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　２９　年　　５　月　２６　日現在

機関番号：　３２６１２
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５３３０３４８

研究課題名（和文）高次構造を考慮した超高速RNA構造アラインメント

研究課題名（英文）Ultra-fast RNA structural alignments with pseudoknots

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）　　3,700,000円

研究成果の概要（和文）：RNA構造アラインメントは古くから研究されているにも関わらず，未だに計算量が大きいという問題がある．このため，長鎖非コードRNAやRNAウィルスのような比較的長いRNA配列に関しては，「配列を比べる」という基本的な解析すら厳密手法では満足に行えない状況である．本研究では，期待精度最大化と双対分解に基づく革新的なアルゴリズムにより，シュードノットなどの複雑な高次構造を考慮したRNA構造アラインメントを高速かつ高精度に計算する手法を開発した．

研究成果の概要（英文）：Despite the fact that RNA structural alignments have been studied for a long time, there is still a problem that the computational complexity is still large. For this reason, we cannot perform even a basic analysis of "comparing sequences" by exact methods for relatively long RNA sequences such as long non-coding RNAs and RNA viruses. In this research, we developed a fast and accurate method of calculating RNA structural alignments with consideration of complicated higher order structures such as pseudoknots by a novel algorithm based on maximizing the expected accuracy and the dual decomposition.

研究分野：　バイオインフォマティクス

キーワード：　RNA二次構造　構造アラインメント　期待精度最大化　双対分解　整数計画法

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，miRNA や snoRNA など配列長が比較的短い小分子非コード RNA small noncoding RNA; small ncRNA）が中心であった．しかし近年 ENCODE 計画などの成果から，長鎖非コード RNA（long noncoding RNA; lncRNA）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．

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，機能と構造の間には強い相関があると考えられている．そのため，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．

配列アラインメントは，配列情報解析においてもっとも基本的な技術の一つである．機能性 RNA の配列情報解析においては，信頼性が高いアラインメントを得ることができれば，そこから共通二次構造を予測することによって，構造よりも精度がより高い二次構造を予測したり，機能性 RNA 遺伝子や構造モチーフを予測することが可能となる．しかしながら，機能性 RNA 配列は配列相同性が低いために，信頼性が高いアラインメントを得るためには各 RNA 配列の正しい二次構造を知る必要があり，いわゆる「ニワトリとタマゴ」問題となってしまう．そこで RNA 配列情報解析の分野では，アラインメントと二次構造予測を同時に行うアルゴリズム（RNA 構造アラインメント）の研究が古くから行われているが，シュードノットや非正規塩基対が存在しないと仮定しても未だに計算量が大きいという問題がある．そのため，lncRNA のような長い RNA 配列群の多重アラインメントを計算するためには，精度を犠牲にしてすべての構造を無視するか，さらに単純化した構造のみを扱うアルゴリズムを用いる他ないのが現状である．したがって，本研究課題が目指す，拡張二次構造や複合二次構造などの複雑な高次構造を考慮した超高速 RNA 構造アラインメントは挑戦的研究課題であると言える．

２．研究の目的

本研究では，革新的なアルゴリズムにより複雑な高次構造を考慮した RNA 構造アラインメントを高速かつ高精度に計算する手法の開発を行う．具体的には以下を目的とする：まず，先行研究と同様にシュードノットおよび非正規塩基対が存在しないと仮定し，期待精度最大化と双対分解に基づき RNA 構造アラインメントを計算するアルゴリズムを開発する．これを拡張し，シュードノットと非正規塩基対を含む拡張二次構造を考慮するモデルを実装する．さらに，RNA-RNA 相互作用を考慮し，複合二次構造予測と構造アラインメントを同時に行うモデルに拡張する．

３．研究の方法

RNA 構造アラインメントを定式化し，期待精度最大化原理に基づいて目的関数を設計する．双対分解によって RNA 構造アラインメントを配列アラインメントと二次構造予測に分解して独立に最適化し，制約が満たされない部分については反復的に改善していく．

双対分解によって分解した部分問題を，シュードノットおよび非正規塩基対からなる拡張二次構造に対応したモデルに置き換えることによって，複雑な高次構造を考慮した RNA 構造アラインメントを実現する．

RNA-RNA 相互作用（複合二次構造）予測を統合し，複合二次構造予測と構造アラインメントを同時に行うモデルの実現を目指す．

４．研究成果

局所 RNA 構造アラインメントをマルチプルアラインメントに拡張するための定式化を行い，さらにそれに基づく実装を行った．既存手法との比較実験を行い，精度の面で優れていることを示した．

他分子との複合構造を考慮するアラインメントモデルの開発に向けて，RNA-タンパク質間の塩基・残基レベルの相互作用を定式化し，機械学習に基づき最適な複合構造を推定するアルゴリズムの開発を行った．最適化アルゴリズムの変更により，大幅な予測精度の向上が見られた．また，大量の訓練データを用意できる弱ラベル付きデータを用いた機械学習を実現するための定式化を行い，プロトタイプの実装を行った．

RNA-RNA 相互作用予測法 RactIP のアルゴリズムを改良し，配列アクセシビリティを考慮することによってより高精度な相互作用予測を実現し，計算機実験およびウェット実験を実施することにより，改良版 RactIP の性能を確かめた．

derived small RNA のトランスクリプトーム解析に局所 RNA 構造アラインメントを応用する手法の定式化を行い，さらにそれに基づく実装を行なった．これは，RNA 二次構造を考慮しながらトランスクリプトームのカバレッジベクターのアラインメントを実行する画期的な手法である．derived small RNA の解析を行う既存手法との比較実験を行い，精度の面で優れていることを示した．

RNA-RNA 相互作用予測法 RactIP のアルゴリズムを改良し，また，RNA 配列解析を統合的に実行できる Web サーバ Rtools の開発に貢献した．

本課題で開発した技術を遺伝子予測に応用し，RNA-seq によって得られた情報をこれまでの遺伝子予測に統合し，より正確な予測を可能とするアルゴリズムを開発した．

本課題で開発した技術を糖鎖インフォマティクスに応用し，マススペクトルから糖鎖構造を推定する手法の開発を行い，既存の手法に比べて非常に高い精度での推定が可能であることを示した．

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　８件）

[1] Kato, Y., Mori, T., Sato, K., Maegawa, S., Hosokawa, H., Akutsu, T.: An accessibility-incorporated method for accurate inference of RNA-RNA interactions from sequence data, Bioinformatics, 33(2):202-209 (Jan. 2017)**，査読あり**
DOI: 10.1093/bioinformatics/btw603

[2] Hamada, M., Ono, Y., Kiryu, H., Sato, K., Kato, Y., Fukunaga, T. Mori, R., Asai, K.: Rtools: a web server for various secondary structural analyses on single RNA sequences, Nucleic Acids Research, 44:W302-W307 (Jul. 2016)**，査読あり**
DOI: 10.1093/nar/gkw337

[3] Tsuchiya, M., Amano, K., Abe, M., Seki, M., Hase, S., Sato, K., Sakakibara, Y.: SHARAKU: An algorithm for aligning and clustering read mapping profiles of deep sequencing in non-coding RNA processing, Bioinformatics, 32(12):i369-i377 (Jun. 2016)**，査読あり**
DOI: 10.1093/bioinformatics/btw273

[4] Inatsuki, T, Sato, K., Sakakibara, Y.: Prediction of gene structures from RNA-seq data using dual decomposition, IPSJ Transactions on Bioinformatics, 9:1-6 (Mar. 2016)**，査読あり**
DOI: 10.2197/ipsjtbio.9.1

[5] Sato, K., Kuroki, Y., Kumita, W., Fujiyama, A., Toyoda, A., Kawai, J., Iriki, A., Sasaki, E., Okano, H., Sakakibara, Y.: Resequencing of the common marmoset genome improves genome assemblies and gene-coding sequence analysis, Scientific Reports, 5:16894 (Nov. 2015)**，査読あり**
DOI: 10.1038/srep16894

[6] Kumozaki, S., Sato, K., Sakakibara, Y.: A machine learning based approach to de novo sequencing of glycans from tandem mass spectrometry spectrum, IEEE/ACM Transactions on Computatinal Biology and Bioinformatics, 12(6):1267-1274 (Dec. 2015) (*Joint First Authors)**，査読あり**
DOI: 10.1109/TCBB.2015.2430317

[7] Kamada, M., Hase, S., Fujii, K., Miyake, M., Sato, K., Kimura, K., Sakakibara, Y.: Whole-genome sequencing and comparative genome analysis of Bacillus subtilis strains isolated from non-salted fermented soybean foods, PLoS ONE, 10(10):e0141369 (Oct. 2015)**，査読あり**
DOI: 10.1371/journal.pone.0141369

[8] Afiahayati, Sato, K., Sakakibara, Y.: MetaVelvet-SL: An extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning, DNA Research, 29(1):69-77 (Feb. 2015)**，査読あり**
DOI: 10.1093/dnares/dsu041

〔学会発表〕（計１０件）

[1] Hamada, M., Ono, Y., Kiryu, H., Sato, K., Kato, Y., Fukunaga, T., Mori, R., Asai, K.: Rtools: a web server for various secondary structural analyses

on single RNA sequences, The 21st Annual meeting of the RNA Society (RNA2016), 2016 年 6 月 28 日～7 月 2 日，京都国際会館（京都府京都市）

[2] Kato, Y., Mori, T., Sato, K., Maegawa, S., Hosokawa, H., Akutsu, T.: Accurate prediction of RNA-RNA interactions from sequence data incorporating interaction site accessibility, The 21st Annual meeting of the RNA Society (RNA2016), 2016 年 6 月 28 日～7 月 2 日，京都国際会館（京都府京都市）

[3] 穴水拓郎，榊原康文，佐藤健吾: 双対分解によるマルチプルアラインメント，情報処理学会バイオ情報学研究会，情報処理学会バイオ情報学研究会，2015 年 9 月 12 日，慶應義塾大学（神奈川県横浜市）

[4] Aoto, Y., Hachiya, T., Okumura, K., Hase, S., Sato, K., Wakabayashi, Y., Sakakibara, Y.: Development of high-accuracy clustering algorithm based on statistical test results, 情報処理学会バイオ情報学研究会，2015 年 9 月 12 日，慶應義塾大学（神奈川県横浜市）

[5] Sato, K., Kashiwagi, S., Sakakibara, Y.: A max-margin model for predicting residue-base contacts in protein-RNA interactions, GIW/InCoB 2015, 2015 年 9 月 9～11 日，科学未来館（東京都台東区）

[6] Inatsuki, T, Sato, K., Sakakibara, Y.: Prediction of gene structures from RNA-seq data using dual decomposition, 情報処理学会バイオ情報学研究会，2015 年 6 月 4 日，沖縄科学技術大学院大学（沖縄県恩納村）

[7] 佐藤健吾，柏木駿也，榊原康文: 機械学習を用いたタンパク質と RNA のコンタクト予測，第 3 回生命医薬情報学連合大会，2014 年 10 月 2～4 日，仙台国際センター（宮城県仙台市）

[8] 佐藤健吾，柏木駿也，榊原康文: 機械学習を用いたタンパク質と RNA のコンタクト予測，第 16 回日本 RNA 学会年会，2014 年 7 月 23～25 日，ウィンクあいち（愛知県名古屋市）

[9] Kumozaki, S., Sato, K., Sakakibara, Y.: A machine learning based approach to de novo sequencing of glycans from mass spectrometry data, 第 2 回生命医薬情報学連合大会，2013 年 10 月 29～31 日，タワーホール船堀（東京都江戸川区）

[10] 雲崎翔太郎，佐藤健吾，榊原康文: 機械学習を用いたマススペクトルデータからの糖鎖構造推定法の開発，情報処理学会バイオ情報学研究会，2013 年 6 月 27 日，沖縄科学技術大学院大学（沖縄県恩納村）

〔その他〕
ホームページ等
Rtools:
http://rtools.cbrc.jp/

DAFS:
https://github.com/satoken/dafs

RactIP:
https://github.com/satoken/ractip

６．研究組織

(1)研究代表者
佐藤 健吾(SATO KENGO)
慶應義塾大学・理工学部・講師
研究者番号:20365472

(2)研究分担者

該当なし

(3)連携研究者

該当なし