様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　５　月　２４　日現在

機関番号：　８２４０１
研究種目：　若手研究(B)
研究期間：　2013～2015
課題番号：　２５８３０１４８

研究課題名（和文）転写因子結合量から転写量を予測する物理化学的モデルの構築

研究課題名（英文）Model of mRNA transcription and transcriptional factor binding

研究代表者

二階堂　愛（NIKAIDO, Itoshi）

国立研究開発法人理化学研究所・情報基盤センター・ユニットリーダー

研究者番号：００３８３２９０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：細胞は適切なmRNAが適切なタイミング、量で転写されることで、その細胞機能を発揮する。近年、網羅的な転写量と転写因子結合位置・量を定量的に計測できるようになってきた。しかし、転写量と転写因子結合量のデータ統合には、それらの単純な積集合を取る解析しか行われてこなかった。そこで転写量と結合量の統計的・物理化学的な関係をモデル化する手法を開発し、細胞機能を発揮する転写ネットワークを定量的に記述できるようにする。

研究成果の概要（英文）：Functions of cell are defined by temporal and special regulations of transcriptional networks in our body. Recently, massively parallel DNA sequencer realizes that we easily measure global gene expressions and binding of transcriptional factors on a genomic DNA. However, researchers had carried out simple integration of mRNA quantity and transcription factor binding to understand a transcriptional regulation. Thus, I study to develop statistical and physical models of mRNA transcription for quantitative description and understanding of transcriptional networks.

研究分野：　バイオインフォマティクス

キーワード：　バイオインフォマティクス

１．研究開始当初の背景

複雑な生命現象を理解するには、生命が作り出すタンパク質や RNA などの制御関係を観察し、解釈する必要がある。近年、超並列型 DNA シーケンサーの登場で、RNA 転写量を観察する RNA-seq のみならず、転写因子とゲノム DNA 相互作用を網羅的に観察する ChIP-seq によりタンパク結合地図とその結合量をも簡便に得られる。しかしタンパク質が結合しても、標的遺伝子の転写を制御しているとは限らない。そのため、どの結合サイトがどの遺伝子をどの程度、転写を制御したのかを ChIP-seq と RNA-seq データから予測する必要がある (Ohyoung et. al. 2009, Wang et. al. 2012)。しかし、この方法に関してはいまだコンセンサスがあるとは言えない。

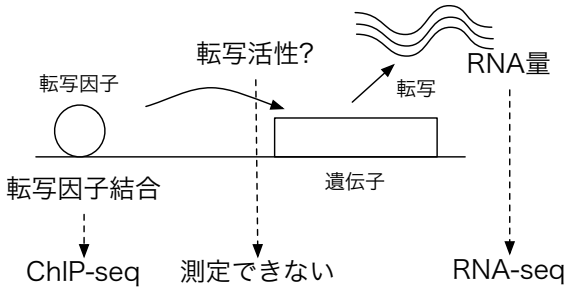


図 1. 転写活性の予測

転写因子の結合量は ChIP-seq で、RNA 量は RNA-seq で測定できるが、転写活性はシーケンサーでは実測できないため、データから予測する必要がある。

結合量と標的遺伝子の対応をつける際に最もよく使われる手段は、転写が変動した遺伝子と結合があった遺伝子の集合積を取りベン図やオイラー図で表わす方法である(図 2 上段)。ところが、転写因子が結合しても、その影響が転写を活性化させる場合と抑制する場合がある。また結合してもコファクターなどがリクルートされなければ転写への影響を示さない場合もある。

そこで転写因子の結合量から転写量を統計的にモデリングする手法が提案されている。これは転写因子の結合量を線形回帰モデルで表現し、RNA-seq で実測された転写量を説明するモデルである。しかし、これらの統計モデルは、結合量の転写への影響が統計的パラメータとして抽象化されており、予測から実験的に測定・操作できるパラメータの情報が得られない。

２．研究の目的

本提案では、ChIP-seq と RNA-seq のデータから、統計モデルを構築する。また、実験的に操作・測定できるパラメータを持った転写量の物理化学モデルの構築を目的とする。この予測モデルから逆に、エンハンサーを設計し、任意の発現パターンを設計する方法論を開発する。

３．研究の方法

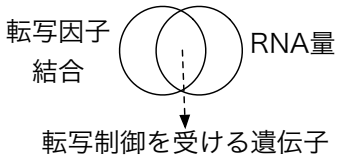
統計・物理化学モデルの構築するために必要な ChIP-seq と RNA-seq データを収集・解析をする。物理化学モデルは RNA 転写量の時間変化を表現する確率微分方程式とするため、時系列 RNA-seq と ChIP-seq データが揃っているデータを収集する。提案者がすでに持っている胚性幹細胞(ES 細胞)を栄養外胚葉細胞(TS 細胞)に分化させた時系列の発現データと Sox2, Esrrb, p300 の時系列 ChIP-seq データの解析を進める。また、より公的データベースから概日時計の転写因子の ChIP データ(タイムポイント数 12-24 程度)を得て解析する。具体的には、すべての実験に関して、生データである FASTQ データから、クオリティチェックやリファレンスゲノムへのマッピング、結合サイトと結合量を定量するピーク発見、DNA 結合モチーフ配列解析などを一通り行う。

転写量の時系列変化を説明する物理化学モデルを構築する。モデルには、RNA 分解項、転写因子による転写制御項、ゆらぎ項からなるランジェバン方程式を作る。転写項はヒル式を使う(Burg J. 2008 を援用)。このモデルに、前年度までに得られた RNA 量と転写因子の結合量をあてはめ、EM アルゴリズムによりパラメータを推定し、クロスバリデーションを行う。モデルの評価には AIC によるモデルの評価を行う。また、提案者が作った非線形回帰モデルや既存の統計モデルとの予測精度の比較を行う。

４．研究成果

胚性幹細胞(ES 細胞)を栄養外胚葉細胞(TS 細胞)に分化させた時系列の発現データと Sox2, Esrrb, p300 の時系列 ChIP-seq データを利用して、統計モデルの構築を行った(図 2)。

集合積による標的遺伝子の発見



統計モデルによる転写のモデル化

$$y_i = a_0 x_0 + a_1 x_1 + \ldots + a_j x_j + \varepsilon_i$$

y_i：遺伝子iの転写量 (RNA-seq)

x_j：転写因子jの結合量 (ChIP-seq)

ε_i：ノイズ項

図 2. 転写標的遺伝子の発見と転写活性の予測

まず、遺伝子ごとのクロマチン状態を考慮に入れた非線形回帰モデルを採用した。まず転写量を周辺のピークの結合量の総和と考えた。ただ、それらのピークの影響は、転写開始地点からの距離に依存して、その効果が変わると考えて、距離が離れると指数的に、その影響が弱まるようにした。

その効果で、データへのあてはまりが、既報の単純な線形回帰モデル(R^2 = 0.6-0.7)よりも提案モデルのほうが良いこと(R^2 = 0.9)を示した。これらから、単純な統計モデルよりも、生物学的に意味のある項を加えると予測精度が挙がることが示唆された(図 3)。

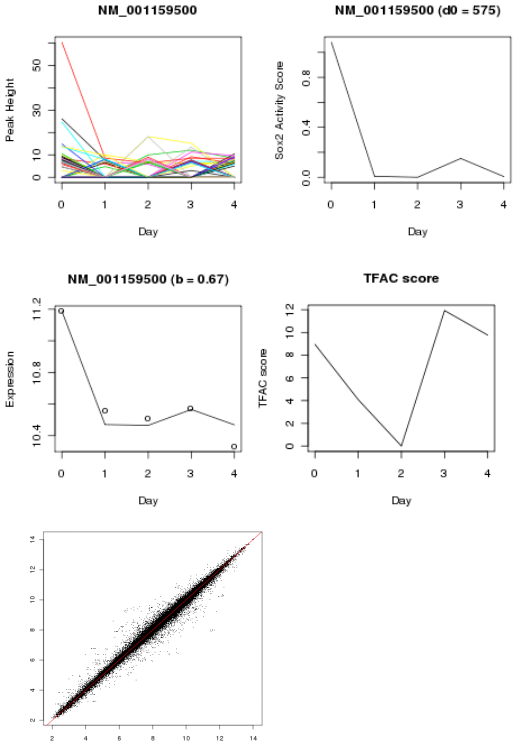


図 3. 提案者の統計モデルによる転写活性の予測. 左上の実線が予測した時系列の転写量、白丸が実測した転写量。横軸が時間、縦軸が転写量。右上はクロマチン状態を考慮しないモデルで予測が悪い。下は全遺伝子の予測転写量と実測転写量の散布図。縦軸が予測、横軸が実測転写量。全体に良く予測できている。

推定したパラメータは、クロスバリデーションにより評価を行った。モデルのパラメータ数については、評価には AIC によるモデルの評価を行い、既報の手法よりも AIC が良いことを示した。

また、ピークと遺伝子の割り当ての方法についても検討した。1 つのピークが周辺遺伝子すべてが影響するか、遺伝子の近傍のピークのみが影響するのか、あるいは、近隣ピークを統合して割り当てるのが良いのか、転写開始点や終了点からのピークの位置を考慮するか、などの方法を比較した。その結果、複数ピークを距離に応じて、近隣の遺伝子に割り当てる方法が、構築した統計モデルで、mRNA 量を説明できることがわかった。

次に、物理化学モデルの構築を行った。まず、統計力学的モデルを構築した。標的遺伝子のプロモーター領域を格子化し、転写因子と基本転写因子の DNA 結合が格子あたり 1 つ結合すると仮定した。さらに、基本転写因子と転写因子は結合すると活性化すると仮定した。また抑制的に働く転写因子は基本転写因子との結合により、転写を抑制すると考えた。これらの 4 つの仮定の結合確率の分配関数からモデルを導出した。人工的に生成した RNA-seq と ChIP-seq のデータを利用し、このモデルへのあてはめを行うプログラムを実装した。

また、確率微分方程式を利用したモデルも構築した。このモデルは、基本転写因子の量、RNA の分解、転写因子結合、ランダム項からなる。転写因子結合項は、ヒル式として表現される。ランダム項は、ガウスノイズを仮定した。このモデルを実装し、最適化問題を解くアルゴリズムでパラメータ推定を行うソフトウェアを実装した。このモデルを、時系列 ChIP-seq と RNA-seq のデータに応用する準備をしているが、予定していた実験データが得られておらず、現在、データを公的データベースなどから収集しているところである。

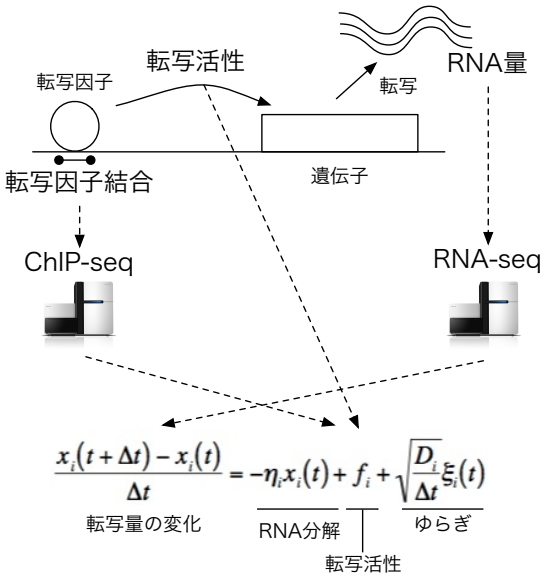


図 4. 転写の物理化学モデル

現在、これらの性能を評価し、論文を準備中である。今後は、DNA ループ構造や高次クロマチン構造をも取り込んだモデルの構築を進め、ChIP-seq や RNA-seq 以外のエピゲノムデータの統合も試みる。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

1. G. Morota, F. Peñagaricano, J. L. Petersen, D. C. Ciobanu, K. Tsuyuzaki and I. Nikaido. An application of MeSH enrichment analysis in livestock. Animal Genetics. 2015, Aug;46(4):381-7. 査読あり.
2. Koki Tsuyuzaki, Gota Morota, Takeru Nakazato, Satoru Miyazaki and Itoshi Nikaido. MeSH ORA framework: R/Bioconductor packages to support MeSH over-representation analysis. BMC Bioinformatics. 2015, Feb 15;16:45:1-17. 査読あり
3. Kenjiro Adachi*, Itoshi Nikaido*, Hiroshi Ohta, Satoshi Ohtsuka, Hiroki Ura, Teruhiko Wakayama, Hiroki R. Ueda & Hitoshi Niwa. Context-Dependent Wiring of Sox2 Regulatory Networks for Self-Renewal of Embryonic and Trophoblast Stem Cells. Molecular Cell. 2013, Nov 7;52(3):380-92. 査読あり. (*Equally contributions)

〔学会発表〕（計 8 件）

1. 二階堂愛. BioDevOps による再現性のあるDNAシーケンス解析環境の構築. NPO 並列生物情報処理イニシアティブシンポジウム. 産業技術研究所･茨城県つくば市. 2016/03/11. (招待講演)
2. 二階堂愛. 生命科学研究を加速するための AI を計測と計算の融合から考える. 第 2 回 理研・産総研共同シンポジウム. 理化学研究所大河内記念ホール・埼玉県和光市. 2016/02/03（招待講演）
3. Koki Tsuyuzaki, Gota Morota, Manabu Ishii, Takeru Nakazato, Satoru Miyazaki and Itoshi Nikaido. MeSH ORA : R/Bioconductor packages for performing MeSH over-representation analysis in model and non-model organism. Oct 29-30. 2015. IIBMP2015 (JSBi2015). 京都大学宇治キャンパスおうばくプラザ･京都府宇治市. (Poster)
4. 二階堂愛. 理研クラウドシステムとBioDevOpsによる再現性のあるシーケンスデータ解析. DDBJ 講習会. 2015/06/12. 科学技術振興機構・東京都. (招待講演)
5. 二階堂愛. R/Bioconductor によるデータ解析パッケージの開発. 第 18 回オープンバイオ研究会. 北陸先端科学技術大学院大学･石川県能美市. 2014/03/21（口頭発表）
6. 二階堂愛. 再現性のあるシーケンスデータ解析環境構築のための BioDevOps. 第54回人工知能学会分子生物情報研究会 (SIG-MBI). 北陸先端科学技術大学院大学･石川県能美市. 2014/03/21（口頭発表）
7. Itoshi NIKAIDO. Estimation of transcriptional activity from ChIP-Seq and transcriptome. Computational Science in Epigenetics. 13-14 Feb. 2014. RIKEN Center for Developmental biology. Kobe, Hyogo. Japan. (Invited Talk)
8. Gota Morota, Koki Tsuyuzaki, Manabu Ishii, Takeru Nakazato, Satoru Miyazaki and Itoshi Nikaido. MeSHR: R/Bioconductor package for finding statistically overrepresented MeSH terms in a set of genes. Annual Bioconductor Conference BioC 2013. July 18-19, Seattle, WA, US (Poster)

〔図書〕（計 3 件）

1. 二階堂愛. NGS データ解析に必要なコンピューティングの基礎. 次世代シークエンス解析スタンダード～NGS のポテンシャルを活かしきる WET&DRY. 羊土社. 2014. 46-59.
2. 二階堂愛. R と Bioconductor で ChIP-seq データを解析する. 二階堂愛（編）. 次世代シークエンス解析スタンダード～NGS のポテンシャルを活かしきる WET&DRY. 羊土社. 2014. 122-130.
3. 團野宏樹. 笹川洋平. 二階堂愛. メチル化結合タンパク質でDNAメチル化を定量する. 二階堂愛（編）. 次世代シークエンス解析スタンダード～NGS のポテンシャルを活かしきる WET&DRY. 羊土社. 2014. 156-164.
4. 荒引健, 石田基広, 高橋康介, 林真広, 二階堂愛. R 言語上級ハンドブック. シーアンドアール研究所. 2014. 518.

〔その他〕

ホームページ等

http://bit.riken.jp/

６．研究組織

(1)研究代表者
二階堂 愛　(NIKAIDO, Itoshi)
理化学研究所・情報基盤センター・ユニットリーダー
研究者番号：00383290