論文内容の要旨

論文題目 Research on Fast and Accurate Comparison of RNA Sequences by Stem Candidate Representation (RNA のステム候補表現による高速かつ正確な比較に関する研究)

氏名　田部井　靖生

Non-coding RNAs (ncRNAs) show a unique evolutionary process where the substitutions of distant bases are correlated in order to conserve the secondary structure of the ncRNA molecule. Although their functions often depend on their 3D-structures rather than their primary sequence, the existence of conserved secondary structures among phylogenetic relatives highlights their functional importance. Therefore, alignment algorithms for RNA sequences should take into account both the primary sequence and the secondary structures.

The Sankoff algorithm simultaneously provides solutions to the structure prediction and alignment problem. However, the original version of the Sankoff algorithm is impractical, because of its prohibitive computational cost. Therefore, an efficient structural alignment algorithm for RNA sequences is required. In this thesis, we propose fast and accurate comparison methods of RNA sequences by stem candidate representation. Stem candidates are a set of potential stems as continuous base-pairs in a secondary structure of an RNA sequence. The computational biology of RNA secondary structure has a long history. First method to predict the secondary structure of an RNA sequence was developed about 40 yeas ago. The methods for analyzing ncRNAs have extremely advanced, recently. Especially, various algorithm designing techniques, machine learning techniques and data mining techniques have been applied to RNA secondary structure prediction and structural alignment of ncRNAs. Combining these techniques, we also show that representing RNA sequences as stem candidates is an effective strategy for designing comparison methods for ncRNAs.

First, we deal with the pairwise alignment problem of RNA sequences. The functions of

ncRNAs are strongly related to their secondary structures, but it is known that a secondary structure prediction of a single sequence is not reliable. Therefore, we have to collect similar RNA sequences with a common secondary structure for the analyses of a new non-coding RNA without knowing the exact secondary structure itself. Because we often want to compare a large number of cDNA sequences or to search similar RNAs in the whole genome sequences, much faster algorithms are required. We propose an efficient pairwise alignment algorithm based on fixed-length stemfragments, implemented in SCARNA (Stem Candidate Aligner for RNAs).

We next deal with the global multiple alignment problem of ncRNAs. Multiple alignments of ncRNAs are useful in order to accurately predict secondary structures of ncRNAs, identify novel ncRNAs from genomic sequences and analyze phylogeny of ncRNAs. We propose a novel global multiple alignment method, implemented in MXSCARNA (Multiplex SCARNA),which is an extension of our pairwise alignment method.

We then deal with the local multiple alignment problem of ncRNAs. Recently, there has been intense focus on multiple alignment investigations for the detection of ncRNAs; however, most of the proposed methods are designed for global multiple alignments. For this reason, these methods are not appropriate to identify locally conserved ncRNAs among genomic sequences. A more efficient local multiple alignment method for the detection of ncRNAs is required. We propose a local multiple alignment method, implemented in SCARNA_LM (SCARNA Local Multiple), as an application of our proposed pairwise alignment algorithm and global multiple alignment method.

Finally, we discuss efficient methods for finding frequent secondary structures from a set of unaligned RNA sequences by means of reverse search techniques. Reverse search is a general scheme for designing efficient algorithms for hard enumeration problems. A state-of-the-art method propose an enumeration method of RNA secondary structures by the gSpan algorithm, which is a graph mining algorithm. However, the computational time of the gSpan algorithm is exponential for the size of graphs. We propose a polynomial time algorithm for this problem.

Through all these topics, we present efficient comparison methods of ncRNAs.