プレスリリース

慶應義塾大学

2021 年 2 月 12 日

報道関係者各位

慶應義塾大学

RNA 二次構造予測で世界最高精度を達成

慶應義塾大学理工学部生命情報学科の佐藤健吾専任講師、榊原康文教授と同大学大学院理工学研究科基礎理工学専攻の秋山真那斗（博士課程 3 年）からなる研究グループは、従来用いられてきた熱力学モデルと深層学習を効果的に組み合わせることによって、高精度かつ頑健な RNA 二次構造予測手法（MXFold2）を開発し、世界最高精度を達成しました。RNA 二次構造※1 が関与する遺伝子発現機構の解明や RNA 創薬などへの応用が期待されます。

本研究成果は、2021 年 2 月 11 日にイギリスの科学雑誌「Nature Communications」のオンライン版に掲載されました。

１．本研究のポイント

- 熱力学モデルと深層学習を組み合わせることによって RNA 二次構造予測の精度を向上させることに成功し、既存手法との比較実験によって世界最高精度であることを確認した。
- 機械学習・深層学習に基づく既存手法のいくつかが訓練データへの過学習による致命的な欠陥があることを示した一方で、本手法は熱力学モデルを組み合わせることで過学習を克服し、高精度なだけでなく頑健な RNA 二次構造予測を実現した。
- 本手法によって計算される RNA 二次構造の評価値が熱力学安定性と強い相関を持つことが示唆された。

２．研究背景

RNA は遺伝情報を伝達する役割がよく知られていますが、それ以外にも遺伝子発現を制御したり、遺伝子の修飾部位をコントロールしたりするなど機能を持ち、さまざまな生命現象に関与している非コード機能性 RNA※2 が近年注目されています。これらの機能の多くは、一本鎖の RNA 分子が自分自身と結合することによって形成される立体構造によってもたらされます。RNA 立体構造の決定は、X 線結晶構造解析、核磁気共鳴法(NMR)、低温電子顕微鏡(cryo-EM)などによって行われますが、解析できる配列長・解像度やコストの問題から、大規模解析にこれらの手法を適用することは現状では困難です。そのため、計算機で予測した RNA 二次構造でこれを代替することが頻繁に行われており、その予測精度が RNA の機能推定に決定的な影響を与えます。

従来、RNA 二次構造予測は熱力学モデルに基づく手法が一般的でした。熱力学モデルでは、事前に計測した熱エネルギーパラメータを用いて、自由エネルギーが最小となる二次構造を予測します。この手法は RNA 二次構造の熱力学安定性を直接評価できる一方で、熱エネルギーパラメータの計測誤差や粗雑さなどの理由で、その予測精度には限界があることが知られています。

そこで近年、機械学習に基づく RNA 二次構造予測手法が開発されています。RNA 配列とその正解二次構造の組を訓練データとして、熱エネルギーパラメータの代替となるパラメータを学習します。既存の熱力学パラメータよりも精緻なパラメータセットを採用することで、機械学習・深層学習を利用した既存手法は予測精度の大幅な向上に成功しています。しかし、精緻すぎるパラメータセットを用いると過学習に陥りやすくなるために、訓練データに近い二次構造を持つ RNA 配列は非常に高精度の

予測が可能である一方、訓練データから離れた未知の RNA 配列に対する予測精度が不十分であることが報告されています。

３．研究内容・成果

本研究では、深層学習を採用した精緻なモデルによる予測精度の向上と、過学習による影響を最小限に抑えて未知 RNA 配列に対する頑健性を両立する手法の開発を目指しました。畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）と長・短期記憶ネットワーク（Long Short-Term Memory; LSTM）からなる深層ネットワークによる熱力学パラメータの精緻化によって高精度化を図りました。一方、深層ネットワークが計算するスコアと既存の熱力学パラメータを統合し、ネットワークの学習の際に正解二次構造と予測二次構造の間で食い違う塩基対の数を最小化する通常の学習に加えて、未知 RNA 配列に対する頑健性を高めるために、統合されたスコアと正解二次構造が持つ自由エネルギーの差を最小化する熱力学的正則化を導入しました。

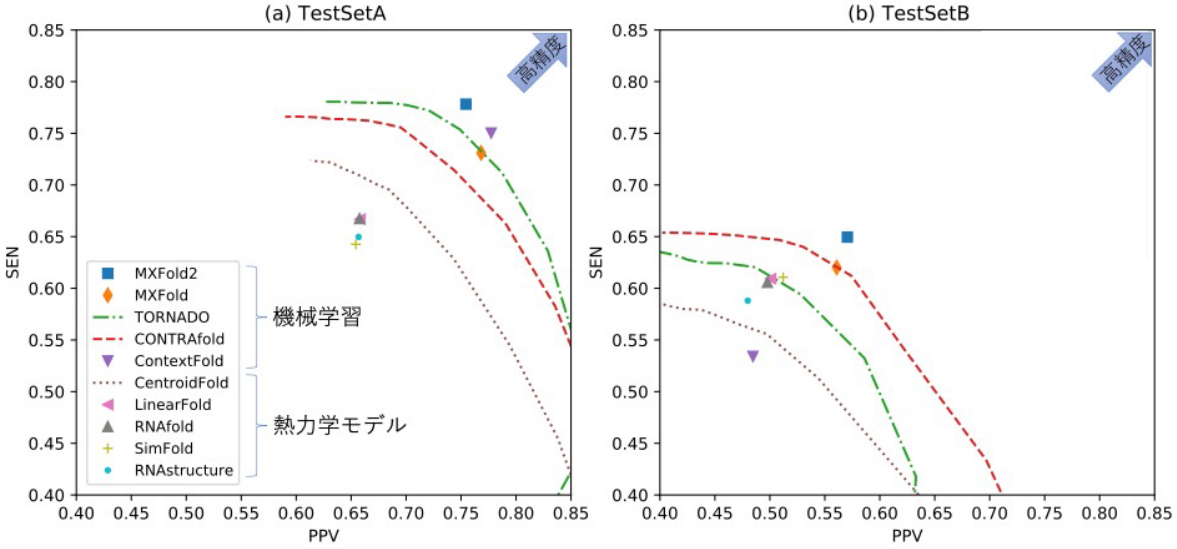

図 1　既存手法との精度比較。PPV は予測した塩基対が正解構造に含まれる割合（正解率）を表し、SEN は正解構造に含まれる塩基対を予測できた割合（網羅率）を表す。

主要な RNA 二次構造予測手法との比較実験では、訓練データと近い二次構造を持つ配列データセット TestSetA と、離れた二次構造を持つデータセット TestSetB の両方において、本手法 MXFold2 が最高の予測精度を示しました（図 1）。機械学習に基づく手法の多くが TestSetB において予測精度を大きく下げているのに対して、本手法 MXFold2 は比較的高い精度を保持しており、このことは未知の RNA 配列に対する頑健性を示しています。さらに、二次構造に加えてその自由エネルギーがわかっているデータセットにおける実験から、本手法が計算する二次構造のスコアと RNA の熱力学安定性が強く相関することが示唆されました。

本研究で開発した MXFold2 は、ウェブサイトよりダウンロード可能なほか、ウェブサーバとして簡単に利用することができます。

ダウンロード用ウェブサイト（GitHub） https://github.com/keio-bioinformatics/mxfold2/

ウェブサーバ http://www.dna.bio.keio.ac.jp/mxfold2/

４．今後の展開

RNA 二次構造予測は計算機による RNA 配列解析において最も基盤となる技術です。本研究で開発した MXFold2 は、既知の RNA 配列のみでなく未知の RNA 配列に対しても頑健な二次構造予測が可能であることを示しました。近年、新規の非コード RNA が次々と発見されており、その機能を推定するために、本手法により実現した高精度かつ頑健な RNA 二次構造予測が有用であると期待されます。また、本研究で開発された二次構造予測の技術は、RNA 二次構造が関与する遺伝子発現機構の解明や RNA 分子をターゲットとするリード化合物探索、RNA 自身が薬となる RNA 創薬などへの応用が期待されます。

※本研究は、日本学術振興会 科学研究費助成事業（19H04210, 19K22897, 18J21767）、文部科学省 科学研究費助成事業 新学術領域「化学コミュニケーションのフロンティア」（17H06410）などの助成や支援を受けて行われました。

＜原論文情報＞

Sato K, Akiyama M, Sakakibara Y (2021) RNA secondary structure prediction using deep learning with thermodynamic integration. Nature Communications.
doi: 10.1038/s41467-021-21194-4

＜用語説明＞

※1 RNA 二次構造
立体構造の骨格を成す塩基対の集合で、A-U、G-C、G-U の規則で結合する。

※2 非コード機能性 RNA
タンパク質に翻訳されずに自身が生体内で活性を持ち、転写・翻訳を制御するなどの機能を持つ RNA の総称。代表的なものとして tRNA や rRNA などが知られている。

※ご取材の際には、事前に下記までご一報くださいますようお願い申し上げます。
※本リリースは文部科学記者会、科学記者会、各社科学部等に送信させていただいております。

・研究内容についてのお問い合わせ先
慶應義塾大学 理工学部 生命情報学科　専任講師　佐藤　健吾（さとう　けんご）
TEL：045-566-1511　FAX：045-566-1789　E-mail： satoken@bio.keio.ac.jp

・本リリースの配信元
慶應義塾広報室（澤野）
TEL：03-5427-1541　FAX：03-5441-7640
Email：m-pr@adst.keio.ac.jp　https://www.keio.ac.jp/