Local Structural Alignment of RNA with Affine Gap Model

Thomas K.F. Wong, Brenda W.Y. Cheung, T.W. Lam, and S.M. Yiu

Department of Computer Science, The University of Hong Kong, Hong Kong
{kfwong,wycheung,twlam,smyiu}@cs.hku.hk

Abstract. Predicting new non-coding RNAs (ncRNAs) of a family can be done by aligning the potential candidate with a member of the family with known sequence and secondary structure. Existing tools either only consider the sequence similarity or cannot handle local alignment with gaps. In this paper, we consider the problem of finding the optimal local structural alignment between a query RNA sequence (with known secondary structure) and a target sequence (with unknown secondary structure) with the affine gap penalty model. We provide the algorithm to solve the problem. Based on a preliminary experiment, we show that there are ncRNA families in which considering local structural alignment with gap penalty model can identify real hits more effectively than using global alignment or local alignment without gap penalty model.

Keywords: Local structural alignment, Affine gap, non-coding RNA.

1 Introduction

A non-coding RNA (ncRNA) is a RNA molecule that does not translate into proteins. It has been shown to be involved in many biological processes [1,2,3,4]. The number of ncRNAs within the human genome was underestimated before, but recently some databases reveal over 212,000 ncRNAs [5] and more than 1,300 ncRNA families [6]. Large discoveries of ncRNAs and their families show the possibilities that ncRNAs may be as diverse as protein molecules [7]. Identifying ncRNAs is an important problem in biological study. However, it is time consuming and there is no effective method to identify ncRNAs in a laboratory, predicting ncRNAs based on known ncRNAs using comparative computational approach is one of the promising directions to identify potential candidates for further verification.

Most of the computational approaches are based on the observation that if two different ncRNA molecules are in the same family (with similar biological functions), they usually exhibit similar sequences as well as secondary structures. One common approach [8,9,10] is as follows. We pick an ncRNA member of a family with known sequence and secondary structure (referred as the query), scan along a genomic sequence and for each possible region (referred as the target), perform an alignment between the query and the target to obtain a similarity measure to decide if the region is a potential ncRNA candidate for that family.
The similarity measure may only base on the sequence or both the sequence and secondary structure (the latter case is referred as \textit{structural alignment}). Along this direction, there are some approaches \cite{11,12,13,14} that make use of secondary structure prediction tools to predict the secondary structure to be formed by the target assuming that it is an ncRNA before performing the alignment. The accuracy may, however, depend on the accuracy of the secondary structure prediction tools.

Instead of using one member of a family, some other approaches \cite{15} use a set of ncRNAs from the same family to train a model (e.g. covariance model). Then, using this model to scan a genomic sequence to identify potential regions that are ncRNA candidates of that family. What information (sequence similarity and/or secondary structure) to be captured from the known ncRNAs depends on how we define the model. However, in some cases, we may not have enough known members in a family to train a model. In this paper, we focus on the problem that uses one known member as the query and align it with a target sequence. We remark that there are also other computational methods that identify ncRNAs without using known members in a family. For example, some try to identify ncRNAs by considering the stability of secondary structures formed by the substrings of a given genome \cite{16}. This method may not be very effective because a random sequence with high GC composition also allows an energetically favorable secondary structure \cite{17}. So, the comparative approach we described in the above is still one of the most popular approaches.

The core idea behind all comparative approaches is to compute the similarity between the query (known member(s)) and the target (each possible region in the genomic sequence to be investigated). Some only consider sequence similarity which may not work well for families in which members do not have high sequence similarity (e.g. members of RF00017 in Rfam 9.1 \cite{6} only have 39% sequence similarity). For example, Gotohscan\cite{8} considers semi-global alignment with affine gap penalty according to the sequence similarity only. For those also consider the similarity of secondary structure, they usually require the whole sequence of the query to be aligned with the whole sequence of the target (referred as \textit{global alignment} in the community) \cite{10}. However, similar to the protein sequence, the ncRNAs in the same family may not have similar sequence or structure for the whole sequence but only for the substrings of them (those supposed to be the functional parts), especially when they belong to species with long evolutionary distance apart. Fig. 1 shows one of these examples. It shows the multiple sequence alignment between some members of the family RF01051 in Rfam 9.1 database. The two circled members (i.e. AAUO01000012 and AAXYO1000014) are not quite similar if we consider the global alignment. Also, for the subregions that they look similar (i.e. the circled region), there exist large insertion/deletion (gaps). There are also evidences that gaps may be common in ncRNA homologs \cite{18}. Considering local structural alignment with gap model seems to be more appropriate for predicting new members for some ncRNA families. \cite{9} consider some restricted cases of local alignment according to the query structure. Another work that also consider local alignment is \cite{11}, but they cannot handle gaps.