Chapter 7
Modeling Conserved Structure Patterns for Functional Noncoding RNA

RNA regulation has been increasingly recognized as a potential and perhaps overlooked genetics of higher organisms. Noncoding RNAs (ncRNA) may play various catalytic and regulatory roles in the genetic operating system, such as RNAi (RNA interference) pathways for inhibiting gene expression and siRNAs silencing pathway leading to the degradation of the target mRNA. Thus, it is critical to develop methods and tools to investigate the featured patterns of RNA structures. Recent studies using comparative genomics and molecular genetics show evidence of the presence of varied ncRNAs. Unlike protein coding genes, there is a lack of comparable information or outstanding signal for ncRNAs. Traditional computational linguistics show limitations in modeling complicated secondary structures and prevent us from identifying structure-function relationships of ncRNAs. This chapter presents a novel approach, based on a set of distance constraints, to model the predicted RNA secondary structures. Further, a filtering schema is presented to identify matched models for the queried secondary structures.

7.1 Introduction

The genetic information was previously viewed as flowing from DNA to proteins via mRNA. It has been assumed that the genetic output is almost completely transacted by proteins in the past decade. This conclusion is true for prokaryotes, in which proteins comprise not only the primary functional and structural components of cells but also the main agents to regulate the cellular dynamics, in combination with cis-regulatory elements and environmental signals [224]. However, the proportion of protein-coding sequences that occupy only a small minority of genome of multicellular organisms is insufficient to perform complex cellular functions. It is thus critical to investigate if these functions of complexity are carried out by unknown noncoding RNAs.

As described in [159], structural genes encode proteins, and regulatory genes produce ncRNA. Noncoding RNAs form transcripts that appear to be developmentally regulated instead of encoding protein. A variety of experimental techniques have
been applied to identify the vertebrate transcriptomes, such as tiling arrays \[24, 58, 173\], cDNA cloning \[247\] and unbiased mapping of transcription factor binding sites \[40\]. The results unveil the involvement of ncRNA in the evolution and developmental programming of complex organisms. This demonstrates that these ncRNAs may fulfill some unexpected functions and constitute a critical hidden layer of gene regulation in mammalian biology. Most of the previous studies suggest that only a small fraction (1.2%) of the genome is transcribed and a large fraction (98%) of the transcriptome comprises noncoding RNAs \[222, 223\]. This demands us to understand the mechanisms of transition of genetic information in higher organisms.

A number of computational approaches have been developed to identify noncoding genes \[12, 199, 202\]. Some of these attempts look for signals that might suggest a functional RNAs in the molecule, such as the promising approach of using secondary structure as a signal \[41, 71, 140, 201\]. Other approaches aim to look for the transcription start and similar signals. However, they have had limited success because ncRNAs do not contain common signals that could be identified at the sequence level as protein coding genes.

Recent studies discovered that a large class of ncRNAs, such as rRNAs, tRNAs, small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs), share characteristic structures that are functional and hence are well conserved through evolution \[295\]. The stabilizing selection of secondary structures results in corresponding substitution patterns in the underlying sequences. Consistent and compensatory mutations substitute one type of base pairs by another one in the helices of the molecule. Moreover, the discrepant distribution in the base composition and length between loops \[212\] and stems \[1, 310\] have been reported. Thus, studies of structural information on RNA can be an alternative to understand structure-function relationships in ncRNAs. A number of comparative methods based on conserved RNA structures over evolution have been developed. Moreover, some methods are to address the relatively easier problem of identifying subsequences that are similar in structure and sequence to query, rather than identifying novel ncRNA families. They have been used to find homologs of a specific RNA \[345\] and functional noncoding RNAs in human genome \[295\].

One of the biggest challenges in this field is to explore and discover the functional significance of this abundant non-coding transcription due to computational complexity and insufficient tools for modeling ncRNAs. Unfortunately, most ncRNA databases provide only sequence data \[249\] or separate sequence data from structure data \[295\]. The low quality data prevent us from applying advanced data mining techniques \[127\] to extract interesting knowledge of ncRNAs.

Linguistic methods have been successfully applied to the representation of biological sequences \[279\] since 1980. They use Chomsky-style grammars \[60, 260\] such as context free grammar and context sensitive grammar, to capture not only informational but also structural aspects of molecules. However, it has been criticized in representing complex cellular processes with overlapping, frameshifted coding regions such as the intersection of two stem-loop structures \[254\]. Although it has