Development of RNA informatics for RNA sequence and structure analysis

February 2022

Manato Akiyama

主　論　文　要　旨

No.1

報告番号	㊙甲　乙　第　　号	氏　名	秋山真那斗

主 論 文 題 名 :

Development of RNA informatics for RNA sequence and structure analysis

（RNA 配列と構造解析のための RNA インフォマティクスの構築）

（内容の要旨）

タンパク質に翻訳されない非コード RNA（ncRNA）は、これまで機能を持たないジャンク領域と見なされてきたが、近年、発生や細胞分化の過程から病気の原因に至るまでさまざまな機能を持つことが明らかになっている。 RNA インフォマティクスによる解析をおいて、ncRNA の機能を知るために不可欠なステップとして ncRNA の構造情報の抽出がある。本論文では RNA の構造解析及び配列解析という異なる二つのアプローチにより ncRNA の効果的な表現方法とそれを用いた新しい解析手法を構築した。

本論文の第 1 章では、ncRNA とその情報解析分野である RNA インフォマティクスの重要性について述べるとともに、本研究で扱う二つの課題について概説した。

第 2 章では、高精度な RNA 二次構造予測アルゴリズムの開発について報告した。ncRNA の機能は二次構造と密接に関連しているため、二次構造が分かればその生物学的機能を推測することができる。 RNA 二次構造を予測するための一般的なアプローチは、熱力学的に最も安定した最小自由エネルギー（MFE）構造を見つけるための熱力学的モデルである。さらなる予測精度の改善のために、より詳細な特徴量のモデリングが可能である機械学習に基づくアプローチが開発されてきた。機械学習ベースの詳細な特徴量を持つモデルは、予測精度において非常に高いパフォーマンスを達成したが、学習データに過剰適合するリスクの可能性が報告されている。本論文では、熱力学的アプローチと機械学習ベースのアプローチを統合する RNA 二次構造予測のための新しいアルゴリズムを提案した。ベンチマークテストでは、提案アルゴリズムは大きな過剰適合を起こさず、既存の方法と比較して最高の予測精度を示した。

第 3 章では、RNA 配列及び塩基のベクトル化技術の開発を行った。DNA 配列やアミノ酸配列をベクトル化する埋め込み(embedding)という技術が DNA 配列モチーフの検出やタンパク質機能予測などの品質を向上させる事がわかっている。一方で、RNA 配列の効果的な埋め込み技術はこれまで開発されていない。本論文では、RNA 配列を効果的に埋め込むための事前トレーニングアルゴリズムを採用して、構造情報や配列の文脈情報を豊富に含む RNA 配列の埋め込みベクトルを取得する手法を開発した。事前トレーニングによって得られた埋め込みベクトルの品質を検証するために 2 つの基本的な RNA インフォマティクスの課題（構造アラインメントと遺伝子のクラスタリング）によるテストを実施し、既存の最先端の方法よりも優れた精度を達成した。

第 4 章では、本研究を総括するとともに、開発した RNA 解析手法について今後の応用可能性を議論した.

以上、本論文では RNA 二次構造予測及び RNA 塩基配列のベクトル化という二つのアプローチを用いて ncRNA の効果的な解析手法を開発することに成功した。各アプローチはあらゆる RNA インフォマティクスにおける課題に応用可能である。

Thesis Abstract

No.

Registration Number	☑ "KOU" □ "OTSU" No. *Office use only	Name	Manato Akiyama

Thesis Title

Development of RNA informatics for RNA sequence and structure analysis

Non-coding RNAs (ncRNAs) that are not translated into proteins were formerly considered as junk regions. However, various functions have been revealed in recent years ranging from the process of development and cell differentiation to the cause of disease. Extraction of ncRNA structural information is an indispensable step to understand the function of ncRNA in the analysis by RNA informatics. In this paper, I have developed an effective expression method for ncRNA using two different approaches: RNA structural analysis and sequence analysis. Furthermore, I constructed a new analysis method using these expression methods.

First, I developed a highly accurate RNA secondary structure prediction algorithm. Since the functions of ncRNAs are believed to be closely related to the structures of ncRNAs, it is possible to infer their biological functions from their structures. A popular approach for predicting RNA secondary structure is the thermodynamic nearest-neighbor model that finds a thermodynamically most stable secondary structure with minimum free energy (MFE). For further improvement, an alternative approach based on machine learning technology has been developed that can employ a fine-grained model that includes much richer feature representations. Although a machine learning-based fine-grained model achieved extremely high performance in prediction accuracy, a possibility of the risk of overfitting for such a model has been reported. In this paper, I propose a novel algorithm for RNA secondary structure prediction that integrates the thermodynamic approach and the machine learning-based weighted approach. Our benchmark shows that our algorithm achieves the best prediction accuracy compared with existing methods, and heavy overfitting cannot be observed.

Next, I developed a technique for vectorizing RNA sequences and bases. "Embedding technology" that vectorizes DNA sequences and amino acid sequences is known to be useful for detecting DNA sequence motifs and predicting protein functions. But, RNA sequence embedding technology has not been developed so far. In the paper, I created a pre-training algorithm with the aim of acquiring an embedded vector of an RNA sequence that contains abundant structural information and sequence context information. I performed two basic RNA informatics tasks (structural alignment and gene clustering) to verify the quality of embedding and achieved greater accuracy than existing state-of-the-art methods.

In this paper, I succeeded in obtaining an effective expression of ncRNA using two approaches: RNA secondary structure prediction and RNA base sequence vectorization. Each approach can be applied to analysis in all RNA informatics including RNA-protein interaction and RNA-RNA interaction, and can be expected to have a large spillover effect.