氏　　名	河村 優美
学位(専攻分野)	博士(統計科学)
学 位 記 番 号	総研大甲第 2071 号
学位授与の日付	平成 31年 3 月22日
学位授与の要件	複合科学研究科　統計科学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Bayesian inference for transcription elongation rates by using total RNA sequencing
論 文 審 査 委 員	主　査　教授　上野　玄太 教授　吉田　亮 准教授　小山　慎介 准教授　逸見　昌之 准教授　山口　類　東京大学　医科学研究所

（様式３）

博士論文の要旨

氏　　名 河村　優美

論文題目 Bayesian inference for transcription elongation rates by using total RNA sequencing

Motivation: This subject matter is to solve real scientific problems appeared in the natural sciences, bioinformatics and so on by Bayesian method or statistical inference. I would like to build how to innovate and apply machine learning and statistical science to tackle the challenges involving molecules and materials, to make scientific discoveries.

Transcription elongation rates affect co-transcriptional events such as splicing, termination and genome stability. However, measurement of genome-wide elongation rates and processing are poorly understood. Sequencing total RNA without poly-A selection enables us to obtain a transcriptomic profile of nascent RNAs undergoing transcription with co-transcriptional splicing. In general, the RNA-seq reads exhibit a sawtooth pattern in a gene, which is characterized by a monotonically decreasing gradient across introns in the 5' to 3' direction, and by substantially higher levels of RNA-seq reads present in exonic regions. Such patterns result from the process of underlying transcription elongation by RNA polymerase II (Pol II), which traverses the DNA strand in a 5' to 3' direction as it performs a complex series of mRNA synthesis and processing. Therefore, data of sequenced total RNAs could be utilized to infer the rate of transcription elongation by solving the inverse problem.

Method and Results: Previously the methodology of the rates of transcription elongation relies on actually measurement which is the moving distance of Pol II division by a passing time. These methods are calculated using the moving distance and time, have need of expensive time and cost for individual experiments. The methodology in this thesis have created statistical scheme that estimate hidden variable from reliable existing data in a large dose of sequencing data that have no need of an individual experimental design. According to Bayes' rule, observation read density data let us to estimate unknown variable Pol II, the rate of transcription elongation are estimated.

We devoted this matter by using a signal reconstruction technique based on a sequential Monte Carlo (SMC) algorithm. The objective is to reconstruct the spatial distribution of transcription elongation rates in a gene from a given noisy, sawtooth-like profile. It is necessary to recover the signal source of the elongation rates separately from several types of nuisance factors, such as unobserved modes of co-transcriptionally occurring mRNA splicing, which exert significant influences on the sawtooth shape.

In this thesis, a Bayesian inference and modeling for the matters of concerning transcription that still have not clarified is constructed. We explored the Pol II elongation rates in 659 genes in mouse embryonic stem (ES) cells. After forwardly modeling the given sequenced RNA-seq

reads for unknown rates of elongating Pol II and unknown modes of splicing by state space modeling, the backward prediction was performed according to Bayes' law to inversely predict the unknowns. Suppose transcription elongation is predicted, this will lead to further clarification of transcription.

One difficulty of the inverse problem lies in the fact that splicing variations cause significant deviations from the expected sawtooth pattern as previously shown. Hence, it is essential to infer the splicing patterns simultaneously with the elongation rate through analysis of a given read density. The prior distribution is used in the SMC calculation to sequentially produce unknown splicing sites for which the sites n are removed out from the transcribed RNA. The difficulty is to avoid the occurrence of infeasible splicing patterns during the random generation.

As a proof of principle, the present method was tested using published total RNA-seq data derived from mouse ES cells. We describe the spatial characteristics of the estimated elongation rates, focusing especially on promoter-proximal sites, exons, and introns. We found that the predicted elongation rates are highly correlated with the epigenetic landscape of nucleosome occupancy and histone modification patterns.

Despite the potentially great promise of utilizing total RNA-seq to study transcription elongation, there has been considerably less progress made in statistical methods. In some previous studies, the slope of the read density gradients, for instance, which is obtained using linear regression, was used as the relative elongation speed. However, as described in this study, different splicing modes could bring different slopes to the read density, thereby drawing the wrong conclusion in the absence of inferring the splicing variations. One contribution of this study is to provide a way to estimate unmeasured states of elongation rates and splicing modes simultaneously.

As a by-product of our method, the RS sites could be identified. Although details were not described, quite a lot of valleys, possibly indicating ratchet points of RS, were found in the intronic regions.

Form8・Separate Sheet
（様式８・別紙）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　　名　河村　優美

Title
論文題目　Bayesian inference for transcription elongation rates by using total RNA sequencing

（論文審査結果）［2019　年　1　月　21　日　実施］
2019 年 1 月 21 日午前 10 時，審査委員 5 名（内外部審査委員 1 名）が立ち会い，河村優美氏の博士論文審査委員会を開催した．審査の結果，本論文は博士（統計科学）の学位授与に値すると判定した．

[論文の概要]
論文は全 5 章 52 頁からなる．「転写伸長」と呼ばれる RNA 合成反応における，活性酵素 RNA ポリメラーゼ II（以下 Pol II）の推定を目的とした研究である．

転写伸長過程において，Pol II は，DNA 鎖を一定方向に移動しながら DNA 塩基配列の情報を RNA 分子に写し取っていく．一方，total RNA-sequencing (以下 total RNA-seq) という実験技術を用いることで，細胞中の RNA 分子の総量を計測することができる．得られるデータは,転写伸長過程の様々なステージの転写産物の状態を反映したものである．このため,データの分布には 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．本研究では，Pol II の存在確率と計測データの関係を状態空間モデルで記述し，ベイズ推論で Pol II の存在確率（転写伸長速度の逆数に比例）の推定を行った．ヒストン修飾やクロマチンの状態等，様々な転写伸長の制御因子との関連性を調べることで，推定された転写伸長速度の妥当性を実証した．

第 1 章は，序章である．概要と研究の位置づけが述べられている．第 2 章は，背景知識の章である．転写伸長反応，total RNA-seq，転写伸長速度を計測する既存技術，転写伸長を制御する生化学的因子を解説している．

第 3 章は，提案手法の解説である．Pol II の存在確率を状態変数とし，total RNA-seq のデータ生成過程を記述した観測モデルを提案している．Pol II 存在確率の平滑化事前分布，スプライス部位に対する生物学的知見による事前分布を設計した上で，逐次モンテカルロ法を用いて状態推定を行う．

第 4 章は，提案手法の実証結果をまとめたものである．マウス ES 細胞のデータを用いて転写伸長速度の推定を行い，様々な切り口から推定された速度分布に対する生物学的考察を行っている．本研究によって明らかになったことは，以下の通りである．

(1) 多くの遺伝子で転写の開始地点と終結地点の近傍で伸長速度が急減に遅くなる現象が確認された．

(2) イントロンと比べてエキソン領域の伸長速度が遅い．

(3) ヒストンの化学修飾の状態及びヌクレオソーム占有率のパターンと転写伸長速度の間

に強い相関が確認された.
(4) 他の計測技術（ChIP-seq）で得られた Pol II 分布と推定結果の間に整合性が確認された.
(5) 他の計測技術（GRO-seq）との比較を行い，提案手法は識別性に優れることを示した.
第 5 章は，結論が述べられている.

[論文の評価]
全ゲノムレベルで転写伸長速度を計測する実験技術は現時点において確立されておらず，Pol II の動的特性や生化学的機能の全体像はほぼ未解明といっても過言ではない．本研究は，total RNA-seq のデータに現れる鋸状のパターンが転写伸長のダイナミクスの帰結であるという着眼から出発し，ベイズ推論で逆問題を解くことで転写伸長速度を再構成できることを実証した．RNA シーケンシングのプロトコルを工夫することで，本研究が全ゲノム規模の転写伸長速度分布の解明に貢献できる可能性がある．Total RNA-seq の本来の用途（RNA 分子の総量の計測）とは全く異なる活用方法を新たに見出し，従来の計測技術と比べ，簡便な転写伸長速度の推定を可能にした．本研究の学術的意義は十分に認められる．第 3 章と第 4 章の内容は，生物情報学のトップジャーナル Bioinformatics 誌に掲載されている（査読付き論文，第一著者）．また，本研究は，日本バイオインフォマティクス学会主催の第 6 回生命医薬情報学連合大会（2017）において，研究奨励賞を受賞している．
以上より，博士論文審査委員会は全員一致で出願論文が博士（統計科学）の学位授与に値すると判定した.