末点からの長さとリードレベルが相関していた．イントロンが長く，リードレベルが高いものを選定した．



7. スプライスサイトの推定結果

*Drosophila*のrecursive splice sites (RSS)

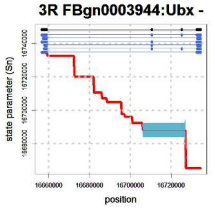


Duff et al. (2015). *Nature* **521**, 376–379. Fig.1

b, Example of total RNA-seq data for the Ubx gene which is known to contain three recursive splice sites.
c, Example of five recursive splice sites identified in luna.

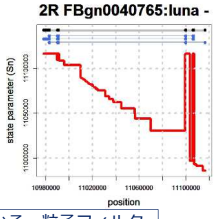
RSSの推定結果 (number of pf = 100,000)

3個のRSSを推定



Estimated

5個のRSSを推定



Drosophila UbxのRSSは3個，lunaは5個が同定されている．粒子フィルタによる状態推定を行ったところ，これらを推定できた．

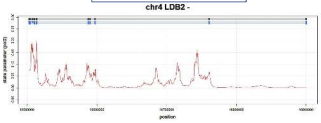
8. Pol II densityの推定結果

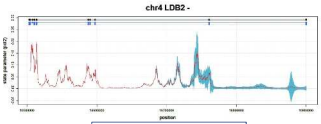
Human fetal (ERR042386)の推定結果
(Chr4 LDB2 -, number of pf = 100,000)

True
Estimated

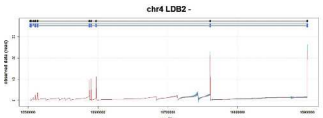
粒子フィルタによる状態推定を行ったところ，pol II densityを推定できた．

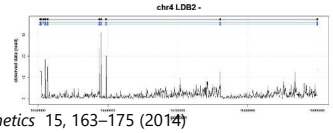
Pol II



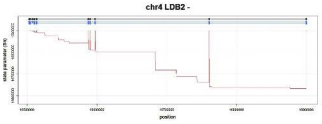


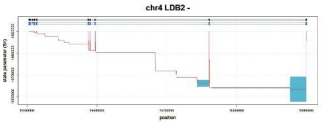
Read level





Splice site





参考文献

- David L. Bentley. Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics* 15, 163–175 (2014)
- Adam Ameur *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural & Molecular Biology* 18, 1435–1440 (2011)
- Churchman LS *et al.* Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469, 368-373 (2011)
- Michael O. Duff *et al.* Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature* 521, 376–379 (2015)
- Christopher R. Sibley *et al.* Recursive splicing in long vertebrate genes. *Nature* 521, 371–375 (2015)
- Iris Jonkers *et al.* Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife Sciences* 3 e02407 (2014)