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　２７　年　　６　月　　４　日現在

機関番号：　１４３０１
研究種目：　若手研究(B)
研究期間：　2012～2014
課題番号：　２４７００２９６

研究課題名（和文）高次構造を考慮した高速ＲＮＡ間相互作用予測

研究課題名（英文）Fast RNA-RNA interaction prediction considering higher order structure

研究代表者
加藤　有己（Kato, Yuki）

京都大学・iPS細胞研究所・特定拠点助教

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、現実に現れる様々な組み合わせ最適化問題を定式化可能な整数計画法を用いることで、高速かつ高精度なRNA間相互作用予測モデルの開発を行った。ここで、分子間相互作用に関与し得る確率を表すアクセシビリティの概念を整数計画問題の定式化に巧妙に導入することにより、アクセシビリティを考慮しない従来モデルよりも大きく予測精度を向上させることができた。提案手法はRNA間の結合部位を塩基対解像度レベルで従来手法よりも高精度に予測可能であるため、今後制御RNAの機能解析などに大きく貢献することが期待される。

研究成果の概要（英文）：In this study, we developed a fast and accurate prediction model for RNA-RNA interactions using integer programming, which can formulate various combinatorial problems appearing in the actual world. Here we incorporated accessibility, probability of being involved in interaction, into the formulation of the integer programming problem, achieving much better predictive performance compared with existing methods that do not consider accessibility. As our proposed method can predict interaction sites in base pair resolution more accurately than the earlier methods can do, this work is expected to greatly contribute to functional analysis of regulatory RNAs.

研究分野：　バイオインフォマティクス

キーワード：　RNA間相互作用　アクセシビリティ　数理計画法

１．研究開始当初の背景

タンパク質へ翻訳されず、遺伝子発現を制御するなどの能動的な役割を持つ RNA（リボ核酸）が注目されている。RNA は多くの場合他の RNA と相互作用（結合）することで何らかの制御機能を発揮することが知られており、RNA の機能解明のための相互作用予測は生命情報科学における重要な課題の１つである。A-U、G-C といった正準塩基対から成る２次構造は立体構造の骨格を形成し、情報科学的観点からのモデル化が比較的容易である。そのため、これまで計算機を用いて２本の RNA 塩基配列から２次構造に基づく相互作用構造または結合部位のみを予測する研究がいくつか行われてきた。中でも小型 RNA と呼ばれるクラスでは、それ自体も内部で２次構造を形成し局所的な配列相補性に基づいて相互作用することが多いため、その正確な予測が難しく計算速度も増加することが分かっている。２次構造に基づく従来法の問題点として、相互作用に関する偽陽性が多い点が挙げられる。あまり行われていない相互作用未知の標的 RNA 分子の網羅的探索では、相互作用部位の正確な予測が特に重要である。また、標的探索は RNA の機能推定を目指す RNA 間相互作用ネットワーク構造の解明につながるという点でも重要である。要約すれば、高精度の相互作用予測および高信頼度のネットワーク推定のためには、２次構造以上のより精緻な情報を考慮して解析することが必要であると考えられる。

RNA の立体構造では塩基の Watson–Crick 端間での正準塩基対のみならず、Hoogsteen 端や Sugar 端での相互作用から成る非正準塩基対も観測され、さらには塩基対間の結合に方向（シス・トランス）の違いがある。つまり、塩基の３種類の相互作用端と塩基間の結合の方向の組み合わせで塩基対は１２種類存在し、入力配列から１２種類全ての塩基対を考慮して塩基対構造（拡張２次構造）を予測することは高次構造予測に向けての重要なステップとなる。

本研究課題申請時以前、申請者は RNA の結合２次構造予測を、精度の期待値を最大化することを目的とした整数計画法で実現することで、圧倒的に高速な計算速度と高い予測精度をもつソフトウェアを開発した経緯がある。整数計画法とは、整数値をとる変数の線形不等式系で表された制約条件の下で、目的関数と呼ばれる線形関数を最大化（あるいは最小化）する問題に定式化して解く方法で、そのモデル記述能力は極めて高い。ここでは、精度の期待値を求めるための結合２次構造全体上で定義された確率分布を直接扱うのではなく、より小単位の構造の確率分布に分解し積近似を行うことで計算量を削減している。さらに、期待精度最大化に貢献し得る解のみを考慮することで、問題のサイズを劇的に減らすことに成功している。そこで、解析に高いコストを要すると思われる拡張２次構造空間で定義される確率分布を、解析が容易な正準塩基対に対する確率分布と非正準塩基対に対する確率分布にうまく分解できれば、高いモデル化能力と柔軟性をあわせもつ整数計画法により期待精度最大化を行うことで、高速高精度な結合拡張２次構造予測法が実現できるのではないかとの着想に至った。さらに、開発手法の高速性を生かすことで、ゲノムワイドな標的 RNA 探索に適用できると考えている。

２．研究の目的

本研究課題の申請時における研究目的は以下の通りである。

⑴ 拡張２次構造の確率分布を解析が容易な分布へ変換し、既知のデータから学習し最適化する。

⑵ 結合拡張２次構造予測問題を解くための適切な整数計画モデルを設計し、計算機に実装する。

⑶ 開発手法を真正細菌の小型 RNA を中心に標的 RNA 探索に適用し、既存の RNA 間相互作用ネットワークと比較して考察を行う。

３．研究の方法

⑴ 結合拡張２次構造の期待精度が最大となるような解を求める問題を整数計画問題として定式化する。期待精度とは、塩基対単位で評価する予測精度を向上させるようにゲイン関数を設計したときの、可能な構造全体にわたるゲイン関数の期待値を指す。このとき、変数の個数や制約式の個数といった問題のサイズがなるべく小さくなるような定式化を考える。また同時に、モデルにフィードバックする情報があるかどうかを考察するため、得られた整数計画問題の数学的な構造についても調べる。

次に、拡張２次構造の確率分布を正準塩基対の確率分布と非正準塩基対の確率分布に分解する。正準塩基対の確率分布は既存研究で開発されている RNAfold などを利用することで対応できる。一方、各々の非正準塩基対について、PDB（Protein Data Bank）データベースから得られる最新のデータセットを利用して、新たに機械学習モデルで学習することで確率分布を決定する。

⑵ 先述の整数計画法を C++言語を用いて計算機に実装する。なお、実際に整数計画問題を解くために、無償のソフトウェア GLPK（GNU Linear Programming Kit）を用いる。GLPK は比較的サイズの小さな問題しか解けないが、本課題では期待精度最大化に貢献し得る解のみを考慮するため、当該問題のサイズが相当小さくなることが予想される。その後、立体構造が実験的に決定されている RNA 複合体を PDB データベースから取り出し、拡張２次構造形式に変換することで、同じ RNA のペアに対する予測結合構造と比較し精度を評価する。また同時に、開発手法の計算速度と有効性の検証、ならびに問題点の検出を行う。

次に、得られた整数計画モデルを改良し、予測精度の向上を目指す。具体的には、現在のパラメータを使った場合の予測結果と正解構造を比較して、その構造的な差異が最小となるような最適化問題を新たに定式化する。この最適化問題の解は元の問題のパラメータであるため、問題自体は連続の最適化問題となるが、アカデミックフリーである高性能最適化ソルバーCPLEX を用いることで対応できると考えている。

⑶ 高速性を損なわないように予測モデルを標的 RNA 探索用に特化する変換を行う。そして、新規標的探索をゲノムワイドに行うことで、RNA 間相互作用ネットワークの考察を行う。ここでは、生物種として真正細菌を選び、小型 RNA（sRNA）の標的を網羅的に探索することを考える。現在例えば大腸菌ではある程度までネットワークの形状が推定されているが、本研究での精緻な高次構造予測に基づく標的探索により、これまで指摘されなかった標的との相互作用関係を新たに提案できる可能性があると考えている。これは高いコストを要する生物学実験で相互作用を検証する際に当たりを付けるという意味での実験支援になることが期待される。

４．研究成果

⑴ 本研究で開発した計算機に基づく RNA 間相互作用予測法 RactIPAce（後述の⑶参照）の旧バージョンとも言うべき RactIP、お

よびシュードノットと呼ばれる複雑な部分構造を考慮した RNA２次構造予測法 IPknot を、商用目的以外でインターネットを介して自由に利用できるようにするため、Web サーバーRtips を開発し公開した（図１参照）（http://rtips.dna.bio.keio.ac.jp/）。

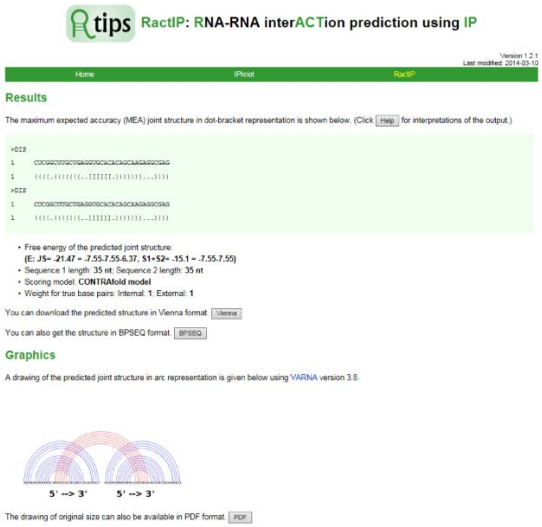

図１　RactIP サーバーの出力例

(2) 次に、相互作用に関与する制御 RNA をゲノム配列から探索する問題に向けて、構造類似 RNA を複数の配列間での比較解析により求める計算手法の開発に着手した。まず、ゲノムから得られる RNA 配列が与えられたと仮定し、２本の RNA 配列から配列類似性と構造類似性の両者を同時に考慮しながら２次構造を考慮した対応付け（アラインメント）を求めるアルゴリズムを開発した。ここで、整数計画法によるモデル化と双対分解と呼ばれる技術を組み合わせることで、高速に２次構造アラインメントを計算することが可能となった。提案手法は他の類似の既存手法と比べ、最も精度良く共通２次構造を予測できることを確認した。

(3) 申請時に立案した拡張２次構造のための必要情報（非正準塩基対の確率分布）を得るのが想定より困難であったため、方針を転換し、相互作用に関与し得る確率を表すアクセシビリティの概念を、RNA 間相互作用予測法 RactIP に導入することを考えた。具体的に、RactIP を構成する整数計画モデルにおいて、アクセシビリティに関する情報をモデルの制約式に反映させるとともに、従来の RactIP の結合部位モデルを、２次構造を考慮したモデルに変更することで、新規予測モデル RactIPAce を開発した（図２参照）。これまでは結合部位が既知であるごく限られたデータを用いて手法の精度検証を行っていたが、今回新たに既知結合部位を持つ相互作用データを、真正細菌を中心に大幅に増加し、検証実験を行った。その結果、アクセシビリティに基づく他の既存手法と比べて、提案手法の予測精度は最良値を達成した。また、ネットワーク探索のためには相互作用有無の判定が行えることも重要であるため、大腸菌、サルモネラ菌からなる相互作用有無に関する正例、負例データを先行研究から取得し、結合エネルギーを指標とした網羅的な識別実験を行った。ここでは、RactIPAce を含めたアクセシビリティに基づく予測手法のコンセンサスを取ることで、良い識別率を達成することができた。換言すれば、相互作用の有無にはアクセシビリティに基づく手法のコンセンサスを取り、その後の結合部位予測では提案手法を単体で用いることが有効であると結論付けることができる。

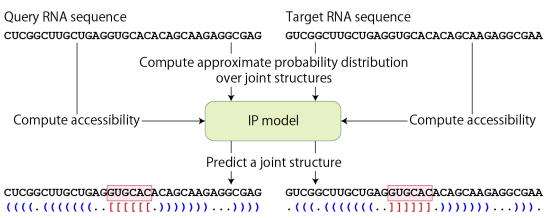

図２　RactIPAce の概念図。図中の IP model は整数計画モデルを表す。

(4) DNA や RNA などの核酸配列には G（グアニン）に富んだ部分配列が存在し、G４重鎖と呼ばれる特殊な高次構造を形成するこ

とが分かっている。この種の構造を解析することは本研究課題に貢献し得ると考え、G４重鎖領域をゲノム配列から探索する手法を開発した。ここでは、G４重鎖領域を隠れマルコフモデルによりモデル化した。G４重鎖に関する正例、負例データを用いた識別実験において、提案手法は高い識別能力を示した。また、ゲノムワイド G４重鎖探索実験では、提案手法を用いることで、正規表現に基づく従来手法により予測された偽陽性の G4 重鎖を除去できる可能性を示した。

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] Masato Yano and Yuki Kato*, **Using hidden Markov models to investigate G-quadruplex motifs in genomic sequences**, *BMC Genomics*, vol. 15(Suppl 9), S15, Dec. 2014, 査読有 (*: corresponding author).
DOI: 10.1186/1471-2164-15-S9-S15

[2] Kengo Sato, Yuki Kato, Tatsuya Akutsu, Kiyoshi Asai and Yasubumi Sakakibara, **DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition**, *Bioinformatics*, vol. 28, no. 24, pp. 3218–3224, Dec. 2012, 査読有.
DOI: 10.1093/bioinformatics/bts612

[3] Yuki Kato, Kengo Sato, Kiyoshi Asai and Tatsuya Akutsu, **Rtips: fast and accurate tools for RNA 2D structure prediction using integer programming**, *Nucleic Acids Research*, vol. 40, Web Server issue, pp. W29–W34, Jul. 2012, 査読有.
DOI: 10.1093/nar/gks412

〔学会発表〕（計 19 件）

[1] Yuki Kato, Jakob Hull Havgaard and Jan Gorodkin, **Using coarse-grained dot plots to screen for structurally similar RNAs in genomic sequences**, The Thirteenth Asia Pacific Bioinformatics Conference (APBC2015), Poster, P-13, 台湾 新竹, 2015 年 1 月 21～23 日.

[2] Kunie Sakurai, Junko Yamane, Kenta Kobayashi, Koji Yamanegi, Takeaki Taniguchi, Yuki Kato and Wataru Fujibuchi, **SHOGoiN cell database: a framework for a new repository on single cell assay data and diverse knowledge of human cells**, The 18th Takeda Science Foundation Symposium on Bioscience, Poster, 070, 大阪府吹田市 武田薬品研修所, 2015 年 1 月 15～17 日.

[3] Kunie Sakurai, Junko Yamane, Kenta Kobayashi, Koji Yamanegi, Takeaki Taniguchi, Yuki Kato and Wataru Fujibuchi, **Stem Cell Informatics Database: a framework for a new repository on single cell assay data and diverse knowledge of human cells**, The 25th International Conference on Genome Informatics (GIW/ISCB-Asia 2014), Poster, 91, 東京都江東区 東京国際交流館プラザ平成, 2014 年 12 月 15～17 日.

[4] Takeaki Taniguchi, Yuki Kato, Susumu Goto and Wataru Fujibuchi, **Development of a pipeline for analysis of meta- and single cell genomic sequences**, The 25th International Conference on Genome Informatics (GIW/ISCB-Asia 2014), Poster, 41, 東京都江東区 東京国際交流館プラザ平成, 2014 年 12 月 15～17 日.

[5] 桜井 都衣, 山根 順子, 小林 健太, 山根木 康嗣, 谷口 丈晃, 加藤 有己, 藤渕 航, **Stem Cell Informatics Database: 多様なヒト細胞情報の統合化を目指したデータベースの構築**, 第 39 回情報処理学会バイオ情報学研究会, 大阪府吹田市 大阪大学, 2014 年 9 月 19 日.

[6] 加藤 有己, **ゲノムワイド RNA 配列比較に向けて**, 2014 年度 RNA インフォマティクス道場, 北海道札幌市 産業技術総合研究所北海道センター, 2014 年 8 月 26～28 日.

[7] Masato Yano and Yuki Kato, **Using hidden Markov models to investigate G-quadruplex motifs in genomic sequences**, 13th International Conference on Bioinformatics (InCoB2014), オーストラリア シドニー, 2014 年 7 月 31 日～8 月 2 日.

[8] Yuki Kato, Jakob Hull Havgaard and Jan Gorodkin, **Fast RNA structural comparison using coarse-grained base-pairing probabilities**, 第 16 回日本 RNA 学会年会, Poster, P-91, 愛知県名古屋市 ウィンクあいち, 2014 年 7 月 23～25 日.

[9] Yuki Kato, Jakob Hull Havgaard and Jan Gorodkin, **Fast RNA structural comparison using coarse-grained base-pairing probabilities**, The 24th International Conference on Genome

Informatics (GIW2013), Poster, 29, シンガポール バイオポリス, 2013 年 12 月 16～18 日.

[10] Yuki Kato, Tomoya Mori, Kengo Sato, Shingo Maegawa, Hiroshi Hosokawa and Tatsuya Akutsu, **Evaluating effectiveness of accessibility to infer RNA–RNA interactions**, 日本バイオインフォマティクス学会 2013 年年会, Poster, 68, 東京都江戸川区 タワーホール船堀, 2013 年 10 月 29～31 日.

[11] Masato Yano and Yuki Kato, **Modeling DNA G-quadruplexes by hidden Markov models**, 日本バイオインフォマティクス学会 2013 年年会, Poster, 6, 東京都江戸川区 タワーホール船堀, 2013 年 10 月 29～31 日.

[12] Yuki Kato, Tomoya Mori, Kengo Sato, Shingo Maegawa, Hiroshi Hosokawa and Tatsuya Akutsu, **Evaluating effectiveness of accessibility to infer RNA–RNA interactions**, 第 35 回情報処理学会バイオ情報学研究会, 北海道札幌市 北海道大学, 2013 年 9 月 19～20 日.

[13] 加藤 有己, **RNA 間相互作用推定におけるアクセシビリティの有効性について**, RNA アルゴリズム研究会 2013, 東京都江東区 生命情報工学研究センター, 2013 年 8 月 10 日.

[14] Yuki Kato, Kengo Sato, Kiyoshi Asai and Tatsuya Akutsu, **Rtips: fast and accurate tools for RNA 2D structure prediction using integer programming**, 第 15 回日本 RNA 学会年会, Poster, P-98, 愛媛県松山市 ひめぎんホール, 2013 年 7 月 24～26 日.

[15] Yuki Kato, Tomoya Mori, Kengo Sato, Shingo Maegawa, Hiroshi Hosokawa and Tatsuya Akutsu, **Evaluating effectiveness of accessibility to infer RNA–RNA interactions**, 第 15 回日本 RNA 学会年会, Poster, P-93, 愛媛県松山市 ひめぎんホール, 2013 年 7 月 24～26 日.

[16] Kengo Sato, Yuki Kato, Tatsuya Akutsu, Kiyoshi Asai and Yasubumi Sakakibara, **DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition**, International Symposium on Genome Science "Expanding Frontiers of Genome Science," Poster, P34, 東京都文京区 東京大学, 2013 年 1 月 9～10 日.

[17] Yuki Kato, **RNA structural alignment using dual decomposition**, Herbstseminar der Bioinformatik (2012), チェコ ドウビツェ, 2012 年 10 月 2～7 日.

[18] Yuki Kato, **Fast and accurate prediction of RNA–RNA interactions using integer programming**, International Workshop on RNA, スペイン ベナスケ, 2012 年 7 月 22 日～8 月 3 日.

[19] Yuki Kato, **Fast and accurate prediction of RNA pseudoknotted structures using integer programming**, スペイン ベナスケ, 2012 年 7 月 22 日～8 月 3 日.

〔図書〕（計 1 件）
[1] 加藤 有己, 桜井 都衣, 藤渕 航, ビッグデータの収集、調査、分析と活用事例, 第 7 章第 2 節「ヒト細胞からのビッグデータの情報管理と情報解析技術」, 技術情報協会, pp. 249–254, 2014 年.

〔その他〕
報道関連情報
マイナビニュース，最大で 4 万倍の高速化 - NAIST など，RNA の 2 次構造予測を行う新計算法を開発，2012 年 8 月 31 日
http://news.mynavi.jp/news/2012/08/31/075/

ホームページ情報
http://rtips.dna.bio.keio.ac.jp/

６．研究組織

(1)研究代表者

加藤 有己（KATO Yuki）

京都大学・iPS 細胞研究所・特定拠点助教

研究者番号：10511280

(2)研究分担者

（ ）

研究者番号：

(3)連携研究者

（ ）

研究者番号：