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

平成　３０　年　　５　月　２３　日現在

機関番号：　１４４０１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００４０１

研究課題名（和文）構造情報の粗視化による高速ゲノムワイドRNA遺伝子発見

研究課題名（英文）Fast genome-wide detection of RNA genes using coarse-grained structural information

研究代表者

加藤　有己（KATO, Yuki）

大阪大学・医学系研究科・助教

研究者番号：１０５１１２８０

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、未だ不完全な非コードRNAのゲノムアノテーションのため、RNAの形成し得る構造情報を粗視化し比較することで、２つの生物のゲノム配列に共通して存在する構造RNAの遺伝子を検出する計算手法を開発した。シミュレーションデータおよび実際のヒトとマウスの染色体配列を用いた計算機実験により、提案手法は遺伝子検出に関して高い感度を達成することが示された。これにより、提案手法は構造RNAの比較ゲノムスクリーニングとして有用なツールであると考えられる。

研究成果の概要（英文）：For the purpose of incomplete genomic annotation of non-coding RNAs, here we presented a computational method for detecting structured RNA genes that two organisms share in common in their genomic sequences, which aims to compare their coarse-grained structures. Computational experiments with both simulated and authentic chromosomal data indicate that the proposed method achieves high sensitivity for the detection of structured RNA genes. We conclude that our approach can be useful as a comparative genomic pre-screening of structured RNAs.

研究分野：　バイオインフォマティクス

キーワード：　RNA　遺伝子発見　２次構造

１．研究開始当初の背景

タンパク質に翻訳されず遺伝子発現を調節するなどの役割を持つ非コード RNA (non-coding RNA) の機能に注目が集まっている。非コード RNA の多くはその機能を発揮する際、何らかの構造を形成する場合が多く、構造と機能には強い相関関係があると考えられている。主に塩基の A と U、G と C といった相補塩基対から構成される折り畳み構造（２次構造）は立体構造の骨格を形成し、計算機による解析が比較的容易である。そのため、機能解析の足掛かりとして、RNA 配列解析における研究の初期の頃は多くの２次構造予測法が競って開発された。近年では、高性能な２次構造予測法と利用可能なゲノム情報の増加に伴って、ゲノム配列上で転写される非コード RNA（以下 RNA と略す）の有力なシグナルと考えられる２次構造情報を探索することで RNA 遺伝子を発見する試みがなされている。

初期の RNA２次構造予測法では、単一の RNA 配列が取り得る２次構造の自由エネルギーが最小のときに安定状態になると考え、単一配列から最小自由エネルギー構造を計算することを目的としていた。後に、複数の RNA 配列上で２次構造の保存に着目した比較配列解析による手法、すなわち共通２次構造を求める手法が構造予測の精度向上につながることが示され、今日では比較解析がゲノム上の RNA 遺伝子探索において重要な役割を果たすものと考えられている。ゲノム上で比較配列解析を行う場合、最初から「厳密な」２次構造を考慮して領域同士の対応付け（アラインメント）を行うのは計算量の観点で困難であることが知られている。そこでこれまでは、ゲノム配列において局所的な配列類似度で相同配列群を検索した後、得られた領域間の共通２次構造を予測し、そのエネルギーなどを評価して遺伝子候補を選別することが行われてきた。ただし、この手法では最初に構造情報を考慮していないため、配列としての類似度は低いが構造レベルでは類似する領域を見逃してしまう。RNA は進化の過程で構造は保存されやすいが配列は変異しやすい傾向にあるため、この問題は特に考慮すべき点である。そのため、各ゲノム配列において「粗い」構造を考慮して RNA 遺伝子候補領域を絞り込み、次に配列間にまたがる領域同士を、構造を考慮して高速に比較する必要がある。

本研究課題申請時以前、申請者は RNA の２次構造予測問題に端を発し、２つの RNA 間に起こる相互作用を予測する問題や、RNA の配列と構造を同時にアラインメントする問題に対し、最適化手法の一種である整数計画法による巧妙な数理モデル化を行い、いずれも当該分野最高水準の予測精度と非常に高速な計算速度を達成した経緯がある。ここでは、２次構造１つのみを考慮するのではなく、可能な２次構造全体の集合（アンサンブル）を考慮して予測精度を向上させている。具体的には、RNA のアンサンブル上で任意の２塩基が塩基対を形成する確率（塩基対確率）を考え、目的関数に導入して最適化（最大化）を行った。

本研究では、ゲノム配列において、効率良く算出した塩基対確率とウインドウ走査を利用して、定量的に RNA 遺伝子候補領域を推定することを試みる。次に、推定した領域集合で構造を考慮したローカルアラインメントを行う必要があるが、ゲノム網羅的に実行するためにはある種の大胆な近似が必要である。そこで、塩基対確率の情報を、整合性を保ちながら文字列情報に圧縮し、それら文字列間のアラインメントスコアを組み込んだときのローカルアラインメントを行うことで、計算量を削減でき、ゲノム配列にも十分適用できるのではないかと考えた。

２．研究の目的

本研究では、粗視化された構造情報をもとに、初めから構造を考慮して高速にゲノム網羅的 RNA 配列比較を行うことで、従来法よりも高い精度で新規 RNA 遺伝子の発見を行うことを目指す。本研究課題の申請時における当初の研究目的は以下の通りである。

(1) 塩基対確率の粗視化による高速 RNA 配列比較法の開発

(2) ２次構造を考慮したゲノム網羅的ペアワイズアラインメント

(3) ゲノム網羅的マルチプルアラインメントへの拡張と遺伝子機能推定

３．研究の方法

(1) ２つの RNA 配列が与えられたとき、構造類似度スコアを高速に計算するアルゴリズムを開発する。ここでは、２次構造比較の計算量の削減のため、各配列において２次元の構造情報（塩基対確率行列）を１次元の文字列情報（２進列）に圧縮し、２本の文字列間のアラインメントを行うことで類似度スコアを算出する。開発アルゴリズムは C++言語を用いて実装する。なお、塩基対確率の計算は ViennaRNA パッケージ（Lorenz *et al.*, *Algorithms Mol. Biol.*, 2011）を用いることで対応可能である。その後、粗視化で用いる閾値や、グローバルアラインメントで用いるパラメータを、公共データベースの１つである Rfam に登録されている実際の RNA２次構造データを用いて調節する。

次に、RNA 配列が与えられるという仮定を取り除き、既知の RNA 遺伝子を含むゲノム領域と、そのシャッフルゲノム領域を人為的に作成する。次に、それらを組み合わせた配列の対において、固定長の推定 RNA 領域（ウインドウ）を動かしながら、構造粗視化に基づく類似度スコアの評価を行う。このとき、なるべく多くの既知の類似遺伝子間のアラインメントスコアと、シャッフル配列とのアラインメントスコアが区別できるようなスコアの閾値を調節する。

(2) 各ゲノム配列にて、ウインドウサイズを固定したときのウインドウ内外での局所的な塩基対確率をもとに、RNA 遺伝子候補を計算する。具体的に、ウインドウ内部での塩基対確率の和が、ウインドウの内外にわたる塩基対確率の和よりも大きくなる領域をゲノム網羅的に探索する。なお、ゲノム全体にわたって計算するため、類似の計算を一度にまとめるなどの、実装上の高速化の工夫をする。また、既知 RNA 遺伝子と比較し、ウインドウサイズなどを最適化する。

次に、上記で得られた遺伝子候補と構造粗視化によるスコア付け体系を、ローカルアラインメントアルゴリズムに組み込むことで、ゲノム網羅的ペアワイズローカルアラインメント法を開発する。その後、既存研究で適用された実際のヒトとマウスのゲノム配列上で開発手法によるローカルアラインメントを行い、有効性の検証ならびに問題点の検出を行う。

(3) 上記で得られたペアワイズアラインメントを３本以上の配列でも扱えるように拡張する。このとき、ゲノムマルチプルアラインメント法 MultiZ (Blanchette *et al.*, *Genome Res.*, 2004) を応用し、本研究で得られた構造局所領域について、ゲノム上での出現順序を考慮したブロックアラインメントを計算することで、マルチプルローカルアラインメントを構築する。次に、既存手法で適用された生物ゲノムである、ヒト、チンパンジー、イヌ、マウス、ラット、ニワトリ、ゼブラフィッシュ、フグの８種に適用し、既存手法と比較した場合の検出遺伝子の差異に関する考察と、遺伝子機能推定を行う。

４．研究成果

比較構造情報に基づくゲノムワイドな構造 RNA 領域（遺伝子）発見のための高速かつ高感度なアルゴリズム DotcodeR（DOT plot enCODEr for RNA）を開発した。先述の研究開始当初の背景にあるように、２つの RNA の構造を厳密に比較することは時間計算量を相当要するため、ゲノムワイドに実施するためには、あらかじめ対象配列間でローカルアラインメントを実行し、候補領域を限定する必要があった。そこで本研究では、ゲノムワイドな構造比較を可能にするため、厳密性を犠牲にし、各推定 RNA 領域の構造情報を２進列に変換した後、それらの内積を計算するアプローチを採用した。提案手法の高速性を利用することにより、従来広く利用されてきた配列アラインメントを実施することなく、対象領域間での全対全比較が可能となっている。換言すれば、これまで発見されなかった RNA 遺伝子を新規に発見できる可能性があるといえる。

ベンチマークデータとして既知の RNA ファミリーから作製された仮想ゲノム配列の組に対して、図１の ROC（receiver operating characteristic）曲線が示すように、DotcodeR は高い予測性能を示した。次に、ヒト 21 番染色体とマウス 19 番染色体の組に対し、DotcodeR を網羅的に実行した結果、既知の構造 RNA 遺伝子を平均して 83%の感度で検出し、探索空間は 97%ほど縮小することに成功した。

以上により、粗視化構造情報を利用した DotcodeR は、構造 RNA の比較ゲノムスクリーニングにおいて、プレフィルターの役割として有用であると結論付けることができる。すなわち、さらなる遺伝子候補選択のための、時間コストが大きい高精細の比較構造解析を行うための入力として利用することが可能である。今回は時間の都合上、ヒト、マウス以外の生物での計算機実験を行うことができなかったが、進化的に保存されている領域を持つ生物であれば、類似の結果が得られる可能性が高いものと考える。なお、本研究の手法を世界に向けて広く発信するため、DotcodeR のプログラムを公開している (https://github.com/ykat0/dotcoder)。

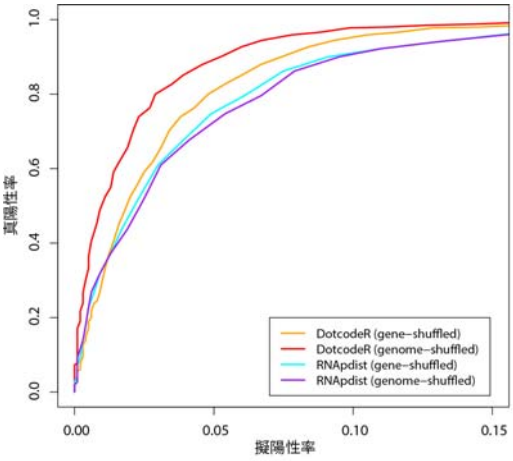

図1　仮想ゲノム配列上での予測性能を表すROC曲線

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] Yuki Kato*, Jan Gorodkin and Jakob Hull Havgaard*, “Alignment-free comparative genomic screen for structured RNAs using coarse-grained secondary structure dot plots,” ***BMC Genomics***, vol. 18, 935, Dec. 2017, 査読有 (*: corresponding author).
DOI: 10.1186/s12864-017-4309-y

[2] Yuki Kato*, Tomoya Mori, Kengo Sato, Shingo Maegawa*, Hiroshi Hosokawa and Tatsuya Akutsu, “An accessibility-incorporated method for accurate prediction of RNA–RNA interactions from sequence data,” ***Bioinformatics***, vol. 33, no. 2, pp. 202–209, Jan. 2017, 査読有 (*: corresponding author).
DOI: 10.1093/bioinformatics/btw603

[3] Michiaki Hamada, Yukiteru Ono,

Hisanori Kiryu, Kengo Sato, Yuki Kato, Tsukasa Fukunaga, Ryota Mori and Kiyoshi Asai, “Rtools: a web server for various secondary structural analyses on single RNA sequences,” ***Nucleic Acids Research***, vol. 44, Web Server issue, pp. W302–W307, Jul. 2016, 査読有.
DOI: 10.1093/nar/gkw337

〔学会発表〕（計 13 件）

[1] 新村 啓介, 加藤 有己, 河原 行郎, 「ビット列のソートに基づくリード直接比較による高速省メモリゲノム構造変異解析」, 第 53 回情報処理学会バイオ情報学研究会, 2018-BIO-53 (12), 石川県能美市, Mar. 2018.

[2] 蒲原 純, 加藤 有己, 河原 行郎, 「スペクトラルグラフ理論を応用した 1 細胞 RNA-seq データのトップダウン型クラスタリング」, 第 53 回情報処理学会バイオ情報学研究会, 2018-BIO-53 (11), 石川県能美市, Mar. 2018.

[3] Yuki Kato, Jan Gorodkin and Jakob Hull Havgaard, “Fast and efficient alignment-free comparative genomic screen for structured RNAs with DotcodeR,” The 25th Conference on Intelligent Systems for Molecular Biology and the 16th European Conference on Computational Biology (ISMB/ECCB2017), Poster, A-215, Prague, Czech Republic, Jul. 2017.

[4] Yuki Kato, Tomoya Mori, Kengo Sato, Shingo Maegawa, Hiroshi Hosokawa and Tatsuya Akutsu, “Accurate prediction of RNA–RNA interactions from sequence data incorporating interaction site accessibility,” The 2016 joint annual meeting of the RNA Society and the RNA Society of Japan (RNA2016), Poster, 349, Kyoto, Japan, Jun. 2016.

[5] Michiaki Hamada, Yukiteru Ono, Hisanori Kiryu, Kengo Sato, Yuki Kato, Tsukasa Fukunaga, Ryota Mori and Kiyoshi Asai, “Rtools: a web server for various secondary structural analyses on single RNA sequences,” The 2016 joint annual meeting of the RNA Society and the RNA Society of Japan (RNA2016), Poster, 342, Kyoto, Japan, Jun. 2016.

[6] Yuki Kato and Wataru Fujibuchi, “A computational method for dissecting single-cell populations based on transcriptomic data,” CiRA/ISSCR 2016 International Symposium, Poster, P1-003, Kyoto, Japan, Mar. 2016.

[7] Takamasa Okanishi, Yuki Kato, Hiroyuki Fujimura, Yoshikatsu Nakano, Shoichiro Suda, Hideko Sone and Wataru Fujibuchi, “Network analysis of marine environmental factors based on MCMC sampling,” 生命医薬情報学連合大会 2015 年大会, Poster, P57, 京都府宇治市, Oct. 2015.

[8] Kenta Kobayashi, Midori Yuji, Yuki Kato, Takeaki Taniguchi, Toru Maruyama, Michihiro Ito, Susumu Goto, Haruko Takeyama and Wataru Fujibuchi, “Development of a pipeline for analysis of meta- and single-cell genomic sequences,” 生命医薬情報学連合大会 2015 年大会, Poster, P51, 京都府宇治市, Oct. 2015.

[9] Yuki Kato, Jakob Hull Havgaard and Jan Gorodkin, “Genomic screen for structured RNAs using coarse-grained dot plots,” 生命医薬情報学連合大会 2015 年大会, Poster/Lightning Talk, P37 (F04), 京都府宇治市, Oct. 2015.

[10] Yuki Kato, Jakob Hull Havgaard and Jan Gorodkin, “Genomic screen for structured RNAs using coarse-grained dot plots,” The First Joint Conference between GIW and InCoB (GIW/InCoB2015), Poster, 159, Tokyo, Japan, Sep. 2015.

[11] Yuki Kato, “De novo prediction of structured RNAs from genomes using coarse-grained dot plots,” Talk at International Workshop on Computational Analysis of RNA Structure and Function, Benasque, Spain, Jul. 2015.

[12] 小林 健太, 加藤 有己, 谷口 丈晃, 丸山 徹, 伊藤 通浩, 五斗 進, 竹山 春子, 藤渕 航, 「海洋環境解析に向けたメタゲノムおよび 1 細胞配列データ解析用パイプラインの開発」, 第 42 回情報処理学会バイオ情報学研究会, 2015-BIO-42 (58), 沖縄県国頭郡恩納村, Jun. 2015.

[13] 岡西 孝真, 加藤 有己, 藤村 弘行, 中野 義勝, 須田 彰一郎, 曽根 秀子, 藤渕 航, 「MCMC サンプリングに基づく海洋環境因子ネットワーク解析」, 第 42 回情報処理学会バイオ情報学研究会, 2015-BIO-42 (35), 沖縄県国頭郡恩納村, Jun. 2015.

〔その他〕

ホームページ等

https://github.com/ykat0/dotcoder

６．研究組織

(1)研究代表者

加藤　有己（KATO, Yuki）

大阪大学・大学院医学系研究科・助教

研究者番号：10511280