様式 C-19

科学研究費補助金研究成果報告書

平成２２年５月３１日現在

研究種目：若手研究（スタートアップ）
研究期間：2008～2009
課題番号：20800023
研究課題名（和文）　形式文法に基づく 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 formal grammars

研究代表者
加藤　有己（KATO YUKI）
京都大学・化学研究所・研究員（科学研究）
研究者番号：10511280

研究成果の概要（和文）：本研究ではまず、形式文法を用いた RNA-RNA 相互作用予測を行った。次に、形式文法によるタンパク質の２次構造のモデル化と、タンパク質配列における２次構造領域予測のための高速なアルゴリズムを開発した。さらに、RNA 配列における結合部位予測のための、プロファイルに基づく効率的なアルゴリズムの開発を行った。いずれの手法に対しても実際の生物データを用いて予測性能評価を行い、既存研究に勝るとも劣らない予測精度を上げることに成功した。

研究成果の概要（英文）：In this study, we first performed RNA-RNA interaction prediction using formal grammars. We then modeled protein secondary structures and developed a fast algorithm for secondary structure prediction given protein sequences. Furthermore, we presented an efficient algorithm for binding site prediction in RNA sequences, based on binding profiles. All of these methods were validated on real biological data and achieved good prediction accuracy at least comparable to that of earlier methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,160,000	348,000	1,508,000
2009 年度	1,040,000	312,000	1,352,000
年度			
年度			
年度			
総　計	2,200,000	660,000	2,860,000

研究分野：バイオインフォマティクス
科研費の分科・細目：情報学・生体生命情報学
キーワード：RNA２次構造、RNA-RNA 相互作用、形式文法、タンパク質２次構造、β シート、動的計画法、RNA-タンパク質相互作用

１．研究開始当初の背景

ポストゲノム時代において、多くの noncoding RNA（以下単に RNA と書く）は触媒機能や遺伝子の転写後調節機能を持つ能動的な役割を果たす分子として認識されている。RNA は折り畳み構造をとることにより、その機能を発現することが多い。分子生物学の経験則から、構造と機能の間には相関があると言われており、機能推定のためには構造を知ることが重要となる。しかしながら、RNA の立体構造を実験的に決定することは、分子内の運動性の高さのために容易ではない。そのため、離れた塩基間の水素結合の情報のみを表す２次構造を、配列データから情報科学的観点に基づいて予測する様々な手法が提案されている。一方、生体内でタンパク質も折り畳み構造をとり、多くの RNA はタンパク質と結合することで機能を発現することが知られている。ところが、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数少ないと思われる。

RNA 認識モチーフ（RRM）と呼ばれるタンパク質中のドメインは、RNA の認識において必要不可欠であると考えられている。RRM ドメインは約 90 個のアミノ酸残基からなり、RNP1、RNP2 と呼ばれる保存された短いモチーフを含む。RRM はαヘリックス（α）とβストランド（β）が組み合わさった構造をとり、そのトポロジーは配列上の順で$\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$となる。立体構造は$\beta_4$ $\beta_1\beta_3\beta_2$の順に並んだストランド４本からなる逆平行βシート、及びそれらに挟まれる２本のαヘリックスα_1、α_2から構成される。RNP1 と RNP2 はそれぞれβ_3、β_1に位置しており、その中の芳香族残基が RNA との結合に重要な役割を果たすと考えられている。RRM が認識する RNA は一本鎖で構造を持たないものと、ヘアピンループ構造や分岐構造をとるものがある。この RNA 認識機構はタンパク質の種類によって変則的になるものの、ある程度普遍的であると考えられており、相互作用予測ではタンパク質や RNA の２次構造解析が鍵を握っていると言える。

本研究の着想に至った経緯は次の通りである。これまで研究代表者は、文脈自由文法では表現できないシュードノットと呼ばれる構造を考慮した RNA の２次構造を、多重文脈自由文法（MCFG）を用いてモデル化し、その構文解析アルゴリズムにより２次構造予測を行った経緯がある。MCFG は文脈自由文法の自然な拡張であり、パラメータを調節することで文法の表現能力を変化させることが可能な柔軟性の高い文法モデルである。ここで、MCFG は任意項の文字列を並行導出可能である点に着目し、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、MCFG による相互作用のモデル化が可能ではないかとの着想に至った。

２．研究の目的

(1) 形式言語理論の観点から、RNA タンパク質相互作用のモデル化に最適な MCFG の部分クラスを同定する。

(2) 予測のための多項式時間の構文解析アルゴリズムを設計する。

(3) 開発したアルゴリズムを計算機に実装し、既知の相互作用データを用いて提案手法の性能評価を行う。

３．研究の方法

(1) 分岐のないヘアピンループ構造をとるRNAとタンパク質との相互作用をモデル化する。また、計算の効率化のため、タンパク質の２次構造予測においてαヘリックス領域を無視し、βシート領域のみを考慮する。以上の制約を持つRNA-タンパク質複合体をモデル化することができるMCFGの部分クラスを同定する。具体的には、文法の非終端記号から並行に導出される文字列の次元や、規則の右辺の非終端記号の個数など、調節可能なパラメータを検討し、文法モデルが上記のRNA-タンパク質相互作用のクラスをモデル化するのに必要十分な表現能力を備えているか否かを理論的に考察する。

(2) 形式文法に基づく本研究のモデルでは、RNAとタンパク質の構造及びその相互作用に対応する規則に、生物学的に意味のある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、塩基対のスタッキングエネルギーに対応する確率を割り当てることが考えられる。タンパク質βシートに対しては、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、既存の塩基アミノ酸相互作用スコアなどを利用することを考えている。その後、確率的拡張モデルに対する多項式時間の構文解析アルゴリズムを動的計画法に基づいて設計する。

(3) 上記構文解析アルゴリズムをC++言語を用いて計算機に実装する。その後、構造既知のRNA-タンパク質複合体とアルゴリズムが出力した予測構造を比較し、提案手法の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などを行う。

４．研究成果

(1) 研究開始当初では、当時進行中であった形式文法の構文解析技術に基づくRNA-RNA相互作用予測を引き続き行った。これは形式文法に基づくRNA-タンパク質相互作用予測に対する基盤を与えるものとして重要な意味を持つ。ここでは、多重文脈自由文法（MCFG）を用いてキッシングヘアピン（図１参照）と呼ばれる複雑な結合２次構造をモデル化する文法RNA-RNA相互作用文法（RIG）を提案した。そして、RIGの確率モデルに対する $O(n^6)$ 時間（n は２本のうち長い配列の長さ）の構文解析アルゴリズムを設計し、結合構造既知の複数のRNA-RNA対を入力として予測実験を行い、提案手法の有効性の検証を行った。その結果、配列内部の塩基対および配列外部の塩基対に基づく予測精度は平均にして約88%と高い数値を上げることができた。これは当時の動的計画法に基づく既存予測法と同等の性能である。なお、RIGに基づく実験結果は、その後複数の国際論文に引用され、RNA間相互作用予測実験のベンチマークとして使われたことは特筆すべきである。

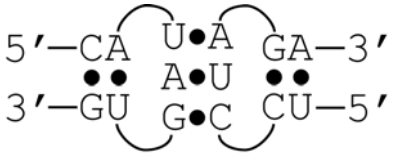

図１　キッシングヘアピン構造。

(2) 次に、タンパク質のβシートはRNA

との相互作用と密接に関係しているため、多重文脈自由文法（MCFG）に基づくβシートに特化した構造のモデル化、ならびにβシート領域を予測する動的計画法（DP）を開発した。ここで扱うβシートのトポロジーとして、図２にあるようなアップダウン型の逆平行βシートおよびβバレルに焦点を当てた。計算機実験が示す DP による平均予測精度は、アミノ酸残基ごとの評価で約 79%、２次構造要素の重なり度合いを評価して約 85%となり、既存手法と同等以上の性能を示した。さらに、任意の平面的βシートを予測する問題は計算量的に困難な問題（NP 困難）であることを証明した。

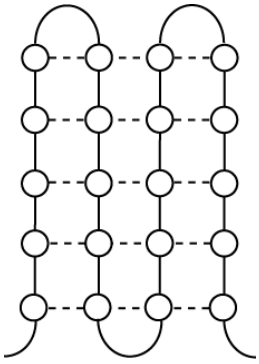

図２　アップダウンβシート。白丸はアミノ酸を表し、破線は水素結合を表す。

(3) 上記の研究結果を踏まえ、RNA と、それに複雑なトポロジーで結合するβシートの両者の構造を多重文脈自由文法（MCFG）の枠組みの中でモデル化することを行った。しかしながら、構文解析に基づく予測に要する時間計算量が入力配列の長さの８乗のオーダーとなり、実際の構造予測には不向きであることが明らかとなった。これは複雑な折り畳み構造を含意する両配列をアルゴリズムに入力して同時に構造を予測することには大きな負荷がかかることを示唆している。そのため、異なる観点から見て計算量の少ない RNA-タンパク質相互作用のモデル化を行う必要に迫られた。

(4) そこで、研究期間後半では視点を変え、構造予測の対象を RNA 配列のみに限定し、相互作用する相手（タンパク質など）はプロファイルという形で情報を縮退させることを考えた。この考え方は新規標的の発見に応用可能であるという意味で重要である。まず、RNA-RNA 相互作用予測に向けて、相互作用することが知られている RNA の結合部位の情報をもとにプロファイルを作成し（図３参照）、２次構造予測の動的計画法の漸化式に組み込むことで、$O(n^3)$ 時間（n は入力配列の長さ）の結合部位予測法を開発した。これは先述の $O(n^6)$ 時間の文法モデル（RIG）の半分の計算量になっており、実際に計算機実験を行った結果、高速にアルゴリズムが動作することを確認できた。また、結合２次構造既知のデータに対し予測実験を行った結果、既存の手法よりも精度向上が見られた。さらに、提案手法を用いて、特定の RNA に対する推測標的 mRNA（の結合部位）をいくつか示唆し、複数の標的を持つ RNA 調節機構に関する洞察を与えた。残念ながら研究期間内に、タンパク質のプロファイルを作成して RNA 配列におけるタンパク質結合部位予測の計算機実験を行えなかったが、このプロファイルに基づく予測手法が、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。

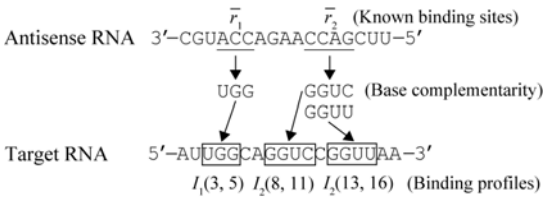

図３　結合プロファイルの例。

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 3 件）

[1] Yuki Kato, Tatsuya Akutsu and Hiroyuki Seki, **Dynamic programming algorithms and grammatical modeling for protein beta-sheet prediction**, *Journal of Computational Biology*, vol. 16, no. 7, pp. 945-957, Jul. 2009, 査読有り.

[2] Yuki Kato, Tatsuya Akutsu and Hiroyuki Seki, **A grammatical approach to RNA-RNA interaction prediction**, *Pattern Recognition*, vol. 42, issue 4, pp. 531-538, Apr. 2009, 査読有り.

[3] Unyanee Poolsap, Yuki Kato and Tatsuya Akutsu, **Prediction of RNA secondary structure with pseudoknots using integer programming**, *BMC Bioinformatics*, vol. 10, suppl. 1, Jan. 2009, 査読有り.

〔学会発表〕（計 13 件）

[1] Unyanee Poolsap, Yuki Kato and Tatsuya Akutsu, **Dynamic programming algorithms for RNA structure prediction with binding sites**, The Fifteenth Pacific Symposium on Biocomputing (PSB2010), 2010 年 1 月 7 日, ハワイ.

[2] Nobuyoshi Mizoguchi, Yuki Kato and Hiroyuki Seki, **Pairwise RNA pseudoknotted structure prediction based on stochastic grammar**, The 20th International Conference on Genome Informatics (GIW2009), 2009 年 12 月 15 日, 神奈川.

[3] Unyanee Poolsap, Yuki Kato and Tatsuya Akutsu, **Dynamic programming algorithms for RNA structure prediction with binding sites**, The 20th International Conference on Genome Informatics (GIW2009), 2009 年 12 月 14 日, 神奈川.

[4] Yuki Kato, **Prediction of interacting RNA secondary structures including binding sites**, International Workshop on Computational methods for RNA analysis, 2009 年 8 月 5 日, スペイン.

[5] Yuki Kato, **RNA pseudoknot prediction based on integer programming**, International Workshop on Computational methods for RNA analysis, 2009 年 7 月 29 日, スペイン.

[6] Unyanee Poolsap, Yuki Kato and Tatsuya Akutsu, **Prediction of RNA secondary structures with binding sites using dynamic programming algorithm**, 情報処理学会バイオ情報学研究会, 2009 年 5 月 26 日, 沖縄.

[7] 田中 翔, 加藤 有己, 関 浩之, **ペア確率多重文脈自由文法によるシュードノットつき RNA２次構造予測**, 情報処理学会バイオ情報学研究会, 2009 年 3 月 6 日, 東京.

[8] Unyanee Poolsap, Yuki Kato and Tatsuya Akutsu, **Prediction of RNA secondary structure with pseudoknots using integer programming**, The Seventh Asia Pacific Bioinformatics Conference (APBC2009), 2009 年 1 月 16 日, 中国.

[9] Yuki Kato, Tatsuya Akutsu and Hiroyuki Seki, **Dynamic programming versus grammatical approach for protein beta-sheet prediction**, The 2008 Annual Conference of the Japanese Society for Bioinformatics (JSBi2008), 2008 年 12 月 15 日, 大阪.

[10] Unyanee Poolsap, Yuki Kato and Tatsuya Akutsu, **An integer programming-based method of predicting RNA secondary structure with pseudoknots and its relation to simple linear tree adjoining grammar**, The 1st Thailand-Japan International Academic Conference 2008 (TJIA2008), 2008 年 11 月 21 日, 東京.

[11] Yuki Kato, Tatsuya Akutsu and Hiroyuki Seki, **Prediction of protein beta-sheets: dynamic programming versus grammatical approach**, Third IAPR Conference on Pattern Recognition in Bioinformatics (PRIB2008), 2008年10月15

日, オーストラリア.

[12] Yuki Kato, Tatsuya Akutsu and Hiroyuki Seki, **Prediction of protein beta-sheets: dynamic programming versus grammatical approach**, 情報処理学会バイオ情報学研究会, 2008 年 6 月 27 日, 沖縄.

[13] Unyanee Poolsap, Yuki Kato and Tatsuya Akutsu, **Prediction of RNA secondary structure with pseudoknots using integer programming**, The 8th International Workshop on Bioinformatics and Systems Biology (IBSB2008), 2008 年 6 月 9 日, ドイツ.

６．研究組織

(1)研究代表者

加藤　有己（KATO　YUKI）

京都大学・化学研究所・研究員（科学研究）

研究者番号：10511280

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：