様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ５ 月 ２８ 日現在

機関番号：３２６１２
研究種目：若手研究
研究期間：2019～2020
課題番号：１９Ｋ１６１１２

研究課題名（和文）遺伝子数増大に耐える高精度遺伝子制御ネットワーク推定法の提案

研究課題名（英文）Development of a highly accurate method for inferring gene regulatory networks that can withstand an increase in the number of genes

研究代表者
山田　貴大（Yamada, Takahiro）

慶應義塾大学・理工学部（矢上）・助教

研究者番号：２０８３７７３６

交付決定額（研究期間全体）：（直接経費）　　1,000,000円

研究成果の概要（和文）：遺伝子発現量を網羅的に測定可能なRNA-seq技術の発展に伴い、様々な生命現象と遺伝子の間の関係性を明らかにする研究が行われてきた。これまでに時系列の遺伝子発現量から遺伝子間の制御関係を推定するネットワーク推定手法が提案されてきたが、従来手法では遺伝子制御ネットワークに見られるスパース性を考慮しないため偽陽性を多数検出するという問題が存在した。
そこで本研究では、スパース性を考慮したネットワーク推定手法を数理的に構築し、偽陽性を抑えたネットワーク推定手法の定式化を目指した。
その結果、スパースなネットワークの推定には成功したものの、発現変動が大きな遺伝子に起因する偽陽性の抑制には到れなかった。

研究成果の学術的意義や社会的意義
遺伝子制御ネットワークは生命現象の解明や創薬に繋がる貢献を果たしてきた。これを簡便に行うための遺伝子発現量からネットワークを推定する手法が提案されてきたが、多数の偽陽性を生み出すことから創薬などへの応用に限界が存在した。
本研究では、ネットワークが持つスパース性を考慮した偽陽性抑制法を提案し、スパースなネットワークの推定を可能とすることに成功した。さらに従来法で置かれていた遺伝子発現変動の大きさこそが制御関係に影響を持つという仮定が、偽陽性の温床となっていることを明らかにした。今後はこの仮定を除いた手法の考案により社会的実装に耐えうる方法論としてネットワーク推定を昇華できると期待される。

研究成果の概要（英文）：With the development of RNA-seq technology, which enables the comprehensive measurement of gene expression levels, research has been conducted to elucidate the relationships between genes and various biological phenomena. However, the conventional methods do not take into account the sparsity of gene regulatory networks, which results in the detection of many false positives.
In this study, a network inference method that takes sparsity into account was mathematically constructed and aimed to formulate a network inference method to suppress false positives.
As a result, inferring a sparse network was succeeded, but it was failed to suppress false positives caused by genes with large expression variation.

研究分野：システム生物学

キーワード：遺伝子制御ネットワーク　ネットワーク推定　L1正則化　Omicsデータ解析

１．研究開始当初の背景

次世代シーケンサを用いた網羅的遺伝子発現測定技術である RNA-seq は今日の生命科学研究において必須の技術となっている．当該技術を用いて様々な生命現象，疾患に対する遺伝子発現測定が行われ，主にサンプル間で発現が有意に変化する遺伝子を検出する発現変動遺伝子(Differentially Expressed Gene : DEG)解析によりこれら生命現象，疾患に関与する遺伝子の同定が広く行われている．DEG 解析は関連遺伝子の同定においては優れた効力を発揮するものの，近年では多くの生命現象，疾患が多数の遺伝子間における複雑な制御関係により構成されることが報告されてきたことから，このような事象を捉える上では DEG 解析だけでは RNA-seq の下流解析としては不十分となってきた．

これに対して，対象とする生命現象や疾患における時系列での RNA-seq データを取得し，多数の遺伝子間の制御関係をデータ解析的に導く手法であるネットワーク推定の研究が近年広く提案されてきた[1]．多くのネットワーク推定手法は各遺伝子の時間的発現変化の前後関係を元に，ある遺伝子が発現変動を示した後に別の遺伝子が発現変動を示す場合に前者の遺伝子から後者の遺伝子に制御関係があると予測することで首尾一貫している．ネットワーク推定のベンチマークである DREAM4[2]などのデータセットを用いたこれら手法の精度評価により少数の遺伝子を対象とした問題 (10 遺伝子程度)においては高精度を達成することが報告されてきた一方で，多数の遺伝子を対象とした問題 (100 遺伝子程度)においては本来制御関係がないはずの箇所に制御関係を見出してしまう偽陽性過多により利用可能性が大幅に制限され，網羅的遺伝子発現測定が可能な RNA-seq 技術との相性が極めて悪いものとなっていた[1]．これは，ネットワーク推定においては全遺伝子からほか全ての遺伝子への制御関係があることを前提とした推定によりあらゆる箇所に制御関係が見出される潜在性を有する一方で，現実に見られる遺伝子制御ネットワークにおいては候補制御関係数に対する実制御関係数が極端に低い (公知の大腸菌における遺伝子制御ネットワークでは候補制御関係数に対する実制御関係数の比率はたった 0.14%)スパース性を有する[3]という，推定手法の前提と現実との乖離が主な要因であると推測された．

そこで本研究では，このスパース性を考慮した遺伝子制御ネットワーク推定アルゴリズムの構築により，既存手法が抱える偽陽性を高度に抑制することを目指し，RNA-seq の下流解析手法としてネットワーク推定が現実的に利用可能な方法論とすることを目標とした．

２．研究の目的

本研究では従来のネットワーク推定法において無視されていた遺伝子制御ネットワークの持つスパース性を考慮した推定手法を数理的に設計，構築することによりスパースなネットワークを推定し，偽陽性を抑えることのできるアルゴリズムの開発を目指した．スパース性を考慮させるためには，従来の推定手法における遺伝子発現の前後関係を検証する回帰の過程において，十分に回帰精度を上げない遺伝子に対する罰則を課す正則化手法の数理的誘導が必要である．これを行い，プログラムとして実装することで多数の遺伝子を対象としたベンチマークデータセットに対して従来法と比較して偽陽性の抑制を行うことを目的とした．

３．研究の方法

(1) ネットワーク推定における制御関係に対する正則化手法の考案と実装

ネットワーク推定における主なプロセスは以下の 2 つに大きく分けることができる．

① 制御対象遺伝子の時系列発現変動の候補制御遺伝子の時系列発現変動に対する回帰

② ①で構築した回帰モデルに基づく候補制御遺伝子からの実制御遺伝子の検出

①の過程では以下の時系列モデルを元に回帰を行う．

$$e_t^i = f_i\left(e_{t-1}^{-i}\right) + \varepsilon_{t-1}$$

ここで，e_t^iは時刻tでの制御対象遺伝子iの発現量，f_iは制御対象遺伝子iの発現量を，制御対象遺伝子i以外の遺伝子全ての発現量e_{t-1}^{-i}を入力として説明することができる関数，ε_{t-1}は制御対象遺伝子iの$t-1$時点での発現のばらつきを示す．ここでf_iは陽には書き下すことのできない遺伝子制御を表した関数であり，少数遺伝子に対するネットワーク推定において高精度を達成した JUMP3[1]や BiXGBoost[4]などではこの関数を学習アルゴリズムの一つである回帰木を用いて推定する．この回帰木の学習は以下の誤差関数Lを用い，誤差関数を最小化する関数内パラメータwを推定することにより行われる．

$$L(w) = \sum_i \frac{1}{2}(\hat{e}_t^i(w) - e_t^i)^2$$

ここで$\hat{e}_t^i(w)$はe_{t-1}^{-i}を入力としてパラメータwを持つ回帰木を用いて推定したe_t^iの予測値である．

学習においては，$L(w)$を与えられたwに対する二次関数として近似し，平方完成により$L(w)$を極小化するwを求めることを逐次的に行うことで$L(w)$ を最小化し学習を進める．学習が終了し，十分にe_t^iを予測することが可能になった回帰木を用いて②のプロセスとして$\hat{e}_t^i(w)$への影響の大きい遺伝子を Mean Decrease Impurity (MDI)[5]をスコアとして選出することで遺伝子iを制御する遺伝子を検出する．ここでスパースな遺伝子制御ネットワーク推定を可能とするためには，①のプロセスにおける誤差関数Lを以下のように変更する必要がある．

$$L(w) = \sum_i \frac{1}{2}(\hat{e}_t^i(w) - e_t^i)^2 + \lambda|\theta(w)|_1$$

$\theta(w)$は遺伝子iに対する各遺伝子からの制御関係を示す②より得られる MDI 指標の値であり，制御関係の有無を表すスコアに対応する．また，λは調整パラメータ，$|.|_1$は L1 ノルムを表す．これは制御関係のスコアを上げるほど誤差関数$L(w)$の値が増加し，容易に制御関係が検出されないことによりスパースなネットワークの推定を可能とすることができる．しかしながら，$|\theta(w)|_1$は MDI に由来することから遺伝子発現量の分散などを含むwに対して複雑な関数であり，これを陽に$L(w)$最小化を行うためのwの更新式に導くことは困難である．そこで，本研究では回帰木による$\hat{e}_t^i(w)$の逐次的な予測過程において，予測に用いられる遺伝子の選択基準に逐次的に求めた$\theta(w)$を加算することで遺伝子発現変動の回帰に強く寄与する遺伝子を優先的に選択するように変更することで，上記の過程を近似したアルゴリズムを構築した．従来のネットワーク推定では①のプロセスを完全に終了した後に②のプロセスに移りネットワーク推定を行うことに対して，本アルゴリズムでは①のプロセスにおける逐次的な過程で②を行い，その都度回帰に寄与する，すなわち遺伝子制御に有望な候補制御遺伝子の情報を①の過程にフィードバックすることで，有望な候補制御遺伝子を選出しスパースなネットワーク推定を可能とする方法と言うことができる．

上記で構築したアルゴリズムをプログラムとして実装した．この際に，ネットワーク推定において高精度を示す BiXGBoost の例を参考に，回帰木のフレームワークである XGBoost[6]の Plugin(XGBOOST_REGISTER_TREE_UPDATER, https://github.com/dmlc/xgboost/tree/master/plugin)として上記のアルゴリズムを実装した．

(2) 当該アルゴリズムの精度評価

本アルゴリズムの精度評価にあたり，まずデータセットとしてはベンチマークデータセットである DREAM4 In Silico Challenge にて公開されている遺伝子数 10, 100 で構成されるそれぞれ 5 つの時系列遺伝子発現量とネットワーク構造を用いた．この時系列遺伝子発現量を本アルゴリズムの入力として与え，出力された遺伝子制御ネットワークを用いて 1) 出力された遺伝子制御ネットワークのスパース性，2) 正解として用意されている遺伝子制御ネットワークと比較した際の偽陽性の割合を評価した．1)については推定された遺伝子制御ネットワークの各制御関係のスコアに基づき特定の制御関係のみを重視できているかを評価するためにシャノンエントロピーを算出した(図 1)．2)についてはスコアに応じた平均偽陽性率である Area Under Precision Recall (AUPR)を算出した．また同一のデータセットを用いて先行研究で少数の遺伝子に対して高精度を達成した BiXGBoost にて 1), 2)を行い比較評価を行った．

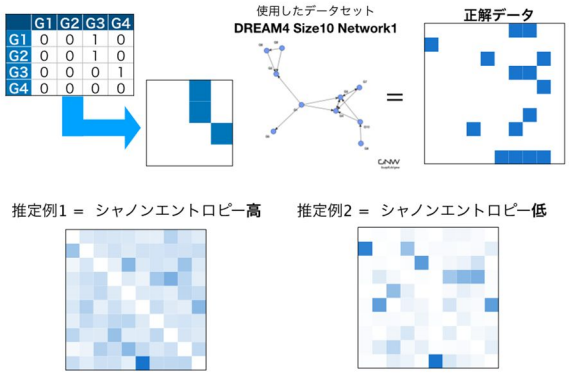

図 1 ネットワーク推定におけるシャノンエントロピーによるスパース性の評価に関する概念図．(上部)ネットワークを接続関係に対するヒートマップとして表した概念図．(下部) 推定したネットワークが密な場合にはシャノンエントロピーが高く(左図), スパースな場合には低くなる(右図)

4．研究成果

(1) スパースなネットワーク出力可否の検証

スパースな遺伝子制御ネットワークが推定できているかを評価するためには，特定の制御関係のみに対して高いスコアを算出できているかを評価すれば良い．これを評価するために本研究では推定された遺伝子制御ネットワークのスコアに対してシャノンエントロピーを算出した．シャノンエントロピーは事象の起こりやすさの乱雑さを示す指標であり，今回の場合，特定の制御関係が特異にスコアが高い場合にシャノンエントロピーは低くなる，すなわちスパースなネットワークを推定できていると

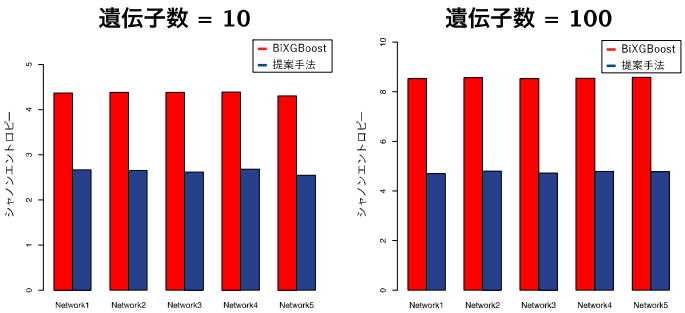

図 2 : 提案手法と従来推定手法(BiXGBoost)によるネットワーク推定結果に対するシャノンエントロピーの評価結果

言い換えることができる．本アルゴリズムと BiXGBoost によって推定した遺伝子制御ネットワークのスコアに対してシャノンエントロピーを算出したところ，本アルゴリズムにより推定された遺伝子制御ネットワークは 10, 100 遺伝子を対象とした問題共に顕著に低いシャノンエントロピーを示すことから，本アルゴリズムによってスパースなネットワークの推定に成功したと言える (図 2).

(2) 偽陽性抑制効果の検証

このスパースなネットワーク推定の達成によって偽陽性の抑制が達成されたかを検証するために，ベンチマークデータセットに含まれる正解の遺伝子制御ネットワークと推定結果を比較することで AUPR を算出した．AUPR は真の制御関係に対するスコアをより高く，本来制御関係がない箇所に対するスコアをより低くするというように，スコアの大小関係とネットワークの制御関係がより対応するほどに高い値を示し，偽陽性の発生度合いを評価することができる．本アルゴリズムと BiXGBoost で AUPR を評価した結果，AUPR の増加は見られず，顕著な偽陽性抑制効果は得られなかった (図 3).

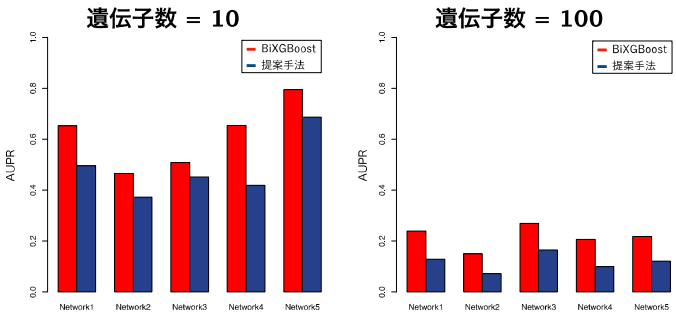

図 3 提案手法と従来推定手法(BiXGBoost)によるネットワーク推定結果に対する AUPR の評価結果

(3)偽陽性発生原因の検証結果

遺伝子制御ネットワークが持つスパース性を再現するネットワークの推定が可能になったにも関わらず偽陽性を抑制できなかったことを鑑みて，今回偽陽性として検出された遺伝子間の時系列発現変動を確認することで，真陽性と偽陽性の間の関係を確認した (図 4).

その結果，提案手法では候補制御遺伝子の絶対的な遺伝子発現の変動幅が大きいものを選択的に制御関係として推定する傾向があることが見出された. 一方で，制御対象遺伝子に対する制御を行う遺伝子の発現変動は分子生物学的には関係がなく (制御対象遺伝子の発現変動は制御を行う遺伝子の発現量ではなくプロモータの活性度合いによって決まるため), ネットワーク推定法に求められるものは時間的な発現変動のタイミングのずれによるものである. すなわち，本アルゴリズムにおける遺伝子発現変動の絶対的な変動幅による過度な制御関係の選出を抑え，時間的な発現変動の前後関係により注目するようにアルゴリズムを改良することによって，偽陽性の抑制が可能になると考えられる.

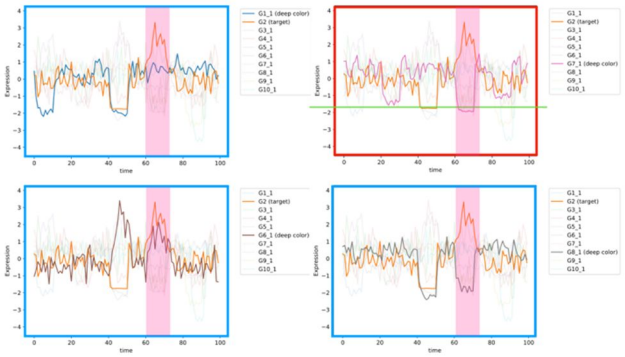

図 4 制御対象遺伝子 (橙)と候補制御遺伝子の時系列発現変動の可視化結果．赤枠および青枠の結果はそれぞれ本来制御関係がない関係において提案手法で高いスコアを算出した結果，および制御関係がある関係において相対的に低いスコアを算出した結果を示す．ピンク箇所において制御対象遺伝子が最も発現変動を示す時刻で最も発現変動の大きな遺伝子のスコアを過大評価する傾向がある.

<引用文献>

[1] Huynh-Thu, Vân Anh, and Guido Sanguinetti. *Bioinformatics* 31.10 (2015): 1614-1622.
[2] Marbach, Daniel, et al. *Nature methods* 9.8 (2012): 796-804.
[3] Huerta, Araceli M., et al. *Nucleic acids research* 26.1 (1998): 55-59.
[4] Zheng, Ruiqing, et al. *Bioinformatics* 35.11 (2019): 1893-1900.
[5] Louppe, Gilles, et al. *Advances in neural information processing systems* 26 (2013).
[6] Chen, Tianqi, and Carlos Guestrin. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016.

5．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　1件／うち国際学会　0件）

1．発表者名 山田 貴大
2．発表標題 限界を超えた能力を生命に付与したい ー設計図志向な生物学を目指してー
3．学会等名 The 40th Scienc-ome（招待講演）
4．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

—

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	舟橋　啓 (Funahashi Akira) (70324548)	慶應義塾大学・理工学部・准教授 (32612)	
研究協力者	比企　佑介 (Hiki Yusuke)	慶應義塾大学・理工学部・修士課程学生 (32612)	
研究協力者	牧野　荘太 (Makino Sota)	慶應義塾大学・理工学部・修士課程学生 (32612)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関