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　６ 年　 ５ 月 ２７ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2021～2023
課題番号：２１Ｋ１５０７５

研究課題名（和文）RNA2次構造の確率分布を内包した機械学習アルゴリズムの開発

研究課題名（英文）Development of Machine Learning Algorithm Encapsulating Probability Distribution of RNA Secondary Structure

研究代表者
寺井　悟朗（Terai, Goro）

東京大学・大学院新領域創成科学研究科・特任准教授

研究者番号：４０７８５３７５

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：近年の実験技術の進歩により、RNA2次構造が関与する生命現象についての大規模データが得られるようになった。そこで、本研究ではRNAの塩基配列と活性に関するデータから重要な２次構造を抽出したり予測モデルを学習したりする手法の開発を行った。そして、この手法を性質の異なる複数のデータセットに対して適用することにより、本手法の汎用性を示した。

研究成果の学術的意義や社会的意義
RNA配列と活性に関するデータを統一的な枠組みで解析するためのアルゴリズムは世界的に見ても類がなく新規性が高い。提案手法を用いることにより、従来は個々の研究者が別々に手法開発を行っていたデータの解析を、統一的な枠組みで実施できるようになる。本手法の開発により、RNA配列と活性に関するデータを取得した研究者が、RNA2次構造を精密に考慮した特徴抽出や予測モデルの開発を簡便に実施出来るよになることが期待される。

研究成果の概要（英文）：Due to recent advancements in experimental techniques, large-scale data on biological phenomena involving RNA secondary structures have become available. In this study, we developed methods to extract important secondary structures and train predictive models from data consisting of RNA sequences and their activity. By applying these methods to multiple datasets with different properties, we demonstrated the versatility of our approach.

研究分野：バイオインフォマティクス

キーワード：RNA2次構造　特徴抽出　回帰モデル　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

RNA２次構造は、細胞内でさまざまな機能を担い、多様な生命現象に関与している。それら生命現象の動作原理を理解したり、人為的にコントロールしたりするためには、RNA２次構造の役割を正確に知ることが重要である。次世代シーケンサーや DNA 合成技術の発展により、RNA２次構造が関与する生命現象についての大規模データが得られるようになった。しかしながら、それら大規模データから２次構造的特徴を抽出したり予測モデルを学習したりする汎用的な方法は存在せず、古典的な２次構造予測アルゴリズムと従来からある統計解析や機械学習の組み合わせがその分析に利用されていた。

２．研究の目的

本研究では RNA とその活性に関する大規模データを解析するための汎用的な枠組みを開発する。具体的には、RNA の塩基配列とその活性に関するデータに基づき、特徴抽出や予測モデルの構築を行う新しい機械学習アルゴリズムを開発する。RNA は細胞の中で 1 つの決まった２次構造を取るだけでなく、複数の構造を確率的に取ると考えられる。そこで、RNA2 次構造の確率的振る舞いを考慮しつつ、高精度な特徴抽出や予測モデルの構築を行うアルゴリズムの開発を目指す。開発した手法を様々なデータへ適用することにより、その有用性と汎用性を示すと共に、未発見の２次構造的特徴の抽出を行う。

３．研究の方法

（１）RNA 配列と活性に関するデータの解析技術の開発

RNA２次構造の確率的な揺らぎを考慮し、RNA の活性を予測したり、活性に影響を与える２次構造を抽出したりするアリゴリズムの開発を行なった。具体的には、RNA に含まれる各塩基が形成する２次構造（塩基対、ヘアピンループ、バルジループなど）の確率を考慮し、RNA の塩基配列から活性値を予測する回帰アルゴリズムの開発を行なった。学習の結果として得られた回帰モデルのパラメータには、どのような 2 次構造が活性に重要な役割を持つのかが反映される。

（２）提案手法を用いたデータ解析

開発したアルゴリズムを性質の異なる複数のデータセットに対して適用した。データ数や配列多様性などに関して、性質の異なるさまざまなタイプのデータに対して本アルゴリズムを適用することにより、その有用性や汎用性を見極めるとともに、データの解析の際に必要な前処理やパラメータチューニングなどに関する問題点の洗い出し行なった。そして、得られた知見のフィードバックに基づき、データの前処理やアルゴリズムの改良を行なった。

４．研究成果

（１）RNA の活性に直結する２次構造の抽出

多くの RNA においては、その２次構造が活性に重要な役割を果たすことが知られている（図１）。しかしながら、RNA２次構造を実験的かつ網羅的に観測することは一般に困難であり、観測可能な塩基配列や活性に関する情報から２次構造の寄与を推定する必要がある。そこで、本研究では計算機による２次構造予測により得られる情報を利用して、RNA の活性に直結する２次構造のパターンを抽出するための機械学習アルゴリズムの開発を行なった。

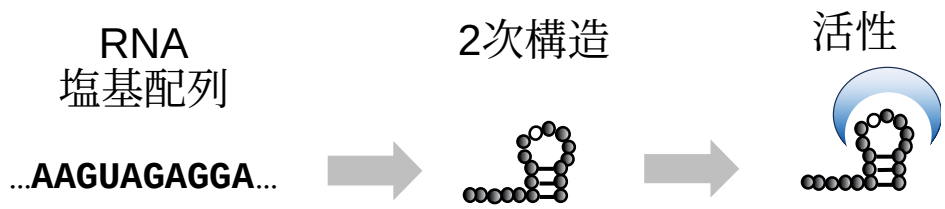

図１　RNA 塩基配列、2 次構造、活性の関係

また、RNA２次構造は確率的な揺らぎを持ち、細胞内で複数の２次構造を形成することが知られている。図２に２種類の構造をとる RNA とその確率の例を示す。この図のように、ある RNA が構造 A を 80%の確率で形成する場合、構造 A が構造 B よりも大きく RNA 活性に寄与する。我々の提案するアルゴリズムでは、計算機により推定した RNA の確率的な揺らぎを考慮し、RNA の活性に直結する２次構造を抽出する。具体的には、RNA の２次構造を図３に示すような６種類の要素に分類する。そして、RNA に含まれる各塩基がある２次構造要素に含まれる確率を計算機により推定する。こうして求めた確率を特徴量として用いることにより、図１に示した確率的な揺らぎを考慮しつつ、RNA の活性に重要な影響を及ぼす可能性のある２次構造のパターンを抽出することができる。抽出した２次構造パターンがどのように表現されるかは、（２）で具体例を示す。

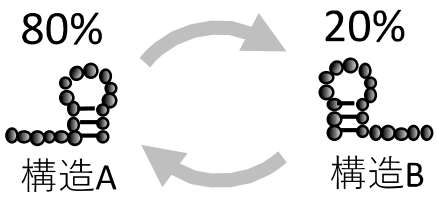

図２　２次構造の揺らぎと、その確率

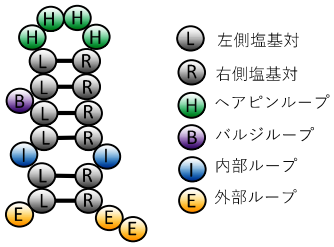

図３　６種類の２次構造要素

（２）公開データへの適用結果

（１）で開発したアルゴリズムの有用性と汎用性を評価するため、RNA とその活性値に関する５種類のデータセットに対して提案アルゴリズムを適用した。表１にデータセットとその性質を示す。これらのデータセットは、RNA 配列の数、長さ、類似度などが大きく異なることに着目してほしい。例えば、データセット１では学習データとして用いる RNA 配列の類似度が非常に低いが、データセット２では類似度が高い（データセット２に含まれる RNA の塩基配列は平均して 90%以上の類似度がある）。本研究では、（１）の提案アルゴリズムを表１のすべてのデータに適用し、性質の異なるデータに対しても２次構造に関する特徴を精度よく抽出できることを示した[文献１]。以下にデータセット１と２に対する解析結果を示す。

表１　データセットの概要

データセット	サンプル数	配列長(nt)	配列類似度[a]	活性値の平均[b]
1 翻訳開始領域	242269	120	47.3	0.49
2 Twisterリボザイム	5778	57	93.3	0.30
3 シャインダルガノ配列	3070	20	77.9	0.29
4 ドナー部位(GU)	16216	15	65.0	0.02
5 ドナー部位(CU)	972	15	65.1	0.04

a）２つのRNA配列の一致割合の平均、b）0-1に規格化した活性値の平均

・データセット１（翻訳開始領域）の解析結果

多くの原核生物において、mRNA の翻訳開始点付近は２次構造が不安定であることが知られている。Cambray らは約 24,4000 種類もの mRNA 配列（翻訳開始点周辺）とその翻訳効率に関するデータを測定した[文献 2]。本研究では提案アルゴリズムを用いて、翻訳効率に直結する２次構造の抽出を行なった。図４に翻訳効率に直結すると予測された２次構造の抽出結果を示す。図の横軸は翻訳開始点からの相対位置（ポジション）を示している。縦軸のアルファベットは図３で示した２次構造要素を示している。図４の数値（色）は翻訳効率を増加させるか、低下させるかを示しており、大きいほど翻訳を増加させることを示している。たとえば、ポジション-12 から-6 が L（左側塩基対）の時と、ポジション４から９が R（右側塩基対）の時に数値が低くなっている。このことは、翻訳開始点をまたぐような２次構造が翻訳効率を低下させる効果があること強く示唆している。

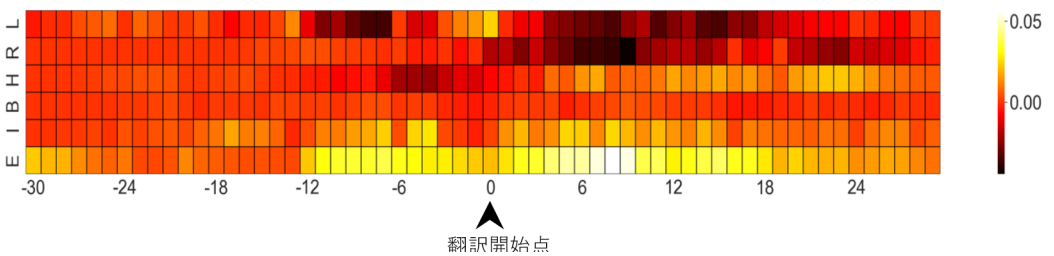

図４　翻訳開始効率と関連する２次構造

・データセット２（Twister リボザイム）の解析結果

Twister リボザイムは、自分自身を切断する活性（自己切断活性）をもつ RNA の１種であり、

自然界にひろく分布していることが知られている。Kobori らは Twister リボザイムにランダム変異を導入することにより 10000 種類以上のリボザイム変異体を作成し、その活性を網羅的に測定した[文献 3]。本研究では 5778 種類のリボザイム変異体とその活性に関するデータを選択し、活性に直結する２次構造の抽出を行なった（文献３の中で RNA 配列と活性の因果原因が分析された変異体は解析から除外した）。図５に自己切断活性に直結すると予測された 2 次構造の抽出結果を示す。横軸は、リボザイムの 5'末端を 0 とした時のポジションを示す（縦軸と色の意味は図４と同じである）。この図から、ポジション 26 と 33 が塩基対を形成する場合には、Twister リボザイムの活性が大きく低下することがわかる。また、自己切断部位の下流２塩基（ポジション 9 と 10）が I（インターナルループ）の時に自己切断活性が上がる傾向があることが示されている。

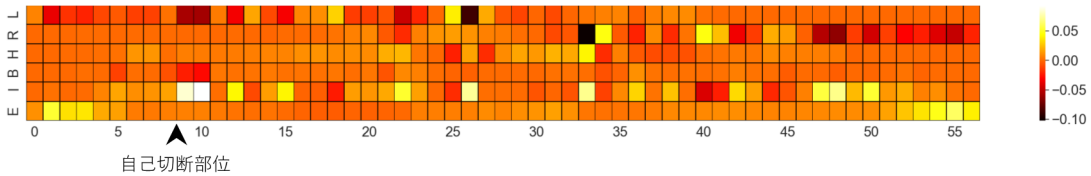

図５　リボザイム活性と関連する 2 次構造

（３）国内外における位置付け

RNA２次構造に関する情報解析の歴史は長く、２次構造の予測だけでなく、２次構造を考慮した配列アライメントや相同性検索などが提案されている。しかしながら、大量のデータから機械学習的な思想で２次構造的特徴を抽出したり予測モデルを学習したりするという研究は多くはない。特に RNA の塩基配列とその活性値に関する大規模データを統一的な枠組みで解析するためのアルゴリズムは世界的に見ても類がなく新規性が高い。

（４）今後の展望

本研究課題で開発したアルゴリズムは、図４や５で示されるように特定のポジションにおける２次構造と RNA 活性の関連を分析するのに適した方法である。しかしながら、たとえば、２つの異なる RNA 配列 x と y において、x の 5'側と y の 3'側に活性に影響する共通の特徴（タンパク質結合部位など）がある場合には、その特徴を考慮して活性を予測したり、２次構造パターンを抽出したりすることが難しい。しかしながら、異なるポジションに出現する重要な２次構造パターンは多くの RNA に含まれる可能性がある。本手法の適用範囲を拡大するためには、異なる位置に存在する 2 次構造の影響を捉えられるようにすることが重要と考えられる。

これを実現するための方向性として、深層学習を利用した特徴抽出及び活性予測法の開発が挙げられる。近年、深層学習は自然言語や画像の生成など従来の機械学習では困難であった複雑なタスクに対して高精度な結果を達成しており、さまざまな分野においてその利用が広がっている。上記のような異なる RNA の部位に存在する特徴を考慮することは深層学習を用いれば実現することができる。しかしながら、深層学習を用いることにより RNA２次構造の揺らぎを正確に考慮することができなくなる可能性がある。そこで、RNA２次構造の揺らぎをなるべく正確に考慮しつつ深層学習の利点を活かす方法を模索している。提案アルゴリズムの利点をできる限り保ちながら深層学習と統合することにより、より汎用性の高い枠組みを開発することを目指す。

[参考文献]

1) G. Terai, K. Asai, QRNAstruct: a method for extracting secondary structural features of RNA via regression with biological activity. *Nucleic Acids Res.* **50**, e73 (2022).
2) G. Cambray, J. C. Guimaraes, A. P. Arkin, Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nat. Biotechnol.* **36**, 1005-1015 (2018).
3) S. Kobori, Y. Yokobayashi, High-Throughput Mutational Analysis of a Twister Ribozyme. *Angew. Chem. Int. Ed. Engl.* **55**, 10354-10357 (2016).

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Terai Goro、Asai Kiyoshi	４．巻 50
２．論文標題 QRNAstruct: a method for extracting secondary structural features of RNA via regression with biological activity	５．発行年 2022年
３．雑誌名 Nucleic Acids Research	６．最初と最後の頁 e73
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nar/gkac220	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 Goro Terai and Kiyoshi Asai
２．発表標題 QRNAstruct: a method for extracting secondary structural features of RNA via regression with biological activity
３．学会等名 第11回生命医薬情報学連合大会（IIBMP2022）
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

RNA機能に直結する2次構造を予測する汎用的な手法を開発
https://www.k.u-tokyo.ac.jp/information/category/press/9412.html

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

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