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　２ 年　 ５ 月 ２７ 日現在

機関番号：３２６１２
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００４０４

研究課題名（和文）次世代シークエンシングデータを利用した機械学習によるRNA二次構造予測の高精度化

研究課題名（英文）Improving the accuracy of RNA secondary structure prediction by machine learning based on next-generation sequencing data

研究代表者
佐藤　健吾（Sato, Kengo）

慶應義塾大学・理工学部（矢上）・講師

研究者番号：２０３６５４７２

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）：部分的な構造情報である二次構造プロファイルを弱レベル学習データとして利用可能とする機械学習アルゴリズムを開発し，既存手法よりも精密な二次構造モデルを大量の二次構造プロファイルから学習することによって，過学習を回避しつつRNA二次構造予測の精度向上を目指す．まず，既存のTurner熱力学モデルに基づく自由エネルギー最小化法と構造化SVMによるパラメータ学習法を融合することによってより頑健かつ高精度なRNA二次構造予測手法の開発を行った．計算機実験の結果，既存の手法に見られる過学習は観測されず，予測精度の向上が確認された．

研究成果の学術的意義や社会的意義
RNA 二次構造予測は古くから研究されているが，長い配列に対する予測精度は未だに十分とは言えない．本研究によりRNA二次構造予測の精度が向上し，生体内におけるRNAの機能を推定する手がかりが得られることが期待される．さらに，RNAウィルスをターゲットとする創薬などに応用することが可能である．

研究成果の概要（英文）：We have developed a machine learning algorithm that makes it possible to use secondary structure profiles, which are partial structural information, as weak-level learning data, and aims to improve the accuracy of RNA secondary structure prediction without overfitting by learning a large number of secondary structure models that are more precise than existing methods. First, we developed a more robust and accurate method for RNA secondary structure prediction by integrating the free energy minimization method based on the existing Turner thermodynamic model with the machine learning method using a structured SVM. The results of the computer experiments showed that no overfitting was observed, unlike in the existing methods, and the prediction accuracy was improved.

研究分野：バイオインフォマティクス

キーワード：バイオインフォマティクス　RNA二次構造予測　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

これまで機能性 RNA に関する研究は，配列長が比較的短い（200 塩基未満）小分子非コード RNA が中心であった．しかし近年，配列長が 200 塩基を大きく越える長鎖非コード RNA が転写制御，スプライシング，翻訳制御，エピジェネティックな制御など様々な機構に関与していることが明らかになった．このことから，配列長が長い機能性 RNA に関する研究に注目が集まっている．

機能性 RNA の配列情報解析においては，配列のみでなく構造を考慮する必要がある．タンパク質コード領域は，コドン使用頻度やフレームシフトを起こさないといった選択圧があるために，配列自体が高度に保存される傾向にある一方，機能性 RNA の多くはそのような制約がないために，同じ機能を持つもの同士であっても，配列相同性はあまり高くない．このような機能性 RNA は，立体構造を形成することによって機能を発揮し，そのため機能と構造の間には強い相関があると考えられている．

多くの RNA 立体構造決定手法は高コストかつ低スループットであるため，現実的な時間で計算可能な計算機による RNA 二次構造予測が広く用いられてきた．RNA 二次構造予測法は，自由エネルギー最小化（Minimum Free Energy; MFE）に基づく手法と機械学習に基づく手法に大別される．近年の RNA 二次構造予測技術の進歩により，短い配列に関しては高精度の予測が可能となったものの，長い RNA 配列からの二次構造予測には精度の限界があり，配列長が 500 塩基を超えたあたりから予測精度が著しく悪化する傾向にある．その原因は，MFE に基づく手法では，部分構造の自由エネルギーの計測精度が十分でないこと，機械学習に基づく手法では，学習に用いる既知 RNA 二次構造の数が少なく，とくに配列長が長いものについては不十分であるために，正確なスコアパラメータを学習することが困難であることによる．さらに，機械学習に基づく手法においては，二次構造モデルをより精密にすることによってより高精度な予測を期待できる一方で，過度な精密化によるパラメータ数の増大は逆に過学習による予測精度の低下を招く恐れがあることが指摘されている．

近年，次世代シークエンサー（NGS）の普及により，大量の DNA 配列を高速かつ低コストで解読することが可能となった．RNA 二次構造に関しても NGS を利用した解析手法がいくつか開発されている．PCR 伸長反応を阻害する化学修飾をループ部位の塩基のみに付加後 NGS で配列を読むことによって，塩基対を形成しにくい部位のシグナル（二次構造プロファイル）を得ることができる．これは完全な RNA 二次構造ではないため，二次構造プロファイルを手がかりに計算機で RNA 二次構造を推定する．このようなパイプラインによって，トランスクリプトームワイドでより正確な RNA 二次構造の推定を可能にしている．しかし逆に，NGS による RNA 二次構造プロファイルがなければこの手法は適用できないため，二次構造の予測精度を常に改善できるわけではない．

２．研究の目的

本研究では，部分的な構造情報である二次構造プロファイルを学習データとして利用可能とする機械学習アルゴリズムを開発し，既存手法よりも精密な二次構造モデルを大量の二次構造プロファイルから学習することによって，過学習を回避しつつ RNA 二次構造予測の精度向上を目指す．これにより，二次構造予測をベースにした機能性 RNA の機能・構造解析の精度向上を実現する．

RNA 二次構造予測は古くから研究されているが，長い配列に対する予測精度は未だに十分とは言えない．NGS による二次構造プロファイルを二次構造予測の手がかりとして利用する手法はこれまでいくつか提案されているが，二次構造プロファイルがない配列には適用できない.それに対して，大量に得ることが可能な二次構造プロファイルをパラメータ最適化の学習データとして用いることによって，二次構造プロファイルがない配列に対しても予測精度を改善する点が本研究の特色である．長鎖非コード RNA の二次構造プロファイルによってパラメータを最適化し，長鎖非コード RNA の二次構造予測の精度向上が達成されれば，これまでよくわかっていない長鎖非コード RNA の機能解析に大きな貢献が可能となる.さらに，二次構造予測をベースにしたアルゴリズム（RNA 構造アラインメント，RNA-RNA 相互作用予測など）に容易に組み込むことが可能であり，機能性 RNA の機能・構造解析の精度や効率を向上させることができると期待される．

３．研究の方法

機械学習に基づく RNA 二次構造予測においては，二次構造モデルおよびパラメータ最適化手法の選択がその性能に大きな影響を与える．二次構造のモデル化は，二次構造をどのように特徴的な部分構造に分割し，それぞれの特徴量（パラメータ）をどのように割り当てるかで決定される．より精密なモデル化を行えば精度向上が期待できる一方で，最適化すべきパラメータがより多く必要となり，学習が難しくなる．この現象は「次元の呪い」と呼ばれている．実際，(Zakov et al., *J. Comput. Biol.*, 2011)では，それまでの手法に比べて圧倒的に多くのパラメータを用いたモデル化により予測精度を向上させているものの，(Rivas et al., *RNA*, 2012)において過学習の可能性が指摘されている．

そこで本研究では，L1 正則化による特徴選択を実装し，学習データ数に応じて最適なパラメータ数を選択する枠組みによって，次元の呪いによる過学習を回避する．さらに，既存の Turner

熱力学モデルに基づく自由エネルギー最小化法と構造化 SVM によるパラメータ学習法を融合することによってより頑健かつ高精度な RNA 二次構造予測手法の開発を行う．

パラメータ最適化は，頑健な最適化が可能な構造化 SVM (Tsochantaridis et al., *J. Machine Learning Res.*, 2005)によって実装する．学習データとして RNA 配列x_iとその既知二次構造y_iが与えられた時 $(i = 1, \ldots, N)$，構造化 SVM はN組の(x_i, y_i)に対して次の処理を繰り返す:(i)現在のパラメータ$\theta^{(t)}$を用いて配列x_iの二次構造 $\hat{y} = f(x_i; \theta^{(t)})$を予測する．(ii)予測構造$\hat{y}$と正解構造$y_i$を比較し，一致しない部分構造に対応するパラメータを適切に更新して$\theta^{(t+1)}$とする．

本研究では，部分的な構造情報である二次構造プロファイルを完全な二次構造の代わりに学習データとして利用できるように，上記の構造化 SVM を拡張した手法を開発する．すなわち，RNA 配列x_iとその二次構造プロファイルz_iが与えられた時$(i = N + 1, \ldots, N + M)$，$M$組の$(x_i, z_i)$に対して次の処理を繰り返す:(i')現在のパラメータ$\theta^{(t)}$を用いて配列x_iの二次構造$\hat{y} = f(x_i; \theta^{(t)})$を予測する．(ii')予測構造$\hat{y}$の二次構造プロファイル$\hat{z}$と正解二次構造プロファイル$z_i$を比較し，一致しない部位に対応するパラメータを適切に更新して$\theta^{(t+1)}$とする．

４．研究成果

上記のアルゴリズム mxfold を実装し，計算機実験により予測精度を確かめた．データセットは，(Rivas et al., *RNA*, 2012)で使用されているデータセットを用いた．これに含まれる TrainSetA/TestSetA と TrainSetB/TestSetB はそれぞれ異なるソースから取得し，かつ構造的に離れた二次構造となっている．

既存の Turner 熱力学モデルに基づく自由エネルギー最小化法と構造化 SVM によるパラメータ学習法を融合することによってより頑健かつ高精度な RNA 二次構造予測が実現されているかを確認するための実験を行なった．

表 1 は，熱力学モデル (TM)，機械学習モデル (ML)および統合モデル (TM+ML) について，TestSetA と TestSetB で二次構造予測精度の評価を行なった結果である．ここで，ML と TM+ML は TrainSetA で学習を行った．TM+ML モデルが最も正確な予測を行った．TestSetA では，ML のみのモデルよりもわずかに優れていた．TrainSetA と構造的に異なる RNA を含む TestSetB では，TM+ML と ML の精度の差が大きくなっていることがわかる．これはより頑健な予測が可能であることを示している．

さらに，他の RNA 二次構造予測手法との比較を行った (図 1)．TestSetA では，ContextFold (F=0.742)は，TrainSetA で学習したビタビ復号を用いた mxfold (F=0.737)に比べて，わずかに優れてる．一方，TestSetB では，ContextFold (F=0.496)は，TrainSetA で学習したビタビ復号を用いた mxfold(F=0.590)よりもはるかに悪い結果となった．これは，精密なモデルを採用している ContextFold は(Rivas et al., *RNA*, 2012)で指摘されているように過学習に陥っており，一方 mxfold は同じように精密なモデルを採用しているにもかかわらず，熱力学モデルを統合することによって過学習を回避していることを示している．さらに，両データセットから学習したビタビ復号を用いた mxfold が最も正確な予測が得られていることがわかる(F=0.626)．

表 1 熱力学モデル (TM)，機械学習モデル (ML)，統合モデル (TM+ML) による二次構造予測の精度

Model	TestSetA			TestSetB		
	SEN	PPV	F	SEN	PPV	F
TM	0.682	0.659	0.670	0.598	0.485	0.536
ML	0.703	0.764	0.732	0.575	0.550	0.563
TM+ML	0.715	0.761	0.737	0.617	0.565	0.590

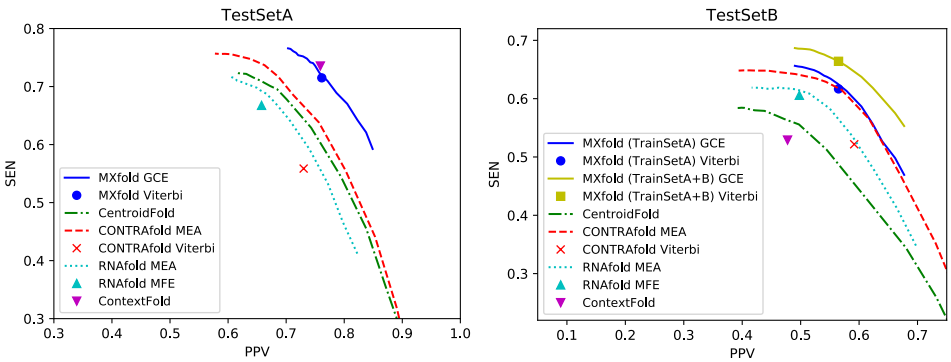

図 1　他の手法との精度比較

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Akiyama Manato、Sato Kengo、Sakakibara Yasubumi	４．巻 16
２．論文標題 A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model	５．発行年 2018年
３．雑誌名 Journal of Bioinformatics and Computational Biology	６．最初と最後の頁 1840025～1840025
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S0219720018400255	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Hirohara Maya、Saito Yutaka、Koda Yuki、Sato Kengo、Sakakibara Yasubumi	４．巻 19
２．論文標題 Convolutional neural network based on SMILES representation of compounds for detecting chemical motif	５．発行年 2018年
３．雑誌名 BMC Bioinformatics	６．最初と最後の頁 526
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12859-018-2523-5	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 －

１．著者名 Mizuguchi Tatsuya、Ito Shino、Sato Kengo、Sakakibara Yasubumi	４．巻 33
２．論文標題 Extension of Question-Answering Program to Automatically Answering the Medical Licensing Examination to Drug Related Questions	５．発行年 2018年
３．雑誌名 Transactions of the Japanese Society for Artificial Intelligence	６．最初と最後の頁 E～I58_1-10
掲載論文のDOI（デジタルオブジェクト識別子） 10.1527/tjsai.e-i58	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7件（うち招待講演　1件／うち国際学会　0件）

１．発表者名 佐藤健吾
２．発表標題 機械学習を用いたRNA二次構造予測
３．学会等名 日本バイオインフォマティクス学会九州地域部会セミナー（招待講演）
４．発表年 2018年

１．発表者名
加藤有己，佐藤健吾，Jakob Hull Havgaard，河原行郎

２．発表標題
深層学習に基づくRNAグアニン４重鎖構造識別法の検討

３．学会等名
第20回日本RNA学会年会

４．発表年
2018年

１．発表者名
Akiyama，M.，Sakakibara，Y.，Sato，K.

２．発表標題
RNA secondary structure prediction using deep learning

３．学会等名
第6回生命医薬情報学連合大会，日本バイオインフォマティクス学会2017年年会

４．発表年
2017年

１．発表者名
青木言太，土谷麻里子，小坂威雄，長谷純崇，佐藤健吾，水野隆一，大家基嗣，榊原康文

２．発表標題
がん細胞株における derived RNA のプロファイル解析

３．学会等名
第19回日本RNA学会年会

４．発表年
2017年

１．発表者名
秋山真那斗，榊原康文，佐藤健吾

２．発表標題
深層学習によるRNA二次構造予測

３．学会等名
第19回日本RNA学会年会

４．発表年
2017年

１．発表者名
Akiyama, M., Sakakibara, Y., Sato, K.

２．発表標題
Improving RNA secondary structure prediction with weak label learning from NGS data

３．学会等名
第5回生命医薬情報学連合大会，日本バイオインフォ マティクス学会2016年年会

４．発表年
2016年

１．発表者名
Taneda, A., Sato, K.

２．発表標題
Inverse folding of two interacting RNA molecules

３．学会等名
第5回生命医薬情報学連合大会，日本バイオインフォ マティクス学会2016年年会

４．発表年
2016年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
MXfold: the max-margin based RNA folding algorithm
https://github.com/keio-bioinformatics/mxfold
MXfold: the max-margin based RNA folding algorithm
https://github.com/keio-bioinformatics/mxfold
Neuralfold
https://github.com/keio-bioinformatics/Neuralfold

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