2018年6月15日　統計数理研究所　オープンハウス

Statistical modeling and inference of the rate of RNA polymerase II elongation by total RNA sequencing

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3年

1. 概要

Total RNA-sequencing without poly(A) selection（以下，Total RNA-seq）という手法を用いることで，転写と共役したスプライシングによる新生転写産物のプロファイリングを得られる．Total RNA-seqの観測データは，プロセシングの様々な段階にある新生転写物の状態を捉えたものである．これにより，リード分布にはRNA Polymerase II (Pol II)の移動方向に沿った減少勾配とエキソン領域の突起（鋸状のパターン）が形成される（Fig1）．本研究では，Pol IIの存在確率とリードの分布の関係を状態空間モデルで表現し，ベイズ推論に則りPol IIの存在確率（転写伸長速度の逆数に比例）とスプライシングのパターンを，sequential Monte Carlo algorithmとスプライシングパターンのモデリングにより同時に推定する．既知の手法では，薬剤処理による多くの時系列データの実験デザインと実験コストが必要であったが，提案手法により，1回の実験データだけで推定ができる．

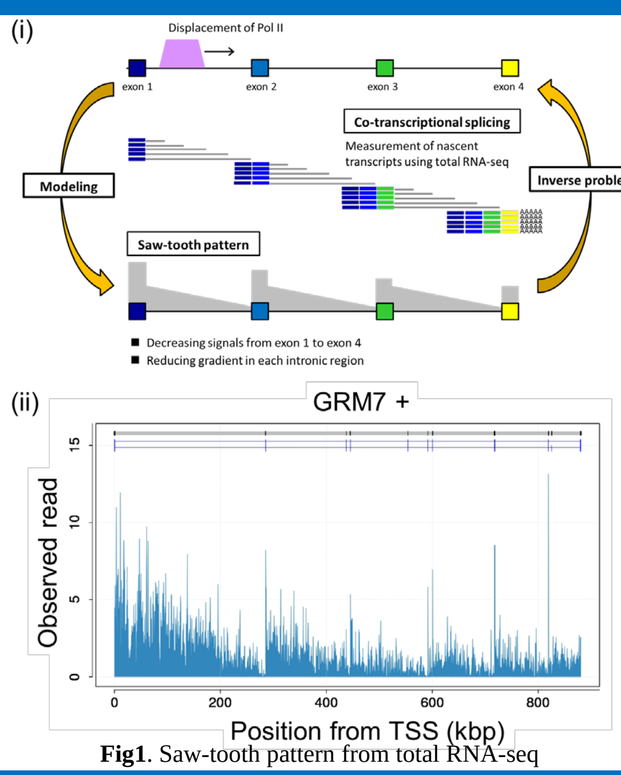


Fig1. Saw-tooth pattern from total RNA-seq

2. イントロダクション

Total RNA-seqのデータは，細胞中に存在するRNA分子の総量を計測したものである．データには，伸長途中のRNAも含まれる．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(Fig.1 下)．Total RNA-seqの場合，リードの分布には，鋸歯状の分布が現れることが知られている．イントロン領域には，5′から3′方向に減少勾配が生じる．エキソン領域では，リード数がイントロンに比べて大きくなるため，スパイクが出現する．さらに，エキソン領域のリード数も一般的には5′から3′方向にかけて減少する．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(Fig.1 上)．

Pol II による転写伸長速度はまだ多くが測定できてはいない．転写伸長速度の測定方法のひとつとして，新生転写物を生成するポリメラーゼ（Pol II）の移動距離と経過時間を定量するものがある．転写活性を阻害する薬剤処理と時系列データから，移動距離と経過時間を測定するこの方法は，genome-wide run-on sequencing (GRO-seq) やprecision run-on sequencing (PRO-seq)などがある(Jonkers *et al.*, 2015).

Elongation rate: $v(t) = 1 / x(t)$

Transformation formula

$$r(t) = \begin{cases} \int_t^{t_{\text{intron } k}} x(t)\mathrm{d}t & t \in I_k \\ \int_t^{t_{\text{last exon}}} x(t)\mathrm{d}t & t \in E_k \end{cases}$$

3. スプライスサイトのモデル化

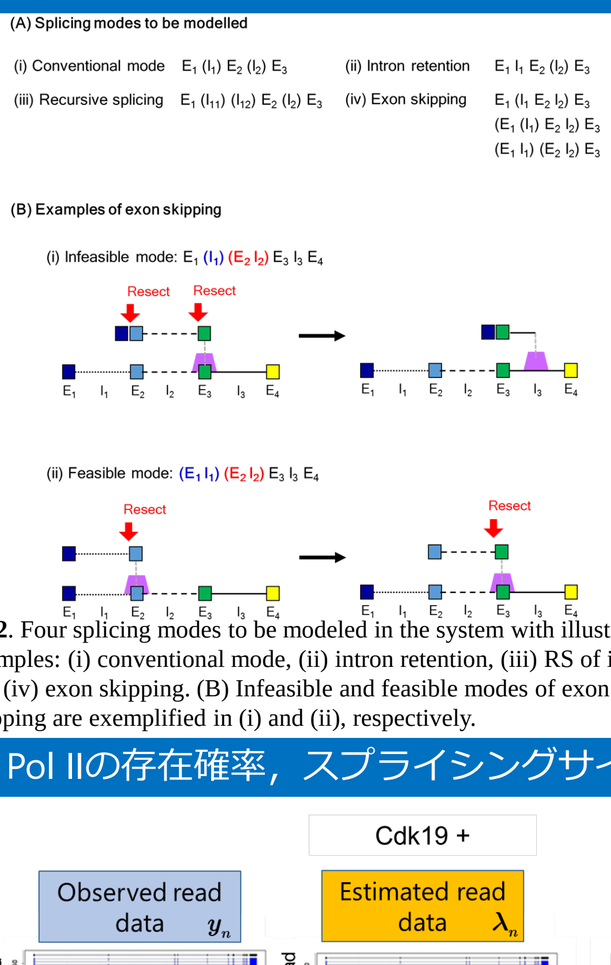


Fig2. Four splicing modes to be modeled in the system with illustrative examples: (i) conventional mode, (ii) intron retention, (iii) RS of introns, and (iv) exon skipping. (B) Infeasible and feasible modes of exon skipping are exemplified in (i) and (ii), respectively.

スプライシングイベントは次の3つ
(i)　エキソンスキッピング
(ii)　イントロンリテンション
(iii)　イントロン内のリカーシブスプライシング (RS)

例えば

$E_1 I_1 E_2 I_2 E_3$

E_2　がエキソンスキッピングする場合

$E_1(I_1 E_2 I_2)E_3$　　起こり得る

$E_1(I_1)(E_2 I_2)E_3$　　起こり得ない

括弧 () 単位でスプライスアウト

4. ベイズ推定によるモデル化

$n(=1,\cdots,N)$: 位置（RNAの各核酸塩基）　3′ ⇒ 5′ (逆向き)

x_n : Pol II density　　推定対象

s_n : Splice site (位置 n の塩基が除去される位置)

y_n : Read density　　観測データ

状態空間表現

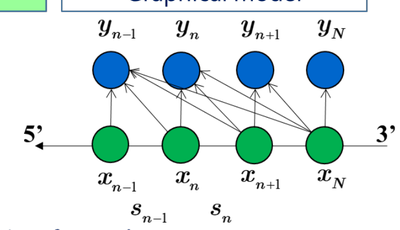
■　観測モデル

$\log y_n = \log \lambda_n + \text{noise},$

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

Pol II ⇒ read densityのtransformation formula

Graphical model



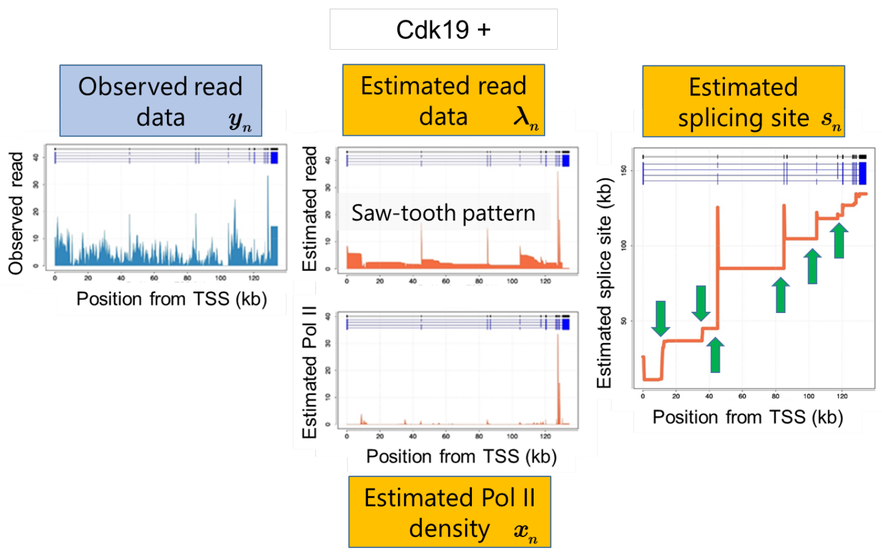
■　システムモデル

Prior: $\log x_n = \log x_{n-1} + \text{noise},\ s_n \sim p(s_n \mid s_{1:n-1})$,　Pol IIの存在確率は滑らかに変化

■　ベイズ推定（粒子フィルタ Particle filter）

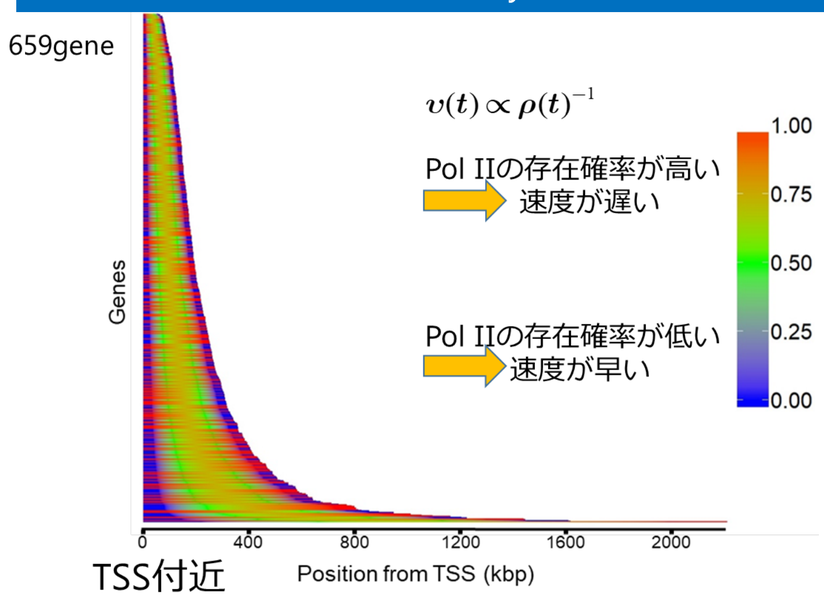
Posterior: $(x, s) \sim p(x, s \mid Y)$　x_n, s_n の同時推定

5. Pol IIの存在確率，スプライシングサイトの推定結果



我々のモデルを使って観測データである左のリードの分布からPol IIの存在確率，スプライシングサイトの推定が得られた．横軸がTSSからのポジション．縦軸が推定できたPol II density. スプライスサイトになる．下図のPol II densityからsaw-tooth patternが得られた．スプライスサイトは矢印で示す階段上のポジションで示している．

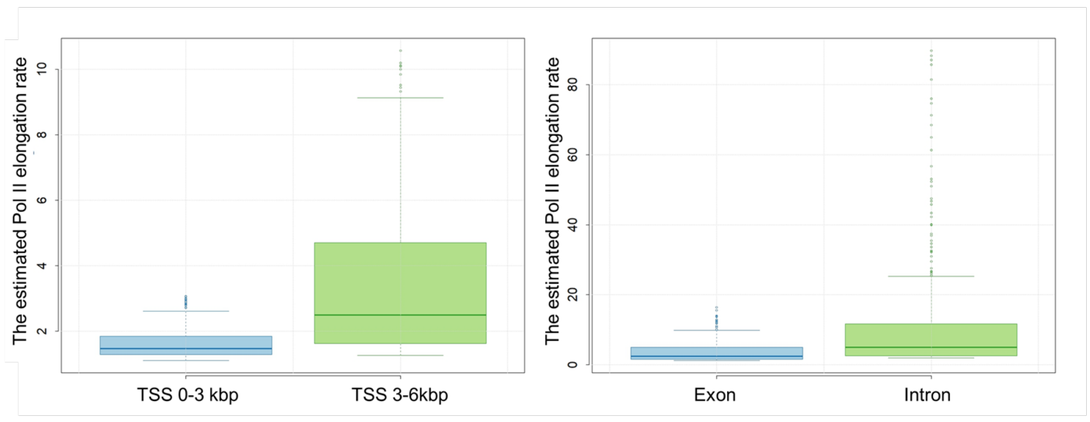
6. マウスES細胞のPol II densityの推定結果



マウスES細胞から選定できた659遺伝子のPol IIの存在確率（転写伸長の相対速度）の推定結果である．横軸が遺伝子の長さ，縦軸が各遺伝子になる．ヒートマップは，Pol IIの存在確率が高く，すなわち速度が遅いところ，Pol IIの存在確率が低く，すなわち速度が早いところを示している．多くの遺伝子はTranscription start site (TSS)付近で速度が遅い．これは転写のメカニズムと一致している．

7. Pol II densityの推定は転写のメカニズムと一致

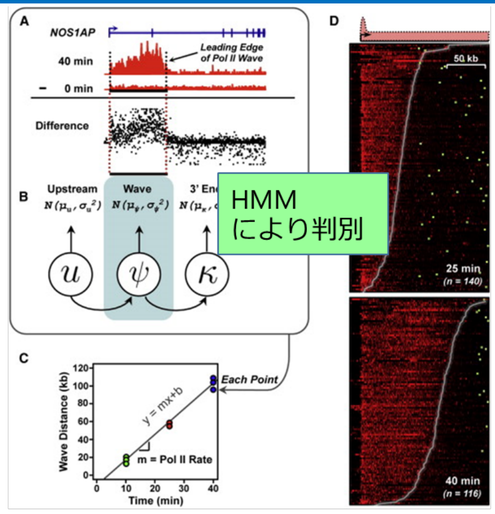
■　TSS付近で速度が遅い
■　エキソンでの速度はイントロンでの速度より遅い　⇒　転写のメカニズムと一致



（メカニズム）TSS付近ではNELFによりPol IIが一時停止し，伸長速度が遅い
（推定結果）TSS付近で約1.75倍速度が遅い

（メカニズム）エキソンでは様々なプロセシングにより転写伸長速度が遅い
（推定結果）エキソンでの速度はイントロンでの速度より約3倍遅い

8. GRO-seqによる転写伸長速度推定との比較



groHMM (Danko *et al*, 2015)
■　Pol II活性化阻害剤によってPol II waveが出現
■　時系列データをとり，Pol IIの移動距離 / 経過時間を定量

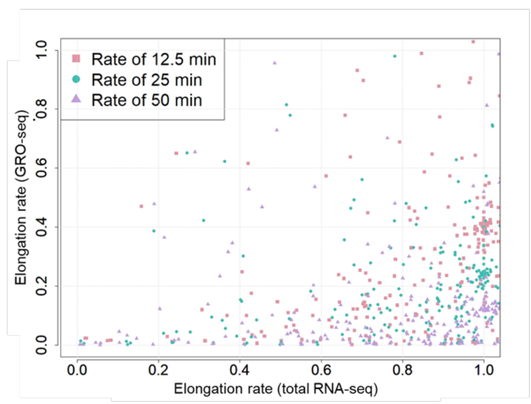


Fig. 3. Comparison between the estimated elongation rates using total poly-A(-) RNA-seq and GRO-seq with the hidden Markov model.

我々のモデルによる転写伸長速度とGRO-seqによる結果を比較したところ，相関は得られなかった

参考文献
- 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)
- 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)