様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成 25 年 6 月 2 日現在

機関番号：14603
研究種目：挑戦的萌芽研究
研究期間：2011～2012
課題番号：23650153
研究課題名（和文）　機械学習と最適化に基づく RNA タンパク質相互作用予測
研究課題名（英文）　RNA-protein interaction prediction based on machine learning and optimization

研究代表者
関　浩之（SEKI HIROYUKI）
奈良先端科学技術大学院大学・情報科学研究科・教授
研究者番号：80196948

研究成果の概要（和文）：本研究では RNA 配列情報解析の視点から、まず、単一 RNA 塩基配列が与えられた時の複雑な折り畳み構造を予測する数理的手法を開発した。次に、複数の配列が与えられた場合に、配列間の対応付けと各構造を同時に予測する手法の開発も行った。これら開発手法について実際の生物データを用いて予測性能評価を行い、既存手法に勝るとも劣らない計算速度と予測精度を上げることに成功した。この技術は実用的な RNA タンパク質相互作用予測への十分な基盤を与えるものと期待される。

研究成果の概要（英文）：From a viewpoint of RNA sequence analysis, we first developed a mathematical method for predicting complex folding structures of an RNA when its single sequence is given. We then developed a computational method for aligning and folding multiple RNA sequences simultaneously. All of these methods were validated on real biological data, achieving fast run-time and good prediction accuracy at least comparable to those of earlier methods. The proposed technologies are expected to provide a good practical foundation for RNA-protein interaction prediction.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	2,800,000	840,000	3,640,000

研究分野：情報学
科研費の分科・細目：生体生命情報学
キーワード：バイオインフォマティクス，RNA-タンパク質相互作用，RNA２次構造，RNA 間相互作用，RNA 構造アラインメント，機械学習，最適化

１．研究開始当初の背景

RNA とタンパク質の相互作用解析の重要性にもかかわらず、情報科学における手法を用いてモデル化する研究は極めて少ない。一方で、計算機に基づく個々の分子内および同一種の分子間の相互作用予測法には種々のアプローチが存在する。これまで研究代表者らは、一般的な結合２次構造からなる RNA 間相互作用の予測を形式文法に基づくアプローチにより行った[1]。また、タンパク質配列中の

βシート部位を予測するために、動的計画法に基づく高速なアルゴリズムを開発した[2]。さらに、期待精度最大化原理と整数計画法を融合させた RNA 間相互作用予測法を開発し、世界最高レベルの計算速度と予測精度を達成した[3]。ここで、整数計画法のモデル記述能力が高い点と、期待精度最大化に基づく解空間の大幅な縮小が可能なことから、[3]を発展させ、入力として RNA 配列とタンパク質配列の組を与えた場合にも、RNA-タンパク質相互作用のモデル化および高速な予測法の開発が可能ではないかとの着想に至った。

２．研究の目的

生体内に存在する機能性 RNA（以下、RNA と略）は多くの場合、タンパク質などの他の生体分子と結合することでその機能を発現することが知られている。また、両者とも折り畳み構造を持つことが多く、その構造の取り方によって相互作用の方法に影響を与えるものと考えられている。RNA とタンパク質の相互作用予測は、その取り得る結合構造の複雑さやバリエーションの多さなどのため、生体生命情報学では未だ確立された予測手法がなく、発展途上かつ挑戦的研究課題であると言える。本研究では、研究代表者らがこれまで蓄積してきた離散最適化法に基づく RNA およびタンパク質の２次構造予測法を発展させ、RNA-タンパク質相互作用のモデル化を組織的に行い、精度の良い相互作用予測法の開発を目指す。

本研究では、RNA-タンパク質相互作用予測に対して、理論的基盤の構築及び実用的アルゴリズムの開発を目指す。具体的には以下を目標とする。

- RNA の塩基とタンパク質のアミノ酸の相互作用傾向を定量的に計算する機械学習モデルを開発する。
- 相互作用予測に要する計算コストの削減が可能な数理計画（整数計画）モデルを設計する。
- 開発手法を計算機に実装し、その性能評価を行う。さらに、提案手法を Web サーバーに実装し、商用目的以外には自由に利用できるようにすることを目指す。

多くの RNA は標的となるタンパク質と相互作用することにより機能を発現するため、上記の目標を達成することにより、計算機を用いた生体高分子の機能推定の発展に貢献できると考えている。

３．研究の方法

本研究では初めに、RNA-タンパク質相互作用を定量化するスコア関数を機械学習モデルにより設計する。次に、相互作用予測問題を記述する数学モデルを設計する。これは、整数計画問題への定式化を表すが、このとき問題の数学的構造をなるべく簡略化することを目指す。次に、得られたモデルの目的関数に機械学習で得られたスコアを組み合わせて計算機上に実装し、既知の相互作用データを用いて提案手法の有効性を検証する。なお、実際に整数計画問題を解く際は商用の最適化ソルバーのライブラリを利用する。計算機実験を通して適宜改良を行い、十分な予測精度と実行効率が得られた段階で、成果を広く発信するため Web サーバーを開発し公開する。

3.1 理論的検討と実験的評価

3.1.1 相互作用スコアの計算

RNA の１塩基とタンパク質の１アミノ酸の相互作用スコアを計算するため、機械学習に基づくモデルを開発する。これは相互作用部位を予測する上で重要なステップとなる。具体的に、既知の RNA-タンパク質複合体から相互作用スコアを自動的に計算するようなシス

テムを開発する。ただし、完全な自動化を目指すのではなく、人間による知見や経験則などを組み込めるように工夫する。

3.1.2 整数計画問題によるモデル化

提案手法[3]のときと同様に、RNA-タンパク質相互作用予測に対して計算コストの削減が可能となるような定式化を考える。具体的には、２本の配列の２次構造の形状に関する制約条件を、整合性を保ちながら可能な限り除去し、目的関数の係数に集約させる。なお、予測結合構造の期待精度を最大化することが目標となる。そのため、RNA とタンパク質の構造及びその相互作用に対応するスコアに、適切な確率を割り当てる必要がある。RNA２次構造に対しては、その配列で任意の塩基対を形成する確率を表す塩基対確率行列[4]を利用することを考える。タンパク質βシートに対しては、研究目的の項目の文献[2]で用いたアミノ酸残基間のコンタクトポテンシャルを確率に変換することで対応できると思われる。また、相互作用に対しては、手順１のスコアを用いることにする。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。

3.1.3 計算機実験

上述の整数計画問題を Java 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、その性能の高さから世界中で広く使用されている商用の最適化ソルバー CPLEX のライブラリを用いる。その後、相互作用することが実験的に確かめられているRNA-タンパク質複合体と提案手法が出力した予測構造を比較し、予測精度や計算速度などを評価することで、有効性の検証や問題点の検出を行う。

3.2 提案手法の改良

3.2.1 整数計画モデルの改良

前年度で得られた整数計画モデルを改良し、計算速度の向上を目指す。最適化問題を直接解くのが困難な場合に用いる手法の１つとして Lagrange 緩和があり、制約条件の一部の制約式に重みを乗じて、制約条件から除去し目的関数に追加する。もし解がその制約を破れば、ペナルティとして目的関数の最大化に影響する仕組みである。Lagrange 緩和は生体生命情報学に現れる最適化問題おいて、文献[5]を皮切りにその有用性が示されている。どの制約条件を緩和するのかを決定することは難しいかもしれないが、後述の計算機実験と並行して行って、計算速度の向上が最も顕著な緩和制約式を特定する。

3.2.2 計算機実験

前年度と同様に、改良モデルを計算機に実装し、既知の相互作用データを用いて提案手法の有効性を検証する。このとき、予測精度の向上が見られないときは、スコア計算のための機械学習アルゴリズムの改良を適宜行う。また、準最適解も計算できるようなモデルに改良することで対応する。

４．研究成果

4.1 理論的検討と実験的評価

初年度では、基盤予測法として１本鎖 RNA 配列の構造解析に焦点を当て、昨年度分担者が開発した RNA 間相互作用予測法 RactIP の方法論を応用し、シュードノットと呼ばれる複雑な２次構造を考慮した高速な予測法を開発した。まず、シュードノットを含む２次構造の事後確率分布を、シュードノットを含まない２次構造の確率分布の積へ分解を行った。次に、最適化における目的関数の設定において、予測２次構造の期待精度の最大化に主眼を置き、期待精度最大化問題を閾値カット付き整数計画問題として実現した。さら

に、多重配列アラインメントが与えられたとき、その共通２次構造を予測するようにモデルの拡張を行った。提案手法（IPknot）の性能評価を構造既知の配列データセットを用いて行ったところ、複数の既存手法と比べて精度の点では同等以上、速度の点では桁違いの高速性を実現した。最後に、研究成果を広く世界に向けて発信するため、IPknot の Web サーバーを開発し、同時期に補完した RactIP のサーバーと統合することで、世界最高レベルの高速性を実現する RNA 構造解析ツールセットを公開した。これにより、RNA 配列情報解析の立場から生体分子間相互作用予測を行うための基盤技術が確立された。

4.2 提案手法の改良

２年目（平成２４年度）は、前年度末に課題に挙げた、塩基アミノ酸間のスコア関数の定量化のためには、進化の過程での配列間の保存情報が有力な手掛かりを与えると考え、まずは RNA 配列のみに焦点を絞り、RNA 配列の構造アラインメント問題を効率的に解く手法の開発に着手した。具体的には、２本以上の複数の RNA 配列同士で、各配列が取り得る構造を考慮しながら同時に配列間の対応付けをとる、いわゆる同時構造アラインメント問題に対して、効率の良い計算手法 DAFS を開発した。DAFS では、構造アラインメント空間上で定義される確率分布を、各２次構造上で定義される確率分布とアラインメント空間上で定義される確率分布の積で近似することで、計算効率の改善に寄与している。また、予測構造アラインメントの精度が最大となるように整数計画問題として定式化を行い、さらにそれを双対分解と呼ばれる最適化法の技術を用いて、整数計画問題を直接解くことよりも高速に解くことが可能となった。既存の構造アラインメントデータ上で網羅的に計算機実験を行った結果、DAFS の予測精度は既存の手法と比べて同等以上、そしてより高速に動作することを実証した。アラインメントに基づく配列解析技術は、RNA-タンパク質間のように、単一配列解析が難しい場合に有効な手法と考えられ、前年度に得られた RNA 配列解析基盤技術と併せて、本研究の成果が実用的な RNA タンパク質相互作用予測への十分な基盤を与えるものと期待される。

参考文献

[1] Y. Kato, T. Akutsu and H. Seki, A grammatical approach to RNA-RNA interaction prediction, *Pattern Recognit.*, 42, 531-538, 2009.

[2] Y. Kato, T. Akutsu and H. Seki, Dynamic programming algorithms and grammatical modeling for protein beta-sheet prediction, *J. Comput. Biol.*, 16, 945-957, 2009.

[3] Y. Kato, K. Sato, M. Hamada, Y. Watanabe, K. Asai and T. Akutsu, RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming, *Bioinformatics*, 26, i460-i466, 2010.

[4] J.S. McCaskill, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers*, 29, 1105-1119, 1990.

[5] A. Caprara *et al.*, 1001 optimal PDB structure alignments: integer programming methods for finding the maximum contact map overlap, *J. Comput. Biol.*, 11, 27-52, 2004.

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４件）

[1] Kengo Sato, Yuki Kato, Tatsuya Akutsu, Kiyoshi Asai and Yasubumi Sakakibara, DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition, Bioinformatics, 28, 3218-3224, 2012, 査読有

[2] 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, 40, W29-W34, 2012, 査読有

[3] Unyanee Poolsap, Yuki Kato, Kengo Sato and Tatsuya Akutsu, Using binding profiles to predict binding sites of target RNAs, Journal of Bioinformatics and Computational Biology, 9(6), 697-713, 2011, 査読有

[4] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu and Kiyoshi Asai, IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming, Bioinformatics, 27(13), i85-i93, 2011, 査読有

〔学会発表〕（計１４件）

[1] 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", 2013.1.9, 東京都文京区

[2] Yuki Kato, RNA structural alignment using dual decomposition, Herbstseminar der Bioinformatik (2012), 2012.10.4, チェコ ドウビツェ

[3] Yuki Kato, Fast and accurate prediction of RNA pseudoknotted structures using Yuki Kato integer programming, International Workshop on RNA (Benasque2012), 2012.7.27, スペイン　ベナスケ

[4] Yuki Kato, Fast and accurate prediction of RNA-RNA interactions using integer programming, International Workshop on RNA (Benasque2012), 2012.7.27, スペイン　ベナスケ

[5] 佐藤 健吾, 加藤 有己, 阿久津 達也, 浅井 潔, 榊原 康文, 双対分解によるRNA構造アラインメント, 第 28 回情報処理学会バイオ情報学研究会, 2012.3.28, 仙台市

[6] Kengo Sato, Yuki Kato, Tatsuya Akutsu and Kiyoshi Asai, Rtips: fast and accurate tools for RNA 2D structure prediction using integer programming, 第 34 回日本分子生物学会年会, 2011.12.14, 横浜市

[7] Yuki Kato, Kengo Sato, Kiyoshi Asai and Tatsuya Akutsu, Rtips: fast and accurate tools for RNA 2D structure prediction using integer programming, CBI/JSBi2011 合同大会, 2011.11.8, 神戸市

[8] Kengo Sato, Yuki Kato, Tatsuya Akutsu, Kiyoshi Asai and Yasubumi Sakakibara, Simultaneous aligning and folding of RNA sequences by dual decomposition, CBI/JSBi2011 合同大会, 2011.11.8, 神戸市

[9] Yuki Kato, Kengo Sato, Kiyoshi Asai and Tatsuya Akutsu, Rtips: fast and accurate tools for RNA 2D structure prediction using integer programming, ICR Symposium to Celebrate the Bioinformatics Center's 10 Year Anniversary and New Restructuring, 2011.8.29, 京都府宇治市

[10] 加藤有己, 複雑な RNA２次構造予測のための高速計算ソフトウェアの開発と今後の展開, 次世代バイオインフォマティクス研究会（招待講演）, 2011.8.2, 札幌市

[11] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu and Kiyoshi Asai, IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming, 19th Annual International Conference on Intelligent Systems for Molecular Biology and 10th European Conference on Computational Biology (ISMB/ECCB2011), 2011.7.17, オーストリア ウィーン

[12] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu and Kiyoshi Asai, IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming, Sixteenth Annual Meeting of the RNA Society (RNA2011), 2011.6.17, 京都市

[13] Yuki Kato, Kengo Sato, Michiaki Hamada, Yoshihide Watanabe, Kiyoshi Asai and Tatsuya Akutsu, RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming, Sixteenth Annual Meeting of the RNA Society (RNA2011), 2011.6.15, 京都市

[14] Hiroyuki Seki, A comparative approach to RNA pseudoknotted structure prediction based on multiple context-free grammar, 8th Asian Workshop on Foundation of Software, 2011.5.13, 中国 上海

〔図書〕（計１件）

[1] 加藤 有己, 実験医学増刊号, 使えるデータベース・ウェブツール, 第４章第８節「二次構造に基づく RNA 配列解析ソフトウェアの進展」, 229-236, 2011, 羊土社

〔その他〕

ホームページ等

Rtips: RNA structure prediction using IP scheme
http://rna.naist.jp/

６．研究組織

(1)研究代表者

関　浩之（SEKI HIROYUKI）
奈良先端科学技術大学院大学・情報科学研究科・教授
研究者番号：80196948

(2)研究分担者
加藤　有己（KATO YUKI）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：10511280