東京大学大学院新領域創成科学研究科
情報生命科学専攻

2013年 3月

修士論文

ハイブリッドモデルによる
半教師付き学習を用いたRNA二次構造予測

指導教員 浅井 潔 教授
47-116914

米本 悠

論文要旨

RNA は細胞内のさまざまな過程に関与しており，その機能と立体構造には深い関わりがあることが知られているそのため，RNA の機能を解明する上で，二次構造の解明は非常に重要なテーマである．RNA の立体構造を実験的に決定する方法としては，X 線による結晶解析や，NMR による解析などの方法があげられるが，時間やコストの関係から，コンピュータを用いて二次構造を予測する方法が広く用いられている．

コンピュータを用いた RNA の二次構造予測では，エネルギーモデルを用いて予測する方法が長く用いられてきた．このモデルでは，実験的に決定されたモチーフのエネルギーパラメータを用いて最適な構造を予測する．この手法の後登場したのが機械学習的なアプローチを用いた予測手法で，これは二次構造既知のデータを用いて推定された確率パラメータを用いて最適な構造を予測する．我々は主に後者の手法の改良を目的として予測手法の開発を行っている．

構造既知データのみを利用するパラメータ推定法は,「教師付き学習」と呼ばれ，現存する確率モデルを扱う RNA 二次構造予測のソフトウェアはほとんどこの方法を用いてパラメータを推定している．この教師付き学習の問題点は，パラメータ推定に用いる構造既知の RNA 配列データが十分に得られない場合，もしくは，配列データの多様性が低いような場合には二次構造の予測精度が向上しない点である．先にも述べたとおり，RNA 立体構造を実験的に決定するには時間とコストがかかるため，実験による網羅的な二次構造同定は現段階では現実的ではない．そこで我々は構造既知のデータに加え，構造未知のデータも用いる「半教師つき学習」により二次構造予測に用いるパラメータを推定する手法を提案した．この半教師付き学習は，自然言語処理などの他の問題で利用される手法で，一般に，限られた教師付きデータしか得られないような場合には予測性能が向上する．

RNA 二次構造の確率モデルは，生成モデルと識別モデルと呼ばれるモデルに大別することができる．入力としての配列を x，出力としての二次構造を y としたとき，同時確率 $p(x, y)$ をモデル化するものを生成モデル，事後確率 $p(y|x)$ をモデル化するものを識別モデルと呼ぶ．識別モデルは事後確率を直接モデル化するため，一般的に生成モデルに比べ予測性能は高いと言われている．

生成モデルのパラメータ推定では，EM アルゴリズムを使うことで構造未知のデータを容易に扱うことができる．例えば，Baum-Welch アルゴリズムは，隠れマルコフモデルのパラメータ推定によく用いられる EM アルゴリズムであり，RNA の確率モデルとして用いられる SCFG のパラメータを推定する際にも同様に用いることができる．一方識別モデルのパラメータ推定では，構造未知データを扱うことは容易ではない．例えば，条件付き確率モデルのパラメータ推定に構造未知データを直接用いて周辺化確率を最大化するようにパラメータを推定しようとしても，そのデータによる効果は目的関数から消えてしまうためである．

このような背景から，Suzuki らは，識別モデルと生成モデルを組み合わせ，パラメータ推定の際構造未知データと構造既知データを効率的にとり扱うことができる確率モデルを提唱した [1]. Suzuki らは，生成モデルに隠れマルコフモデル（HMM）識別モデルに隠れマルコフモデルから導出した CLLM を用いて，この確率モデルを固有表現抽出 (Named-entityrecognition) とチャンキング (Chunking) と呼ばれる構文解析のタスクに適用し， 高い性能を持つことを示した．

本研究では，このハイブリッドモデルを RNA の二次構造予測に応用することにより，教師なしデータが利用できないという問題を解決することを試みた．前述の通り，ハイブリッドモデルは生成モデルと識別モデルを組み合わせたモデルであり，我々は，生成モデルに HMM の拡張である SCFG を，識別モデルには SCFG から導出した CLLM を用いた．このハイブリッドモデルは，教師付き学習の部分に用いた配列データと構造的に似たデータセットに対しての予測性能は既存の確率モデルと同程度であるが，教師なし学習の部分に用いた配列データと構造的に似たデータセットに対しての予測性能は，既存の確率モデルよりも高いことが示された．

キーワード

RNA 二次構造予測 半教師付き学習

参考文献

[1] J. Suzuki, A. Fujino, and H. Isozaki. Semi-supervised structured output learning based on a hybrid generative and discriminative approach. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 791–800, 2007.