Chapter 7. Conclusions

Although several comparative ncRNA-finding algorithms had been claimed to be effective in ncRNA finding, their abilities to find ncRNAs from genome-wide alignments had not yet at the time of preparation of this thesis been appropriately assessed. In the first part of this thesis, I assessed the two factors, the abundance of covariations between synteny-conserved ncRNAs, and the synteny-conservation ratios of ncRNAs, which may determine the performance of comparative algorithms in genome-wide ncRNA finding.

In chapter 2, I showed that only a few compensatory mutations could be found in the alignments of orthologous ncRNAs in vertebrate genomes. In general, orthologous ncRNAs in vertebrates are so conserved that their alignments cannot provide sufficiently strong signals to indicate the existence of structural motifs in ncRNAs. In addition, I showed that, when applied to real genome alignments, existing comparative algorithms suffered from a high false negative rate. Based on these results, I conclude that existing comparative algorithms are not ideal for finding ncRNAs in vertebrate genomes. This conclusion is consistent with the recent paper using comparative algorithms to attempt to find structural ncRNAs in the ENCODE regions of the human genome, where a false discovery rate as high as 50% ~ 70% was reported (Washietl et al. 2007).

In chapter 2, I also showed that the synteny-conservation ratios of mammalian ncRNA categories varies between 1% and 74%. In the second part of chapter 2, I examined the gene-order conservation of the tRNA-gene loci in the human and mouse genomes in detail to explore the evolutionary processes leading to this non-synteny. Interestingly, I found that there are repetitive multi-tRNA-gene blocks, suggesting that duplication may play a major role in the evolution of tRNA gene loci in mammalian genomes.

In chapter 3, I explored possible rules that can be used to distinguish functional ncRNAs from pseudogenes. There were two interesting findings in this work. Firstly, the low-scoring peak of the bi-modal distribution of the bit scores of Rfam-predicted tRNA genes were found to be likely to be nuclear mitochondrial tRNAs (numt-tRNAs), which appear to be pseudogenes. Secondly, I found that circumstantial evidence that clustering might be an important factor associated with the functions of tRNA-gene loci. Low-scoring tRNA genes are enriched with non-clustered tRNA genes in the human genome. Besides, clustered human tRNA genes can cover the required anticodons for translating proteins in eukaryotic cells.

To address the problem of genome-wide ncRNA finding, it is useful to consider complementary structure-independent approaches, in addition to structure-dependent algorithms. In chapter 4, the methods that were later used to model the transcription regulatory regions were introduced. Then, in chapter 5, the Eponine system was used as a quick approach to learn a new model for selectively predicting tRNAs, as well as novel ncRNA genes transcribed by RNA polymerase III (pol III genes), in the mammalian genomes. However, the results from modelling of the TSSs of mammalian pol III type II genes were not clear. Numerous TSSs predicted using the Eponine Anchored Sequence (EAS) pol III type II model overlapped with MIR repetitive elements. No evidence could be found to support the suggestion that these MIRs might generate functional transcripts.

The other strand of this project was the development of the Eponine RNA-motif extension. With the methods introduced in chapter 4, the capabilities of both the EAS and Eponine Windowed Sequence (EWS) models were extended to model consensus RNA motifs from sets of related but unaligned sequences. I demonstrated, in chapter 6, that EAS mixed models could find consensus primary-sequence and secondary-sequence motifs in a set of unaligned sequences when reference points, such as TSSs and TTSs, were available. I also demonstrated that the EWS mixed model could still find consensus RNA motifs even when no

reference points were assigned to training sequences, although with poorer specificity. Potential future work involves trying to build generalized ncRNA models using the EWS mixed model approach, which may prove useful for finding undiscovered ncRNAs in mammalian genomes.