論 文 内 容 の 要 旨

博士論文題目

Formal Grammars for Describing RNA Pseudoknotted Structure and Their Application to Structure Analysis

(RNA シュードノット構造記述向き形式文法とその構造解析への応用)

氏 名　加藤 有己

(論文内容の要旨)

近年，核酸やタンパク質などの生物学的に重要な分子の構造解析が注目を浴びている．これらの構造は階層的に１次構造，２次構造，３次構造と分類される．本論文では，多くが Watson-Crick 相補塩基対間の相互作用で決定される，RNA（リボ核酸）の２次構造に焦点を当てている．典型的な RNA では塩基対が互いに入れ子になって現れるため，RNA２次構造を文脈自由文法（CFG）でモデル化し，２次構造予測を文法の構文解析に置き換える試みが行われてきた．一方で，いくつかの塩基対が交差して現れる，シュードノットと呼ばれる部分構造が存在し，CFG では表現できないことが知られている．そのため，線形接木文法（SLTAG），拡張 SLTAG（ESLTAG），RNA シュードノット文法（RPG）などの，シュードノットを含む RNA の２次構造を記述する形式文法がいくつか提案された．しかしながら，現在までこれらの文法の生成能力間の関係は明らかではなかった．

本論文の最初の目的は，上記文法の生成能力を比較することである．そのために，上記文法クラスを CFG の自然な拡張である多重文脈自由文法（MCFG）の部分クラスとして同定している．具体的には以下が示されている：

(1) RPG が生成する言語のクラスは，次元が２以下，ランクが２以下の MCFG が生成する言語のクラスに一致する，

(2) ESLTAG が生成する言語のクラス（ESLTAL）は，次元が２以下，ランクが２以下，自由度が５以下の MCFG が生成する言語のクラスに一致する，

(3) ESLTAL は，SLTAG が生成する言語のクラス（SLTAL）と CFG が生成す

る言語のクラスとの和集合を真に含む,

(4) SLTAL は full trio である,

(5) ESLTAL は代入のもとで閉じた full AFL である.

これらの結果を考慮し，現在知られている形式文法の中でシュードノットを表現できる生成能力最小の文法は ESLTAG であることが述べられている.

本論文の後半では，ESLTAG に対応する MCFG の部分クラスを用いて，シュードノットを含む RNA 2 次構造を解析する手法について論じられている．構造予測を文法の構文解析とみなすとき，入力 RNA 配列に対して一般に多くの導出木が存在するという問題に直面する．そこで実用性を考慮し，文法を確率モデルに拡張し，確率最大の導出木を求めるアプローチが取られている．本論文では，上記 MCFG の部分クラスを確率 MCFG（SMCFG）と呼ばれる確率モデルに拡張している．次に，多項式時間で確率最大の導出木を求める構文解析アルゴリズム及び EM アルゴリズムに基づく確率パラメータ推定アルゴリズムが与えられている．また，SMCFG の構文解析アルゴリズムを用いて，ウイルス性 RNA に対して 2 次構造予測を行った結果が示されている．さらに，上記構文解析アルゴリズムに基づく走査アルゴリズムを用いて，シュードノットを持つ RNA 遺伝子が含まれているいくつかのゲノム配列に対して RNA 遺伝子発見が行われている．これらの実験結果は 100%に近い予測精度を示している.

氏　名	加藤　有己

（論文審査結果の要旨）

近年，RNAやタンパク質などの生物系列（または生物配列．以降，生物系列と書く）の高次構造予測に，形式文法の構文解析法が応用されている．これまで主として文脈自由文法(CFG)の構文解析法が用いられてきたが，生物系列の高次構造には，RNA２次構造におけるシュードノットのように，CFG では原理的に表現することのできない部分構造が存在する．そこで最近，CFG より生成能力の大きい文法を用いたRNA２次構造予測法がいくつか提案されているが，それらの文法間の関係は不明であった．一方，従来より，計算言語学では，CFG が自然言語の構文記述には能力が不足していることが指摘されており，CFG より生成能力が大きく，かつ，多項式時間で構文解析可能な文法としてTAG(接木文法)などが提案されてきた．

このような背景の下，本論文では，シュードノットを含むRNA２次構造記述向き形式文法として，MCFG（多重文脈自由文法)に着目し，以下の（１），（２）の成果がまとめられている．

（１） MCFG はCFG の自然な拡張としてKasami らによって導入された形式文法であり，多項式時間認識可能性など，CFG の良い性質を受け継いでいる．本論文ではまず，既存のRNA２次構造記述向き形式文法をMCFG の部分クラスとして同定し，合わせてそれらの文法クラスの性質も精査している．具体的に，Rivas と Eddy らの RNA シュードノット文法と(2,2)-MCFG の生成能力が等しいこと，Uemura らの ESLTAG と自由度５以下の(2,2)-MCFG の生成能力が等しいことなどを示しており，さらに，これらのクラスの言語演算に対する閉包性も明らかにしている．

（２） 上で言及した自由度 5 以下の(2,2)-MCFG を確率付モデルに拡張し（SMCFG とよぶ)，SMCFG に対する確率最大の導出木を求めるCYK 型のアルゴリズム，および，EM 法に基づく確率パラメータ推定アルゴリズムを提案している．アルゴリズムの時間計算量は入力系列長 n に対して $O(n^5)$ であるが，動的計画法で用いるデータ構造をハッシュ表によって効率的に実装することにより，実用的な予測プログラムの開発に成功している．具体的に，RNA２次構造データベー

スRfamに登録されている３つのRNAファミリーについて２次構造予測実験を行った結果，適合率，再現率ともに平均99%以上の高い精度を得ている．これは先行研究であるMatsuiらのPSTAGを用いた木アラインメントに基づく予測法と同等以上の結果であり，その有用性が実証されている．

本研究は形式言語理論に基づく理論的考察を行う一方で，構造予測アルゴリズムを実装し，実際の生物系列に対して提案手法が高い精度の構造予測を行えることを実証している．以上の通り，本論文で提案する手法と得られた結果は，形式言語理論，とりわけ文脈自由文法より生成能力の大きい形式文法の生物系列解析への応用に関する重要な知見を与えており，博士（工学）の学位論文として価値あるものと認める．