Computational methods for accurate and efficient identification of noncoding RNAs

September 2012

Yutaka Saito

主　論　文　要　旨

報告番号	㊙甲 乙 第　　号	氏 名	齋藤 裕

主 論 文 題 目 :

Computational methods for accurate and efficient identification of noncoding RNAs

（非コード RNA を高精度かつ高速に同定するための計算手法）

（内容の要旨）

高等生物のゲノム配列の大部分はタンパク質をコードしていない。非コード RNA は、これらのゲノム領域から発現する遺伝子であり、従来のタンパク質を中心とした生物学に革命を起こそうとしている。非コード RNA のアノテーションや機能ファミリの発見は、タンパク質と比べて大幅に遅れている。したがって、非コード RNA を高精度かつ高速に同定する計算手法の開発は、生命情報科学における重要な課題となっている。

本研究は、非コード RNA の同定における 2 つの問題に対して、新しい計算手法を提案した。第 1 の手法は、入力の RNA を既知のファミリへ分類するアノテーションの問題を扱う。第 2 の手法は、どの既知ファミリにも属していない RNA の集合から新規ファミリの候補を発見する問題を扱う。これらの問題では、2 つの RNA の類似度を評価する過程が手法の性能を決定付ける。第 1 の手法は、入力 RNA と既知ファミリメンバとの類似度に基づいて分類を行い、第 2 の手法は、RNA の集合から内部の類似度が高い部分集合を発見する。そのため、本研究の独自性は、RNA 間の類似度指標の開発へと集約される。

本論文の第 1 章では、非コード RNA 同定の重要性について述べた。

第 2 章では、既知ファミリへの分類問題について述べた。既存研究においては、入力 RNA として単一の配列ではなく近縁種とのアラインメントデータを利用する手法が主流である。そこで、本研究では、RNA のアラインメントデータ間の類似度指標を開発した。本手法は、アラインメントデータから抽出される種間のプロファイル情報を利用することにより、既存手法よりも高い分類精度を達成した。さらに、本研究では、アラインメントデータに含まれるエラーが識別精度に与える影響について、詳細な検証実験を行った。これにより、本手法のエラーに対する高い頑健性が示された。また、既存手法がエラーに対して極めて脆弱であることが初めて明らかになった。本手法は miRNA、snoRNA などの様々なファミリのアノテーションに適用することができる。

第 3 章では、新規ファミリの発見問題について述べた。既存研究は、この問題を RNA 配列のクラスタリングとみなし、クラスタリングにおいて必要となる配列間の類似度指標を提案している。しかし、既存の類似度指標は計算量が大きく、クラスタリングの結果も不正確である。そこで、本研究では、高精度かつ高速な RNA 配列間の類似度指標を開発した。本手法は、非コード RNA の機能と密接に関連している 2 次構造に着目して、その類似度を近似的なアルゴリズムによって計算する。これにより、既存手法に対して約 1000 倍の高速化を実現しながら、非常に高精度なクラスタリングを達成した。特に、検出対象のファミリが大きな配列多様性を有している場合、本手法は既存手法よりも解釈しやすい明確なファミリ候補の検出に成功した。

第 4 章では、本研究を総括するとともに、提案した類似度指標について他の類似度検索問題への応用可能性を議論した。

SUMMARY OF Ph.D. DISSERTATION

School	Student Identification Number	SURNAME, First name
School of Fundamental Science and Technology	81045196	SAITO, Yutaka

Title

Computational methods for accurate and efficient identification of noncoding RNAs

Abstract

The vast majority of genomic sequences in higher organisms do not code for proteins. Noncoding RNAs (ncRNAs) are genes expressed from these genomic regions, and revolutionizing the traditional "protein-centered" view of functional genomics. The annotation and the establishment of ncRNA families are far from complete when compared to those of proteins. Therefore, accurate and efficient identification of ncRNAs is a major goal of computational biology.

In this dissertation, we present computational methods that address two different problems for the identification of ncRNAs. First, we proposed a method for classifying an input RNA into a known family, i.e., annotation. Second, we proposed a method for finding candidates of novel families from a set of unannotated RNAs. In these problems, the evaluation of the similarity between two RNAs is essential for the performance of the proposed methods. Thus, the originality of our studies is the design of similarity measures for RNAs.

In Chapter 1, the importance of the identification of ncRNAs was introduced.

In Chapter 2, a method for classifying into a known family was described. In previous studies, one common approach was to take alignment data rather than single sequences as input RNAs. Hence, we developed a similarity measure between two alignment data. By utilizing the profile information contained in alignment data, our method can achieve better accuracy than existing methods. Our method can be applicable to a wide range of families including micro RNAs (miRNAs) and small nucleolar RNAs (snoRNAs).

In Chapter 3, a method for finding novel families was described. We considered this problem as clustering of RNA sequences, and developed a similarity measure for clustering procedure. Our method incorporates secondary structure information of RNAs, and evaluates the similarity using efficient algorithms. Our method is about 1000 times as fast as the previous state-of-the-art, while achieving better accuracy.

In Chapter 4, this study was summarized and future works were discussed.