Computational detection of non-coding RNAs in genomes

Yen-Hua Huang

This dissertation is submitted for the degree of Doctor of Philosophy

The Wellcome Trust Sanger Institute and Churchill College, Cambridge

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

The work in this thesis has not been submitted in whole, or in part, for a degree, diploma, or any other qualification at any other university.

Yen-Hua Huang

March 2008, Cambridge, UK

Abstract

Noncoding RNAs (ncRNAs) have become implicated in a variety of regulatory mechanisms as well as structural roles, suggesting that functional ncRNAs may be more prevalent in genomes than previously supposed. Nonetheless, *in silico* ncRNA finding is difficult, even though a mass of genome sequence is publicly available. Few computational approaches are really reliable for genome-wide ncRNA finding. This thesis is devoted to assessing available approaches and trying new solutions for finding ncRNAs in genomes.

In the first half of this thesis, reasons that may contribute to the slow progress of genome-wide ncRNA finding are explored. A comprehensive analysis on a genome-wide scale of the credibility of currently used signals for classifying ncRNAs is conducted. Two factors, conservation of ncRNAs in human-mouse syntenic regions and abundance of covariations between human-mouse synteny-conserved ncRNAs, are evaluated. The result reveals that current comparative-genomics-based methods may not be able to find ncRNAs effectively in mammalian genomes. In addition, possible genomic features that could distinguish real ncRNAs from pseudogenes are investigated. Two different criteria, distribution of bit scores and physical clustering in genomes, are applied to filter out tRNA pseudogenes and to enrich *bona-fide* tRNA genes. Physiological roles of the tRNA genes in human-mouse synteny-conserved clusters are discussed and the degradation patterns of tRNA pseudogenes are analyzed.

In the second half of this thesis, computational techniques are applied to model signals that may be potentially useful for genome-wide ncRNA finding. A sparse Bayesian learning algorithm, Eponine, is applied to model the transcription start sites of mammalian ncRNA genes that are transcribed by RNA polymerase III. In addition to modelling *cis*-regulatory elements for transcription, a new computational module, which extends the capability of

Eponine to learn motifs consisting of both primary sequences and RNA secondary structures, is created. The capability of this new module is demonstrated by applying it to analyze several known cases of ncRNA motifs. The strength and the weakness of applying this new computational approach for finding ncRNAs are discussed.

Acknowledgements

I thank everyone who has influenced me during this project. I am particularly grateful to my supervisor, Dr. Tim Hubbard, for giving me this opportunity to work on this project in the Wellcome Trust Sanger Institute. I greatly appreciate his support and advice, which have helped me extend many originally hopeless paths of this project.

Being a bioinformatics PhD student with a pure background of biomedical science, I encountered many difficulties in understanding the mathematics required for my project. I am so fortunate that I worked with Thomas Down. His guidance enabled me to apply machine-learning techniques to my project. I greatly benefited from the discussion with him and the development of the Eponine RNA-motif extension was based on the Eponine and related codes he has created.

I thank Aroul Selvam Ramadass for the stimulus he provided. I was therefore inspired to investigate human-mouse synteny-conserved tRNA gene clusters and to create the Eponine RNA-motif extension. Without the discussions with him, the related sections might not exist.

I thank Jenny Mattison, Thomas Down, Mutlu Dogruel, and Jing Su for reading and checking this manuscript.

I thank Andreas Prlic for his helps on my project and for making the work interesting.

I thank the Wellcome Trust Sanger Institute for the funding and Churchill College, Cambridge, for academic support.

Finally, I would like to express my best gratitude to my wife, Yu-Yen, who accompanies me with love through all the difficulties and joys these years.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Over the past decade, numerous novel non-coding RNAs (ncRNAs) have been discovered. As opposed to classic ncRNAs including transfer RNAs (tRNA), and ribosomal RNAs (rRNA), these novel ncRNAs are not directly involved in producing proteins. Instead, they are implicated in a wide variety of regulatory mechanisms, including transcriptional regulation, chromosome replication, RNA processing and modification, modulation of messenger RNA stability and translation, and even protein degradation and translocation (for review see Storz 2002).

Although a vast amount of genomic sequence is publicly available, it is unknown how many ncRNAs there are in different organisms. Much evidence suggests that there are still many unannotated ncRNA genes in mammalian genomes. For example, a survey on human chromosomes 21 and 22 suggests that much of the human transcriptome could be transcripts of ncRNA genes (Kampa et al. 2004). Based on functional annotation of experimentally defined transcription units, it was claimed that as much as one-third of the mammalian transcriptome might consist of ncRNA genes (Okazaki et al. 2002). In addition to ncRNA genes, there might be other functional RNA elements that are hitherto undiscovered. For example, some *cis*-regulatory RNA motifs are known to regulate prokaryotic and eukaryotic gene expression at the post-transcriptional level, however their abundance, distribution, and possible classifications are generally unknown (for review see Kozak 2005).

Systematic ncRNA finding in complex organisms such as vertebrates is difficult. Although experimental approaches can collect thousands of transcripts efficiently, ncRNAs, as well as mRNAs, with low expression levels or with temporal expression patterns may be absent from experimental preparations. At the same time, most gene finding algorithms have been designed to predict protein-coding genes, not ncRNAs. Algorithms for *ab initio* prediction of protein-coding genes take advantage of propensities in base composition of protein-coding

regions. These propensities, including usage of amino acids, usage of synonymous codons, and usage of hexamers (for review see Rogic et al. 2001), cannot be used to distinguish ncRNAs from random genomic sequences. Although signals that are not specific to protein-coding genes, such as patterns of splice sites and polyadenylation signals, have also been used by many *ab initio* gene finders, many of these signals do not exist in genomic loci of single-exon ncRNAs, non-polymerase-II transcribed ncRNAs, and non-polyadenylated ncRNAs. Recently attempts have been made to use the information from comparative genomics to boost the accuracy of *ab initio* gene finding in vertebrate genomes (for review see Brent 2005). However, the development of similarity-based gene finders has also focused on the prediction of protein-coding genes.

Compared to computational protein-coding gene finding, computational ncRNA finding has been a relatively neglected field until recently. Before discussing the reasons that may contribute to the slow progress of genome-wide ncRNA finding (see section 1.4.), some basic knowledge of the biological importance of ncRNAs is required and is therefore introduced in the next section.

1.1. What are ncRNAs

An RNA (ribonucleic acid) molecule is a chain of ribonucleosides that are covalently linked. The only compositional difference between RNA and DNA (deoxyribonucleic acid) molecules is the use of ribose sugar in RNA, instead of 2'-deoxyribose sugar in DNA (Figure 1-1), and for one of the four bases the use of uracil instead of thymine.

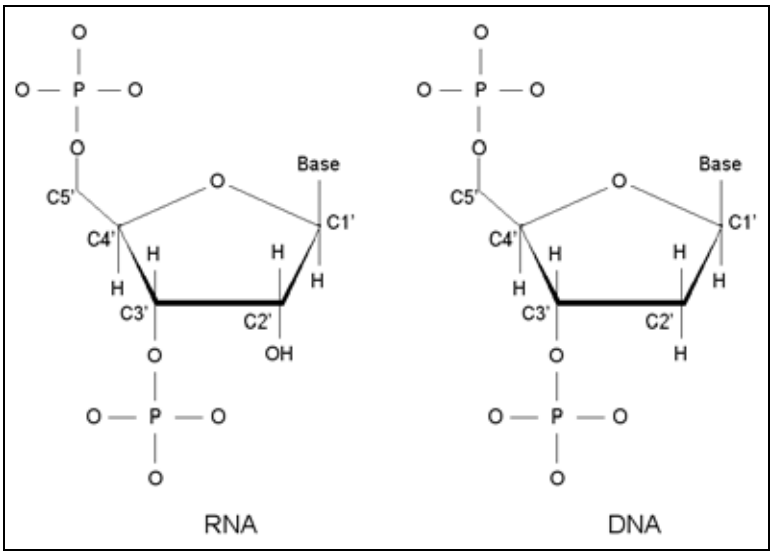


Figure 1-1. Organization of repeating units in RNA and DNA respectively.

As early as the 1960s, it was known that cells contained RNA genes that did not code for proteins. The transcripts of these RNA genes are called ncRNAs. Classic ncRNAs, such as tRNAs and rRNAs, were considered as adaptors and scaffolds respectively for protein production. For a long time, DNA attracted much more attention than RNA, because the latter did not seem to possess specifically useful features. For example, RNA molecules are more easily degraded in solution than DNA molecules. In addition, an initial impression was that RNA might not provide as much structural flexibility as DNA, since RNA helices appear to be more rigid than DNA helices due to the physical constraints rendered by the 2'-hydroxyl group of the ribose sugar (see Varani and Pardi 1994).

Nonetheless, RNA-unique features do enable ncRNAs to be functionally active molecules. Firstly, the 2'-hydroxyl group on the ribose sugar, which is the culprit for RNA's easy degradation in solution, blesses RNA with high chemical reactivity. As a result, RNAs can catalyse chemical reactions without the assistance of proteins. For example, group I and II introns can perform the functions of spliceosomes by RNA alone (Cech et al. 1981; Kruger et al. 1982). The ability of RNA to catalyze chemical reactions has made many people believe that

there was an ancient RNA world before the current DNA-and-protein-dominant world (for review see Joyce 2002). Recent evidence also suggests that ncRNAs may be responsible for core mechanisms, such as catalyzing the formation of peptide bonds in protein synthesis in all organisms (Nissen et al. 2000; Schmeing et al. 2002), and catalyzing the splicing of pre-mRNAs in eukaryotes (For review see Will and Luhrmann 2001).

Secondly, single-stranded RNA molecules can fold into high-order structures (see section 1.2. for details). Some people believe that the complexity of RNA structures is comparable to that of proteins (see Klosterman et al. 2004). A variety of regions in RNA molecules can be functional elements that interact with other molecules. For instance, both the double-stranded regions and single-stranded regions in folded RNA molecules have been reported as important protein-binding motifs (see Varani and Pardi 1994).

In recent years, novel regulatory functions have been found to be associated with ncRNAs. For example, conservation of a microRNA (miRNA), *let-7*, and conservation of its targets were found in diverse animals (Pasquinelli et al. 2000; Slack et al. 2000). miRNAs, which are 20-26 bases in length, can regulate expression of other genes by inducing translation repression or degradation of target mRNAs (for review see Bartel 2004). With pure experimental approaches and also strategies assisted by *in silico* comparative genomics, many novel miRNAs have been discovered (see Grosshans and Slack 2002; see Bentwich et al. 2005) and the number of unique miRNAs is still growing (Griffiths-Jones et al. 2006).

One stereotype about ncRNA genes is that they are much shorter than protein-coding genes, because the lengths of all classic ncRNA genes are shorter than 400 bases. The same rule seems applicable to other novel ncRNAs such as miRNAs. Nonetheless, evidence suggests that short ncRNA genes might not cover all the hidden ncRNA mass in mammalian genomes. In addition to short and structural ncRNA genes, thousands of mRNA-like ncRNAs (nc-mRNAs) have been found (Okazaki et al. 2002; Ota et al. 2004; Carninci et al. 2005; Ravasi et al. 2006). These

nc-mRNAs can be several kilo bases in length and their gene structures may contain introns. Little is known about their functions except that they do not appear to code for proteins. Existing evidence suggests that nc-mRNAs may be implicated in important regulatory mechanisms. One example is H19, which encodes a 2.3-kb nc-mRNA that appears to influence growth (for review see Arney 2003) and may behave as a putative tumour suppressor gene (Matouk et al. 2007). Besides, some mammalian nc-mRNAs, which have been shown to be antisense to normal transcripts of protein-coding genes (Katayama et al. 2005), seem capable of interfering with transcription or mRNA stability of protein-coding genes. However, it is still unknown whether these noncoding transcripts can escape the surveillance of the nonsense-mediated decay (NMD) system which can eliminate aberrant transcripts with premature stop codons (for review see Weischenfeldt et al. 2005). The discovery of nc-mRNA transcripts has brought us more questions than answers to the roles of ncRNAs in vertebrates.

In addition to recently discovered regulatory roles of many ncRNA genes, RNA motifs in transcripts have long been known as important regulators of gene expression. *Cis*-regulatory RNA motifs can regulate transcription termination, mRNA decay (for review see Steege 2000), translation regulation (for review see Kozak 2005), *etc*. For example, *rho*-independent transcriptional terminators, which are believed to be composed of a stable hairpin and a uridine-rich region, can determine the 3' boundaries of polycistronic transcription units in *E. coli* and in *B. subtilis* (Farnham and Platt 1981; Ingham et al. 1999). Recently, novel ncRNA motifs in bacterial transcripts have also been found to form switch controls of gene expression, which can respond to concentration changes of small metabolites (Mandal et al. 2003; Nahvi et al. 2004). *Cis*-regulatory RNA motifs are also implicated in the efficiency of translation initiation (for review see Lopez-Lastra et al. 2005) and the decay of mRNAs (Ringner and Krogh 2005) in eukaryotes. The word ncRNA is actually a common name for diverse classes of non-protein-coding genes and versatile functional elements in transcripts. For simplicity, both

ncRNA genes and intragenic RNA motifs are generally referred to as ncRNAs in the rest of this thesis.

1.2. RNA structures

One of the most important characteristics of many ncRNAs is their capability to fold into high-order structures. It is widely believed that conservation of structure is more important than of primary-sequence motifs for ncRNA function. Features of RNA structures, such as folding stability and multi-species conservation of structures, have been used for genome-wide ncRNA finding (Rivas et al. 2001; di Bernardo et al. 2003; Coventry et al. 2004; Washietl et al. 2005). Consequently, before further discussion of the current status of genome-wide ncRNA finding (see section 1.4. for details), it is necessary to give an overview of RNA structures and available algorithms for RNA structure prediction.

RNA folding seems to be a hierarchical process: initially secondary-structure motifs form in the primary sequence, and then tertiary structures are formed through interactions between secondary-structure motifs (see Onoa and Tinoco 2004). Although the details of RNA folding may require further refinement, this hierarchical view has been a useful guideline for studying and predicting RNA structures. RNA secondary-structure motifs are introduced in subsection 1.2.1. and RNA tertiary-structure motifs are introduced in subsection 1.2.2. Algorithms for predicting RNA structures are introduced in section 1.3.

1.2.1. RNA secondary-structure motifs

Similar to DNA double helices, RNA can form anti-parallel helices (see Westhof and Michel 1994). By and large, RNA helices are held together by the hydrogen bonds formed between Watson-Crick base pairs. In addition to standard types of A-U and G-C pairs, G-U type pairs are frequently seen in RNA helices and are regarded as valid wobble pairs. Base pairs other

than A-U, G-C or G-U are regarded as non-canonical in RNA helices. Non-canonical base pairs are not completely prohibited from real-world RNA secondary structures and may play key roles in tertiary interactions (for review see Gutell et al. 1994). They may also serve as specialized sites for interacting with other macromolecules, such as proteins (for review see Hermann and Westhof 1999).

Whereas DNA double helices preferably adopt B-form structures in solution, RNA helices adopt mainly A-form structures. Due to the presence of a 2'-hydroxyl group of each RNA ribose sugar, each ribose should assume the 3'-endo conformation to avoid steric clashes beween the 2'-hydroxyl group and the C8 atom (of the purine) or C6 atom (of the pyrimidine) that are attached to the ribose (see Neidle 2002). No B-form RNA helices have ever been reported. Consequently, the thermodynamic parameters for RNA helices are different from those of DNA helices.

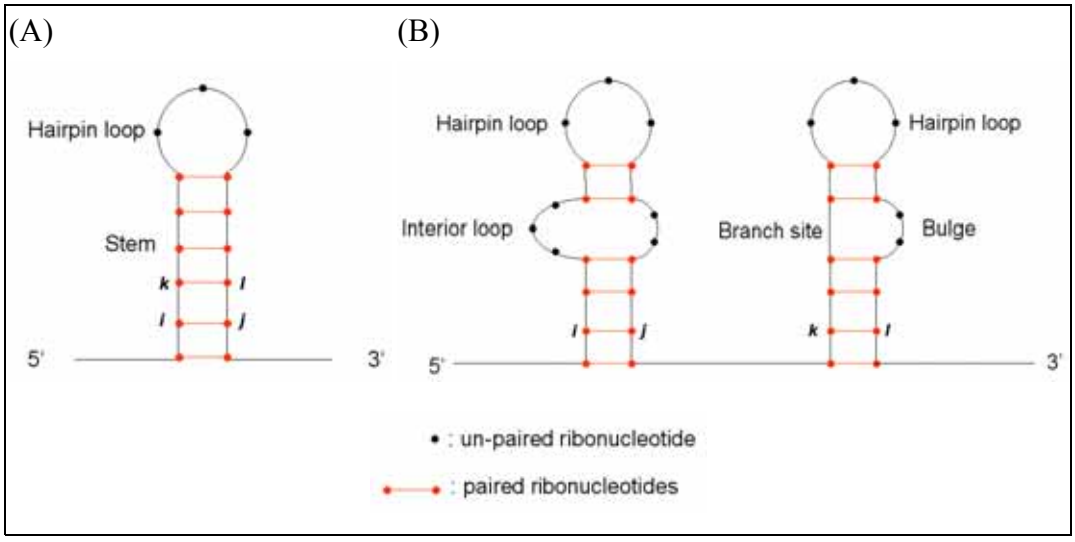


Figure 1-2. Elements of RNA secondary structures

RNA helices can be formed either intra-molecularly or inter-molecularly, although inter-molecular helices are not further discussed in this thesis. Only the features of the secondary structures formed intra-molecularly are of interest, because inter-molecular interactions are currently not used for genome-wide ncRNA finding.

When an RNA molecule fold back on itself, a number of paired regions may form. All the base pairs formed intra-molecularly at the secondary-structure level are supposed to obey the nested rule: for any two base pairs, i-j and k-l, where $i < j$, $k < l$, and, $i < k$, the order of the 4 bases should be either $i < k < l < j$ (Figure 1-2, A) or $i < j < k < l$ (Figure 1-2, B). A region of continuous base pairs in an RNA secondary structure is referred to as a stem.

For the unpaired regions in an RNA secondary structure, a series of names can be used to describe them according to their respective relations to the nearest neighbouring stems. A "hairpin loop" is the terminal unpaired region of a stem (Figure 1-2, hairpin loop). A "bulge loop" is a region where at least one unpaired ribonucleotide is on one strand of a stem, while all ribonucleotides on the opposite strand are base paired (Figure 1-2, bulge loop). An "interior loop", which linearly separates two stems, is formed when there is at least one unpaired ribonucleotide on each strand (Figure 1-2, interior loop).

A hairpin loop together with its nearest stem is referred to as a hairpin. The formation of hairpins is possibly one of the most fascinating features of ncRNAs. One of the best known examples of hairpins is that of tRNA which has a canonical cloverleaf-like secondary structure (Figure 1-3).

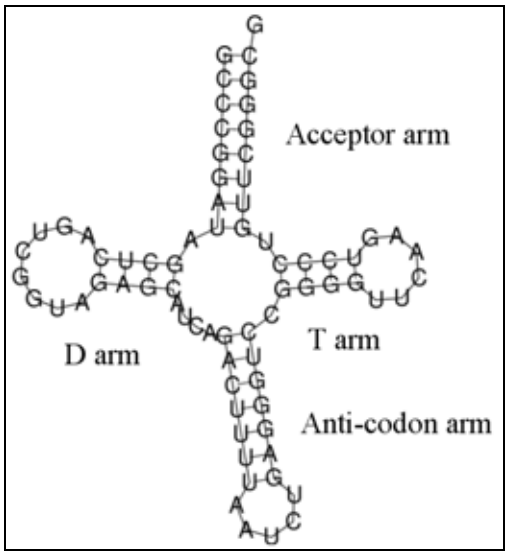


Figure 1-3. The cloverleaf-like secondary structure of a tRNA

This diagram of the cloverleaf-like secondary structure of a human Lys-tRNA is plotted by RNAplot of ViennaRNA package (Hofacker 2006). The human Lys-tRNA sequence is retrieved from NCBI35:Chr11:59080478-59080550.

1.2.2. RNA tertiary structures

Specific combinations of RNA secondary-structure motifs are necessary for RNA molecules to fold into functional tertiary structures. Well known RNA tertiary-structure motifs include base triples, kissing hairpin loops, ribose zippers, *etc*. (see Tamura et al. 2004). Predicting the complete tertiary structure of ncRNAs is not investigated in this thesis, because determining it using pure computational approaches is very difficult and it is not essential for the algorithms devoted to simply finding ncRNAs in genomes.

There are a number of reasons for the prediction of ncRNA tertiary structures being difficult. Firstly, the interactions between interacting strands of RNA molecules do not always adhere to the Watson-Crick base-pairing rule (for review see Leontis and Westhof 2003). Secondly, the interaction rules governing the formation of tertiary-structure motifs have still not been studied in detail. Thirdly, the computational complexity of predicting RNA tertiary structures is much higher than that of predicting RNA secondary structures (see subsection 1.3.3.3.). Therefore

only those tertiary-structure motifs that can be simultaneously predicted by existing secondary-structure prediction algorithms are covered in the next two subsections (1.2.2.1. and 1.2.2.2.).

1.2.2.1. Co-axial stacking

A *quasi*-continuous helix can be formed when two adjacent stems stack co-axially. For instance, in the final inverted L-shaped conformation of tRNAs, there are two co-axial stackings: one is between the acceptor arm and the T arm (Figure 1-3) and the other is between the D-arm and the anticodon arm (Figure 1-3).

Co-axial stacking is an important force to guide secondary-structure motifs of an RNA molecule to fold into functional tertiary structures. Co-axial stacking proved to enhance the stability of RNA secondary structures (Walter et al. 1994). Besides, co-axial stacking may be important for stabilizing the multi-loop junctions in RNA secondary structures (Walter et al. 1994). Evidence suggests that taking the co-axial stacking into consideration can be useful for improving the predictions of RNA secondary structures (Walter et al. 1994).

1.2.2.2. Pseudoknots

Figure 1-4. Non-nested base pairs in a pseudoknot

A pseudoknot is defined as a double-stranded region, which is formed between the loop

region of a hairpin and the single-stranded region outside this loop (Figure 1-4). The first experimental example of pseudoknots was found at the 3' end of turnip yellow mosaic virus (TYMV) RNA (Rietveld et al. 1982). The nested rule of base pairs in stems at the secondary-structure level (for details see subsection 1.2.1.) is broken by the formation of base pairs in pseudoknots. Developing prediction algorithms that consider pseudoknots is considerably harder because of this. A pseudoknot is sometimes categorized as a secondary-structure motif, because it can be decomposed into individual hairpins. However, due to the relationships between base pairs in a pseudoknot, pseudoknots are sometimes classified as tertiary-structure motifs.

Pseudoknots have been found to play diverse and important roles, such as forming the catalytic core of ribozymes, binding of regulators for translation, and inducing ribosomal frameshifting in many viruses (see Staple and Butcher 2005).

1.2.3. The dynamic aspect of RNA structures

Instead of regarding RNAs as static molecules consisting of static stem-loop structures, a "dynamic" view should be considered. One RNA molecule can potentially fold into various conformations (see Flamm et al. 2000). In response to certain circumstances, such as fluctuations of ligand concentrations (Mandal et al. 2003), or particular ionic strength (Olson et al. 1976; Rangan and Woodson 2003), RNA molecules may fold into alternative structures. Besides, interaction of RNA molecules with other macromolecules can induce conformational changes (Rould et al. 1991; Cavarelli et al. 1993). Post-transcriptional modification of ncRNAs can also affect the stability of RNA structures (for review see Helm 2006). Prediction strategies for ncRNAs should therefore take into account the potential for RNA molecules to adopt alternative structures under different conditions. This is considered further when developing loop-dependent rules for predicting RNA secondary structures (for details see subsection 1.3.1.2.) and in

locating local hairpins for creating models of RNA motifs (for details see subsection 4.2.1.1.).

1.2.4. The definition of "RNA motifs" used in this thesis

In the remainder of this thesis, "RNA motifs" are used to describe combinations of primary-sequence motifs and stem-loop structures, where stem structures consist mainly of Watson-Crick base pairs. However, it should be noted that the exact meaning of this term might not be consistent across all research fields. For example, "RNA motifs" in structural biology specifically refer to combinations of non-Watson-Crick base pairs that enable the phosphodiester backbones of interacting RNA strands to form distinctive folds (see Leontis and Westhof 2003).

1.3. Prediction of RNA structures

Although experimental approaches are available for determining structures of RNA molecules (for review see Neidle 2002), there are certain limitations. For example, X-ray crystallography can provide high-resolution structural information, however the process of crystallization is a slow process and not very predictable (see Ke and Doudna 2004). Besides, ncRNAs can be larger than the size at which current nuclear magnetic resonance (NMR) methods can work effectively (see Lukavsky and Puglisi 2005).

Given these limitations, computational methods can be valuable, especially when the lengths of the ncRNAs of interest are longer than 100 bases, which is the upper limit for NMR RNA structure determination (for review see Riek et al. 2000). The prediction of RNA structures is often narrowed down through first predicting RNA secondary structures. One reason is that RNA tertiary structures seem to be held by tertiary interactions between secondary-structure motifs. It is generally believed that with reliable predictions of secondary structures, it should be possible to infer the tertiary structures, although as discussed in 1.2.2. predicting complete RNA tertiary structures is not the objective of this thesis.

Intuitively, predicting RNA secondary structure is similar to finding the alignments between two nucleic acid sequences, except that in this case the aligned strand is composed of complementary bases rather than identical or similar bases. Various algorithms have been designed for predicting RNA secondary structures. These algorithms can be generally categorized into three classes: minimization of free energy, phylogenetic comparative analysis, and probabilistic models. These algorithms are introduced in subsections 1.3.1. , 1.3.2. , and 1.3.3.

1.3.1. Minimization of free energy (MFE)

1.3.1.1. Base-pair dependent energy rule

Energy minimization is one of the favourite *ab initio* methods for predicting RNA secondary structures. The first algorithm that was introduced is the base-pair dependent energy rule (Nussinov and Jacobson 1980). In this energy model, formation of hydrogen bonds for each base pair is assumed to be independent from its neighbouring base pairs. The overall energy is expressed as of the sum of energies of individual base pairs in an RNA molecule:

$$E(\mathrm{S}) = \sum_{i,j \text{ in } S} e(i, j) \qquad [1.1]$$

The optimal solution can be found by using a dynamic programming algorithm. The recursion for this can be written as

$$W(i, j) = optimal \begin{cases} W(i+1, j-1) + e(i, j) \\ W(i, k-1) + W(k, j), \quad i < k \le j \end{cases} \qquad [1.2]$$

where $W(i, j)$ is the minimum folding energy for the region from base i to base j in a given RNA sequence. In [1.2], if base i can pair with base j, $e(i, j)$ returns the pairing energy (presumably some negative values), positive infinity otherwise. "k" is sometimes called the branching site, because sequence i to j is divided into two parts: i to $k - 1$, and k to j. In real hairpins, short-range base pairs are not permitted due to sterical hindrance. If $(j - i)$ is smaller

than 4, $W(i, j)$ returns positive infinity. The time complexity of the recursion is $O(N^3)$, where N is the length of each sequence. $W(i, j)$ can also be used to find the structure with the maximum number of base pairs for any given RNA molecule, if used with an energy function $e(i, j)$ that returns 1 when base i and base j are paired, and 0 otherwise.

However, based on biochemical data, it has been generally accepted that the thermodynamic stability of a base pair depends on the identity of nearest neighbours (for review see Borer et al. 1974). This rule is also termed as the individual nearest-neighbour (INN) rule (Gray 1997). Clearly, the base-pair dependent energy rule is not compatible with the INN rule, because the energy term, $e(i, j)$, considers only the energy contributed by formation of hydrogen bonds between base i and j, but not the energy contributed by the stacking of neighbouring bases.

1.3.1.2. Loop-dependent rule

The first free-energy formulation that takes dependence of base pair energy on nearest neighbours into consideration is the loop-dependent rule. The main idea is to decompose an RNA secondary structure into combinations of individual hairpins (Zuker and Stiegler 1981):

$$E(\mathrm{S}) = \sum_{i.j \quad in \quad S} e(i, j) + e(L_{ext}) \qquad [1.3]$$

, where L_{ext} is the structure that may fold by sequence outside the range between i and j.

The optimal solution can be found by using a dynamic programming algorithm. The recursion is:

$$W(i,j) = optimal \begin{cases} W(i+1, j) \\ W(i, j-1) \\ V(i, j) \\ \underset{i \le k < j}{optimal} \; W(i,k) + W(k+1, j) \end{cases} \qquad [1.4]$$

$$V(i,j) = optimal \begin{cases} h(i,j) \\ s(i,j) + V(i+1, j-1) \\ VBI(i,j) \\ VM(i,j) \end{cases} \quad [1.5]$$

$$VBI(i,j) = \underset{\substack{i<k<l<j \\ k-i+j-l>2}}{optimal} \; ebi(i,j,k,l) + V(k,l) \quad [1.6]$$

$$VM(i,j) = a + \underset{i<k<j-1}{optimal} \; W(i+1,k) + W(k+1, j-1) \quad [1.7]$$

$W(i, j)$ is similar to the energy term in the recursion for the base-pair dependent energy rule (see subsection 1.3.1.1.). $V(i, j)$ is the minimum energy for sequence i to j, when base i can pair with base j. There are several cases for $V(i, j)$: 1) base pair *i-j* closes a hairpin loop and h is the energy for this loop; 2) base pair *i-j* stacks on base pair $(i+1)$-$(j-1)$ and s is the stacking energy; 3) base pair *i-j* closes a bulge or internal loop and the energy for this loop is *VBI*; 4) base pair *i-j* closes a multi-loop and *VM* is the energy for this situation, where a is the energy penalty for opening a multi-loop. In *VBI* [1.6], *ebi* denotes the loop region closed by base pair *i-j* and containing base pair *k-l*.

The computational complexity of [1.7] is $O(N^3)$, and the complexity of [1.6] is $O(N^4)$. In order to limit the time complexity of [1.6], an additional constraint, where $(k - i + j - l)$ must be no greater than some fixed number, can be added. Lots of extensions have been made to include additional energy terms, such as single-base stacking, mismatched pair stacking, coaxial helix stacking (Walter et al. 1994; Rivas and Eddy 1999), empirical rules, and pseudoknots (Rivas and Eddy 1999).

The general problem of predicting pseudoknots has been proven to a non-deterministic polynomial (NP-complete) problem (Lyngso and Pedersen 2000). Several algorithms are now

available for predicting optimal pseudoknot-inclusive structures under certain constraints (Rivas and Eddy 1999; Dirks and Pierce 2003; Matsui et al. 2004). However using these algorithms, predictions of some complex cases, such as interlaced pseudoknots, are not guaranteed to be optimal. Besides this, the computational complexities in time and space can be as high as $O(N^5)$ and $O(N^4)$ respectively. Therefore, only simple pseudoknots in short RNA sequences can be predicted within a reasonable period of time using these approaches.

1.3.1.3. Considerations when using MFE based approaches

One concern about using MFE based approaches to predict RNA secondary structures is its high error rate. It is suggested that only 50% – 70% of base pairs in RNA secondary structures can be correctly predicted by using minimization of free energy (Eddy 2004). Several reasons account for this situation. Firstly, thermodynamic parameters are not complete. Not all possible combinations of sequences in loops, stacked bases, *etc*. have been experimentally evaluated. Secondly, structures with minimal free energies are not necessarily the biologically functional ones (Konings and Gutell 1995; Fields and Gutell 1996). In order to address this problem of alternative structures, programs such as MFOLD (Zuker 1989) were designed to predict multiple alternative, but less stable, secondary structures for one RNA molecule. MFOLD can also use experimental results as folding constraints (Zuker 1989). Further experiments can be designed to test predictions and feed back into the prediction process. This iterative process is very useful in the determination of RNA secondary structures.

1.3.2. Phylogenetic covariation analysis

Unlike MFE based methods, which can be used on a single sequence, phylogenetic covariation analysis depends on alignments of multiple related sequences. These could be either expressed ncRNA or genome sequence and could be from different species or from paralogous regions within a single genome. The approach takes compensatory mutations (covariations)

found within these alignments as indicators of conserved double-stranded regions. The basic assumption is that the functions of ncRNAs depend more on high-order structures than on primary sequences. Therefore compensatory mutations that preserve the pairing potential in helices can support the existence of conserved structures. Conversely, if the mutations that are found in naturally existing homologues can destabilize the putative helical regions, the structures are unlikely to be truly functional *in vivo*.

Phylogenetic covariation analyses have been successfully applied to the elucidation of the structures of rRNAs, class I and class II introns, and snRNAs (James et al. 1989). Putative covariations can also be used as constraints in running programs using MFE to refine the predicted structure (Shanab and Maxwell 1991). This approach has been demonstrated to be one effective approach for determining the higher-order structures of large RNAs (Gutell et al. 1994)

A phylogenetic covariation analysis for RNA secondary structure prediction depends on appropriate alignments of homologous sequences. If functionally related ncRNAs are really divergent, too many mutations may prevent us from obtaining optimal alignments for structure predictions. On the other hand, if the number of covariations in ncRNA homologues is small, the information content may not be sufficient to validate putative stem regions. This paradox is also applicable to other algorithms that use comparative genomics for ncRNA finding. The suitability of using comparative genomics for genome-wide ncRNA finding is further investigated in subsection 1.4.2. and in chapter 2.

1.3.3. Grammatical approaches for RNA sequence analysis

Ideas from computational linguistics have been applied to RNA secondary structure analysis. One important example is the application of stochastic context-free grammars to RNA structure (RNA SCFGs) (Eddy and Durbin 1994; Sakakibara et al. 1994), which provide a way to perform probabilistic modelling of RNA secondary structures. SCFGs are a stochastic version

of context-free grammars, which correspond to the second level of the Chomsky hierarchy of transformational grammars (Chomsky 1959). Other grammar-based approaches have also been proposed to model limited types of RNA tertiary-structure motifs. Before further discussing grammar-based RNA analysis, I introduce some basics of computational linguistics.

In computational linguistics, an important task is to determine whether an observed string is grammatically correct. The Chomsky hierarchy of transformational grammars (Chomsky 1959) provides a general theory for modelling strings of symbols. A transformation grammar can be considered as a device that can generate strings of symbols. A transformational grammar consists of several components: 1) a finite set of terminal symbols; 2) a finite set of nonterminal symbols; 3) a finite set of production rules. Terminal symbols correspond to the actual symbols that may appear in a string that can be observed in a particular language. Nonterminals can be transformed, by a production rule, into a new string of terminals and/or nonterminals. Transformational grammars are also called generative grammars because of their capability of generating strings of symbols. Here is an example of a simple generative grammar in which there is only one production rule:

$$S \rightarrow aS \mid \varepsilon .$$

S is a nonterminal; a is a terminal; ε is a special terminal to represent an empty string; "$\rightarrow$" means transformation; a vertical bar means "or". This production rule says that a nonterminal S can be transformed into aS or ε. Such a simple generative grammar is capable of generating strings consisting of a's of any length.

By incorporating more nonterminals and more terminals into a generative grammar, a string of symbols with a more complicated structure can be modelled. An important feature of the Chomsky hierarchy is its capability to model a variety of strings with different levels of structural complexities. In computational linguistics, "structure" is used to indicate the

correlations between different symbols in a string. In order to model structures of different complexities, Chomsky described four levels of restrictions on the production rules. Accordingly, transformational grammars are classified into four classes, which form the Chomsky hierarchy of transformational grammars. The Chomsky hierarchy can be expressed in a set inclusion form:

$$\text{regular} \subset \text{context-free} \subset \text{context-sensitive} \subset \text{unrestricted}.$$

The ordering in this hierarchy indicates the relative descriptive power of the grammars. The grammars on the left-hand side are more restricted than the ones that are on the right-hand side. Regular grammars, which are the most restricted and lowest level of the Chomsky hierarchy, allows production rules only in the form of "$W \rightarrow aS$", "$W \rightarrow a$", or "$W \rightarrow \varepsilon$", where W and S can be any nonterminals and terminals, respectively. ε is an empty string. Regular grammars can generate any strings. However, regular grammars are unsuitable for describing high-order correlations, such as the nested pairwise correlations (Figure 1-5 A) in the secondary structures that can be folded in an RNA molecule. In the next two subsections, I introduce the grammar-based approaches for determining RNA secondary structures and for finding related RNA sequences in sequence databases, respectively.

(A)

(B)

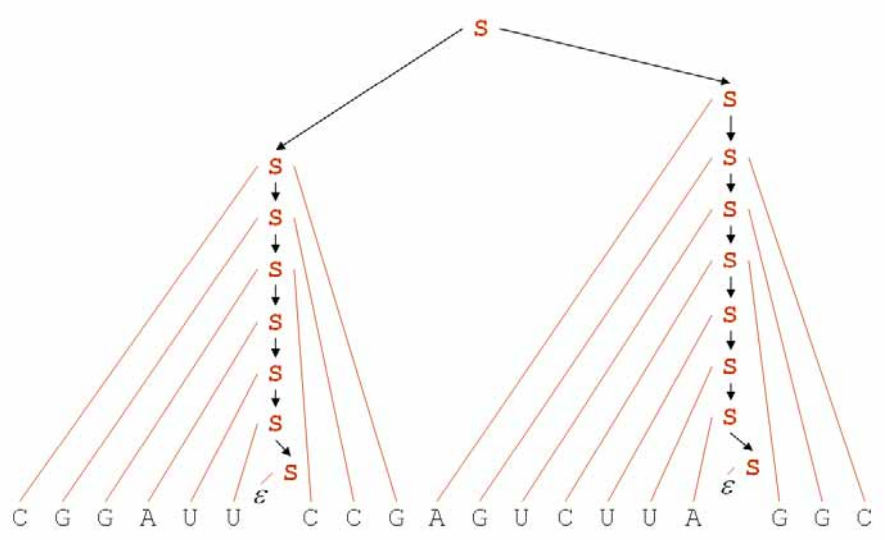


Figure 1-5 Two representations of the pairwise correlations in an RNA molecule with two non-interlaced hairpins

(A) The nested pairwise correlations formed in an RNA molecule with two hairpins (B) The parse tree of the nested pairwise correlations in (A)

1.3.3.1. SCFG-based RNA secondary structure analysis

Context-free grammars, which are a higher level in the Chomsky hierarchy than are regular grammars, have been used to model the RNA secondary structures. For instance, any stems in RNA secondary structures, such as the arms in figure 1-3, can be generated by the following production rule that adheres to CFGs:

$$S \rightarrow aSu \mid cSg \mid gSc \mid uSa \mid gSu \mid uSg \mid \varepsilon \text{ . (paired production)}$$

Bulges or loops in RNA secondary structures can be generated by

$$S \rightarrow aS \mid cS \mid gS \mid uS, \text{ or (left unpaired production)}$$

$$S \rightarrow Sa \mid Sc \mid Sg \mid Su. \text{ (right unpaired production)}$$

Taking the RNA secondary structures in Figure 1-2 as the example, the hairpin loops can be generated by left unpaired productions; the bulge shown on the right-hand side of Figure 1-2 B can be generated by right unpaired productions.

For the cases where there are multiple hairpins folded by an RNA molecule, as in the case in Figure 1-5 A, a rule of bifurcation is required:

$$S \rightarrow SS. \text{ (bifurcation)}$$

The secondary structure of an RNA molecule can be represented as a so-called parse tree (Figure 1-5 B).

The RNA CFG described above essentially follows the base-pair dependent rule, which is used in the Nussinov's algorithm for predicting RNA secondary structures. In terms of predicting the RNA secondary structure for an RNA sequence, a better energy rule, as suggested at the end of subsection 1.3.1.1, is the individual nearest-neighbour rule. An RNA CFG can also be extended to follow the INN rule by incorporating more nonterminals and modifying the original production rules (Durbin et al. 1998).

One problem with using an RNA CFG is that it is only possible to decide whether an RNA sequence can be generated by this grammar. In the cases where many parse trees exist for an RNA sequence given an RNA CFG, it is impossible to determine which tree (*i.e.* secondary structure) is the most probable one. One solution to improve this situation is using a stochastic form of RNA CFGs. In stochastic SCFGs, probabilities can be assigned to different production rules. For instance, in an RNA SCFG, non-Watson-Crick G-U pairs are accepted in RNA helices but should be generated with a lower frequency than Watson-Crick G-C and A-U pairs are. The probabilities of different production rules, including bifurcations, paired production, and unpaired productions, can be estimated from the known secondary structures folded in well-studied RNA sequences.

In order to use an RNA SCFG to determine RNA secondary structures, we need algorithms that can align sequences to the grammar. The relevant algorithms include the Cocke-Younger-Kasami (CYK) algorithm, the inside-outside algorithm, *etc*. The CYK algorithm (Durbin et al. 1998) can be used to find the most probable parse tree for a sequence given a SCFG. The inside-outside algorithm (Durbin et al. 1998) can be used to calculate the probability of a sequence with an RNA SCFG. For predicting RNA secondary structures, both the CYK and inside-outside algorithms have the same the algorithmic complexity as the Zuker's algorithm does (see subsection 1.3.1.2).

The score of a sequence X is often given as a log-odds ratio, $\log (P(X, \hat{\tau} | \theta)/P(X|\phi))$ (Durbin et al. 1998). $P(X, \hat{\tau} | \theta)$ is the probability of a sequence and the best alignment given an RNA SCFG. This probability, $P(X, \hat{\tau} | \theta)$, is calculated by multiplying together the probabilities of the productions chosen to generate the best alignment ($\hat{\tau}$) of X to the RNA SCFG θ. $P(X | \phi)$, is the probability of generating X by a null (random) model ϕ. When base-2 logarithms are used to calculate the log-odds ratios, scores are reported in bits and are so called bit scores.

1.3.3.2. RNA covariance models

SCFGs can be applied to searching for the homologous members of a family of related RNAs in a sequence database. One approach is the "covariance model" (CM) (Eddy and Durbin 1994), which is so named because it can describe the compensatory mutations (covariations) in the consensus secondary structure of homologous ncRNAs.

Given an alignment of related RNAs that share a common structure like the one in Figure 1-5 A, a very simple CM can be written as an ordered list of production rules to model this RNA family:

	Stem 1	**Stem 2**
$S_0 \rightarrow S_1S_8$	$S_1 \rightarrow cS_2g$ …	$S_8 \rightarrow aS_9$…
	$S_2 \rightarrow gS_3c$…	$S_9 \rightarrow gS_{10}c$…
	$S_3 \rightarrow gS_4c$…	$S_{10} \rightarrow uS_{11}g$…
	$S_4 \rightarrow aS_5$…	$S_{10} \rightarrow cS_{11}g$…
	$S_5 \rightarrow uS_6$…	$S_{12} \rightarrow uS_{13}$…
	$S_6 \rightarrow aS_7$ …	$S_{13} \rightarrow uS_{14}$…
	$S_7 \rightarrow \varepsilon$	$S_{14} \rightarrow aS_{15}$…
		$S_{15} \rightarrow \varepsilon$

In a CM, one nonterminal is needed for each singlet base and one nonterminal is needed for each base pair. Therefore the number of nonterminals in a CM is about linearly proportional to the length of the alignment. A pairwise production that is in the form "$V \rightarrow a W b$" should have 16 pair emission probabilities; a leftwise or rightwise production, such as "$V \rightarrow aW$" or "$V \rightarrow Wa$", should have 4 singlet emission probabilities. In the rules above, only one production per production rule is listed and other possible productions are omitted (as indicated by "…") for simplicity. In a practical CM that can be used to search for RNAs in a sequence database, further modification of the production rules is required. For example, additional nonterminals and productions for modelling insertions and deletions may be required in either pairwise production rules or singlet production rules.

The parameters of a CM can be estimated from a curated RNA sequence alignment, which should reveal the consensus secondary structure of a family of related RNAs. For instance, the probabilities of different singlet bases and base pairs are calculated per column in the sequence alignment, and are used as the parameters in the production rules of a CM.

1.3.3.3. Modelling high-order RNA structures using grammar-based approaches

SCFGs are suitable for modelling the nested base pairs in RNA secondary structures. However, in higher-order RNA structures, the interactions between bases may not follow the nested rule.

In RNA tertiary structures, there may be crossing interactions such as:

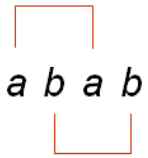


Figure 1-6. A crossing interaction that may be found in RNA tertiary structures

One example is RNA pseudoknots, as the one shown in Figure 1-4. In the standard forms of the grammars from the Chomsky hierarchy, context-sensitive grammars (CSGs) are required to model such structures. CSGs can reorder the nonterminals according to their local context and thus can generate strings of symbols that contain crossing dependence. However, the general problem of parsing strings that are generated by CSGs is a nondeterministic polynomial problem (*NP*-complete problem) (Durbin et al. 1998).

Attempts have been made to apply grammars, whose computational complexity lies between CFGs and CSGs, to the modelling of RNA pseudoknots and some limited forms of RNA tertiary-structure motifs. Crossed-interaction grammars (CIGs) (Rivas and Eddy 2000) are an example. In addition to the production rules of CFGs, a CIG also has a set of rearrangement rules. It is the set of rules that make CIGs different from CFGs. The rearrangement rules apply to reorder the terminals only after all the conventional CFG-compatible nonterminals have been used to generate terminals. A rearrangement rule consists of a zero-length hole string $\wedge$ and a set of special nonterminals. The hole string $\wedge$ is used to indicate the possible points that can be inserted by another string. Special nonterminals, including $\times$, (, and), are used to specify how symbols should be rearranged.

Here is an example of how a complicated pseudoknotted structure can be derived ("$\underset{R}{\Longrightarrow}$" is used to represent a rearrangement.):

$$((a \wedge a) \times (b \wedge b \times a \wedge a)) \underset{R}{\Longrightarrow}$$

$$a \wedge a \times ba \wedge ba \underset{R}{\Longrightarrow}$$

$$aba \wedge aba.$$

CIGs are not the only grammars that can be used to model high-order RNA structures. In recent years, the variant forms of tree adjoining grammars (TAGs) (Uemura et al. 1999; Matsui et al. 2004; Chiang et al. 2006) have also been applied to RNA sequence analysis.

A major consideration in applying these grammars to genome-wide RNA analysis is high computational complexity. The time complexity and storage complexity of parsing the CIG above is $O(n^6)$ and $O(n^4)$, respectively, where n is the length of the string. The time complexity of parsing a TAG variant, which has the capability of modelling RNA secondary structures including pseudoknots, is $O(n^5)$ (Uemura et al. 1999). If more complicated crossed interactions are allowed, the required computational complexity can be even higher (Rivas and Eddy 2000; Chiang et al. 2006).

1.4. Current state of genome-wide ncRNA finding

Computational detection of ncRNAs in genomes is not a completely new field. Based on RNA secondary structure prediction algorithms described above (section 1.3.), many a*d hoc* ncRNA finders have been designed to predict specific classes of ncRNAs in genomes. One of the most successful cases is genome-wide tRNA finding. For example, tRNAscanSE can identify 99%-100% tRNA genes in genomic sequences with very low false positive rate (Lowe and Eddy

1997). In addition, many programs can predict miRNAs in genomes with impressive specificities and sensitivities. (Ohler et al. 2004; Nam et al. 2005; Xue et al. 2005). In general, once a few sequences of a particular ncRNA family are available, probabilistic models that describe the statistical features of both primary-sequence and structural motifs can be derived (Eddy and Durbin 1994; Sakakibara et al. 1994; Gautheret and Lambert 2001). One widely used probabilistic model of structural motifs is the covariance model (CM) (see subsection 1.3.3.2.). Besides, even when only a single ncRNA sequence is known, some algorithms have been created to search sequence databases for homologs with similar primary-sequence and secondary-structure motifs (Klein and Eddy 2003; Bafna and Zhang 2004; Havgaard et al. 2005).

While genome-wide searches for ncRNAs of known structural features are relatively straightforward, *ab initio* genome-wide ncRNA finding is still very challenging. A probabilistic model of a particular class of ncRNAs is unlikely to be useful for finding other classes of ncRNAs, because different classes of ncRNAs do not seem to have many common structural motifs that can be predicted by available secondary structure prediction algorithms.

Some alternative approaches based on assumptions of RNA structural features have been developed (Rivas and Eddy 2001; di Bernardo et al. 2003; Coventry et al. 2004; Washietl et al. 2005; Pedersen et al. 2006). However, none of them have proved to be effective for finding different classes of ncRNAs in real genomic sequences. For example, a recent report about finding ncRNAs in the human genome indicates that existing algorithms may exhibit fairly high false discovery rates of 50%~70% (Washietl et al. 2007). This situation can be partly attributed to three factors: 1) few statistically useful features have been found that can be used for identifying ncRNAs in genomes; 2) some algorithms have been developed based on assumptions rather than on statistics collected from real data; 3) there are few appropriate data sets of functional ncRNAs for testing and improving algorithms effectively. These three issues are discussed in more details in subsections 1.4.1. , 1.4.2. , and 1.4.3.

1.4.1. Few statistically useful features for classifying ncRNAs

Unlike protein-coding genes, no compositional propensities at primary sequence level have been found to be statistically useful for *ab initio* ncRNA finding in genomes. Intuitively, features associated with synthesis, maturation, or functions of ncRNAs should be useful for identifying ncRNAs, however, mechanisms involved in synthesis and function may vary from one class of ncRNAs to another class of ncRNAs. For example, the transcription of ncRNAs may not use the general machinery required for mRNAs. RNA polymerase II (RNA pol II) is not the only polymerase responsible for the transcription of ncRNAs. Though most snRNAs are transcribed by RNA pol II, U6 snRNA is transcribed by RNA polymerase III (RNA pol III) (Reddy et al. 1987). Also, ncRNAs may not always exist as independent transcription units. Though in vertebrates, the most abundant snoRNAs, U3, U8, and U13 RNAs, are synthesized from independent transcription units by RNA pol II, most of the other known snoRNAs (U14-U22) are encoded within introns of protein-coding genes (Kiss and Filipowicz 1995).

With respect to post-transcriptional processing of ncRNAs, there is again a diversity of mechanisms. Many classes of ncRNAs must be specifically processed in order to perform their unique functions. For example, the nascent transcripts of tRNAs require RNaseP for removing their 5' leader sequences, endonucleases for cutting the middle of their 3' trailer sequences, and exonucleases for removing their residual 3' trailer sequences (for review see Nakanishi and Nureki 2005). For structural ncRNAs that are transcribed by RNA pol II, it has been shown that some of these ncRNAs require unique (non-polyadenylation) mechanisms for their 3' end maturation. For example, snoRNAs may not undergo the standard mechanism required for 3' end maturation of snRNAs (Fatica et al. 2000; Morlando et al. 2002). miRNA precursors must be processed by RNase-III enzymes, including Drosha and Dicer, in order to generate mature miRNAs (Lee et al. 2003).

In summary, biogenesis of ncRNAs does not seem to give as many common and useful signals for *ab initio* ncRNA finding in genomes as for protein-coding genes, which makes the development of algorithms more difficult and complex.

1.4.2. Assumptions made in previous work

The ability to fold into high-order structures is undisputedly the most obvious feature shared by most structural ncRNAs. Several structure-based assumptions have been used to develop algorithms for genome-wide ncRNA finding. Firstly, if stable structures were preferred for ncRNA functions, maybe evolutionary stresses would select ncRNAs with significantly lower folding energies than random sequences with similar sequence compositions. Secondly, if secondary structures, instead of primary sequences, were more important for ncRNA function, covariations should be numerous. Hypothetically, if sufficient covariations could be found, it should be possible to infer conserved secondary structures in syntenic regions between different genomes.

The first assumption, *i.e.* that stable structures are preferred in evolution, is not universally applicable to all classes of ncRNAs. It is now generally believed that the stability of RNA secondary structures is insufficient for classifying ncRNAs in genomes (Rivas and Eddy 2000). Conversely, the second assumption, *i.e.* there are numerous covariations, has been widely applied to genome-wide ncRNA finding (Rivas and Eddy 2001; di Bernardo et al. 2003; Coventry et al. 2004). Although two comparative algorithms, RNAz and EvoFold, do not explicitly depend on existence of covariations (Washietl et al. 2005; Pedersen et al. 2006), the abundance of covariations still matters. When there are very few mutations in a set of alignments, it is difficult to distinguish conservation of high-order structures from other kinds of functional constraints. In the worst cases where there are no mutations at all, the information content of a multiple-sequence alignment is equivalent to only one sequence.

Practical issues emerge when these algorithms are used to find ncRNAs in real genomes. Genomic alignments taken by these ncRNA-finding algorithms are generally generated by using primary-sequence alignment algorithms, but seldom by using structural alignment algorithms. However, primary-sequence alignment algorithms may mis-align sequences containing RNA secondary structures. There is no guarantee that these alignments (frequently generated by ClustalW) can reveal covariations correctly. In addition, no comprehensive survey has been performed to investigate whether covariations among orthologous ncRNAs contain sufficient information to be useful in prediction. In particular, the abundance of covariations between orthologous ncRNAs in vertebrate genomes is unknown. A comprehensive survey of covariations is therefore performed in chapter 2.

1.4.3. Few appropriate data sets for training ncRNA-finding algorithms

Creating ncRNA-finding algorithms is often hindered by the lack of decent training and test data sets. tRNA finding is an extremely fortunate case, since there are hundreds of experimentally verified tRNAs (Sprinzl and Vassilenko 2005); however, there are many classes of ncRNAs where only a few verified sequences are available. For example, *rho*-independent transcription terminators have been reported for two decades (Brendel et al. 1986); however, of the data set of 148 sequences that are frequently used for training and testing new algorithms (d'Aubenton Carafa et al. 1990; Ermolaeva et al. 2000; Lesnik et al. 2001; de Hoon et al. 2005), only 66 have been checked by either biochemical or genetic approaches (d'Aubenton Carafa et al. 1990). In addition, the creation of sets of mammalian ncRNAs is complicated by abundant ncRNA-like repetitive elements in genomes. For example, there are hundreds of U6 snRNA-like sequences in the human genome (Giles et al. 2004), but it is likely that only a few of them are truly functional (Domitrovich and Kunkel 2003). In fact, no obviously effective rules have been

developed to distinguish functional ncRNAs from pseudogenes in mammalian genomes.

Sometimes there are insufficient appropriate ncRNAs, even where there are numerous experimentally verified ncRNAs. For example, some genome-wide ncRNA-finding algorithms, such as RNAz and MSARI, take only ncRNA alignments with sequence identities greater than 50% and 60% respectively for both training and testing (Coventry et al. 2004; Washietl et al. 2005). These algorithms should work properly if they are used to scan genomic alignments with at least 50% identity. However, there can be substantially less test data for classes of ncRNAs that are more divergent at primary sequence level. It turns out that the trained algorithms are evaluated on biased test data and their performance on certain classes of ncRNAs, for which only divergent sequences are available, is not well assessed.

1.5. Objectives of this project

There are several issues that can be investigated with the aim of improving genome-wide ncRNA finding:

- Signals that have been widely adopted by existing algorithms can be evaluated using data sets from real genomes to better assess their value.
- Promising signals, other than structural features, for finding ncRNAs in real genomes can be tested.
- Attempts can be made to develop new algorithms combining primary-sequence and structural features.

In chapter 2, I conduct a comprehensive analysis on a genome-wide scale of the utility of signals currently used for identifying ncRNAs. I assess two factors: the conservation of ncRNAs in syntenic regions and the abundance of covariations between the synteny-conserved ncRNAs (for the definition see the introduction of chapter 2). Besides, the conservation of the

arrangement of tRNA-gene loci in mammalian genomes is explored. This study should provide useful information about the evolution of tRNA genes in mammalian genomes, and thus may guide us to choose suitable strategies for genome-wide ncRNA finding.

The synteny-conservation ratios of ncRNAs may determine the performance of the ncRNA finding methods based on a comparative strategy. In chapter 3, I explore the criteria that could potentially be useful for distinguishing functional ncRNAs from pseudogenes, Two different criteria, the distribution of bit scores and the physical clustering of tRNA genes in the human genome, are used to separate Rfam-predicted tRNAs into distinct groups, where the functionality of the tRNAs in each group are assessed.

Modelling the *cis*-regulatory elements for the transcription of ncRNAs is another strategy potentially useful for genome-wide ncRNA finding. In the first part of chapter 4, I introduce the machine learning approaches that may be useful for modelling the transcription regulatory regions of ncRNAs. In chapter 5, a sparse Bayesian learning system, Eponine, is applied to modelling the transcription start sites (TSSs) of pol III type II ncRNAs.

How many ncRNAs are still undiscovered in genomes? Given the huge number of genomic sequences, there is clearly a need for algorithms that can learn common structural motifs in a set of related sequences, which could then be used to construct probabilistic models of ncRNAs. Such algorithms might have potential for *ab initio* ncRNA finding. In the second part of chapter 4, a new module is created to extend the capability of Eponine to learn motifs consisting of both primary-sequence and RNA structural motifs. In chapter 6, real applications of this new module are demonstrated and its strength and weakness are discussed.

Chapter 2. Constraints from comparative genomics for ncRNA finding

Among various approaches for *ab initio* ncRNA finding, comparative algorithms have been claimed to have good performance in identifying structural ncRNAs in test data sets (Rivas and Eddy 2001; di Bernardo et al. 2003; Coventry et al. 2004; Washietl et al. 2005; Pedersen et al. 2006) and simple genomes, such as bacteria and yeasts (Rivas et al. 2001). One algorithm, RNAz, was also claimed to perform well in identifying structural ncRNAs in mammalian genomes (Washietl et al. 2005). One requirement for using these comparative algorithms is that the input data must be sequence alignments.

Recently, some of these comparative algorithms have been applied to finding ncRNAs in vertebrate genomes (Washietl et al. 2005; Pedersen et al. 2006), where the alignments used for prediction were mainly derived from syntenic regions of multiple vertebrate genomes. In this thesis, such type of alignments is referred to as synteny alignments. However, the properties of synteny alignments that may contain ncRNAs are not necessarily comparable to the test data sets used to assess these comparative algorithms. This makes it uncertain whether these algorithms will have the same performance in finding ncRNAs, when synteny alignments are used.

For convenience, some terms are defined here. “Synteny-conserved ncRNAs” is used to indicate ncRNAs, in one organism, that are conserved in the corresponding syntenic regions of other genomes; if an ncRNA is not synteny-conserved, it is referred to as “synteny-non-conserved”; “synteny-conservation ratio” of ncRNAs refers to the ratio of one organism’s ncRNAs that are “synteny-conserved ncRNAs” to the total number.

There are several considerations when using synteny alignments as the target for

genome-wide ncRNA finding. Firstly, if many functional ncRNAs are synteny-non-conserved in the genomes under investigation, finding ncRNAs using only synteny alignments would risk missing a significant number of ncRNAs. To date, the synteny-conservation ratio of different classes of ncRNAs in vertebrate genomes has not been comprehensively surveyed. One obstacle in carrying out such a survey is that classic ncRNAs, which are frequently related to repetitive elements in vertebrate genomes, have generally been removed before building synteny data sets (Schwartz et al. 2003; Frazer et al. 2004; Siepel et al. 2005).

Secondly, if orthologous ncRNAs in the genomes under investigation are so conserved that only a few covariations are found, it may be difficult to determine whether the sequence conservation means the existence of RNA high-order structures or simply of primary-sequence motifs. The number of covariations in alignments of the orthologous ncRNAs may be expected to be greater for more distantly related organisms. This is why the sequence identity of a primary-sequence alignment is usually required to be within certain ranges for comparative ncRNA finding algorithms. For instance, the desired ranges of sequence identity for running QRNA and ddbRNA are 65%-85% (Rivas and Eddy 2001) and 60%-80% (di Bernardo et al. 2003), respectively. Likewise, RNAz implicitly requires that the sequences of orthologous ncRNAs are divergent to a certain extent, because the false positive rate of RNAz was reported to increase when alignments of high identities were used (Washietl et al. 2005). However, so far, no systematic survey has been performed to estimate the abundance of covariations in the orthologous ncRNAs in vertebrate genomes.

This chapter is therefore dedicated to investigating the conservation patterns of ncRNAs in vertebrate genomes, especially in mammalian genomes. A detailed survey of the conservation patterns of both classic (such as tRNAs, rRNAs, and snRNAs) and non-classic (such as miRNAs, snoRNAs, *etc*) ncRNAs in mammalian genomes was performed, in order to provide a solid basis for using the mammalian synteny alignments in genome-wide ncRNA

finding. The conservation patterns explored in this chapter include:

- The synteny-conservation ratios of ncRNAs.
- The abundance of covariations between orthologous ncRNAs.

In the first section of this chapter (section 2.1), a protein-coding gene based strategy for locating the respective syntenic regions of individual human ncRNAs was used. The conservation patterns of multiple classes of human ncRNAs in these human-mouse syntenic regions were then investigated. The synteny-conservation ratios, as well as the abundance of covariations, of the ncRNAs in the human genome with respect to the mouse genome were then calculated. A survey of the abundance of covariations was also performed on the human-mouse synteny-conserved ncRNAs with respect to their best homologues in the zebrafish genome. Based on this data, the possible effects of using real genomic alignments of ncRNAs on the performance of several comparative ncRNA finding algorithms was explored.

One caveat with respect to the syntenic-region locating strategy used in the first section of this chapter is the ignorance of gene-order conservation of ncRNAs. This means that, if there are local changes of the ncRNA copy numbers and/or of the ncRNA gene order within syntenic regions, these will be missed. Since the changes caused by evolutionary events may help explain the observed synteny-conservation ratios of ncRNAs, gene-order conservation is of interest.

In section 2.2, I examined the conservation/change of the physical arrangements of tRNA gene loci in mammalian genomes. This study is intended to explore if the pattern of gene-order conservation may give any insight into the origin of the substantial number of synteny-non-conserved ncRNAs observed in mammalian genomes. In particular, the gene-order conservation of clustered tRNA gene loci in mammalian genome is of interest. This idea was motivated from the observations of many clustered ncRNAs in diverse genomes,

from virus (Wilson et al. 1972), bacteria (Fournier et al. 1974), yeast (Beckmann et al. 1977), to primates (Chang et al. 1986). For instance, a tRNA gene cluster consisting of ~150 tRNA gene loci were found on human chromosome 6 (Mungall et al. 2003). The specific issues I intend to address in section 2.2 are as follows:

- Are there synteny-conserved clusters of tRNA gene loci?
- Are there many gene-order changes in the syntenic tRNA gene clusters?

This study is useful to genome-wide ncRNA finding in several ways. First, it may provide a high-resolution view on how tRNA genes have evolved in mammalian genomes, and may therefore give insights on how alignments should be generated for the purpose of genome-wide ncRNA finding. Second, this study may potentially be useful for distinguishing the tRNA gene loci that are functional, from those that have become pseudogenes. Although the rules derived from the case of mammalian tRNA genes may not necessarily be valid for the cases of other classes of ncRNA genes, this study may provide an independent piece of evidence, which is not biased toward protein genes, to the evolution of mammalian genomes.

2.1. The conservation patterns of vertebrate ncRNAs

2.1.1. Materials and Methods

2.1.1.1. Recruiting human ncRNAs

The genomic loci of human tRNAs were retrieved from Ensembl release 29. Ensembl is a software system that aims to provide a comprehensive annotation of selective eukaryotic genomes (Birney et al. 2006). Different releases of Ensembl may use different versions of genome assemblies. The human genome assembly that is used in Ensembl release 29 is NCBI 35, which was released by NCBI in April 2004. (http://www.ncbi.nlm.nih.gov/genome/guide/human/release_notes.html)

The genomic loci of human tRNAs in Ensembl are annotated using tRNAscanSE, which is a tRNA finding pipeline that integrates several tRNA finding algorithms (Lowe and Eddy 1997). The algorithms used by tRNAscanSE include tRNAscan (Fichant and Burks 1991), eufindtRNA (Pavesi et al. 1994), covels (Eddy and Durbin 1994), and coves (Eddy and Durbin 1994). tRNAscan is a hierarchical and rule-based system to identify intragenic promoters and consensus secondary structures of tRNAs. eufindtRNA was designed to find intragenic promoters of tRNAs. Covels is a search algorithm that uses a covariance model (CM) (see subsection 1.3.3.2.) to detect both primary-sequence and secondary-structure motifs with high specificity in genomes, although it is very slow. In the tRNAscanSE pipeline, both the outputs of tRNAscan and eufindtRNA are combined into one set of candidate tRNA genes, which are further assessed by covels in order to remove false positives. The criterion for deciding true positives is the degree of conservation at both primary-sequence and secondary-structure levels (Lowe and Eddy 1997). The final structural alignments are generated by coves. In Ensembl release 29, there are 498 tRNA genes in the human genome, after excluding pseudogenes and the tRNAs with undetermined codon types.

Other human ncRNAs were retrieved from Rfam 6.1 (Griffiths-Jones et al. 2005). Rfam is a database of curated sequence alignments and CMs of different classes of ncRNAs. The CMs created by Rfam are also used to search for novel ncRNAs in the EMBL nucleotide sequence database (Kanz et al. 2005), which includes sequences of the human genome and the mouse genome. The sequences and the ncRNAs so predicted are also deposited in Rfam. Infernal (a system for “INFERence of RNA ALignment”, http://infernal.janelia.org/) is the software package used by Rfam to build CMs and to find ncRNA-like sequences in the sequence database (Griffiths-Jones et al. 2005).

The coordinates of Rfam ncRNAs in the human genomic contigs were retrieved from Rfam.full, which was downloaded from the Rfam ftp site (ftp://ftp.sanger.ac.uk/pub/databases/

Rfam/). The coordinates were converted to human chromosomal coordinates using software libraries provided by the Ensembl Project written in the Perl programming language referred to as Application Programming Interfaces (APIs). Although there have been newer releases of Ensembl since the analyses in this thesis were performed, NCBI 35 has continued to be used by a number of later releases of Ensembl (releases 30 ~ 36). This procedure of mapping ncRNAs to the human genome is exactly the same as that used for generating the ncRNA annotation of Ensembl releases 30 ~ 36.

2.1.1.2. Searching for human-mouse synteny-conserved ncRNAs

The alignments of human-mouse syntenic regions were retrieved from Ensembl Compara release 29 (Clamp et al. 2003) using the Ensembl Compara Perl APIs. The Ensembl Compara database is the component of Ensembl that contains comparative genomic information, including predictions of orthology relationships between protein-coding genes and synteny alignments among different genomes. The genome assemblies used by Ensembl Compara release 29 include human NCBI 35 and mouse NCBI M33 (http://www.ncbi.nlm.nih.gov/genome/seq/NCBIContigInfo.html).

The existence of synteny-conserved ncRNAs in candidate alignments was searched using cmsearch and Rfam CMs. cmsearch is a program of the Infernal package that can use a Rfam CM trained using a particular type of ncRNAs to search for new occurrences of ncRNAs of the same type. Given a sequence, cmsearch can align it to a Rfam CM and return high scoring matches. cmsearch reports matches with bit scores (for more details about bit scores see subsection 1.3.3.1). The regions with bit scores higher than corresponding family-specific thresholds pre-determined by Rfam (Griffiths-Jones et al. 2003) were considered to be ncRNA loci.

In order to correctly include classic ncRNAs in genomic regions that are missing from available resources of genome-wide alignments, an approach was adopted which takes

advantage of the syntenic regions defined by human-mouse orthologous protein-coding genes. This approach allows the identification of missing synteny-conserved ncRNAs in initially unaligned syntenic regions. The basic idea is that, if the relation of a particular ncRNA to its 5' and 3' flanking protein-coding genes has been preserved in evolution, a synteny-conserved ncRNA may also be found in the corresponding syntenic region defined by synteny-conserved protein-coding genes in the other genome (Figure 2-1, a).

One issue when using this strategy to find the synteny-conserved ncRNAs is the ambiguity in assigning orthology to protein-coding genes retrieved from different genomes. For instance, ambiguity can occur whenever multiple protein-coding genes, which are paralogous to each other in one organism, appear orthologous to a particular gene in the other organism. Such many-to-one or even many-to-many relationships between protein-coding genes may cause difficulties in determining unique human-mouse syntenic regions for individual human ncRNAs. In order to control the complexity of finding the appropriate syntenic regions, best reciprocal protein homologs (UBRHs), where there is only one uniquely best hit in both directions between two genomes, were used in the following analyses. Each pair of UBRHs (UBRHP) consists of two homologous members from the human and mouse genomes, respectively. All UBRHPs between these two genomes were retrieved from Ensembl Compara release 29. The 5' and 3' flanking protein-coding genes nearest to a particular human ncRNA, which are also the members of two consecutive UBRHPs, were used to define the boundaries of the corresponding mouse syntenic region (Figure 2-1, a).

Synteny-conserved counterparts of human ncRNAs in the mouse (UBRHPs-bound) syntenic regions were obtained by using WU-BLAST alignment algorithm to scan the UBRHP-bound mouse genome sequence with the human ncRNA sequence. The threshold used for filtering alignment hits was set to be at least 40% identity. Certainly, the cost of this heuristic is an inevitable decrease in sensitivity; however hits with low percent identities (<

50%) are also unsuitable for using existing algorithms for *ab initio* ncRNA finding. The existence of synteny-conserved ncRNAs was further verified using Infernal and Rfam CMs. Human ncRNAs that were found to be conserved in the syntenic regions were labelled as "synteny-conserved ncRNAs"; otherwise they were labelled as "synteny-non-conserved ncRNAs". It should be noted that the set of UBRHPs, and accordingly, UBRHPs-bound syntenic regions, can change between releases of Ensembl, even if exactly the same genome assemblies were used. Such changes result from improvements in the annotations of protein-coding genes in Ensembl. However, the annotation of genes in the mouse genome (NCBI M33) was constant through Ensembl releases 29 ~ 31, so there were essentially no major changes in the set of UBRHPs-bound syntenic regions in the Ensembl Compara database of these Ensembl releases.

Several complicated situations could be encountered when using the UBRHPs based approach to find synteny-conserved ncRNAs: 1) ncRNAs at either end of chromosomes may not be flanked by members of UBRHPs (Figure 2-1, b); 2) the members of two consecutive UBRHPs may be partitioned into two different chromosomes (Figure 2-1, c); 3) the relationships of UBRHPs-bound blocks between two genomes may be inconsistent due to some unknown evolutionary events (Figure 2-1, d). Each of these three situations makes the search process more difficult, and might thus cause false negatives in determining synteny-conserved ncRNAs.

In order to reduce the false negatives caused by the first and the second situations, either the 5' or the 3' member of the flanking UBRHP of a particular ncRNA was used as the anchoring point to extend the candidate sequence blocks for searching for a synteny-conserved ncRNA in the second genome. The cases of the second situation are marked as "inter-chromosomal translocation" (Figure 2-1, c).

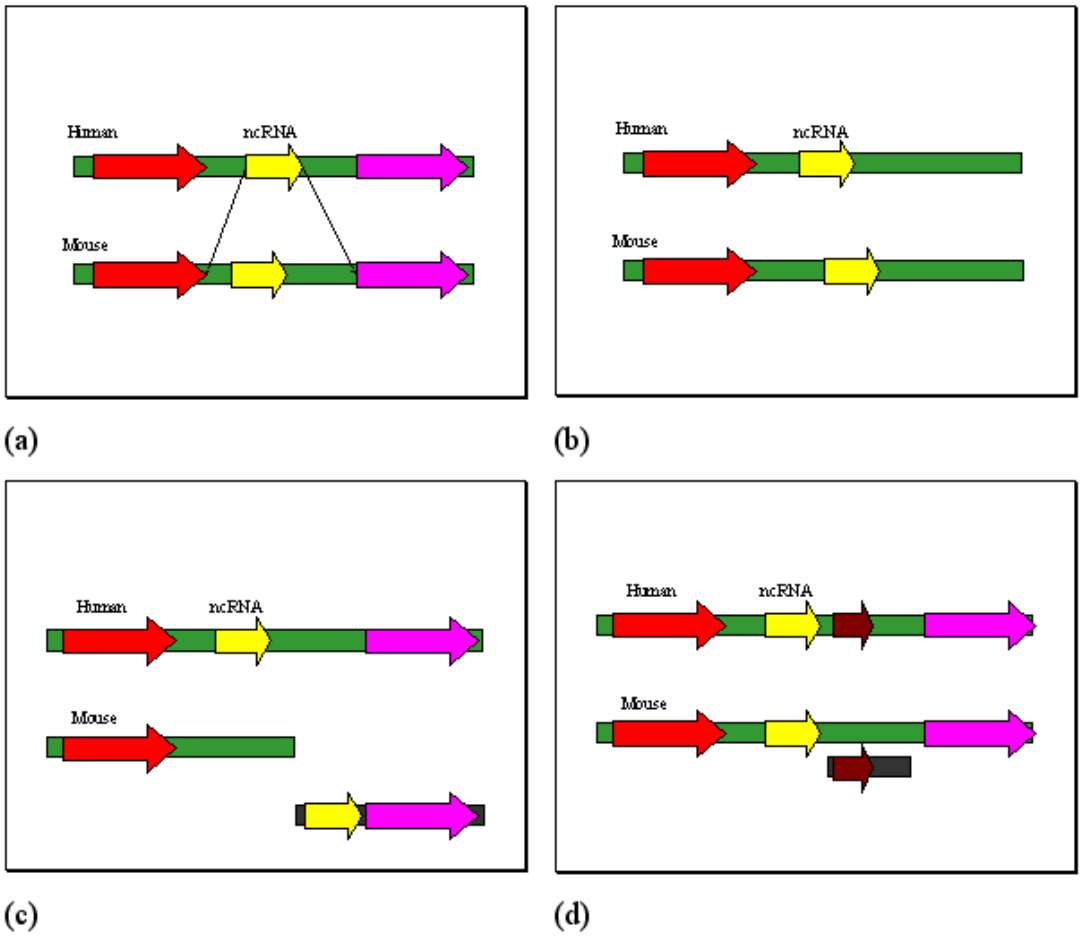


Figure 2-1. Physical relations of human and mouse synteny-conserved ncRNAs to UBRHPs-bound syntenic regions

Red arrows: one pair of unique best reciprocal protein homologues (UBRHP) in the 5' flanking region of one ncRNA. Magenta arrows: one UBRHP in the 3' flanking region of one ncRNA. Yellow arrows: synteny-conserved ncRNAs. (a) The mouse members of two consecutive UBRHPs are on the same chromosome. (b) ncRNAs that are near the ends of chromosomes are flanked by only one UBRHPs (either in the 5' or in the 3' flanking region). (c) The mouse members of two consecutive UBRHPs are separated into two chromosomes. (d) The relationship of UBRHPs-bound blocks becomes incompatible between two genomes due to unknown evolutionary events.

For the third situation, however, it is unknown how to determine the real evolutionary event leading to the finding of pairs of protein-coding genes that are out of order (Figure 2-1, d, the brown arrows). It is possible that, in these regions, there might have been inter-chromosomal rearrangements, pseudogenisations of duplicated genes, *etc*. Consequently, it is difficult to define a clear rule to avoid possible false negatives in such complicated cases. To partially address this problem, one additional measure was adopted. In recruiting two consecutive UBRHPs to define a suitable syntenic block for one ncRNA, next adjacent

UBRHPs was tried (Figure 2-1, d, the magenta arrows) if the initial UBRHPs was not on the same chromosome as their 5' and 3' flanking UBRHPs (Figure 2-1, d, compare the red and brown arrows). These regions were marked as "complicated regions".

In addition, I also considered cases where there might be segmental inversions in the UBRHPs-bound syntenic regions. I took the incompatibility of the strand combinations of the UBRHPs in different genomes as an indicator of segmental inversions. The argument is that, when there are no segmental inversions, the strand combination of the respective members from 5' and 3' UBRHPs in the first genome should be consistent with the strand combination in the second genome.

2.1.1.3. Determination of covariations between orthologous ncRNAs

To determine covariations between orthologous ncRNAs, cmalign was used. cmalign is a program of the Infernal package that can simultaneously align multiple sequences to a Rfam CM corresponding to a particular type of ncRNAs. Given a set of ncRNAs of the same type, cmalign returns an alignment augmented with secondary-structure annotation, as shown in Figure 2-2. Such an output was then processed to determine types of mismatches, which can either be covariations or just unpaired changes, between stem regions of orthologous ncRNAs.

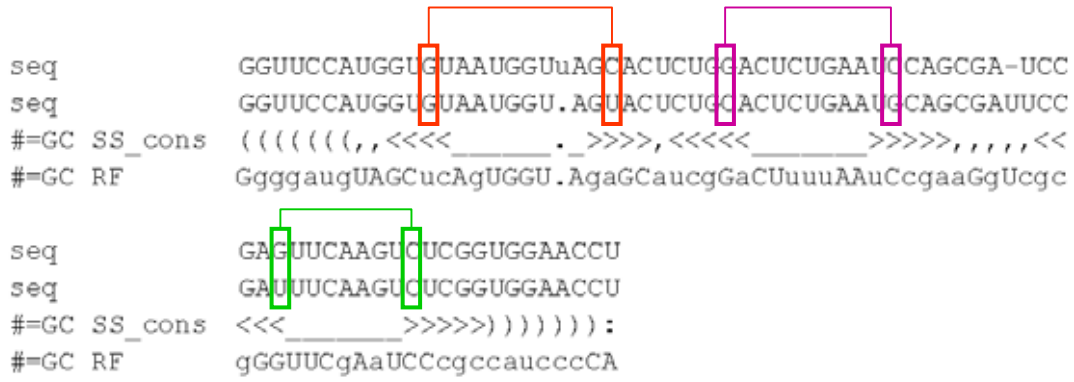


Figure 2-2. A multi-sequence secondary-structure alignment generated by cmalign

Mismatches in double-stranded regions were further categorized into three subtypes. An incomplete covariation is a case where only one base was changed at a base-paired position, such that a conversion occurs between a non-canonical pairing (G-U) and a canonical pairing (G-C or A-U) (*e.g.* red boxes in Figure 2-2). A complete covariation is a case where paired bases were simultaneously mutated to other types of valid pairing, such as G-C to C-G (*e.g.* magenta boxes in Figure 2-2), A-U, U-G, or U-A. A base change that results in a non-canonical and non G-U pairing is referred to as an unpaired change (*e.g.* green boxes in Figure 2-2).

The reason for separating incomplete covariations from complete covariations is that the former type of covariation is a weaker signal for indicating the existence of secondary structures than the latter type. For instance, when the information of covariations is calculated using the standard mutual information (MI) measure (Chiu and Kolodziejczak 1991; Gutell et al. 1992), covariations consisting only of GC and GU pairings do not contribute. However, incomplete covariations still provide useful information for RNA secondary structure prediction (Hofacker et al. 2002; Lindgreen et al. 2006), and should be included in covariation analysis. Thus, in this thesis, the numbers of incomplete covariations and complete covariations were counted separately.

2.1.2. Evaluating different approaches for finding human-mouse synteny-conserved ncRNAs

2.1.2.1. Using the synteny alignments retrieved from public-domain resources

By using the human-mouse syntenic regions that were retrieved from Ensembl Compara release 19, only 26.7% (133/498) of human tRNA genes predicted by tRNAscanSE were found to have synteny-conserved counterparts in the mouse genome (NCBI M30). By using the later releases of the Ensembl Compara database (19-31) where different assemblies of

human (NCBI 35) and mouse (NCBI M32 and NCBI M33) genomes were used, even fewer synteny-conserved tRNA genes could be found. The differences caused by using different Ensembl Compara database releases were due to the changes of strategies for building synteny used by Ensembl. One reason for these changes was to avoid Ensembl Compara containing alignment artefacts caused by repetitive elements. These results show that using existing resources for comparative genomics cannot be relied upon to give a correct estimate of the synteny-conservation ratios of classic ncRNAs between mammalian genomes.

Fortunately, a useful insight was gained from the investigation of tRNA gene clusters in mammalian genomes. A relevant finding is the identification of multiple human-mouse synteny-conserved tRNA gene clusters (for details see section 2.2). As many as ~68% (338/498) of human tRNA genes predicted by tRNAscanSE were found to be in the human-mouse synteny-conserved tRNA gene clusters, although some of their respective synteny-conserved counterparts in the mouse genome might have been lost in evolution.

These results suggest that the real synteny-conservation ratio of human and mouse tRNA genes is much higher than the highest number (26.7%) derived from syntenic alignments retrieved from the Ensembl Compara database alone. Using other public-domain resources of comparative genomics would be unlikely to make much difference, because the algorithms used for creating syntenic alignments in the different releases of the Ensembl Compara database have also been used by these other resources (Schwartz et al. 2003; Frazer et al. 2004). I concluded that the synteny alignments provided by public-domain databases were inadequate for the purpose of generating a comprehensive set of human-mouse synteny-conserved ncRNAs.

2.1.2.2. Using the UBRHPs-bound syntenic regions

Using the UBRHPs-based approach, 74.5% (371/498) of the human tRNA genes that are predicted by tRNAscanSE were found to be conserved in the mouse syntenic regions. These

results suggest that, for finding the human-mouse syntenic regions of classic ncRNAs, the UBRHPs-based approach is likely to be much more effective than using the syntenic regions retrieved from public-domain resources (such as the Ensembl Compara release 29) of comparative genomics.

2.1.3. Results

2.1.3.1. The synteny-conservation ratios of human ncRNAs from Rfam

Since the UBRHPs-bound syntenic regions strategy for finding human-mouse synteny-conserved tRNA genes proved successful, it was further used to identify other human-mouse synteny-conserved ncRNAs. 4,201 unique human ncRNA genomic loci were recruited from Rfam 6.1 for analysing their patterns of conservation in human-mouse syntenic regions. These ncRNAs correspond to 157 classes of ncRNAs (41% of 379 classes of ncRNAs in Rfam 6.1).

Analysing the patterns of conservation of these ncRNAs in human-mouse syntenic regions revealed that the synteny-conservation ratios vary greatly among the different classes. For example, 73.6% of human miRNAs were found to be synteny-conserved; however, only 1.1% of miscellaneous ncRNAs were synteny-conserved (Table 2-1). Overall, 78.1% of the human ncRNAs identified by Rfam6.1 were not found to be conserved in the corresponding mouse syntenic regions. The overall initial estimated synteny-conservation ratio for human ncRNAs is only 21.9%.

In order to evaluate whether the calculated synteny-conservation ratios of human and mouse ncRNAs might be affected by the quality of the mouse genome assembly, the assembly status for the UBRHPs-bound syntenic region corresponding to each human ncRNA was determined. 63.8% of the mouse UBRHPs-bound syntenic regions, where the synteny-non-conserved ncRNAs are supposed to reside, were found to contain genome

sequence fragments labelled either unfinished regions (UR) or whole genome shotgun (WGS) (Table 2-2). It was found that in these UR- or WGS-containing regions there were more synteny-non-conserved ncRNAs than synteny-conserved ncRNAs (compare Table 2-3 with Table 2-2). On average, 63.8% of the synteny-non-conserved ncRNAs and 59.8% of the synteny-conserved ncRNAs are in mouse UR-WGS-containing syntenic regions. The P-value (*Chi*-square test) is far less than 0.001. This result suggests that there is an association between the inability to detect synteny-conserved ncRNAs and the quality of the mouse genome assembly. Consequently, the synteny-conservation ratio for the human ncRNAs that were retrieved from Rfam should be higher than ~22%, because some synteny-conserved ncRNAs will have been missed in mouse UR-WGA regions.

class	mapped to NCBI 35	synteny-conserved	synteny-non-conserved
IRES	8	3 (37.5%)	5 (62.5%)
ribozyme	3	2 (66.7%)	1 (33.3%)
miRNA	87	64 (73.6%)	23 (26.4%)
snoRNA	390	199 (51.0%)	191 (49.0%)
cis-reg	194	96 (49.5%)	98 (50.5%)
tRNA	842	370 (43.9%)	472 (56.1%)
rRNA	350	13 (3.7%)	337 (96.3%)
misc ncRNA	924	10 (1.1%)	914 (98.9%)
snRNA	1403	163 (11.6%)	1240 (88.4%)
Total	4201	920 (21.9%)	3281 (78.1%)

Table 2-1. Conservation of different classes of Rfam human ncRNAs in human-mouse syntenic regions

"IRES" consists of IRES_Bag1, IRES_Bip, IRES_c-myc, IRES_FGF, IRES_L-myc, and IRES_n-myc. "ribozyme" consists of RNaseP_nuc and RNase_MRP. "rRNA" includes 5S_rRNA, 5_8S_rRNA, and SSU_rRNA_5. "cis-reg" consists of Antizyme_FSE, CAESAR, G-CSF_SLDE, GAIT, Histone3, IFN_gamma, IRE, REN-SRE, RRE, SECIS, Spi-1, TAR, and Vimentin3. snRNA consists of U1, U2, U4, U5, U6, U7, U12, and U14. Other ncRNAs, including 7SK, S15, SRP_euk_arch, Telomerase-vert, Vault, and Y., are grouped into "misc ncRNA" (miscellaneous ncRNA).

class	synteny-non-conserved in mouse finished contigs	synteny-non-conserved in mouse UR or WGS
IRES	3 (60.0%)	2 (40%)
ribozyme	0 (0.0%)	1 (100%)
miRNA	6 (26.1%)	17 (73.9%)
snoRNA	61 (31.9%)	130 (68.1%)
cis-reg	37 (37.8%)	61 (62.2%)
tRNA	167 (35.4%)	305 (64.6%)
rRNA	104 (30.9%)	233 (69.1%)
misc ncRNA	346 (37.9%)	568 (62.1%)
snRNA	464 (37.4%)	776 (62.6%)
Total	1188 (36.2%)	2093 (63.8%)

Table 2-2. Distribution of the human synteny-non-conserved ncRNAs in the regions corresponding to mouse finished contigs or UR-WGS-containing regions (regions with unfinished gaps in contig-base sequencing and regions from whole genome shotgun sequencing)

class	synteny-conserved in mouse finished contigs	synteny-conserved in mouse UR or WGS
IRES	2 (66.7%)	1 (33.3%)
ribozyme	1 (50%)	1 (50%)
miRNA	29 (45.3%)	35 (54.7%)
snoRNA	70 (35.2%)	129 (64.8%)
cis-reg	66 (68.8%)	30 (31.3%)
rRNA	6 (46.2%)	7 (53.8%)
tRNA	165 (44.6%)	205 (55.4%)
misc ncRNA	0 (0%)	10 (100%)
snRNA	31 (19%)	132 (81%)
Total	370 (40.2%)	550 (59.8%)

Table 2-3. Distribution of human synteny-conserved ncRNAs in the regions corresponding to mouse finished contigs or UR-WGS-containing regions (regions with unfinished gaps in contig-base sequencing and regions from whole genome shotgun sequencing)

These results show that human ncRNAs are more likely to be synteny conserved in mouse syntenic regions containing only mouse finished contig based sequence (FCS) than in regions that are unfinished (UR) or whole genome shotgun (WGS), but that the effect is small. The average synteny-conservation ratio only increases from ~22% (920/4201) to ~24% (370/1558) when only FCS is considered (see the statistics in the context of mouse finished

contigs in Table 2-2 and Table 2-3). There is a much bigger variation of synteny-conservation ratio between categories. When ncRNAs are considered by category, an inverse correlation was found between the average copy numbers and the synteny-conservation ratios (Figure 2-3).

The previous comparison considers the effect of sequence quality on the apparent ncRNA synteny-conservation ratio. Another factor is assembly completeness. Among the ncRNAs that were investigated, surprisingly low synteny-conservation ratios were found between human and mouse 5S rRNA genes (5S rDNAs). One concern is that the mouse genome assembly (NCBI M33) may have missed *bona fide* 5S rDNAs. Prior to the large-scale sequencing of the human and the mouse genomes, 5S rDNAs were known to be exist as tandem repeats in both genomes (Little and Braaten 1989; Suzuki et al. 1994). It is possible that the strategy of whole genome shotgun sequencing may lead to the omission of tandem repeats, such as 5S rDNAs.

In order to clarify if there are tandemly arranged 5S rDNAs in the mouse genome assembly used in this chapter, a reliable mouse 5S rDNA (GenBank accession number: X71804) was used to search for all 5S rDNAs in NCBI M33. This mouse 5S rDNA sequence, which was published before any large-scale genome sequencing projects were finished, is one unit of the 5S rDNA tandem repeats in the mouse genome (Hallenberg et al. 1994). The result indicates that no such tandem repeats can be found in NCBI M33, while the 5S rDNA tandem repeats can be found in the human genome assembly NCBI 35. In addition, this mouse 5S rDNA is perfectly identical (100%) to the human 5S rDNA. Consequently, the evidence does not suggest that functional 5S rDNAs become synteny-non-conserved after the primate-rodent split. The apparent low synteny-conservation ratio of human and mouse 5S rDNAs is most likely an artefact caused by the missing of *bona fide* 5S rDNAs in NCBI M33.

During the preparation of this thesis, a new mouse genome assembly NCBI M36 is available and the 5S rDNA tandem repeats can be found in this genome assembly. This result

suggests that the quality of the mouse genome assembly has been improved since the release of NCBI M33. NCBI M36 may be a suitable genome assembly for re-estimating the synteny-conservation ratios of human and mouse ncRNAs.

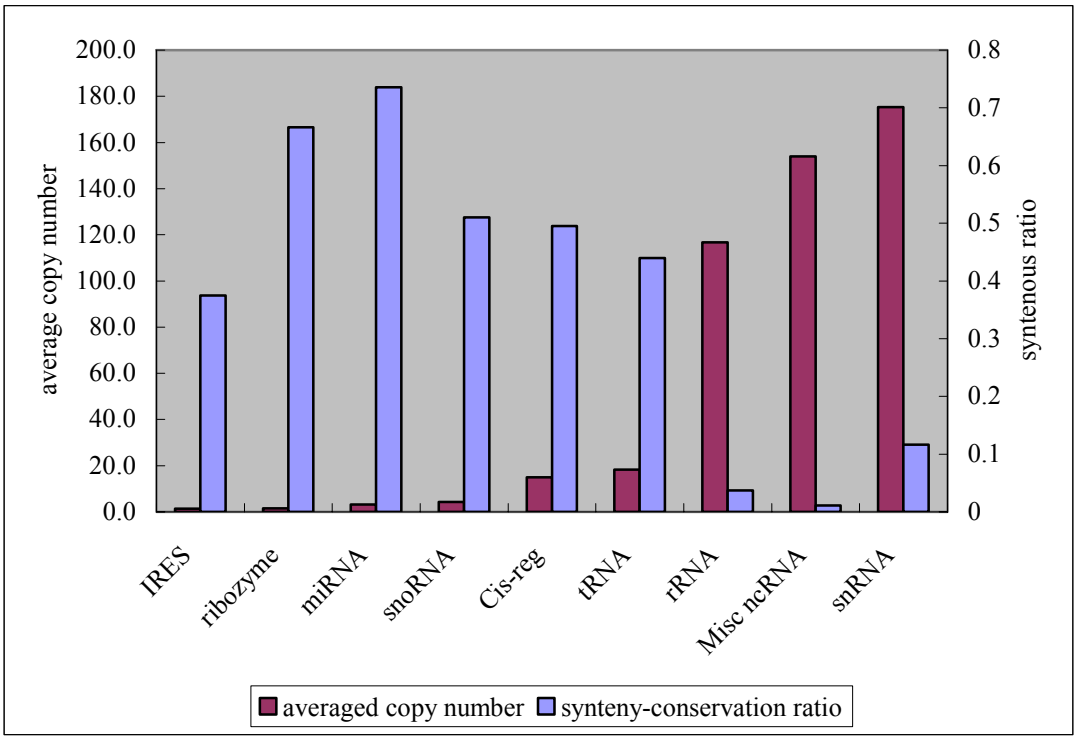


Figure 2-3. Synteny-conservation ratios and average copy numbers for different categories of human ncRNAs (mapped by Rfam)

2.1.3.2. Effect of genome rearrangements on synteny conservation

In order to assess any relationship between genome rearrangements and the estimated synteny-conservation ratios of ncRNAs, chromosome-compatibility and strand-compatibility were taken as the indicators of inter-chromosomal rearrangement and intra-chromosomal rearrangement (for the method see subsection 2.1.1.2.). In the cases where the gene orders and the strand-relationship of the ncRNAs and their flanking genes have been conserved, the syntenic regions were assigned as evolutionary-intact regions.

From this analysis, the syntenic blocks between human and mouse were categorised as

intact regions, segmental inversions, inter-chromosomal translocations, and 'complicated' regions, *i.e.* where evolutionary processes are unclear (Figure 2-1, d). The number of synteny-conserved and synteny-non-conserved ncRNAs in each of these regions is listed in Table 2-4. The synteny-conservation ratio of ncRNAs in the syntenic blocks with inter-chromosomal translocations is not significantly different from that in the intact syntenic blocks (*Chi*-square test, P-value >> 0.1). The synteny-conservation ratio of ncRNAs in the syntenic blocks of the complicated type appears significantly lower than that in the intact syntenic blocks (*Chi*-square test, P-value << 0.001), however this could be an artefact where some synteny-conserved ncRNAs were missed in these regions due to difficulties with the UBRHPs-based method in such regions. It is possible that the method used in this chapter to find synteny-conserved ncRNAs was vulnerable to certain types of genome rearrangements. For instance, if an event of genome rearrangement has changed the linear order of a ncRNA with respect to its flanking synteny landmarks (*i.e.* the protein-coding genes that can be used to define syntenic blocks), this ncRNA may be mistakenly classified as a synteny-non-conserved one. It can be inferred that the calculated synteny-conservation ratios of ncRNAs might be underestimated due to genome rearrangements in "complicated" regions.

The synteny-conservation ratio of ncRNAs in the syntenic blocks with segmental inversions, which are a type of intra-chromosomal rearrangements, is much higher than that in the intact syntenic blocks (*Chi*-square test, P-value << 0.001). No obvious explanation could be found to explain this surprising observation, however such an affect has been reported before. Inversions were found to reduce recombination dramatically (for review see Hoffmann et al. 2004).

synteny conditions	synteny-conserved	synteny-non-conserved	subtotal
evolutionary-intact	579 (24%)	1800 (76%)	2379
segmental inversion	131 (48%)	141 (52%)	272
inter-chromosomal translocation	51 (24%)	163 (76%)	214
complicated	153 (11%)	1183 (89%)	1336

Table 2-4. Numbers of the human-mouse synteny-conserved and the synteny-non-conserved ncRNAs in regions which have undergone different evolutionary events

2.1.3.3. Few covariations in human-mouse synteny-conserved ncRNAs

The aligned sequences of the set of human-mouse synteny-conserved ncRNAs were assessed for covariations as previously defined (see subsection 2.1.1.3.). 64% of human-mouse synteny-conserved tRNAs and 54% of human-mouse orthologous snRNAs were found to not contain any covariations. In addition, no covariations could be found in 70% of human-mouse synteny-conserved miRNAs and in 51% of human-mouse synteny-conserved snoRNAs. Since incomplete covariations are weaker signals than complete ones (see subsection 2.1.1.3.), the cases with only one incomplete covariation were combined with exactly conserved ones (*i.e.* these with no mutations in stem regions), as shown in columns "0-1" base involved in covariations in the following tables (see Table 2-5 and Table 2-7).

On average, 73% of human-mouse synteny-conserved ncRNAs do not provide useful number of covariations (Table 2-5). These results suggest that the alignments of human-mouse synteny-conserved ncRNAs do not contain sufficient covariations for ncRNA finding. Even though the average identity of human-mouse synteny-conserved ncRNAs is 86%, which is only slightly greater than the upper limit of identities requested by some algorithms (*i.e.* ddbRNA and QRNA), covariations are not enriched in the mismatches between the members of each orthologous ncRNA pair. Much of the primary-sequence difference between human-mouse synteny-conserved ncRNAs is attributed to mutations that were found in the

single-stranded regions, and to mutations that may destabilize the stem regions.

Bases in covariations	0-1	2-10	11-23	Subtotal
cis-reg	83 (86%)	13 (14%)	0 (0%)	96 (100%)
misc ncRNA	5 (50%)	4 (40%)	1 (10%)	10 (100%)
IRES	0 (0%)	2 (67%)	1 (33%)	3 (100%)
miRNA	54 (84%)	10 (16%)	0 (0%)	64 (100%)
ribozymes	0 (0%)	2 (100%)	0 (0%)	2 (100%)
rRNA	0 (0%)	12 (92%)	1 (8%)	13 (100%)
snoRNA	139 (70%)	60 (30%)	0 (0%)	199 (100%)
snRNA	110 (67%)	41 (25%)	12 (7%)	163 (100%)
tRNA	282 (76%)	74 (20%)	14 (4%)	370 (100%)
Subtotal	673 (73%)	218 (24%)	29 (3%)	920 (100%)

Table 2-5. Numbers of the human-mouse synteny-conserved ncRNAs that contain various numbers of covariations

	Human-mouse	Human-zebrafish
cis-reg	0.6 (96)	0 (1)
misc ncRNA	3.6 (10)	33.0 (2)
IRES	7.7 (3)	N/A
miRNA	0.7 (64)	3.2 (20)
ribozyme	5.5 (2)	N/A
rRNA	6.2 (13)	9.0 (4)
snoRNA	1.2 (199)	3.5 (2)
snRNA	2.2 (163)	2.1 (79)
tRNA	1.4 (370)	1.1 (185)

Table 2-6. Average numbers of bases involved in covariations per sequence of the human-mouse synteny-conserved ncRNAs and of the human-zebrafish orthologous ncRNAs

N/A: no synteny-conserved ncRNAs found. Each parenthesized value is the number of sequences for respective category of ncRNAs.

Bases in covariations	0-1	2-10	11-33	Subtotal
cis-reg	1 (100%)	0 (0%)	0 (0%)	1 (100%)
Misc ncRNA	0 (0%)	0 (0%)	2 (100%)	2 (100%)
miRNA	7 (35%)	12 (60%)	1 (5%)	20 (100%)
rRNA	0 (0%)	3 (75%)	1 (25%)	4 (100%)
snoRNA	1 (50%)	1 (50%)	0(0%)	2 (100%)
snRNA	51 (64.6%)	25 (31.6%)	3 (3.8%)	79 (100%)
tRNA	133 (71.9%)	52 (28.1%)	0 (0%)	185 (100%)
Subtotal	193 (65.9%)	93 (31.7%)	7 (2.4%)	293 (100%)

Table 2-7. Numbers of the human-mouse-zebrafish orthologous ncRNAs that contain various numbers of covariations

2.1.3.4. Only a few covariations in the human-zebrafish best-fit ncRNAs

From the conclusion that there are insufficient covariations between human and mouse synteny-conserved ncRNAs (for details see subsection 2.1.3.3.), it is reasonable to infer that successful detection of ncRNAs through using comparative ncRNA finding approaches may require more distantly related species than human and mouse. Zebrafish was therefore used in order to investigate if comparing the human genome with other vertebrate genomes can provide significantly more covariations for the purpose of ncRNA finding.

Initially, the zebrafish ncRNAs that are synteny-conserved to human-mouse synteny-conserved ncRNAs were searched in the human-zebrafish UBRPHs-bound syntenic regions; however, only 110 out of 920 human-mouse synteny-conserved ncRNAs could be matched to 58 non-redundant zebrafish ncRNAs. This is most likely due to the lost of synteny between these distantly related species.

In order to recruit more human-zebrafish orthologous ncRNAs, WU-BLAST (Gish 1996-2004) was used to perform a whole genome search for homologues for individual human-mouse synteny-conserved ncRNAs. The best hit for each ncRNA was used for further analysis. 31.8% (293/920) of 920 human-mouse synteny-conserved ncRNAs matched to 112 non-redundant zebrafish ncRNAs. Taking the number of covariations from human-mouse

synteny-conserved ncRNAs as the reference, the number of covariations was found to increase in the human-zebrafish orthologous miRNAs and snoRNAs. However, there were not significantly more covariations in the human-zebrafish orthologous tRNAs and snRNAs than in the human-mouse synteny-conserved ones (Table 2-6). In fact, there were no useful covariations in 65.9% (193/293) of the human-zebrafish orthologous ncRNAs (Table 2-7).

2.1.3.5. Using real genomic alignments to assess the performances of ncRNA finding algorithms

The credibility of existing comparative ncRNA finding algorithms generally comes from benchmarks against adopted test data sets created by aligning well-curated ncRNAs, and not the alignments of ncRNA-containing genomic sequences. For example, one of the popular data sets is the alignments of ncRNAs retrieved from Rfam. These Rfam ncRNAs are different from real genomic sequences in that their 5' and 3' flanking sequences have been carefully trimmed. It is possible that additional noise may be introduced to complicate the detection of consensus RNA motif, if alignments of real genomic sequences, instead of Rfam seed sequences, are used.

In the following test, pairwise and three-way genomic alignments of human tRNA genes were generated to assess the performances of RNAz, QRNA, and ddbRNA. In particular, an additional 20 bases from both the 5' and 3' flanking regions of human tRNA genes were included when generating the alignments. The reason for including (2 x 20) bases is that, including longer flanking sequences to generate alignments may result in a significant drop of identities and only a few of the generated alignments may have identities within the identity range preferred by the three algorithms under test. On the other hand, including flanking sequences shorter than 20 bases may not introduce noise into alignments and the property of the generated alignments is still similar to that of the alignments of curated tRNAs.

One thousand pairwise alignments and one thousand three-way alignments were generated by using ClustalW 1.83. Three algorithms, RNAz, QRNA, and ddbRNA, were

tested on these alignments using their default parameters. These algorithms are ncRNA classifiers. Given a sequence alignment, they will determine whether the sequences as a whole are ncRNAs or not. The result reveals that the performances of none of these algorithms are as good as claimed in their respective papers (Table 2-8). For example, in the original paper of RNAz, the sensitivity was as high as ~95% for detecting tRNA genes by using alignments of identities within 60% ~ 100%; however, using the genomic alignments of human tRNA genes, the sensitivity is only ~49%, when pairwise alignments of identities no less than 60% are used (Table 2-8). In addition, changing the threshold of alignment identity does not improve the sensitivity of any of the algorithms.

In order to rule out the possibility that the bias of using only human tRNA genes could cause the drop in sensitivities, a positive control was performed by using the alignments of human tRNA genes without the 5' and 3' flanking regions. The sensitivity of RNAz on this positive control data set is 94% (data not shown), which is close to the published value (95%) (Washietl et al. 2005). Consequently, the incorporation of flanking regions of human tRNA genes in the test alignments is the only obvious explanation that contributes to the drop in sensitivity of these ncRNA-finding algorithms. These results clearly indicate that it is much harder to identify ncRNAs from the alignments of real genomic sequences than from the alignments of curated ncRNAs.

	RNAz (three-way)	ddbRNA (three-way)	RNAz (pairwise)	ddbRNA (pairwise)	QRNA (pairwise)
All	64.2% (642/1000)	36.2% (362/1000)	61.1% (611/1000)	36.2% (362/1000)	36.6% (366/1000)
Identities >=50%	75.7% (115/152)	57.9% (88/152)	53.8% (148/275)	42.2% (116/275)	46.5% (128/275)
Identities >=60%	75% (6/8)	37.5% (3/8)	48.8% (20/41)	31.7% (13/41)	36.6% (15/41)
Identities >=70%	NA	NA	44.4% (8/18)	5% (1/18)	27.8% (5/18)

Table 2-8. Estimating sensitivities of ncRNA-finding algorithms by using the alignments of genomic sequences of human tRNA genes

Additional 20 bases from both the 5' and 3' flanking regions of human tRNA genes are included when generating alignments of human paralogous tRNA genes. NA means in 1000 alignments, none of them have identities greater than certain thresholds as indicated in the first column of this table. In parentheses, numerators are the numbers of alignments that are correctly classified as ncRNAs. Denominators are the numbers of alignments with identities within a certain range indicated in the first column of this table.

2.1.4. Discussions

2.1.4.1. Practicality of ncRNA prediction based on comparative genomics

With the results already presented in this section (section 2.1), pairwise and three-way alignments of vertebrate genomes do not appear to be ideal data sets for ncRNA finding algorithms. Firstly, there are limited numbers of covariations between orthologous ncRNAs and high primary sequence conservation (see subsections 2.1.3.3. and 2.1.3.4.). Secondly, algorithms that take alignments as input data may be unable to properly score RNA motifs from genome alignments (see subsection 2.1.3.5.).

The difference between the performance of ncRNA finding algorithms on these data sets and their published performance is due to the different data sets used. Many comparative ncRNA finding algorithms have been trained and tested using alignments of ncRNAs, such as seed sequences used to build the Rfam CMs. These alignments are referred to as synthetic alignments in this thesis, because they are not generated directly by aligning genomic sequences. ncRNA finding algorithms perform better on synthetic alignments than genomic alignments. Also, while few, if any, covariations could be found in human-mouse syntenic ncRNAs, there were larger numbers of covariations in these synthetic alignments. One reason

for the difference is that they were generated from more distantly related organisms. A second reason is that the alignments also contained paralogous ncRNAs. Comparison reveals that paralogous ncRNAs can provide more covariations than comparison of orthologous ncRNAs. Synthetic alignments of ncRNAs from ncRNA databases (such as Rfam) may include paralogous ncRNAs. By contrast, synteny alignments should contain few, if any, paralogous ncRNAs.

Under the situation of few covariations in vertebrate ncRNA alignments, the use of multi-way alignments of more than three genomes is an alternative choice that should be considered. In a recent report, eight-way genome alignments were used for genome-wide ncRNA finding (Pedersen et al. 2006). However, several cases presented by Pedersen *et al*. demonstrated that candidate regions of ncRNAs are very well conserved and only a few putative compensatory mutations could be found. In other words, the evidence presented in Pedersen *et al*.'s report actually indicates good conservation at the primary-sequence level. These cases should therefore be considered only as good candidates for functional elements, but not necessarily good candidates for RNA structural motifs.

I therefore conclude that, although comparative ncRNA finding algorithms have been used to find ncRNA in multiple vertebrate genomes, there are still concerns with the results presented in relevant papers. Further examining the ncRNA conservation patterns in multiple vertebrate genomes may be required, in order to determine the potential of using multi-way alignments of vertebrate genomes for ncRNA finding.

It is possible that multi-way ncRNA alignments from sufficient vertebrate genomes will contain enough variations and covariations for ncRNA finding algorithms to work effectively. However, a serious issue for practical genome-wide ncRNA finding is the quality of genome alignments that must be scanned by these algorithms. Up to now, a significant proportion of existing vertebrate genome assemblies are composed of sequences generated from whole

genome shotgun sequencing (WGS). Compared to genome assembly composed of mainly clone based sequencing, genome assemblies consisting of much WGS may contain more sequence misassignment errors and unfinished regions (Cheung et al. 2003). It can be inferred that WGS may result in missing synteny-conserved ncRNAs (false negatives). Even when finished contig sequences are used, multi-way genome alignments provided by public-domain resources may still miss synteny-conserved ncRNAs. For instance, in the 10-way vertebrate genome alignments generated using the Pecan algorithm, a new comparative-genomics resource provided by Ensembl, only 114 human tRNA gene loci were found to be aligned to their synteny-conserved counterparts in other species (data not shown). This number is much smaller than that found using the UBRHPs-based approach (371 loci, see subsection 2.1.2.2.), even though the mouse genome assembly used to generate Pecan alignments consists mainly of finished contig sequences. The UBRHPs-based approach is useful for evaluating ncRNA conservation, as it has been used here, but cannot be used in *de novo* ncRNA prediction as it relies on the location of ncRNAs in one species already being known. An additional source of false negatives, when using ncRNA finding algorithms that depend on genome alignments, will be ncRNAs which are genuinely synteny-non-conserved. In genomes that are distantly related, numerous ncRNAs may be synteny-non-conserved. Such a situation has been demonstrated by the low synteny-conservation ratio of human and zebrafish ncRNAs (see subsection 2.1.3.4.). A similar situation was also encountered when comparing the human and chicken genomes (Hillier et al. 2004).

When evaluating ncRNA finding algorithm performance on genome alignments, it is also necessary to consider the number of false positives. Recently ncRNA finding algorithms were applied to a high-quality set of 28-way vertebrate genome alignments consisting mainly of finished contig sequences and corresponding to 1% of the human genome sequence (Washietl et al. 2007). This is part of the ENCODE project (The ENCODE Project Consortium 2007).

The ncRNA finding algorithms were found to have successfully detected the small number of known ncRNAs. However with an evaluation using shuffled alignments that preserved the dinucleotide frequency to that of the 28-way genome alignments, Washietl *et al*. estimated that these comparative algorithms for genome-wide ncRNA finding may suffer from a high false positive rate, 50% ~ 70%.

All in all, in the context of using existing vertebrate genome assemblies and their alignments, I conclude that the effectiveness of ncRNA finding algorithms that are based on comparative genomics is limited.

2.1.4.2. Proportion of human ncRNAs which are human-mouse synteny-non-conserved

In the process of collecting synteny-conserved ncRNAs to assess comparative algorithms for genome-wide ncRNA finding, the occurrence of synteny-non-conserved ncRNAs was also established. Synteny-conservation ratios of ncRNAs were calculated from this and were found to vary substantially for ncRNAs in different categories (see subsection 2.1.3.1.). At first sight the ratios for all categories appear substantially lower than published estimates of for protein coding genes (Mouse Genome Sequencing Consortium 2002), which were estimated as high as 96%. However there are substantial differences in the protein and ncRNA data sets from which the synteny-conservation ratios have been calculated which should be considered before any conclusions are drawn. For ncRNA genes in vertebrate genomes it is very difficult to determine which predictions are *bona fide* ncRNAs and which are ncRNA pseudogenes. Estimating synteny-conservation ratios for *bona fide* ncRNAs of various classes in vertebrate genomes is therefore difficult. For protein genes it is much easier to determine which ones are pseudogenes and the figures quoted were calculated after pseudogenes have been excluded, unlike figures for ncRNAs.

If many synteny-non-conserved ncRNAs are pseudogenes, the synteny-conservation ratio of human and mouse ncRNAs may be significantly higher than estimated previously in this

section (2.1). Apart from the effect of pseudogenes, there are several other factors that will contribute to an underestimate of the synteny-conservation ratios of ncRNAs, though only to a small extent. Firstly, some uncertain type(s) of genome rearrangements may potentially cause artefacts in finding synteny-conserved ncRNAs (for details see subsection 2.1.3.2.). However, even if the real synteny-conservation ratio of ncRNAs in "complicated" regions is comparable to that under other evolutionary conditions, the overall synteny-conservation ratio of ncRNAs would only be ~2% higher than previously estimated. Secondly, ~40% of the mouse genome assembly (NCBI M33) used in this section was composed of whole genome shotgun sequencing (WGS). However here too, the effect is small, and estimated to have lowered the synteny-conservation ratio by only ~2% (for details see subsection 2.1.3.1.). The major uncertainty relates to the functionality of synteny-non-conserved ncRNA. This issue is further explored in the next chapter (chapter 3).

2.2. Gene-order conservation of mammalian tRNA genes

2.2.1. Materials and methods

2.2.1.1. Recruiting mammalian tRNA gene loci

The genomic loci of the human and mouse tRNA genes were retrieved from Ensembl release 40. These tRNA gene loci were predicted by tRNAscanSE (Lowe and Eddy 1997). The human and mouse genome assemblies used in the following analysis are NCBI 36 and NCBI M36, respectively. They are the most updated assemblies that have been annotated by Ensembl (April 2007, http://www.ensembl.org/index.html). Unlike the previous mouse genome assemblies that consist of many sequences generated from whole genome shotgun sequencing (WGS), NCBI M36 is a highly polished genome assembly, where most of the sequence is composed of finished contig sequences (http://www.ncbi.nlm.nih.gov/projects/genome/seq/NCBIContigInfo.html). Investigating the gene-order conservation of mammalian

tRNA genes using this higher quality mouse genome assembly should therefore be far less affected by genome assembly artefacts.

One issue when trying to understand the evolution of tRNA genes is that, by comparing two genomes, it is difficult to determine whether a difference (*i.e.* an unaligned tRNA gene symbol, referred to subsequently as a 'gap') in an alignment between them is caused by the deletion and/or degradation of tRNA genes in one genome or the insertion of tRNA genes in the other. One way to try and distinguish between these possibilities is to recruit a set of tRNA gene loci, as an external reference, from a third genome that is an outgroup of the first two. An organism that has split from a common ancestor of placental mammals (including human and mouse) before the primate-rodent split can suffice for this purpose. In the following analysis, opossum was used which is a species of marsupials. Marsupials diverged from placental mammals about 180 millions years ago (Lawn et al. 1997). By using such an external reference, the evolutionary event that led to a gene order difference in human and mouse may possibly be inferred. For instance, when considering alignments of tRNA gene clusters if a symbol insertion found in a human-mouse tRNA symbol alignment remains an insertion in a human-opossum tRNA symbol alignment, this insertion is likely to be the result of a duplication or transposition event that occurred in the genome of the human ancestors. Likewise, a deletion and/or degradation of a tRNA gene locus after the primate-rodent split may also be inferred. The tRNA gene loci of the opossum genome were retrieved from Ensembl release 40 and the opossum genome assembly used in the following analysis is MonDom4.

There is one concern about using the tRNA gene arrangements in the opossum genome. The sequence assembly of the opossum genome consists mainly of the sequences from whole genome shotgun sequencing (http://www.ensembl.org/Monodelphis_domestica/index.html). For this reason, the opossum tRNA gene loci are used only for inferring the evolutionary

history after the primate-rodent split, but not that before the primate-rodent split, *i.e.* apparent differences in gene order unique to opossum were ignored.

2.2.1.2. Identifying the syntenic tRNA gene clusters

The steps for identifying synteny-conserved tRNA gene clusters are presented in the flowchart in Figure 2-4. In comparing the tRNA gene order in the human and mouse genomes, the first genome is the human genome and the second genome is the mouse genome. The tRNA gene loci were sub-grouped into clustered and non-clustered ones (singlets), respectively. A threshold of the maximal distance allowed between the nearest neighbouring tRNA genes in a cluster was defined to be 1 mega bases. This threshold was set as the minimum distance required to ensure the super cluster (*e.g.* 150 tRNA gene loci) that spans several mega bases on human chromosome 6 remained a single unit.

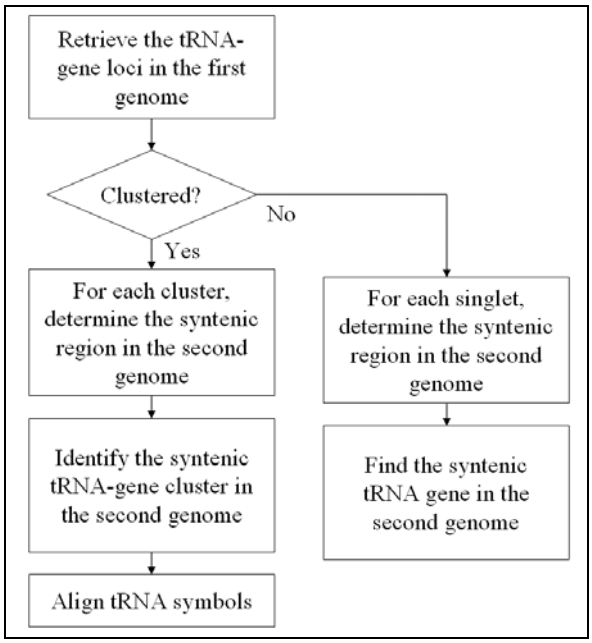


Figure 2-4. The procedure of identifying the syntenic tRNA gene clusters in mammalian genomes

For each human tRNA gene cluster, the syntenic region in the mouse genome was determined using the UBRHPs-based approach (for details see subsection 2.1.2). Each human tRNA gene cluster, together with the corresponding tRNA gene cluster in the syntenic region

in the mouse genome, becomes a pair of synteny-conserved tRNA gene clusters. The conservation of tRNA gene order was investigated by comparing the arrangements of tRNA gene loci in each pair of human-mouse syntenic clusters.

2.2.1.3. Assigning symbols to mammalian tRNA gene loci

A general approach for investigating gene-order rearrangements is to represent genes as symbols and then compare their order (for review see Sankoff and El-Mabrouk 2000). In investigating the tRNA gene-order conservation, I followed a similar strategy. Each tRNA gene locus was thus assigned with a symbol according to its features. These features include the anticodon types and the genomic orientation. For example, there are two different anticodons, GCA and ACA, used by tRNAs for carrying cysteines (tRNA-Cys). Cys1 was used to represent the tRNA-Cys gene loci that have the anticodon GCA. Cys2 was used to represent the tRNA-Cys2 gene loci that have the anticodon ACA. If a Cys1 was on the forward strand of a chromosome, a suffix "F" was added. Conversely, Cys1R was used when a tRNA-Cys1 gene locus was on the reverse strand of a chromosome. A lookup table of the relations between anticodon types and tRNA gene symbols can be found in Table A 1, Appendix A.

There is one consideration in the use of a set of anticodon based tRNA gene symbols. If there are transitions of anticodon types, finding two loci with the same anticodon types does not necessarily mean that both loci should have evolved from a common ancestral locus. Likewise, a mismatch of the anticodon types does not necessarily mean that the two tRNA gene loci should have evolved from two distinct ancestral loci.

In order to compensate for this limitation of the anticodon-type tRNA gene symbols in the gene-order comparison, another set of tRNA gene symbols based on sequence identities was also created. The steps are as follows. Firstly, all human tRNA gene loci were classified according to their anticodons. For example, there are two anticodons, UUU and CUU, for

tRNAs that carry the amino acid lysine. All lysine-tRNA genes, which carry either one of the two anticodons, were grouped together. Secondly, using the TIGR Gene Indices Clustering Tools (TGICL) (TIGR 2002-2003), each group of tRNA genes was further divided into subgroups according to pairwise sequence identities. The grouping was performed by Cap3 (called by TGICL) (Huang and Madan 1999) using default parameters. Subgroup assignments were performed automatically using TIGR. For example, Thr-tRNAs were divided into S_Thr_1, S_Thr_2, and S_Thr_3 subgroups. Forty subgroups were so created. The pairwise sequence identities within individual subgroups range from 94% to 100%. Sequences in each group are fairly homogeneous at the primary-sequence level. Each subgroup was used as a unique sequence type of tRNA genes. For the purpose of comparing the tRNA gene orders in different genomes, each tRNA gene loci in the human, mouse, and opossum genomes was assigned with the best-hit sequence type according to its sequence identities to all sequence types. The sequence-type symbols of tRNA genes were used to find anticodon transitions that may cause the generation of gaps in the anticodon-type symbol alignments.

2.2.1.4. Filtering out possible tRNA-like SINEs

In this tRNA gene-symbol based comparison one issue is filtering out the large number of tRNA-like SINEs which are present mammalian genomes. If too many are included, many false gaps will be generated when comparing the gene orders of two different genomes. In practice, it is very difficult to prepare a comprehensive list of free of the many tRNA-like SINEs. For instance, there are, in the mouse genome, thousands of species-specific SINEs that are related to tRNA genes (Mouse Genome Sequencing Consortium 2002). This is discussed in more depth in the introduction to chapter 3, however for the purposes here, only mouse tRNA genes with tRNAscanSE bit scores greater than 40 were included. There are two reasons for setting this threshold. First, in the set of 2,345 tRNA genes of low scores (tRNAscanSE bit score < 40), 97.3% (2,282) of them overlap with SINEs. Secondly, the

bit-score distribution of the mouse tRNA genes reveals a bi-modal distribution (data not shown), where bit score 40 seems to be a point that can preserve as many normal mouse tRNA genes as possible, while most of the tRNA-like SINEs can be removed. After this filtering, 504 tRNA gene loci in the mouse genome were recruited for this study, while without any particular filtering, there are by coincidence 504 human tRNA gene loci. Only 11.1% (55 / 504) of the high-scoring tRNA gene loci in the mouse genome overlap with SINEs.

For the opossum tRNA gene loci only the simple pseudogene filter by tRNAscanSE was used to clean the data set of the opossum tRNA gene loci. This is due to there being relative little knowledge about repetitive elements in the opossum genome during the preparation of this manuscript.

2.2.1.5. Types of gene-order conservation

The tRNA gene symbols of the human and mouse tRNA gene clusters were initially aligned using a dynamic programming implementation in Biojava (http://biojava.org). Except in the cases of perfect-type conservation, there were gaps in the tRNA symbol alignments of the human-mouse or human-opossum syntenic tRNA gene clusters. According to the source of the unaligned symbols, these gaps were assigned as either insertions or deletions. The unaligned tRNA symbols that were from the human genome were assigned as insertions. Conversely, when the unaligned symbols were from the other genome, either the mouse or opossum genome, the gaps were assigned as deletions. This convention was used only for indicating the source of gaps in symbol alignments, without implying anything about the evolutionary origin of these gaps.

The gene symbols from the two genomes are aligned on both strands to generate two separate alignments, *i.e.* for human and mouse one is the human-forward-strand *versus* mouse-forward-strand alignment; the other is the human-forward-strand *versus* mouse-reverse-strand alignment. The two symbol alignments automatically generated by using

the Biojava were then examined manually. The purpose of this step was to decide which alignment can best explain the evolutionary relationship between the human-mouse synteny-conserved tRNA gene clusters. In some cases, this decisions was not easy to make, especially when there had been chromosomal inversions in the tRNA gene clusters after the primate-rodent split. In cases where there were also synteny-conserved protein-coding genes intervening in the synteny-conserved tRNA gene clusters, these protein-coding genes were used as landmarks. These intervening protein-coding genes could be used to sub-divide tRNA gene clusters into smaller sub-clusters allowing conservation of tRNA gene orders within these sub-clusters.

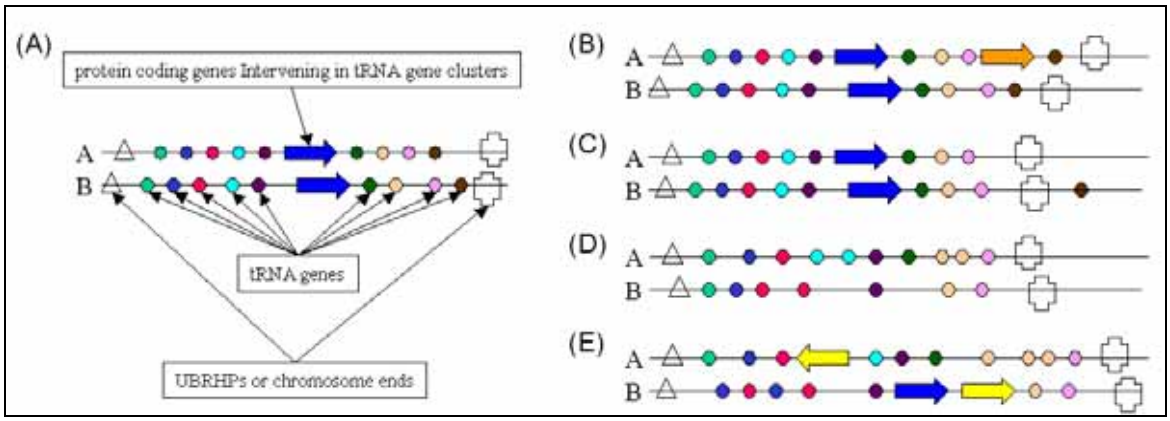


Figure 2-5. Different types of tRNA gene-order conservation

Five types of conservation patterns of the mammalian tRNA genes were defined as follows (see also Figure 2-5):

- "Perfect" conservation (Figure 2-5, A) refers to a pair of syntenic tRNA gene clusters in which the arrangement of all functional elements, including tRNA genes and intervening protein-coding genes, has been completely conserved and all the symbols can be perfectly aligned.

- "Sub-perfect" conservation refers to a pair of synteny-conserved clusters where there are

minor differences between them. "Sub-perfect type-one" conservation (Figure 2-5, B) is used when there is between-syntenic-clusters inconsistency in the physical arrangement of protein-coding genes intervening in the clustered tRNA genes. "Sub-perfect type-two" conservation (Figure 2-5, C) is used when there are non-syntenic tRNA genes at the ends of the syntenic clusters.

- "Gapped" conservation (Figure 2-5, D) refers to a pair of synteny-conserved clusters where a few tRNA gene loci are not aligned.

- "Complicated" conservation (Figure 2-5, E) refers to a pair of synteny-conserved clusters where there may have been multiple genome rearrangements. The existence of a complicated case is inferred when there are multiple gaps in the tRNA symbol alignment. Besides, the linear relations of the protein-coding genes in the neighbourhood of tRNA gene loci may have also changed.

- "Single" conservation refers to the case where, in a tRNA gene cluster, only one synteny-conserved tRNA gene locus was found in the corresponding syntenic region in the second genome.

2.2.1.6. Checking the conservation of the internal promoters of tRNA genes

For the purpose of checking the conservation of the internal promoters in these tRNA genes, eufindtRNA (Pavesi et al. 1994) was used. eufindtRNA is a tRNA-finding algorithm that can recognize the features of important promoting elements, such as A and B boxes, termination signals, and relative spacing between signals, for the transcription of eukaryotic tRNAs. The relaxed mode of eufindtRNA was used here to evaluate only the integrity of intragenic control regions. The stringent mode of eufindtRNA, which can also assess the quality of termination signals, was not used in the following analysis, because evidence suggests that some variations in termination signals are allowed (Gunnery et al. 1999).

2.2.2. Results

2.2.2.1. 32 human-mouse synteny-conserved tRNA gene clusters

Among the 504 tRNA gene loci in the human genome, 92 (18%) loci are not clustered (singlets) (Table 2-9). There are more singlets (27%, 134/504), and also fewer clustered tRNA gene loci in the mouse genome than in the human genome. The significance of this finding is unclear given that we know the data sets used are not entirely clean of loci such as tRNA-like SINEs.

	number of tRNA genes	number of clusters	number of clustered tRNA gene loci	number of non-clustered tRNA gene loci (singlets)
human	504 (100%)	38	412 (82%)	92 (18%)
mouse	504 (100%)	48	370 (73%)	134 (27%)
opossum	991 (100%)	121	597 (60%)	394 (40%)
opossum (bit score >= 40)	546 (100%)	46	408 (75%)	138 (25%)

Table 2-9. The statistics of clustered tRNA gene loci in the human, mouse, and opossum genomes

	human tRNA gene loci in clusters	synteny-conserved clusters	human tRNA gene loci in synteny- non-conserved clusters	human tRNA gene loci in the synteny-conserved clusters
human-mouse	412 (100%)	32	29 (7%)	383 (93%)
human-opossum	412 (100%)	28	181 (44%)	231 (56%)

Table 2-10. The synteny conservation of clustered human tRNA gene loci

Eighty-two percent and seventy-three percent of the tRNA gene loci (Table 2-9) in the human and mouse genomes were grouped into 38 and 48 clusters, respectively (for the detailed lists see Table A 2 and A 3, Appendix A). Thirty-two pairs of human and mouse tRNA gene clusters were found to be synteny-conserved (for a detailed list see Table A 4 in

Appendix A). 93% (383/412) of the tRNA gene loci that are clustered in the human genome are within the human-mouse synteny-conserved tRNA gene clusters (Table 2-10). The conservation of tRNA gene order was then investigated by aligning the symbols of the 32 human-mouse pairs of synteny-conserved tRNA gene clusters. The gene order comparison was performed primarily by using the anticodon-type symbols of tRNA gene loci. The result reveals some unaligned regions in the tRNA symbol alignments (Table 2-12). Among the 383 clustered human tRNA gene loci that reside in the human-mouse synteny-conserved clusters, 230 loci (60%) can be aligned without much uncertainty. A special case is the alignment of human cluster 4.1.36 and mouse cluster 5.1.26. In the initial alignment of this pair of syntenic clusters, only 10 out of the 36 human loci can be aligned. By manual curation, a track of 15 tRNA gene loci that can be aligned in an inverted way was found (Figure 2-6).

Figure 2-6. The conservation pattern of the human tRNA gene clusters 4.1.36 and its syntenic cluster in the mouse genome (see next page)

tRNA gene loci are represented in two ways: (1) the ones in rounded rectangles with symbols indicating the codon type of tRNA genes; (2) the ones that are plotted in red dots, indicating the loci whose evolutionary origins cannot be unambiguously assigned based on sequence identity. Dotted-rounded rectangles are used to indicate the unitary blocks that repeat for multiple times in both the human and mouse genomes. Arrows are used to indicate the orientation of these repetitive blocks, where the red ones are used to indicate the complete unitary blocks, and the cyan and magenta ones are used to indicate the incomplete unitary blocks. Red lines are used to indicate the possible region of a chromosomal inversion.

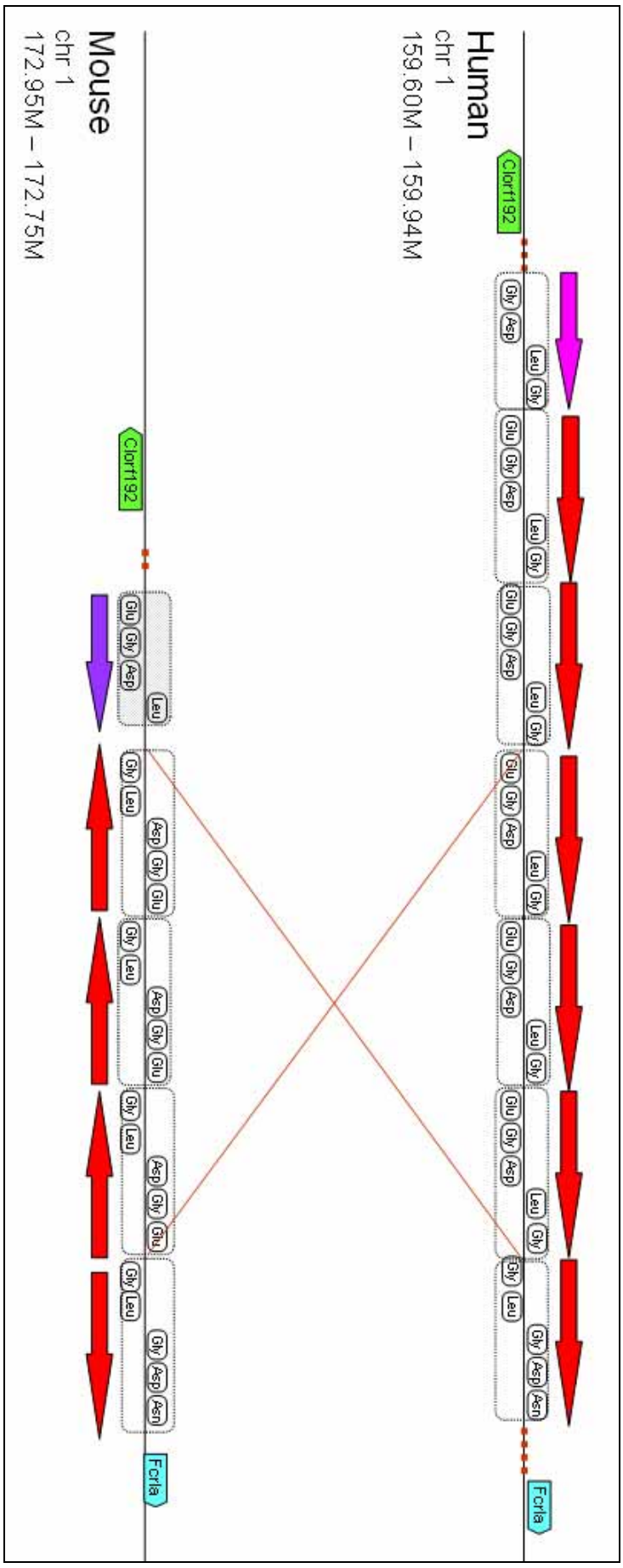


Figure 2-6 (for figure legend see the previous page)

2.2.2.2. Anticodon transitions are rare after the primate-rodent split

The conservation of gene order was also evaluated by comparing the arrangements of the sequence-type symbols. The purpose here was to find if there was any evidence of anticodon transitions that could cause mutated tRNAs to carry different amino acids. The result reveals that, as expected, anticodon transitions in mammalian genomes are very rare. By comparing the human and mouse synteny-conserved tRNA sequence types, only six anticodon transitions were found (Table 2-11). The observed anticodons in these six human tRNA gene loci are not consistent with the expectations inferred from their respective sequence types. The transitions from tRNA-Cys to tRNA-Ser and tRNA-Tyr in human cluster 17.7.20 are also supported by the conserved arrangement of the tRNA gene loci in the corresponding mouse syntenic cluster, in which there are only tRNA gene loci of anticodon type Cys1 and sequence type S_Cys_1.

cluster ID	Coordinate	observed anticodon type	observed sequence type	expected anticodon type	bit score
3.1.42	chromosome:NCBI36:1:147561290:147561360:-1	Val3	S_Gly_1	Gly2/Gly3	60.62
3.1.42	chromosome:NCBI36:1:146185653:146185726:1	Asn2	S_Asn_1	Asn1	52.07
14.6.150	chromosome:NCBI36:6:27379547:27379618:-1	Thr3	S_Met_1	Met1	46.44
14.6.150	chromosome:NCBI36:6:28811185:28811256:-1	Val4	S_Ala_1	Ala3/Ala4	64.08
17.7.20	chromosome:NCBI36:7:148886066:148886138:1	Tyr1	S_Cys_1	Cys1	49.4
17.7.20	chromosome:NCBI36:7:148936400:148936471:1	Ser4	S_Cys_1	Cys1	62.1

Table 2-11. Transitions of the anticodons of tRNA gene loci

2.2.2.3. Numerous gaps between synteny-conserved human and mouse clusters

There were numerous gaps between synteny-conserved human and mouse clusters (Table 2-12). As many as 40% of the human loci in these gene clusters were insertions in symbol alignments. According to the distribution pattern of gaps in the symbol alignments, the synteny-conserved tRNA gene clusters were further grouped into the five conservation types

(Table 2-13) (for the definitions of the five types, see subsection 2.2.1.5. , Materials and Methods). ~65% (267/412) of the human clustered tRNA gene loci are within the human-mouse synteny-conserved clusters where there are multiple gaps in their symbol alignments ("gapped", Table 2-13). Other statistics about the conservation types, aligned loci, *etc*. of the human-mouse synteny-conserved clusters are listed in Table 2-13.

An attempt was made to look for possible relationships between human-mouse non-syntenic tRNA clusters by searching for similarities in the gene order. No significant tRNA gene-order conservation was discovered.

	human tRNA gene loci in the synteny-conserved clusters	insertions in the symbol alignments	aligned human tRNA gene loci in symbol alignment
human-mouse	383 (100%)	153 (40%)	230 (60%)
human-opossum	231 (100%)	104 (45%)	127 (55%)

Table 2-12. The statistics (aligned and inserted regions) of the human-mouse tRNA symbol alignments

conservation type	human tRNA gene clusters	human tRNA-gene loci	aligned loci in the human genome	unaligned loci (insertions)*
perfect	8	17 (4%)	17	0
sub-perfect type one	5	36 (9%)	36	0
sub-perfect type two	4	11 (3%)	9	2
gapped	8	267 (65%)	157	110
complicated	1	42 (10%)	6	36
single	5	10 (2%)	5	5
synteny-non-conserved	7	29 (7%)	0	29
subtotal	38	412 (100%)	230	182

Table 2-13. The statistics of the gene-order conservation of human and mouse tRNA gene clusters

*: There are also 61 deletions in the human-mouse tRNA symbol alignments. Deletions are defined as the additional tRNA symbols in the mouse genome that cannot be aligned to suitable syntenic counterparts in the human genome. Fifty-eight deletions belong to gapped conservation type. Three deletions belong to "single" conservation type.

In addition to clustered tRNA gene loci, some non-clustered tRNA gene loci were also found to be conserved in the corresponding mouse syntenic regions. There are 92 non-clustered tRNA gene loci in the human genome. 37 of them are human-mouse synteny-conserved (see Table A 5, Appendix A).

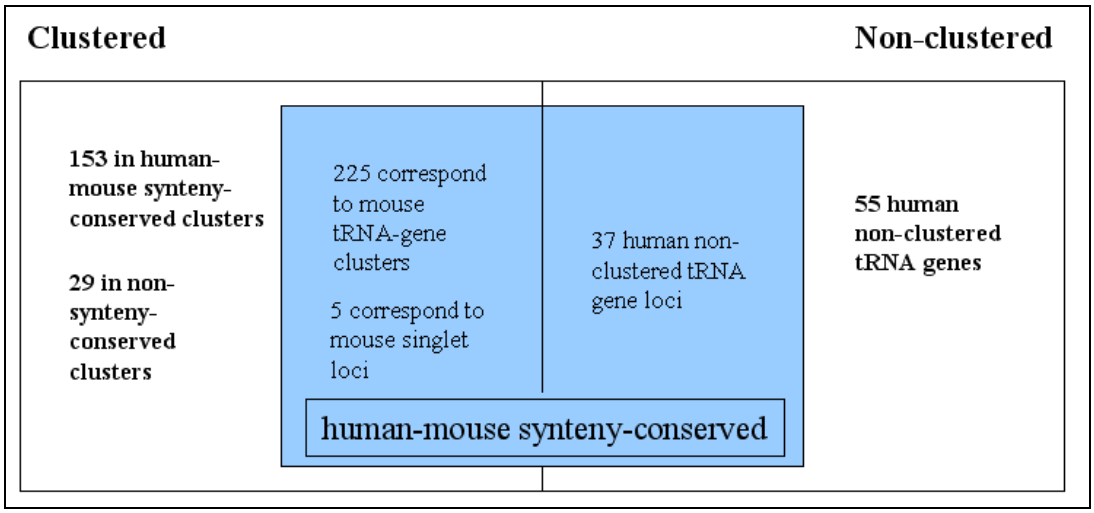


Figure 2-7. Summary of the synteny conservation of human and mouse tRNA gene loci

When the gene order is taken into consideration, only ~53% (267/504) of the human tRNA gene loci are synteny-conserved. This value is much lower, by 21% (74% - 53%), than the previous estimate made under the ignorance of the gene-locus arrangement in each tRNA gene cluster. Obviously, the main source of this big difference is that the arrangements of 153 loci within the synteny-conserved clusters are not conserved (Figure 2-7).

2.2.2.4. The association of the synteny-conservation of tRNA gene clusters with the quality of genome assembly

One factor that may affect the determination of synteny-conservation of tRNA genes is the quality of genome assembly. It is therefore important to explore if the synteny-non-conservation of human tRNA gene loci is associated with unfinished regions or

WGS in the genome assemblies. The investigation reveals that in the synteny-conserved tRNA gene clusters, the gaps in the tRNA symbol alignments are generally not related to the quality of genome assembly (Table 2-14). Within the human-mouse synteny-conserved tRNA gene clusters, all the genomic sequences intervening between each neighbouring tRNA gene loci in the mouse genome are composed of finished contig sequences, but no unfinished contigs nor WGS. Besides, four out of the seven synteny-non-conserved human tRNA gene clusters were found to be in the regions where the genome assembly consists of finished contig sequences.

human tRNA gene clusters	FCS	CSN	WGS*
synteny-conserved clusters	31^{+}	0	0
synteny-non-conserved clusters	4	1	2

Table 2-14. Relation of synteny-conservation of tRNA gene clusters and the quality of the mouse genome assembly
FCS: finished contig sequence; CSN: unfinished contig sequence (with gaps); WGS: whole genome shotgun sequence

*: there are also unfinished gaps in these WGSs.

+: In 3 human-mouse synteny-conserved clusters, the intervening (mouse) genomic sequences between each pair of neighbouring tRNA gene loci are composed of finished contig sequences (FCS), while there are WGSs between the (5' or 3') end tRNA gene loci of a cluster, and the protein-gene boundaries that define the corresponding human-mouse syntenic blocks.

human non-clustered tRNA gene loci	FCS	CSN	WGS
synteny-conserved singlets	36	0	1
synteny-non-conserved singlets	51	1	3

Table 2-15. Relation of synteny-conservation of non-clustered tRNA genes (singlets) and the quality of the mouse genome assembly

The association between the quality of genome assemblies and the synteny conservation of non-clustered tRNA gene loci (singlets) was also evaluated. The inability to find syntenic mouse counterparts to human tRNA gene singlets does not seem to be biased by the quality of genome assembly (Table 2-15). Among the 55 synteny-non-conserved singlets, 51 of the corresponding syntenic regions in the mouse genome are composed of FCS, but no WGS.

These results suggest that the gaps in the synteny-conserved clusters, the synteny non-conservation of at least four human tRNA gene clusters, and the synteny non-conservation of 51 non-clustered human tRNA gene loci, are more likely to be caused by evolutionary events, *i.e.* genome rearrangements, retro-transpositions, degraded genes (pseudogenes), tRNA-related SINEs, *etc*.

2.2.2.5. The information from the tRNA gene loci in the opossum genome

The comparison of the human and opossum tRNA gene loci reveals that there are fewer (28) human-opossum synteny-conserved tRNA gene clusters than human-mouse synteny-conserved clusters (Table 2-10). An example is that no opossum tRNA gene clusters were confirmed to be syntenic counterparts of the super tRNA gene cluster, 14.6.150, which is on human chromosome 6. Besides, more gaps (unaligned human tRNA gene symbols) were found in the human-opossum alignments than in the human-mouse alignments. These findings essentially fit expectations because opossum split from the placental mammals long before the primate-rodent split and the genome assembly quality is much lower.

The arrangement of tRNA genes in the opossum genome provides information that can help us understand tRNA gene evolution in mammalian genomes. The insertions and deletions in the human-mouse tRNA symbol alignments, can be re-categorized by examining the 3-way, human-mouse-opossum, alignments of the tRNA gene symbols and applying the following rules:

- If an inserted tRNA gene symbol is found in opossum in the human-opossum tRNA symbol alignment, this symbol insertion may represent a deletion or degradation of a tRNA gene locus in the mouse genome after the primate-rodent split.
- If an inserted tRNA gene symbol cannot be found in opossum in the human-opossum tRNA symbol alignment, this symbol insertion may represent an insertion of a tRNA gene locus in the human genome after the primate-rodent split.
- If a deleted tRNA gene symbol is also missing from opossum in the human-opossum

tRNA symbol alignment, this symbol deletion may represent an insertion of a tRNA gene locus in the mouse genome after the primate-rodent split.

- If a deleted tRNA gene symbol can be found in opossum in the human-opossum tRNA symbol alignment, this symbol deletion may represent a deletion or degradation of a tRNA gene locus in the human genome after the primate-rodent split.

The re-categorization of gaps in the human-mouse tRNA symbol alignment was performed using the above rules.

human tRNA gene clusters	insertions in the human-mouse alignments	Post primate-rodent-split insertions in the human genome	Post primate-rodent-split deletions/degradations in the mouse genome
6.1.3	1	1	0
13.5.17	10	9	1
16.6.2	1	1	0
17.7.20	2	NA	NA
18.8.4	1	1	0
20.11.2	0	0	0
23.13.2*	2	0	1
24.14.14	9	0	7
26.15.2	1	1	0
30.16.5	2	1	1
33.17.8	2	1	1
37.19.2*	2	0	2
Subtotal	44	15	13

Table 2-16. Evolutionary origin of the insertions in the human-mouse tRNA symbol alignments

NA: not available. The placement of gaps in the alignments is not unique.

*: these tRNA gene clusters are not human-mouse synteny-conserved, but are human-opossum synteny-conserved.

human tRNA gene clusters	deletions in the human-mouse alignments	Post primate-rodent-split insertions in the mouse genome	Post primate-rodent-split deletions/degradations in the human genome
13.5.17	1	1	0
17.7.20	34	NA	NA
18.8.4	1	0	1
20.11.2	1	1	0
Subtotal	38	2	1

Table 2-17. Evolutionary origin of the deletions in the human-mouse tRNA symbol alignments

NA: not available. The placement of many gaps in the alignments is not unique.

Based on the information derived from comparing the human-opossum synteny-conserved tRNA gene clusters, 28 insertions (*i.e.* the unaligned tRNA symbols in the human genome) can be re-classified to 15 post primate-rodent-split insertions of tRNA gene loci in the human genome, and 13 post primate-rodent-split deletions/degradations of tRNA gene loci in the mouse genome (Table 2-16). Two human tRNA gene clusters that are not human-mouse synteny-conserved were found to be human-opossum synteny-conserved (23.13.2 and 37.19.2, Table 2-16). These two clusters may have been deleted/degraded in the mouse genome after the primate-rodent split. Besides, among the deletions in the human-mouse tRNA symbol alignments, there are two post primate-rodent-split insertions of tRNA gene loci in the mouse genome, and one post primate-rodent-split deletion/ degradation of a tRNA gene locus in the human genome (Table 2-17).

2.2.2.6. Duplicated multi-loci blocks in the mammalian tRNA gene clusters

There are several human-mouse synteny-conserved tRNA gene clusters in which gaps in the tRNA symbol alignments cannot be unequivocally placed, due to the existence of so many unaligned regions in the tRNA symbol alignments. Human cluster 3.1.42 is a classic example (Figure 2-8). In the human cluster 3.1.42, not only the arrangement of the tRNA gene loci, but also the relation of the tRNA gene loci to the neighbouring protein-coding genes has changed.

One question that arises from these observations is about the mechanism by which tRNA gene loci in mammalian genomes evolve. Are there any particular rules that govern the changes of tRNA gene orders in these syntenic clusters? Or is the rearrangement of the tRNA gene loci in these synteny-conserved clusters generally random?

Interestingly, the arrangement of the opossum tRNA gene loci provides useful information on this issue. By comparing the arrangements of tRNA gene loci as well as neighbouring protein-coding genes in the human, mouse, and opossum genomes, a vague picture about the evolution of the tRNA gene loci in the human cluster 3.1.42 is revealed (Figure 2-8). My conclusions are summarized as follows:

- The syntenic clusters contain four distinct blocks, A, B, C, and D, of protein-coding genes. The gene order in each block is quite conserved among the human, mouse, and opossum genomes.

- The arrangements of the first three blocks, including A, B, and C, consisting of protein-coding genes, are quite conserved in the mouse and opossum genomes. However, in the human genome, the arrangement of A, B, and C is as C_R-A-B. The subscript "R" indicates that the C block is on the reverse strand. It can be inferred that there might be one segmental inversion in the human genome after the primate-rodent split.

- Between the C and D protein-gene blocks, the arrangements of tRNA gene loci in the human, mouse, and opossum genomes is very different.

- There are multiple species-specific multi-tRNA-loci duplications in each cluster. No common unit blocks of these species-specific duplications were found among the human cluster, 3.1.42, and its syntenic clusters in the mouse and opossum genomes. In the human cluster, 3.1.42, there are two blocks of Gln2-Asn1 tRNA gene loci, two blocks of Gln2-His1 loci, and two duplicated blocks of Asn1-Asn1 loci. In the syntenic tRNA gene cluster in the mouse genome, there are three duplicated blocks of Asn1-His1 tRNA gene loci, two duplicated blocks of Glu1-Gly3 loci. In the syntenic cluster in the opossum genome, there are at least seven types of duplicated blocks, where each distinct type consists of unique combinations of different tRNA gene loci.

- In the human cluster 3.1.42, there are 16 tRNA-Asn1 gene loci which are arranged into

several separated sub-clusters consisting of varied numbers of tRNA-Asn1 gene loci. By contrast, there are 7 tRNA-Asn1 gene loci that are interspersed in the syntenic mouse cluster. 15 out of the human 16 tRNA-Asn1 gene loci were found to have better intra-cluster (other tRNA gene loci in the same cluster, 3.1.42) hits than inter-cluster hits (other tRNA gene loci not in cluster 3.1.42). This means that these tRNA-Asn1 gene loci in the human cluster 3.1.42 are more likely to be generated by intra-cluster duplications than by inter-cluster duplication. In addition to at least three duplicated blocks of two tRNA-Asn1 gene loci, there appear to have been a number of tandem duplications of single tRNA-Asn1 gene loci.

- Some of the single units of duplicated multi-tRNA-loci blocks in one genome cannot be found in the other genome(s). For instance, the Glu1-Gly3 unit of a pair of duplicated blocks in the mouse genome cannot be found in either the human or opossum syntenic cluster.

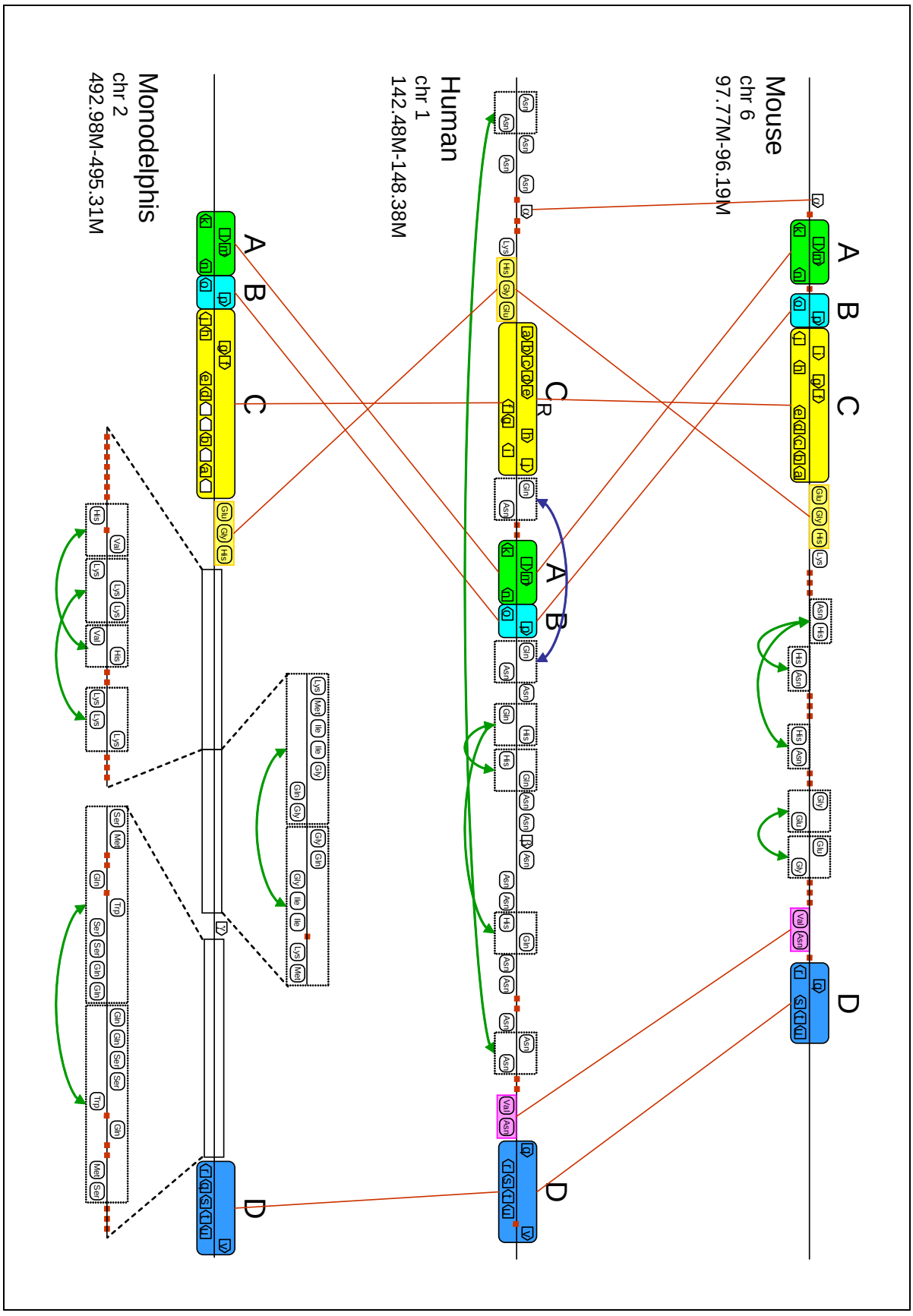


Figure 2-8 (for figure legend see the next page)

Figure 2-8. The conservation pattern of human tRNA gene cluster 3.1.42 and its syntenic clusters in the mouse and opossum genomes

This figure was not prepared to the scale, because it was intended to provide an overview of the putative, both intra-species and inter-species, tRNA gene locus duplications on human chromosome one, 142.48M-148.38M, with respect to the corresponding syntenic regions in the mouse and opossum genomes.

tRNA gene loci are represented in two ways: (1) the ones in rounded rectangles with symbols indicating the codon type of tRNA genes; (2) the ones that are plotted in red dots, indicating the loci whose evolutionary origins cannot be unambiguously assigned based on sequence identity. Color-shaded boxes are used to indicate the inter-species synteny-conserved regions, which are connected by red lines. The dotted boxes around multiple tRNA gene loci are used to indicate the regions that may be involved in intra-species duplications. Curved lines are used to indicate the relation between intra-species duplicated blocks, where the blues ones are used to indicate the blocks of directed duplications, and the green ones are used to indicate the blocks of inverted duplications.

Protein coding genes are represented using arrows. Synteny-non-conserved protein coding genes are represented as open arrows. The symbols for the protein-coding genes used as the landmarks in this figure are as follows:

a	TXNIP	f	NUD17_HUMAN	k	FMO5	p	GJA8	u	ZA20D1
b	LIX1L	g	POLR3C	l	CHD1L	q	BOLA1	v	VPS45A
c	RBM8A	h	ZNF364	m	BCL9	r	HIST2H2AB	α	PDE4DIP
d	ANKRD35	i	CD160	n	ACP6	s	SV2A	β	NP_110423.3
e	PIAS3	j	PDZK1	o	GJA5	t	MTMR11	γ	HIST2H2AA3

2.2.2.7. The synteny conservation of non-clustered tRNA gene loci in mammalian genomes

In addition to the exploration about the evolution of tRNA gene loci in clusters, non-clustered but synteny-conserved tRNA gene loci (singlets) were also investigated in this study. Interestingly, ~78% (29/37) of the human-mouse synteny-conserved tRNA gene singlets were also human-opossum synteny-conserved. All these synteny-conserved tRNA gene singlets were high-scoring (tRNAscanSE bit scores > 64).

2.2.2.8. The association between local duplications and unaligned tRNA gene loci in the human-mouse tRNA symbol alignments

Motivated by the finding of intra-cluster duplicated multi-tRNA gene blocks in the human cluster 3.1.42, and its syntenic clusters in the mouse and opossum genomes, I systematically surveyed the association between local duplications and synteny-non-conserved

tRNA gene loci in mammalian genomes.

The starting point of this survey is to find candidate blocks for local multi-loci duplications. Candidate blocks are defined as repeating multi-loci blocks of 2-6 tRNAs in length that are not necessarily tandemly arranged, e.g. if a 2-locus block re-occurs 4 times, the number of loci involved in the putative duplication is 8, and so forth. If a series of tRNA gene loci of the same anticodon type are tandemly arranged, they are also defined as a type of candidate block. When all human tRNA gene clusters were surveyed, ~20% (108/504) of all human tRNA gene loci were labelled candidate blocks. The existence of local duplications is supported by the observation that, among these 108 loci, ~81% (88/108) have their best (sequence identity) match within the putative regions of human-specific duplications. The remaining ~19% have matches that have only one or two more mismatches than their best hits to the regions outside the putative regions of duplications. The evidence, from the conservation of gene order and the good sequence identities between putative duplicated loci, suggests an association between local duplications and the evolution of tRNA gene loci in mammalian genomes.

Further investigation reveals that local duplications may be implicated in the unaligned tRNA gene loci in synteny-conserved tRNA gene clusters. A substantial proportion of the insertions in the human-mouse tRNA symbol alignments can be explained by species-specific local duplications. ~46% (70) of insertions (153, Table 2-12) overlap with putative human-specific candidate blocks involving multi-tRNA-gene loci; ~16% (25/153) of insertions overlap with human-specific tandem duplications of single tRNA gene locus. In addition, duplications may also associate with the species-specific tRNA gene clusters in mammalian genomes. In the synteny-non-conserved human cluster 1.1.10, there is one pair of candidate blocks, which are arranged in an inverted way. The synteny-non-conserved human cluster, 38.X.3, consists of 3 tRNA-Ile gene loci. The synteny-non-conserved human cluster,

15.6.8, is likely to be the result of a segmental duplication of the human cluster 14.6.150. In summary, 63% of the unaligned tRNA gene loci in the human-mouse tRNA symbol alignments can be explained by local duplications (Table 2-18).

conservation type	unaligned loci (insertions)*	unaligned loci that can be explained by local duplications
sub-perfect type two	2	1 (50%)
gapped	110	66 (60%)
complicated	36	29 (81%)
single	5	0 (0%)
synteny-non-conserved	29	19 (66%)
subtotal	182	115 (63%)

Table 2-18. Local-duplication associated insertions in the human-mouse tRNA symbol alignments

*: The definition of insertion is the same as that in Table 2-13.

2.2.3. Discussions

2.2.3.1. Possible evolutionary events involved in the rearrangements of tRNA gene loci in mammalian genomes

Based on the investigation of gene-order conservation, the human-mouse synteny-conservation ratio of tRNA gene loci is estimated to be only ~53% (see subsection 2.2.2.3. and Figure 2-7). This is lower than the UBRHPs-based estimate of ~74% which did not take into account gene-order and indicates the substantial number of gene-loci whose order is not conserved within tRNA clusters.

One evolutionary event implicated by the low synteny-conservation ratio appears to be local duplication. More than half of the changes between the human-mouse syntenic tRNA gene clusters can be explained as the results of local duplications (see subsection 2.2.2.8. and Table 2-18). In addition to species-specific (post primate-rodent split) duplications, there is

evidence for local duplications before the primate-rodent split. For instance, in the human cluster 4.1.36, three duplicated blocks of five-tRNA-gene loci can be found in both the human and mouse syntenic clusters. Local duplication may be a ubiquitous rule for the evolution of tRNA gene loci in mammalian genomes.

In many cases of putative duplications, the candidate blocks, which may consist of multiple tRNA gene loci, are linked in either a direct or an inverted order. Formally, direct local duplications are called tandem duplications. One mechanism which may generate tandem duplications is unequal crossing-over between sister chromosomes during meiosis (for review see Anderson and Roth 1977). On the other hand, when local duplicated blocks are arranged in an inverted order, the duplications are called inverted duplications. There are at least two possible mechanisms which may generate inverted duplications. First, inverted duplication may be the result of post-tandem-duplication chromosomal inversion. Second, a model with double crossing-overs, which is proposed by Passananti *et al*. (Passananti et al. 1987), can also generate inverted duplications. However, from the investigations already made in this chapter, it is impossible to determine by which mechanism each inverted duplication has been generated. Future work could be to look for evidence to support one of the mechanisms. One possible way to resolve this problem might be to look for existence for replication origins, which is a required feature, proposed by Passananti *et al*., in the generation of inverted duplication.

2.2.3.2. The co-amplification model of the formation of gene clusters

The mechanisms that may lead to gene amplifications through tandem duplications and inverted duplications in one of the daughter strands can also cause the de-amplification of gene loci in the other strand. It has therefore been proposed that local duplications in prokaryotic genomes can act as a dynamic and reversible mechanism that can facilitate adaptation to a variety of environmental conditions (for review see Reams and Neidle 2004).

A co-amplification model has been proposed to explain the generation and maintenance of the clustering of related genes in prokaryotes (Reams and Neidle 2004). One main argument is that clustered genes are more likely to be co-amplified and so equally regulated by gene dosage. Besides, if a gene cluster has been evolutionarily selected by the co-amplification model, the order of genes in this cluster does not need to be strictly conserved.

Interestingly, the differences in tRNA gene order observed between the syntenic counterparts in different mammalian genomes suggest that the co-amplification model may have contributed to the formation and evolution of tRNA gene clusters in mammalian genomes. The findings relevant to the co-amplification model include increases of copy number of tRNA genes through mechanisms leading to local duplications, and the partial conservation tRNA gene orders in mammalian genomes.

One question that remains unanswered is about the advantage to survival conferred by the amplification of tRNA gene loci in mammalian genomes. In prokaryotes, over-expression of gene products caused by gene amplification has been suggested to play a critical role in coping with environmental stresses, such as existence of heavy metals, antibiotics, *etc*. (for review see Romero and Palacios 1997). When a particular selection force disappears, the duplicated loci may be de-amplified through the reversible mechanisms of local duplications. Perhaps, the finding of species-specific duplications of tRNA gene loci in the human, mouse, and opossum genomes, respectively, reflect the differential requirements in the evolution of different mammalian species. Due to local duplications, there is significant difference between the numbers, in the respective genomes, of the tRNA gene loci of particular isoacceptor (anticodon) types. For instance, there are 20 tRNA-Cys1 gene loci in the human cluster, 17.7.20, while there are 52 and 43 loci in the syntenic clusters in the mouse and opossum genomes, respectively.

2.2.3.3. Observations that cannot be explained by the co-amplification model

From the observed synteny-conservation pattern of tRNA gene loci in mammalian genomes, several phenomena were found to be incompatible with the co-amplification model.

Firstly, there are synteny-conserved singlet tRNA gene loci in mammalian genomes. For instance, 29 human non-clustered tRNA gene loci were found to be synteny-conserved in the human-mouse-opossum syntenic regions (Figure 2-9). The synteny conservation of these non-clustered tRNA gene loci strongly suggests they should be functional genes. None of these singlet tRNA gene loci are single copies of respective isoacceptor (anticodon) types. There is also no evidence that these singlets are the degraded remnants of tRNA gene clusters. One question is that, if the co-amplification and clustering is so beneficial to the survival of different mammalian species, why these singlet tRNA gene loci should be still conserved after tens of million years of evolution? During the preparation of this manuscript, no obvious advantages/disadvantages can be proposed to explain this observation.

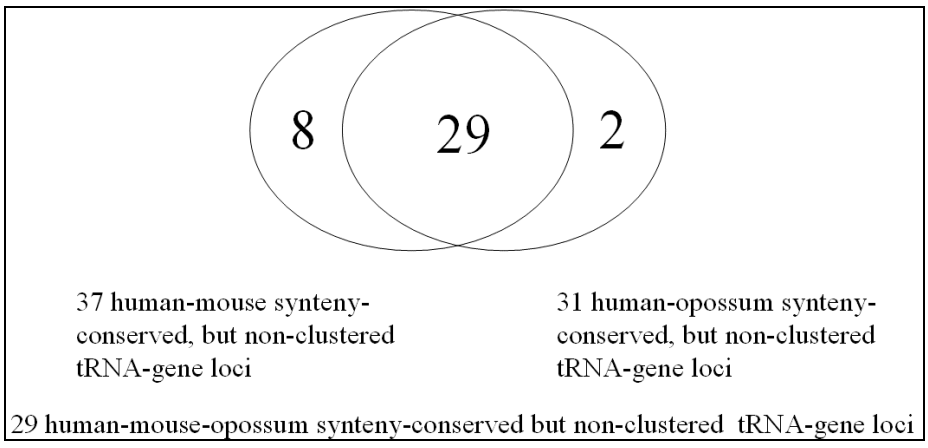


Figure 2-9. the synteny conservation of human non-clustered tRNA gene loci in the syntenic regions of other mammalian genomes

Secondly, there are some synteny-non-conserved human tRNA gene loci, which cannot be explained by local duplication. Possible explanations may include the retro-transpositions,

and the post primate-rodent-split deletions/degradation of tRNA gene loci. These two issues are investigated in the following subsections (2.2.3.4. and 2.2.3.5.).

Finally, recent evidence has implied that the co-amplification model may not be the only plausible mechanism for the clustering of tRNA gene loci in the genomes. In the co-amplification model, clustered genes need not to be co-regulated by a cluster-associated enhancer. However, there is evidence that, under different conditions, the relative expression levels of tRNAs of different isoacceptor types may change (Dittmar et al. 2006). One idea is that the internal promoters may provide a basal-level regulation of tRNA transcription, and the non-promoter regulatory regions may be responsible for controlling the differential expression under different situations. Searching for transcription regulatory elements for clustered tRNA gene loci in mammalian genomes is discussed briefly at the end of chapter 5.

2.2.3.4. Degradation or deletion?

Although the co-amplification model is an appealing hypothesis for interpreting the observed conservation patterns of tRNA gene loci in mammalian genomes, not all unaligned tRNA gene loci can be explained by species-specific local duplications or its reversible process (Table 2-18). In order to find other evolutionary events that may also lead to the unaligned regions in the human-mouse tRNA gene symbol alignments, another possibility, the post primate-rodent-split degradation of the sequences of tRNA gene loci, was therefore explored.

For the non-clustered (singlet) and synteny-non-conserved human tRNA gene loci, the search for the evolutionary remnants in their corresponding syntenic regions in the mouse genome proved to be not very informative. For the 54 synteny-non-conserved singlet tRNA gene loci, only short hits could be found by using WU-BLAST. Most of the e-values are much higher than 0.05, except two cases with borderline significance (0.014 and 0.053). Since the evidence is so weak, it is unclear if there has been pseudogenisation through sequence

degradation of singlet tRNA gene loci in the mouse genome.

Interestingly, for the unaligned tRNA gene loci in the human-mouse syntenic clusters, two putative cases of pseudogenisation through sequence degradations were found. None of the two pseudogenes have previously been annotated by Ensembl (using tRNAscanSE). These cases suggest that sequence degradation is implicated in the evolution of clustered tRNA gene loci in mammalian genomes.

The first case is the degraded remnant in the mouse syntenic region of the Gly1-tRNA gene locus in the human cluster 37.19.2, which is a human-mouse synteny-non-conserved cluster. The e-value of the hit is 2.9e-06 (reported by WU-BLAST). The coordinate of the syntenic tRNA gene locus in the mouse genome is chromosome: NCBIM36: 17: 55852840: 55852911: 1.

```
Human            GCGUUGGUGGUAUAGUGGUuAGCAUAGCUGCCUUCCAAGCAGUUGA
Mouse(degraded)  AUAUUGGUAGAAUAGUGGUuAGgAAAGCUGCCUUCCAAA-AGGUGG
SS_cons          (((((((,,<<<<_______._>>>>,<<<<<________>>>>>,,,

Human            -CCCGGGUUCGAUUCCCGGCCAACGCA
Mouse(degraded)  CCCCGGGUUCUAGUCCCAGAUUGCUUA
SS_cons          ,,<<<<<________>>>>>)))))))):
```

Figure 2-10. The structural alignment of a human tRNA gene locus and its syntenic (but degraded) counterpart in the mouse genome

This previously undiscovered mouse tRNA gene locus does not seem to be a functional one. Firstly, the sequence of the promoter, B box, appears to be degraded. Using eufindtRNA, which is a tRNA-finding algorithm based on the promoter conservation of tRNA genes, this sequence was determined to be a worse promoter than the one in the human orthologous tRNA gene. Secondly, even if this mouse tRNA gene could be transcribed, the secondary structure of the generated tRNAs is likely to be unstable. The putative tRNA product of the degraded gene

locus contains 10 non-Watson-Crick (W-C) and non-GU base pairs in the stem regions (red regions on the mouse strand, Figure 2-10). For comparison, there is only one non-canonical base pair potentially de-stabilizing the secondary structure of the tRNA products transcribed from the orthologous human tRNA gene locus (red regions on the human strand, Figure 2-10).

The second case of pseudogenisation is the degraded locus in the human syntenic region of the Arg4-tRNA gene locus in the mouse cluster 10.3.5, which is the syntenic cluster of the human cluster 18.8.4. The e-value of the hit is 7.8e-09 (reported by WU-BLAST). This previously undiscovered human tRNA gene locus, chromosome: NCBI36: 8: 67187730: 67187802: -1, should be a pseudogene, although the secondary structure of the putative tRNA product have largely been preserved (red regions on the human strand, Figure 2-11). Its promoter, B box, has mutated from GGTTCGACT to GGTCCAGCT (corresponding to the RNA sequences in magenta color on the human and mouse strands, resepctively, Figure 2-11). The degradation of the promoter pattern, which cannot be identified by eufindtRNA, suggests that this degraded tRNA gene locus should be untranscribable. This finding is interesting, because it provides an example of pseudogenisation through promoter-specific degradation. Pseudogenization through promoter-specific degradation is investigated and discussed more generally in chapter 3.

```
Mouse            GGGCCAGUGGCGCAAUGGAuAACGCGUCUGACUACGGAUCAGAAGAUUGU
Human(degraded)  AGGCCAGUGGCGCAAGGGAuAACGUGUCUGACCACGCAUCAGAAGAUUGU
SS_cons          (((((((,,<<<<_______._>>>>,<<<<<________>>>>>,,,,,<<

Mouse            AGGUUCGACUCCUACCUGGCUCG
Human(degraded)  AGGUCCAGCUCCUGCCUGGCUCG
SS_cons          <<<________>>>>>)))))))):
```

Figure 2-11. The structural alignment of a mouse tRNA gene locus and its syntenic (degraded) counterpart in the human genome

One advantage of pseudogenisation through promoter-specific degradation is that it is efficient and safe. If pseudogenisation of a tRNA gene locus proceeded through random mutation, accumulated generation by generation until the functions of the tRNA products were fully abolished, it is possible that some intermediate diseased species of tRNAs would be produced and thus decrease the fitness of the affected organism. By contrast, promoter-specific degradation achieves pseudogenisation by mutating only a few residues in the promoter region of a tRNA gene locus. Although only two cases of promoter-specific degradation were found, it is likely that there are other undiscovered degraded tRNA gene loci. Searching for evidence of old pseudogenes can be very difficult, because without functional constraints, pseudogenes may, after millions of years of evolution, have accumulated so many random mutations that sequence similarity search algorithms cannot find the significant remnants. Consequently, determination of the differential contributions made by sequence degradation and deletions, respectively, to the evolution of tRNA gene loci in mammalian genomes is difficult.

2.2.3.5. Finding pseudogenes through the human-mouse tRNA gene symbol alignments

One purpose of investigating the tRNA gene-order conservation is to search for the evidence which can help us to differentiate functional tRNA gene loci from pseudogenes, a topic more broadly discussed in chapter 3. An appealing argument is that synteny-non-conserved tRNA gene loci will tend to be pseudogenes. In addition to this, the human-mouse tRNA gene symbol alignments of synteny conserved tRNAs provide some other insights relevant to the determination of tRNA pseudogenes.

Firstly, several cases of anticodon transitions were found (Table 2-11) and anticodon transitions may potentially be an indicator of tRNA pseudogenes. In order to realize this argument, a brief introduction to tRNA *identity* is necessary. The term, tRNA identity, refers to the amino acid charging specificity of each tRNA molecule by aminoacyl-tRNA synthetases.

For most tRNAs, the determinants of tRNA identity include the anticodon loop as well as the amino acid accepting stem (for review see Giege et al. 1998). It is unknown if these anticodon transitions would change the tRNA identity of the tRNAs produced from the gene loci in Table 2-11. If the tRNA identities of tRNAs with anticodon transitions remained unchanged, there could be incorrect incorporation of amino acids in protein synthesis. Under the consideration related to tRNA identity, the tRNA gene loci with anticodon transitions should be regarded as potential pseudogenes. An alternative possibility may be errors in the human genome sequence. The significance of these tRNA gene loci with anticodon transitions needs further investigation.

Secondly, the human-mouse tRNA gene symbol alignment also reveals at least one synteny-conserved but low-bit-score tRNA gene locus. Such a locus may also represent a candidate pseudogene. The example is the human tRNA-Asp1 gene locus, chromosome: NCBI36: 1: 159768539: 159768610: 1, which is a member of the human cluster 4.1.36. Its bit-score (reported by tRNAscanSE) is 34.08, which is much lower than that (72.92) of its syntenic counterpart, chromosome: NCBIM36: 1: 172873704: 172873775: -1, in the mouse genome. A putative tRNA product from this gene locus may have an unstable amino-acid accepting stem. In addition, this locus may be untranscribable, since its internal promoters might have degraded (data not shown). This finding is consistent with the pseudogenisation mechanism, promoter-specific degradation, which has also been suggested by previous findings in this section (see the examples of Figure 2-10 and Figure 2-11).

2.2.3.6. Other evolutionary events that may be implicated in the evolution of tRNA gene loci in mammalian genomes

The involvement of various evolutionary events, such as local duplications, inversions, and gene degradation, in the evolution of tRNA gene loci in mammalian genomes have been demonstrated in this section. A question is that, what is the involvement of other evolutionary

events, such as retrotranspositions, transpositions, segmental duplications, gene deletions, or even gene transfer from other organisms? In the following discussions, I consider these possibilities under the following conditions, including the species-specific tRNA gene clusters, species-specific singlet tRNA gene loci, and the unaligned tRNA gene loci in synteny-conserved clusters.

For species-specific tRNA gene clusters evolved after the primate-rodent split, an important feature is the pattern of gene arrangement which should have been generated by local duplications. An example is the human cluster 1.1.10, which contains a duplicated block of four tRNA gene loci. There can be two alternative hypotheses to the formation of this cluster. Firstly, it is possible that this human-specific tRNA gene cluster formed before the primate-rodent or even placental-marsupial split. Perhaps, through independent events of genome rearrangements in the mouse and opossum genomes, respectively, the syntenic clusters in either genome have been deleted. Secondly, the human-specific clusters could have evolved after the primate-rodent split. Theoretically, the second hypothesis should be more likely, since the probability of independent segmental deletions in respective genomes should be low. Besides, in the human cluster 1.1.10, interspersed between the duplicated blocks are the primate-specific protein-coding genes (*e.g.* ENSG00000179571, *etc.*) (based on the annotation made by Ensembl). A similar finding was also observed in the human cluster 38.X.3, where two tRNA-Ile2 gene loci are located within the intronic regions of a pair of duplicated genes (*e.g.* ENSG00000205663), which are also primate-specific. In fact, no other tRNA-Ile2 gene loci can be found in the mouse and opossum genomes.

With the evidence collected in this subsection, it can be concluded that segmental deletions in other mammalian genomes are less likely the reason which can explain the existence of species-specific tRNA gene clusters. However, it is still unclear by which mechanism, either retrotranspositions, transpositions, or segmental duplications, the

human-specific clusters have been formed in new genomic loci. Similar situations were also encountered in investigating the evolutionary origin of the synteny-non-conserved singlet tRNA gene loci, and of some of the unaligned loci in the synteny-conserved tRNA gene clusters. A preliminary result indicates that most of the synteny-non-conserved tRNA gene loci in the human genome are not associated with simple repetitive elements, which might be the evidence of retrotranspositions.

2.3. Summary

In the first part of this chapter, the conservation patterns of the human ncRNAs retrieved from Rfam were investigated. The findings and conclusions relevant to comparative ncRNA finding ncRNA finding approaches are summarized as follows:

- Few covariations are found in either human-mouse synteny-conserved ncRNAs or in the human-zebrafish orthologous ncRNAs.
- ncRNA finding algorithms perform worse when applied to genome synteny alignments than on the single ncRNA gene test alignments they were evaluated.
- Multi-vertebrate synteny alignments can contain more co-variations but the performance of ncRNA finding algorithms on them is similarly affected by alignment quality and completeness, resulting in both false positive and false negative predictions.
- The synteny-conservation ratios of categories of Rfam ncRNAs in the human and mouse genomes vary from ~1% to ~74%.
- ncRNAs with more copies in mammalian genomes appear to be less synteny-conserved.
- Genome assembly quality and artefacts resulting from genome rearrangements

(Figure 2-1, d), have only a small effect on calculations of synteny-conservation ratio of Rfam ncRNAs

In the second part of this chapter, the gene-order conservation of mammalian tRNA genes (predicted by tRNAscanSE) was investigated. My findings include that:

- When gene order is considered, only ~53% of the human tRNA gene loci are human-mouse synteny-conserved (see subsection 2.2.2.3. and Figure 2-7). Besides, 6% (29/504) of human tRNA gene loci are in human-specific clusters (see Table 2-10).
- The low gene-order conservation ratio is not biased by the quality of the mouse genome assembly used in this study (see subsection 2.2.2.4.).
- Tandem duplications and inverted duplications may be important reasons for the low gene-order conservation ratio of tRNA gene loci in mammalian genomes (see subsection 2.2.2.8.).
- Promoter-specific degradation may be involved in the pseudogenisation of mammalian tRNA genes (see subsection 2.2.3.4.).

There are a number of hypotheses with respect to the discovery of numerous synteny-non-conserved ncRNAs in mammalian genomes. Finally, I summarize the evidence for or against each of them:

1. Hypothesis: low quality genome assemblies lead to synteny-conserved ncRNAs being misclassified as synteny non-conserved.
   - Evidence for this hypothesis:
     - Synteny-non-conserved ncRNAs (comparing the human genome assembly NCBI 35 and the mouse genome assembly NCBIM 33) were significantly enriched in regions consisting of whole genome shotgun sequencing or

unfinished regions of clone-based sequencing in the mouse genome (see subsection 2.1.3.1. ,Table 2-2 and Table 2-3).

- ◆ Conclusion:
  - Low quality genome assemblies do lead to some ncRNAs being misclassified as synteny non-conserved, but does not explain the majority.

2. Hypothesis: genome duplication and rearrangement can generate synteny-non-conserved ncRNAs.
   - ◆ Evidence for this hypothesis:
     - There are duplicated multi-loci blocks in the mammalian tRNA gene clusters (see subsection 2.2.2.6.).
     - There might be one segmental inversion in the human tRNA gene clusters after the primate-rodent split (see subsection 2.2.2.6. and Figure 2-6).
   - ◆ Conclusion:
     - Analysis of tRNA clusters is highly suggestive that genome duplication and rearrangement is a mechanism for the generation of synteny-non-conserved ncRNAs.

3. Hypothesis: deletion through degradation can generate synteny-non-conserved ncRNAs.
   - ◆ Evidence for this hypothesis:
     - Degraded remnants of tRNAs can be found that correspond to synteny-non-conserved ncRNAs (see subsection 2.2.3.4.)
   - ◆ Conclusion:
     - There is evidence that some synteny-non-conserved ncRNAs are generated through pseudogenisation, degradation and deletion of the corresponding ncRNA in the other species.

4. Hypothesis: retrotransposition can generate synteny-non-conserved ncRNAs.
   - ◆ Evidence for this hypothesis:

- The generation of species-specific tRNA gene clusters (see subsection 2.2.3.6.) could be explained by retrotransposition, but also by other mechanisms.

◆ Conclusion:

- There is no convincing evidence for or against the mechanisms of retrotransposition.

Chapter 3. Distinguishing functional ncRNAs from pseudogenes in mammalian genomes

The results presented in the previous chapter (chapter 2) suggest that many Rfam human ncRNAs appear to be synteny-non-conserved in the mammalian genome after the primate-rodent split. When considering using comparative methods for genome-wide ncRNA finding, one important question is whether synteny-non-conserved ncRNAs tend to be functional genes or pseudogenes. If a considerable proportion of synteny-non-conserved ncRNAs in the genomes under investigation are functional, the strategies that predict ncRNAs only in the alignments of syntenic regions will fail to predict those functional ncRNAs. Conversely, if most synteny-non-conserved ncRNAs are pseudogenes, methods that depend on alignments derived from synteny may be sufficient for genome-wide ncRNA finding.

Before exploring the likelihood of synteny-non-conserved ncRNAs to be pseudogenes, it is necessary to briefly introduce how pseudogenes might be generated, and how they can be computationally identified. Pseudogenes are believed to be generated by either genome duplication or retrotransposition, followed by non-functionalization of a subset of the duplicated copies (for review see Lynch and Conery 2000). The mechanisms that may lead to genome duplications include unequal crossing-over (for review see Graur and Li 2000), and duplication of a segmental (Gu et al. 2002) or entire chromosome (Van de Peer 2004; Dehal and Boore 2005). In so-called retrotransposition, which is a RNA-mediated process, the RNA transcript of a gene is reverse transcribed into DNA, which is then inserted back into the genome at a new location (Maestre et al. 1995). The pseudogenes that are generated through retrotransposition have usually lost the original gene's intron-exon architecture and thus are often referred to as processed pseudogenes, while the pseudogenes generated through duplications of genomic DNA are referred to as non-processed pseudogenes.

Currently, pseudogenes can be computationally identified by searching protein coding genes for indicators of non-functionality. For instance, a duplicated protein pseudogene can be evolutionarily unconstrained, and hence have accumulated random mutations that may destroy its protein gene-like features; a retrotransposed protein pseudogene can completely lose introns (Figure 3-1 A). Several surveys already performed for exploring pseudogenes in the human genome were based on indicators of functionality derived from features of multi-exon protein coding genes (Ohshima et al. 2003; Torrents et al. 2003; Zhang et al. 2003). In particular, by using the ratio of silent to replacement nucleotide substitutions (K_A/K_S), Torrents *et al*. discovered ~20,000 protein pseudogenes in the human genome, where as many as 70% of them were retrotransposed (Torrents et al. 2003). These results, together with the estimate that ~96% of the human protein genes are mouse-synteny-conserved (Mouse Genome Sequencing Consortium 2002), suggest that a protein coding gene sequence that is synteny-non-conserved in mammalian genomes is very likely to be a pseudogene.

However, since the surveys mentioned above were limited to investigating protein pseudogenes, the tendency of synteny-non-conserved ncRNAs to be pseudogenes is unknown. To date, the functionality of the synteny-non-conserved ncRNAs in mammalian genomes has not been systematically investigated. One reason for this is that in mammalian genomes there are abundant ncRNA-derived short interspersed repetitive elements (SINEs) (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002) which make the determination of ncRNA pseudogenes difficult. SINEs are repetitive elements that are amplified in the genomes through retrotransposition (for review see Smit 1999). Most eukaryotic SINEs have evolved from the ncRNAs that are transcribed by RNA polymerase III. Known evolutionary sources of eukaryotic SINEs include tRNA genes, 7SL genes, 5S rRNA genes (for review see Kramerov and Vassetzky 2005). With respect to ncRNA pseudogene identification some of the SINEs in mammalian genomes are so similar, at both the

primary-sequence and structural levels, to functional ncRNAs that even well tuned ncRNA finding algorithms may falsely predict them as real ones. For instance, about 2,700 tRNA genes, which is more than five times of the tRNA genes annotated in the human genome, were initially predicted in the mouse genome (Mouse Genome Sequencing Consortium 2002). In order to generate a smaller, but more confident, set of functional mouse tRNA genes, the Mouse Genome Consortium has used an additional criterion, non-overlapping with the SINEs identified by RepeatMasker (Smit and Green unpublished), to filter the initial prediction. However, there are at least two considerations with such a criterion. First, it may be too arbitrary to hypothesize that all SINEs are pseudogenes. Second, ncRNA pseudogenes that are unrelated to SINEs can not be filtered out. The above case about filtering out tRNA pseudogenes illustrates the difficulty of distinguishing functional ncRNAs from pseudogenes.

It is possible that some synteny-non-conserved ncRNAs are functional genes. Firstly, a synteny-non-conserved ncRNA might be functional and originally synteny-conserved, but has been deleted in the other lineage. Secondly, a synteny-non-conserved ncRNA may be a functional gene as a result of mechanisms creating a functional copy. Perhaps, due to unique features of certain types of ncRNAs, there is a high tendency for these genes to be synteny-non-conserved in mammalian genomes. One argument is that the mechanisms that generate protein pseudogenes may generate synteny-non-conserved but functional ncRNAs, in addition to ncRNA pseudogenes. While a mechanism of pseudogenisation may effectively cause a newly amplified protein gene to lose the association with its upstream regulatory regions, the same mechanism may not necessarily cause the nonfunctionality of a recently amplified ncRNA locus in the genome.

Retrotransposition appears to be one possible mechanism that can lead to the generation of protein pseudogenes, but new and functional ncRNA loci. Since the transcription regulatory elements in the 5’ flanking regions of the protein genes are not contained in mRNA transcripts,

a retrotransposed protein gene, even if it has retained part of the intron-exon structure, should generally be untranscribable. Therefore, a retrotransposed protein gene may become a pseudogene as soon as the redundant sequence is generated (Figure 3-1 A). Conversely, a retrotransposed ncRNA that is not truncated may remain transcribable, if its intragenic promoters are still intact during the process of generating this redundant copy (Figure 3-1 B).

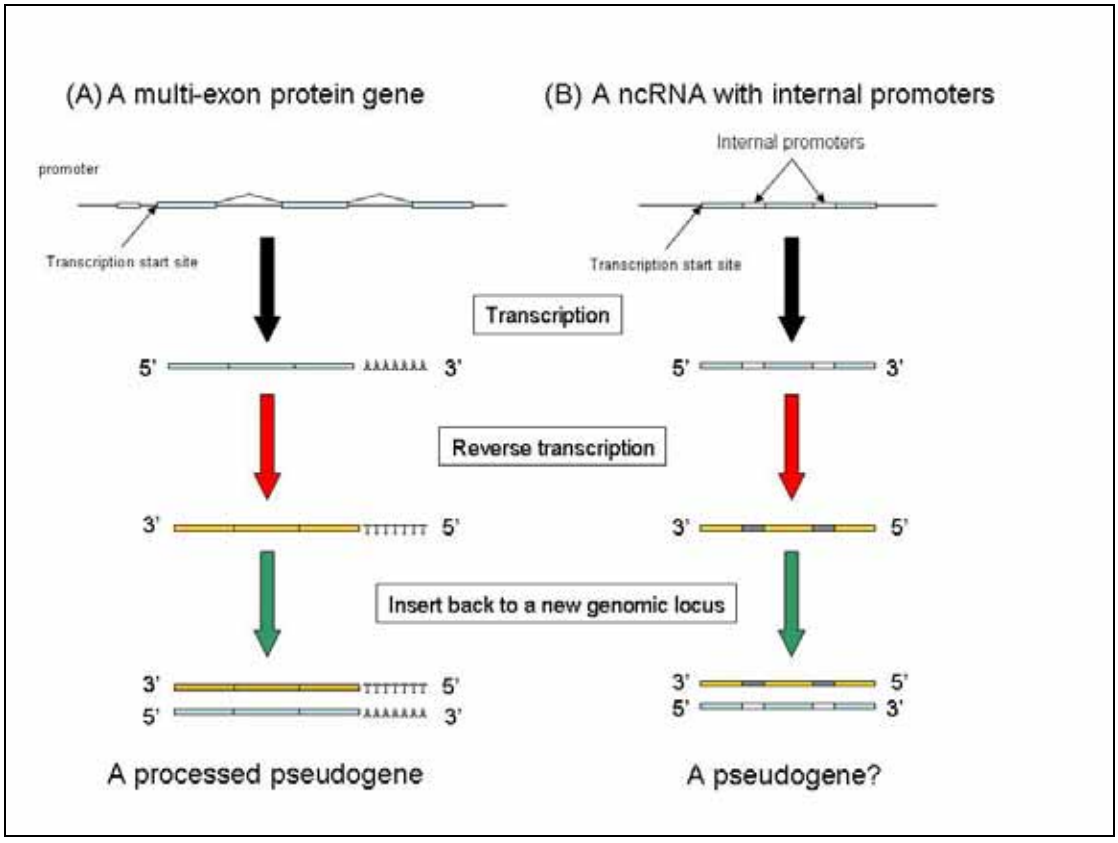


Figure 3-1. Comparison of the gene structures of a retrotransposed protein gene and a hypothetical retrotransposed ncRNA that contain internal promoters.

Therefore, this chapter is dedicated to distinguishing functional ncRNAs from ncRNA pseudogenes in the context of genomic sequences. There are two purposes in this chapter:

- To explore whether human synteny-non-conserved ncRNAs tend to be pseudogenes
- To evaluate novel rules that may be useful for distinguishing functional ncRNAs from ncRNA pseudogenes

Mammalian tRNA genes were chosen for further investigation. One reason for this decision is that many features of functional tRNA genes have been well studied. For example, a tRNA molecule can fold into a cloverleaf-like secondary structure; tRNA genes have internal promoters, which consist of A and B boxes (DeFranco et al. 1980); mammalian tRNA genes tend to cluster in the genomes (Lasser-Weiss et al. 1981). It was therefore hoped that, by integrating the information of sequence similarity, anticodon types, clustering, *etc*., evidence might possibly be found to determine if synteny-non-conserved tRNA genes in the mammalian genomes tend to be pseudogenes.

In the first part of this chapter (section 3.1), I investigate whether the human synteny-non-conserved tRNA genes that were retrieved from Rfam tend to be pseudogenes. The conservation of secondary structures and conservation of promoters, as well as conservation of primary sequences, were used to infer the functionality of the human synteny-non-conserved tRNA genes. The idea is that, if certain tRNA genes are pseudogenes, their sequences may have accumulated mutations which may change the features important for the functionality of tRNAs. The specific questions I address here include:

- Is there a clear-cut difference between the bit-score distributions of synteny-non-conserved tRNA genes and synteny-conserved tRNA genes?
- Do synteny-non-conserved tRNA genes tend to have more unstable structural features than synteny-conserved tRNA genes do?
- Do synteny-non-conserved tRNA genes tend to have degraded internal promoters?

A particular property of tRNA genes is that they frequently exist in synteny conserved clusters, as examined in chapter 2. In the second part of this chapter, I explore whether properties of copies of tRNA genes that are clustered and copies that are un-clustering are different and whether there is any evidence that can relate this to the likelihood of being

pseudogenes. Clustering seems to be an effective strategy to ensure each transcription unit can be accessed with generally equal probability by transcription machinery. Evidence suggests that clustering is important for regulating expression of ncRNAs. It has been demonstrated that clustered miRNA genes tend to be co-expressed (Baskerville and Bartel 2005). Besides, a cluster of 40 miRNA genes has been found in the human imprinted 14q32 domain and only the maternally inherited genes are expressed (Seitz et al. 2004).

I therefore hypothesized that non-clustered tRNA genes tend to be pseudogenes. Two tests were therefore designed to evaluate this hypothesis:

- Is there an enrichment of non-clustered tRNA genes in the low-scoring group which are more likely to be pseudogenes?
- Are clustered tRNA genes sufficient for covering 46 types of anticodons that are necessary for protein translation? If so, this would be evidence that non-clustered tRNA genes are not absolutely required for protein translation, supporting hypothesis that they could be pseudogenes.

3.1. Are Rfam synteny-non-conserved tRNA genes functional?

3.1.1. Materials and methods

The coordinates of human and mouse tRNA genes were retrieved from RFAMSEQ of Rfam 4.1 (Griffiths-Jones et al. 2003) and then converted to chromosomal coordinates in the human and mouse genomes respectively. The reference genome assemblies are human NCBI 33 and mouse NCBI M30. The bit scores of the Rfam tRNA genes were calculated using Infernal and the tRNA covariance model (CM) of Rfam 4.1 (Griffiths-Jones et al. 2003). The

human tRNA genes predicted using tRNAscanSE were retrieved from Ensembl release 19 by using the Ensembl Perl APIs (Birney et al. 2004).

In order to compare the bit-score distributions of the Rfam tRNA genes and the tRNAscanSE-predicted tRNA genes with that of *bona fide* tRNA genes, a trusted set of functional tRNA genes from the human genome is required. However, only a few experimentally verified human tRNA genes are available (Sprinzl and Vassilenko 2005). One consideration is that the bit-score distribution of a small number of tRNA genes may be biased and thus unsuitable for use as the reference distribution. Therefore, I decided to recruit Rfam tRNA genes that are human-mouse synteny-conserved as a trusted set of functional tRNA genes. Since synteny conservation has been widely accepted as a strong indication for the existence of functional elements, the human-mouse synteny-conserved tRNA genes are very likely to be functional tRNA genes. The sequences of these tRNA genes were prepared using the Ensembl Compara Perl APIs to search syntenic regions identified by Ensembl Compara release 19 (Clamp et al. 2003).

The preservation of structural features of tRNA genes was evaluated by using Infernal to align these sequences to Rfam tRNA CM. For the purpose of checking the conservation of the internal promoters in these tRNA genes, eufindtRNA (Pavesi et al. 1994) was used (for a brief introduction of Infernal and eufindtRNA, see materials and methods, section 2.1, chapter 2).

3.1.2. Results

3.1.2.1. Distribution of the Rfam bit scores of tRNA genes

808 human and 452 mouse tRNA genes were retrieved from Rfam (release 4.1). At first glimpse, it seems that there are more tRNA genes in the human genome than in the mouse genome; however, a substantial portion of the mouse genome assembly NCBI M30 is composed of sequences from whole genome shotgun sequencing, which has not been scanned

by Rfam 4.1. The number of the tRNA genes in the mouse genome is therefore an underestimate.

Interestingly, both the bit-score distributions of the human and the mouse tRNA genes sequences are bimodal (Figure 3-2, see "Rfam-human" and "Rfam-mouse" respectively). The bimodal bit-score distribution of the human tRNA genes seems to consist of two well-shaped distributions, which have modes at 65 and at 30 respectively.

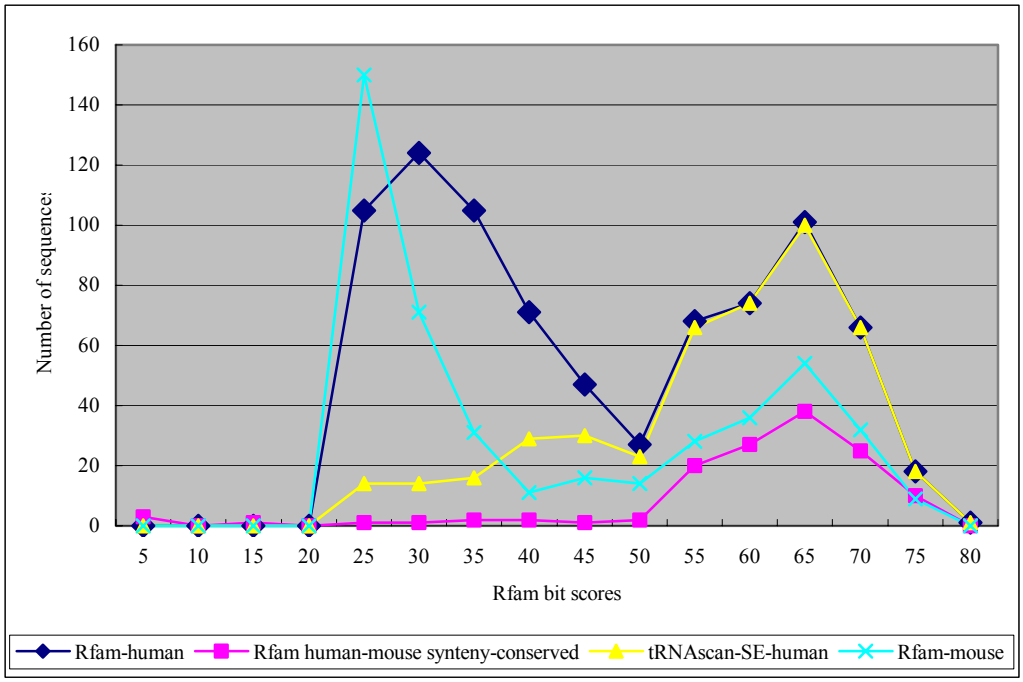


Figure 3-2. Distributions of Rfam bit scores of tRNA genes of different categories

The bin size of Rfam bit scores is 5. Almost no tRNA genes (except the human numt-tRNAs) have bit scores less than 25 because Rfam has used 25 bits as the gathering threshold for tRNA genes.

One interpretation of these results is that the bimodal distribution represents two groups of evolutionarily distinct tRNA genes. This idea is supported by the similarity between the high-scoring part of this bimodal distribution and the bit-score distributions of other sets of tRNA sequences. For example, the contour of the bit-score distribution of the tRNAscanSE-predicted human tRNA genes (Figure 3-2, "tRNAscanSE-human") is very similar to the high-scoring part of the bimodal bit-score distribution. In addition, the bit-score

distribution of the trusted set of *bona fide* tRNA genes (Figure 3-2, "Rfam human-mouse synteny-conserved") is also very similar. Only 9% (12/133) of the trusted *bona fide* tRNA genes have bit scores lower than 50. This comparison suggests that the high-scoring mode represents the bit-score distribution of human *bona fide* tRNA genes.

At this stage, this evidence is not convincing enough to conclude that the low-scoring tRNA genes are more likely to be pseudogenes. For example, the small bump in the distribution for "human-tRNAscanSE" within the range of 35 to 50 suggests that some *bona fide* tRNA genes may have bit scores indistinguishable from what are presumed to be tRNA pseudogenes (Figure 3-2, "tRNAscanSE-human"). In addition, the existence of a prominent low-scoring peak in the bit-score distribution of the tRNA genes predicted by Rfam does not really favour the hypothesis that "the low-scoring tRNA genes are pseudogenes". If the low-scoring tRNA genes are pseudogenes and the descendants of ancient functional tRNA genes, the random drifts caused by neutral mutations would be expected to result in a tail at the left side of the bit-score distribution, rather than generating an obviously bimodal distribution.

Consequently, I evaluated additional information, such as loss of primary-sequence and secondary-structure features, to look for additional evidence that low-scoring tRNA genes might be pseudogenes. Such information cannot be directly inferred from the bit scores of individual tRNA genes. An Rfam bit score for a particular ncRNA is actually a statistical evaluation of its degree of conservation at both primary-sequence and secondary-structure levels. It turns out that two factors can contribute to low bit scores for a tRNA gene: 1) the loss of the capability to fold into cloverleaf-like secondary structure; 2) the loss of the internal promoter which is required for being recognized by RNA polymerase III in order to generate functional tRNAs. These factors are further explored in subsections 3.1.2.2. and in 3.1.2.3. respectively.

3.1.2.2. Moderate preservation of secondary structures in the low-scoring and synteny-non-conserved tRNA genes

The number of non-canonical base pairs in Rfam tRNA predictions, as compared to a reference tRNA structure, is plotted. For the synteny-non-conserved tRNA genes with bit scores lower than 50, the mode of the number of non-canonical base pairs that may make the secondary structures unstable is 3 and the average is ~5 (Figure 3-3). In other words, for a low-scoring tRNA gene, there is on average slightly more than 1 non-canonical base pair per stem region (*i.e.* 4 stems in a tRNA molecule in its functional form).

However, even for the synteny-conserved tRNA genes which are more likely to be *bona fide* tRNA genes, the mode is 2 non-canonical base pairs in their stem regions (Figure 3-3) and the average is 2.6. This suggests that for one stem region of a tRNA, one non-canonical base pair can still be tolerated and its secondary structure can still be preserved. The evidence suggests that there is moderate preservation of structural features in the low-scoring tRNA genes and a moderate level of non-canonical base pairs may be tolerated. The degree of loss of structural features provides only limited support for the view that these low-scoring tRNA genes tend to be pseudogenes.

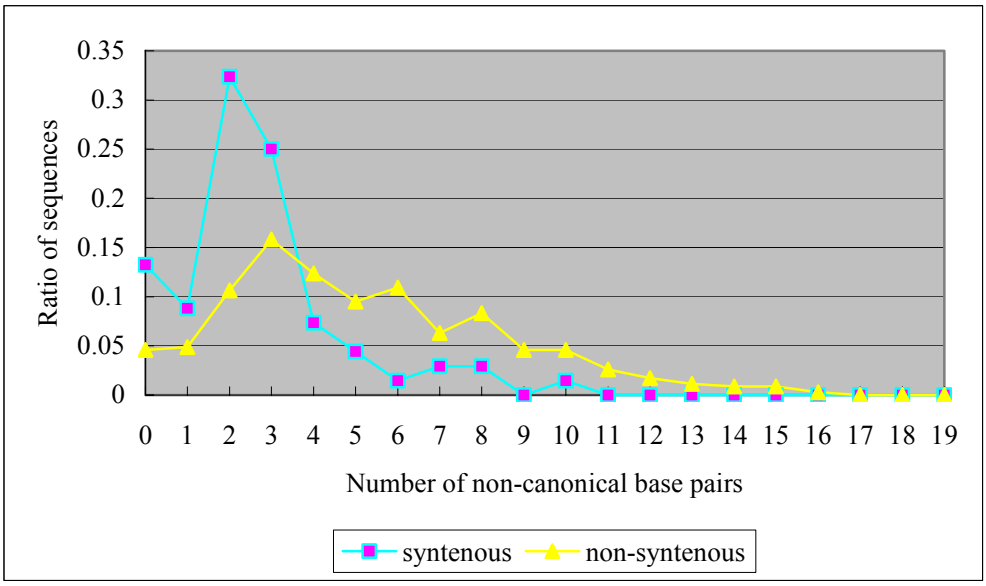


Figure 3-3. Distributions of numbers of the non-canonical base pairs in human tRNA genes

The synteny-conserved and the synteny-non-conserved tRNA genes are aligned to the tRNA consensus structures by using Infernal and the Rfam tRNA CM. Non-canonical base pairs that may destabilize the secondary structures of these tRNA genes are counted, except that G-U base pairs are tolerated.

3.1.2.3. Degradation of the internal promoters in the low-scoring tRNA genes

The genomic loci containing tRNA genes need to be transcribed into tRNA molecules in order to function in cells. If these low-scoring tRNA genes are not transcribable, they are pseudogenes. In order to be transcribable a functional promoter is required. The internal promoters of the tRNA predictions were evaluated using the eufindtRNA algorithm (see methods in subsection 2.2.1.6 of the materials and methods of section 2.2). Previously in subsection 2.2.3.4 in chapter 2, two cases of promoter-specific degradation of synteny-non-conserved tRNAscanSE-predicted tRNA gene loci were found. Here, pseudogenization through promoter-specific degradation is investigated more generally in synteny-non-conserved and low-scoring Rfam tRNA genes.

The results reveal that, about three-quarters (339/441) of the low-scoring tRNA genes do not have intact promoters in their intragenic regions. According to current knowledge, these low-scoring tRNA genes in the human genome cannot be transcribed into tRNAs by

eukaryotic RNA polymerase III. This is good evidence which indicates that the set of low-scoring tRNA genes is enriched with pseudogenes. This result suggests that in the human genome there is a group of tRNA-related pseudogenes, where their internal promoters are degraded, while their secondary structures are moderately conserved.

3.1.2.4. Tracing the evolutionary origins of low-scoring tRNA genes

The finding that the majority of low-scoring tRNA genes appear to have more significantly degraded internal promoters than secondary structures and may be pseudogenes, suggests the hypothesis that mutations that degrade internal promoters have a selective advantage in mammalian evolution. It seems possible that degradation of internal promoters might be the most effective mechanism for disabling tRNA genes, since aberrant tRNA genes with mutations that make RNA secondary structures unstable would be still transcribable and lead to abnormal protein translation and damage the cell.

If selective degradation of synteny-non-conserved tRNA genes were an important mechanism in the human evolution, it would be reasonable that the human genome would contain numerous tRNA genes which have lost functional promoters, but not yet lost their secondary structures. In order to test this hypothesis, it was proposed to demonstrate that random mutations are unlikely to generate tRNA genes, where their internal promoters have degraded and structural features are still moderately preserved.

Consequently, a simulation, where a random mutation model is applied to the ancestors of these low-scoring tRNA genes, was planned. The initial step for preparing this simulation was to find an appropriate ancestral sequence for each low-scoring tRNA gene. The considerations for finding the ancestral sequences of these low-scoring tRNA genes are discussed in the following two subsections (3.1.2.4.1. and 3.1.2.4.2.).

3.1.2.4.1. Weak evolutionary relation of low-scoring tRNA genes with bona fide *human tRNA genes*

A sensible conjecture is that the ancestral sequences of the low-scoring tRNA genes are *bona fide* human tRNA genes. According to the discussions above (for details see subsections 3.1.2.1. , 3.1.2.2. , and 3.1.2.3.), it is conceivable that *bona fide* tRNA genes are enriched in the sets of human-mouse synteny-conserved tRNA genes, the tRNAscanSE-predicted high-scoring tRNA genes, and the tRNA genes in manually-curated tRNA repositories. However, the search for the evolutionary origins of the low-scoring tRNA genes proved difficult. Using WU-BLAST a possible ancestor could be found for less than one-quarter (101/441) of the low-scoring tRNA genes. In addition, less than half of the low-scoring tRNA genes were found to have homologous sequences in the sets of tRNAscanSE-predicted human tRNA genes and of tRNA compilation (Sprinzl et al. 1998).

3.1.2.4.2. Strong evolutionary relation of low-scoring tRNA genes with mitochondrial tRNAs

Because of the failure to find the ancestral sequences for the majority of the low-scoring tRNA genes from the set of *bona fide* human tRNA genes, it was necessary to consider other sources of tRNA genes that might be the evolutionary ancestors of the low-scoring tRNA genes. In eukaryotic cells, the nuclear genome is not the only sequence that contains tRNA genes. Some intracellular organelles, such as mitochondria and chloroplasts, have their own tRNA genes in their organelle genomes. The tRNA genes of these organelles are divergent, at the primary-sequence level, from the vertebrate nuclear tRNA genes. They are another possible origin of the low-scoring tRNA genes.

The sequences of the low-scoring tRNA genes were searched against the genomic sequence of the human mitochondrion (GenBank accession number: NC_001807.4), and better matches were found to human mitochondrial tRNA genes than to trusted tRNA genes in many cases (human-mouse synteny-conserved tRNA genes) (Table 3-1). In addition, 239 of the sequences that did not appear to have any homologous sequence in the set of human tRNA

genes matched human mitochondrial tRNA genes. The average identity of the 280 tentative nuclear mitochondrial tRNA sequences (numt-tRNAs) to human mitochondrial tRNA genes is 84.8%. The average coverage of these alignments to the full length of the mitochondrial tRNA genes is 85.3%. The evidence strongly suggests that many low-scoring tRNA genes in the human nuclear genome are derived from the human mitochondrial tRNA genes, and not from the tRNA genes in the human nuclear genome.

More similar to the human nuclear tRNA genes	128 (29%)
More similar to the human mitochondrial tRNA genes	280* (64%)
None	33 (7%)
All the human low-scoring tRNA genes	441 (100%)

Table 3-1. Numbers of the human low-scoring tRNA genes which are more similar to either the human nuclear tRNA genes or the human mitochondrial tRNA genes.

"None" is used to indicate the low-scoring tRNA genes which are not significantly similar to either human nuclear tRNA genes or mitochondrial tRNA genes. "*" indicates that 239 out of the 280 low-scoring tRNA genes do not have homologous sequences in the set of human tRNA genes.

For the 128 human tRNA genes that are more similar to human nuclear tRNA genes than to mitochondrial tRNA genes, 71.9% (92/128) of them were recognised using eufindtRNA. This means that the majority of human-nuclear-tRNA-derived low-scoring tRNA sequences still preserve their internal promoters to a certain extent. Consequently, the hypothesis which asserts that there might be selection for mutations that degrade the promoters of the tRNA genes in mammals does not appear to apply to tRNA genes derived from other human tRNA genes.

3.1.2.5. Searching for nuclear mitochondrial tRNAs in mammalian genomes

3.1.2.5.1. Finding nuclear mitochondrial tRNA sequences in the human genome

Since the Rfam tRNA CM (covariance model) is not specifically trained for finding nuclear mitochondrial tRNA sequences (numt-tRNAs) in the human genome, there may be

other human numt-tRNAs which were not identified by Rfam. In order to discover as many numt-tRNAs as possible, blastz and the human mitochondrial genome were used to search for nuclear mitochondrial sequences (numt-seqs) in the whole human genome (NCBI 33). Blastz was used since it is well tuned for aligning genomic sequences (Schwartz et al. 2003).

177 human genomic loci were found to be similar to mitochondrial sequences. Many loci contain more than one nuclear mitochondrial genes (numt-genes). The arrangements of mitochondrial genes in these loci are mostly consistent with those of the real mitochondrial genes encoded in the human mitochondrial genome. It is therefore reasonable to infer that the numt-genes of each locus have been co-transferred into the nuclear genome. There are 627 numt-tRNAs in the 177 human loci of numt-seqs. The average identity between these numt-tRNAs and the human mitochondrial tRNA genes is 84.5%. The average coverage of these alignments to the full-length mitochondrial tRNA genes is 85.3%. None of the 627 tRNA genes overlap with known repetitive elements except tRNAs. Only 30 out of the 627 sequences were found to have homologous sequences in the set of trusted *bona fide* human tRNA genes (human-mouse synteny-conserved tRNA genes). By using eufindtRNA, only 33 out of the 627 sequences were found to have RNA Pol III promoters. The discovery of human numt-tRNAs could explain the low-scoring mode in the bimodal score distribution of the human tRNA genes identified by Rfam well (Figure 3-4, human numt-tRNAs). Although the curve for numt-tRNAs does not fit exactly with the low-scoring group of the bimodal distribution, it is almost parallel.

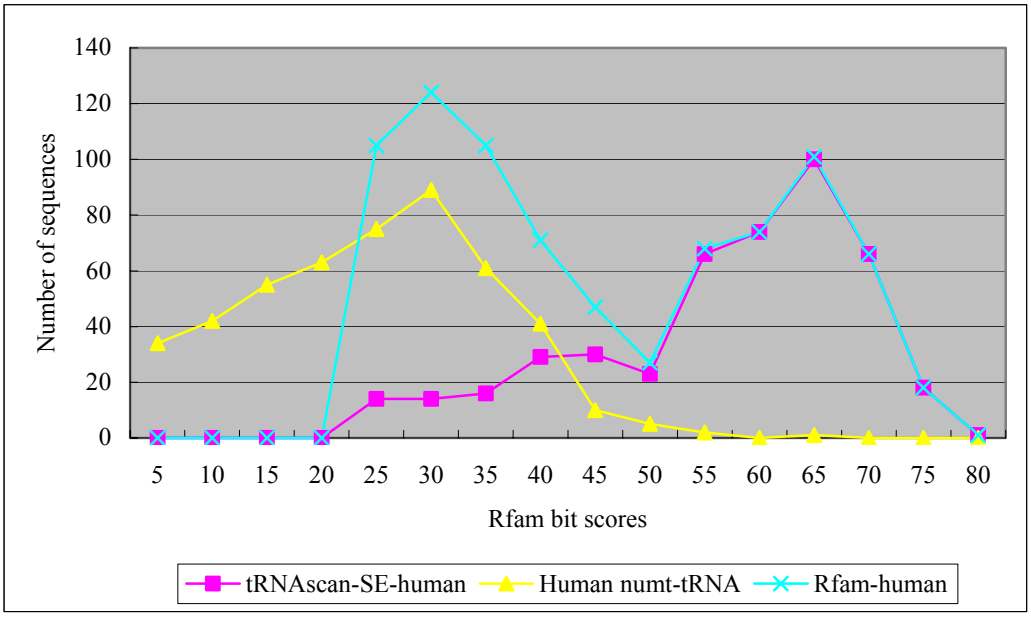


Figure 3-4. Distributions of Rfam bit scores of tRNA genes of human-numt, Rfam-human, and tRNAscanSE tRNA genes.

3.1.2.5.2. Few numt-tRNAs in the mouse genome

Following the discovery of numt-seqs related sequences in the human genome the same analysis was repeated for the mouse genome. In contrast to the discovery of numerous numt-tRNAs in the human genome, far fewer numt-tRNAs could be found in the mouse genome. The bit-score distribution of the mouse low-scoring tRNA genes is obviously different from that of the human low-scoring tRNA genes (Figure 3-2, Rfam-mouse). 86% (217/252) of the mouse low-scoring tRNA genes from Rfam 4.1 are SINEs. Surprisingly, only 64 numt-tRNAs were found in the mouse genome assembly NCBI M30. Not only is the number of numt-seqs smaller than that in the human nuclear genome, but also the average length for each locus of integration is shorter. There are on average 1.7 numt-tRNAs per locus of mouse numt-seq (64 numt-tRNAs / 38 loci), while there are on average 3.5 numt-tRNAs per locus of human numt-seq (627 numt-tRNAs / 177 loci).

There are various hypotheses that might explain the difference between the numbers of numt-tRNAs in the human genome and in the mouse genome. However, before designing

strategies to test these hypotheses, the effect of the quality of the mouse genome assembly on identifying numt-seqs needs to be addressed. Unlike the high coverage of clone-based sequences used in the current human genome assembly, the mouse genome assembly NCBI M30 consists of sequences from both whole genome shotgun (WGS) and high throughput genome sequencing (HTGS). One limitation of WGS sequence assembly is its inability of resolving duplicated regions. If there were numerous recent integrations of the mitochondrial genomic sequence into the mouse nuclear genome, it is possible that the numt-seqs could still be quite similar to one another and thus inappropriately collapsed by WGS sequence assembly. In order to confirm that there is a significant difference between the numbers of the numt-seq loci in the human and mouse genomes respectively, the latter value should be reassessed in the future when more clone-based sequences are used in the mouse genome assembly.

3.1.2.5.3. Effects of numt-tRNAs on finding mammalian tRNAs

The presence of numt-seqs in the human genome has not been considered in the annotation of the human genome. For example, at least five tRNAscanSE-predicted tRNA genes were found within regions of numt-seqs in the human genome. It is unknown whether human numt-tRNAs can be transcribed into functional tRNAs in human cells (for further discussion see subsection 3.1.2.6.). Numt-seqs are also frequently ignored in annotations provided by public-domain genome databases. Unlike the annotation of repetitive elements, consideration of numt-seqs is not part of the procedure in pipelines of genome annotation. In addition, most of the mitochondrial genes are not included in the current release of RepBase (released on 10/09/2004) and there are only two mitochondrial tRNA genes from *G. gallus* in RepBase.

3.1.2.6. Are numt-tRNAs functional?

The existence of numt-seqs in the nuclear genome has been known for some time (Tsuzuki et al. 1983), and their evolutionary dynamics have been discussed in a number of

papers (Mourier et al. 2001; Tourmen et al. 2002; Woischnik and Moraes 2002; Hazkani-Covo et al. 2003; Ricchetti et al. 2004). Most related research suggests that nuclear mitochondrial protein-coding genes (numt protein-coding genes) are pseudogenes. One important factor is that the genetic code of the genes encoded in mitochondrial genomes is different from that of the genes encoded in nuclear genomes. Presumably numt protein-coding genes cannot be translated into functional proteins.

In contrast, the functions of numt-tRNAs have never been explicitly discussed. The arguments, which have been used to infer that numt protein-coding genes should be pseudogenes, may not be applicable to the case of numt-tRNAs. The functions of numt-tRNAs do not depend on being translated into proteins. Numt-tRNA genes could be functional if they were transcribed into tRNA molecules. The following two subsections (3.1.2.6.1. and 3.1.2.6.2.) are therefore dedicated to finding evidence to support the hypothesis that human numt-tRNAs were initially functional while other nuclear mitochondrial sequences (non-tRNA numt-seqs) lost functions upon integration of numt-seqs into nuclear genomes.

3.1.2.6.1. Comparing patterns of mutations of numt-tRNAs and non-tRNA numt-seqs

In order to investigate the possibility that numt-tRNAs were once functional, the patterns of mutation in numt-tRNAs and other non-tRNA numt-seqs were compared. The hypothesis is that, in order to protect the organism from the deleterious effects of transcripts of numt-tRNAs, mutations that disable these genes would accumulate more rapidly than in non-tRNA numt-genes which might be expected to be inactive upon initial insertion. In other words, differences between the patterns of mutations in numt-tRNAs and in non-tRNA numt-seqs might be considered as evidence that either numt-tRNAs or non-tRNA numt-seqs were once functional.

By aligning various human numt-seqs to the human mitochondrial genome, numbers of mutations in human numt-tRNAs and in human non-tRNA numt-seqs were counted separately.

Unexpectedly, on average numt-tRNAs were found to be slightly more conserved than other non-tRNA numt-seqs (Figure 3-5). This result suggests that while evolutionary pressures on human numt-tRNAs and human non-tRNA numt-seqs may be different; overall human numt-tRNAs are not degraded faster than human non-tRNA numt-seqs. In addition, there is no obvious difference between the substitution patterns of the numt-tRNAs and the non-tRNA numt-seqs (Figure 3-6).

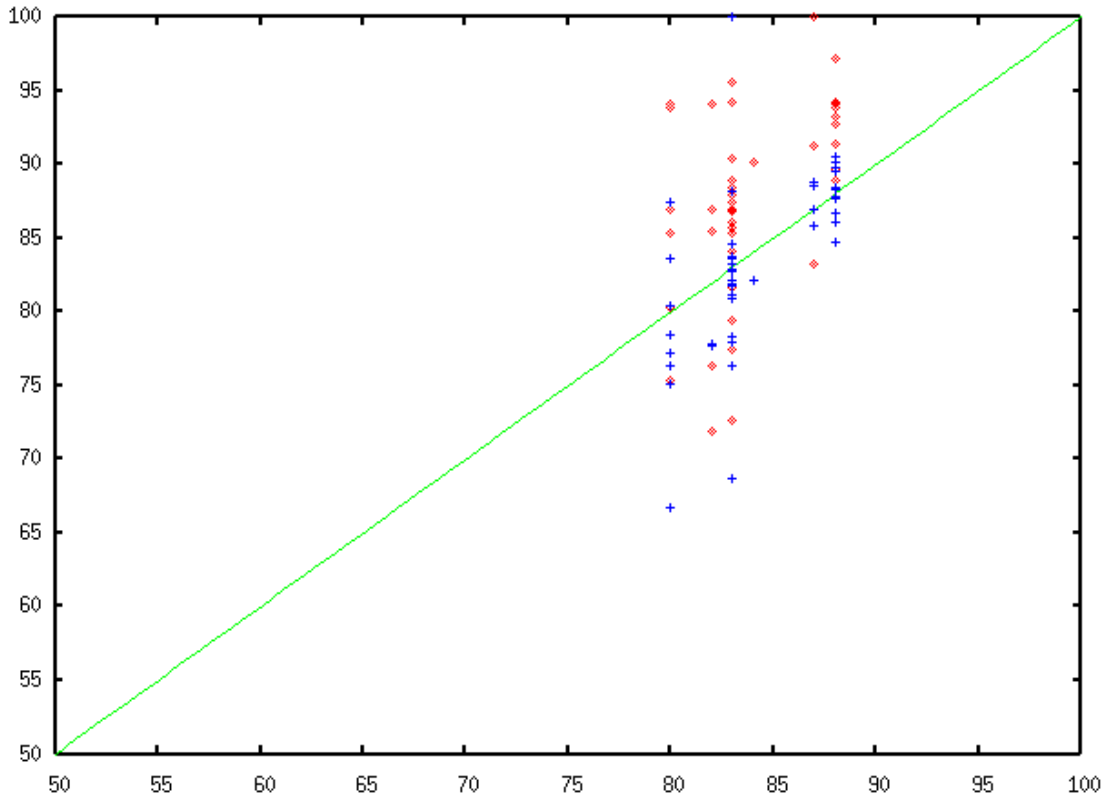


Figure 3-5. Distribution of identities of human numt-tRNAs and human non-tRNA numt-seqs in 80-90 percent identity regions to the human mitochondrial genome

The red points indicate numt-tRNAs and the blue crosses indicate non-tRNA numt-seqs. The green line is the diagonal line (x=y). Numt-tRNAs and non-tRNA numt-seqs were separated from all numt-seqs (found by using blastz) with 80-90 percent identities to the human mitochondrial genome. There are 43 numt-tRNAs and 43 non-tRNA numt-seqs in this plot. The y-axis is the identities of numt-tRNAs or non-tRNA numt-seqs to their corresponding human mitochondrial genes. The x-axis is the identities to the human mitochondrial genome for respective numt-seqs, in which the numt-tRNAs or non-tRNA numt-seqs are embedded.

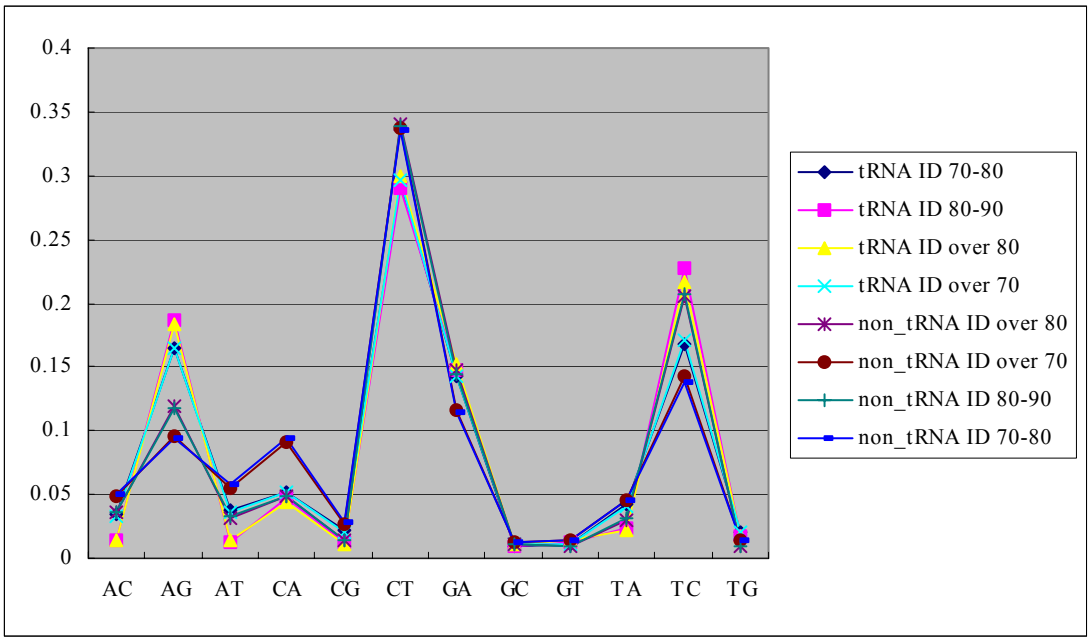


Figure 3-6. Patterns of substitution in the human numt-tRNAs and in the human non-tRNAs embedded in regions with different percent identities to the human mitochondrial genome

"tRNA ID 70-80" indicates the numt-tRNAs embedded in regions with 70-80 percent identities to the human mitochondrial genome and so forth. In the x-axis, "AC" means the base adenosine being substituted with the base cytosine in numt-seqs, and so forth. The y-axis is the normalized ratio of substitutions (*i.e.* number of each type of substitutions normalized by total number of substitutions in each category of numt-tRNAs or non-tRNA numt-seqs).

3.1.2.6.2. Uneven distribution of mutations along human numt-tRNAs

Although the previous results show the overall mutation rate of numt-tRNAs is lower than for non-tRNA numt-seqs, I decided to investigate the distribution of mutations along numt-tRNAs sequences. Given that tRNAs contain internal regulatory elements that promote their transcription, if mutations in numt-tRNAs were found preferentially in positions that could effectively degrade these elements, this would support the hypothesis these numt-tRNAs had initially been active, but subsequently inactivated. Previously counted mutations from alignments between numt-tRNAs and the human mitochondrial genome were therefore counted in bins along the consensus numt-tRNA sequence. The 95% confidence interval for each bin was estimated based on the beta distribution, assuming that the number of mutations was α and that the number of bases in each bin was the sum of α and β.

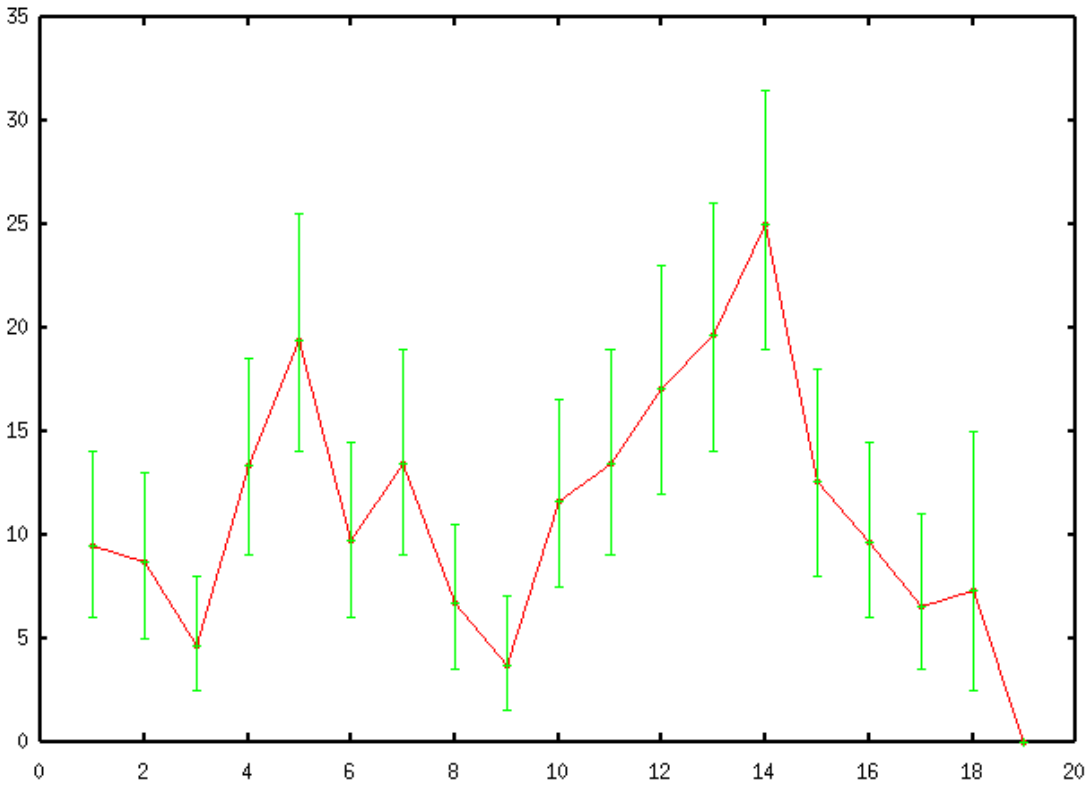


Figure 3-7. Distribution of mutation numbers along human numt-tRNAs

X-axis is the bins along human numt-tRNA sequences. Y-axis is the number of total mutations in each bin. The bin size is 4 bases in length. Forty-three numt-tRNAs are extracted from the numt-seqs with 80 to 90 percent identities to the human mitochondrial genome. The mutations for the first 4 bases for the recruited numt-tRNAs are summed up to give the number of mutations in the first bin and so forth. The green bars are the 95% confidence intervals for bins.

Interestingly, two regions, the 16th to 19th (bin 5) and 52nd to 55th (bin 14) nucleotides, were found to contain significantly more mutations than the 28th to 35th (bin 8 and 9) nucleotide (Figure 3-7). The 95% confidence intervals of mutations for the former two regions do not overlap with those for the 28th to 35th nucleotides. The locations of these two regions are consistent with the positioning of A and B boxes in the nuclear tRNA genes (DeFranco et al. 1980; Galli et al. 1981).

In numt-tRNAs there are significantly more mutations in the positions that correspond to known regulatory regions of human tRNAs and the tRNA promoter finding algorithm eufindtRNA fails to find sequences that score well as promoters. These results might appear consistent with the hypothesis that numt-tRNAs were initially functional when copied into the

mammalian nuclear genomes, but have since become pseudogenes as a result of promoter degradation through selective acceptance of mutations. Unfortunately, proof of this hypothesis needs additional evidence. For example, the mechanism of expression of tRNAs in the mitochondria is different to that of human tRNAs. There is also no evidence to show that expression of mitochondrial tRNAs in the cytoplasm would interfere with the protein synthesis of the genes encoded in nuclear genomes. There are no papers dealing specifically with the fidelity of terminal maturation, aminoacylation, and roles in protein translations if the mitochondrial pre-tRNA transcripts are in the cytoplasm.

3.1.3. Discussion

These results presented in this section (section 3.1) suggest that the 64% of the human synteny-non-conserved tRNA genes retrieved from Rfam are nuclear mitochondrial tRNA genes (numt-tRNAs), whose ancestors are tRNAs in the human mitochondria. With the investigations performed in the previous subsections, these numt-tRNAs should be untranscribable pseudogenes. The pattern of mutation in these numt-tRNAs is interesting and suggestive of pseudogenisation through promoter inactivation. By contrast, the vast majority of the remaining low scoring synteny-non-conserved tRNA genes retrieved from Rfam have sequence similarity to synteny-conserved tRNA genes and ~72% are recognised using eufindtRNA suggesting they have intact promoters and may not be pseudogenes (see subsection 3.1.2.4.2.).

The bit-score distribution appears to be only weakly useful in distinguishing functional tRNA genes from tRNA pseudogenes. The bimodal bit-score distribution observed for low scoring synteny-non-conserved tRNA genes was mainly the result of the special case of numt-tRNAs, however when these were removed any relationships became unclear. This is consistent with the bit-score distributions of other classes of Rfam ncRNAs, where no

particular pattern can be found. With a bit-score distribution that is simply single-modal and heavy-tailed, such as in the case of human U6 snRNA genes identified by Rfam 4.1 (Figure 3-8), it is difficult to choose any clear-cut threshold that might separate functional and non-functional genes. Although ncRNA sequences with higher bit scores are more likely to be synteny-conserved and functional genes, whether ncRNA sequences with lower scores are functional or not cannot be unambiguously determined. Similarly there is little evidence that an ncRNA gene with synteny-non-conserved status is necessarily a pseudogene.

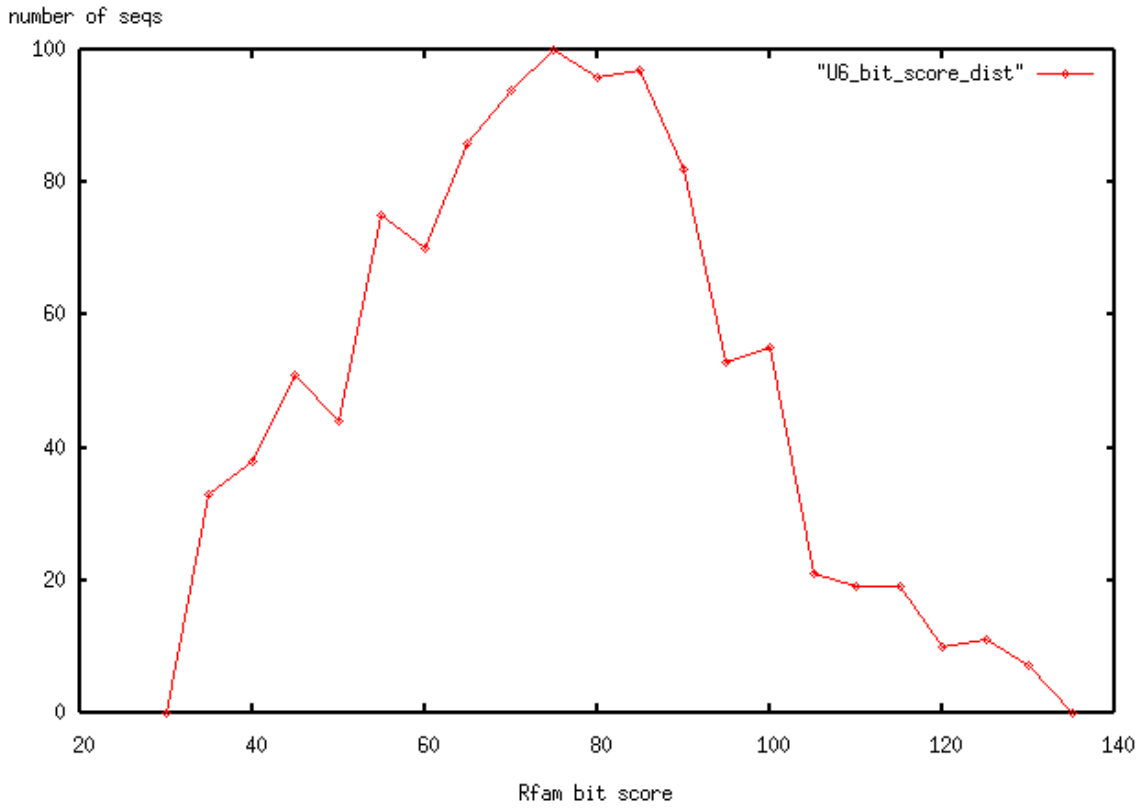


Figure 3-8. Distribution of the Rfam bit scores of the human U6-like sequences identified by Rfam 4.1

The heavy-tailed distributions suggest that, for many classes of ncRNAs in mammalian genomes, the generation of pseudogenes may be a continuous process. It seems that abundant ncRNA pseudogenes in mammalian genomes do not have a strong negative effect on the fitness of organisms. While this is good news for the survival of mammals, it also means that bit score distributions cannot be very helpful in filtering out ncRNA pseudogenes in ncRNA

finding. More specific signals are necessary for distinguishing *bona fide* ncRNAs from ncRNA pseudogenes. One such signal might be whether an ncRNA retains a recognisable internal promoter, however verification of the computational evidence presented here is needed.

3.2. Clustering – a useful criterion for filtering out ncRNA pseudogenes?

3.2.1. Materials and methods

3.2.1.1. Recruiting human and mouse tRNA genes

The human and mouse tRNA genes used in this section were retrieved from Ensembl release 29 by using Ensembl Perl APIs. These genes were predicted by using tRNAscanSE.

3.2.1.2. Defining tRNA-gene clusters

In assessing the features of clustered tRNA genes, one issue concerns deciding a suitable distance criterion, *i.e.* the maximal distance allowed between the nearest neighbouring tRNA genes, for defining tRNA gene clusters. If the selected distance is longer than necessary, more potentially non-clustered tRNAs may be included into clusters. On the other hand, if the selected distance is too short, some clustered *bona fide* tRNA genes may be incorrectly grouped or classified as non-clustered. Several different distances, such as 5-kilo bases and 10-kilo bases, were therefore tried to define tRNA-gene clusters.

3.2.1.3. Comparing the ratios of non-clustered tRNA genes within different bit-score ranges

All human tRNA genes are categorized into five bins according to their bit scores: 20-55, 56-65, 66-75, 76-85, and 86-95. The ratio of tRNA genes that are clustered was calculated separately for each bin. The enrichment of clustered tRNA genes in each bin is determined by comparing the ratios in different bins. The 95% confidence intervals for individual ratios were

estimated based on the beta distribution, assuming that each numerator was α and that each denominator was the sum of α and β.

3.2.1.4. The anticodons required for protein translation

It is known that not all 61 types of anticodons are required for protein translation in eukaryotic cells. Because the interactions between codons and anticodons allow wobble pairs in the third positions (of codons), some codons can share recognition by the same tRNA. Guthrie and Abelson estimated that 46 types of tRNAs that have 45 unique anticodons are sufficient for translation (for review see Guthrie and Abelson 1982). Two types of tRNAs with exactly the same anticodon are used for carrying Met_m and Met_i respectively ("i" indicates translation initiation codon "m" indicates a general non-initiation codon for methionine).

3.2.2. Results

3.2.2.1. Enrichment of mammalian non-clustered tRNA genes in the low-scoring group

A 10-kb distance threshold was initially used to subgroup all human tRNA genes into clustered and non-clustered ones. Among the 608 human tRNA genes predicted by tRNAscanSE, ~65% (125/192) of the tRNA sequences with scores 20-55 were found to be non-clustered. By contrast, ~27% (16/59) with scores 55-65, ~30% (45/152) with scores 65-75, ~25% (42/171) with scores 75-85, and ~26% (9/35) with scores 85-95, are non-clustered (Figure 3-9). These results suggest that non-clustered tRNA genes are enriched in the low-scoring group. There is also a similar finding when the clusters were defined by using the 5-kb distance threshold (data not shown).

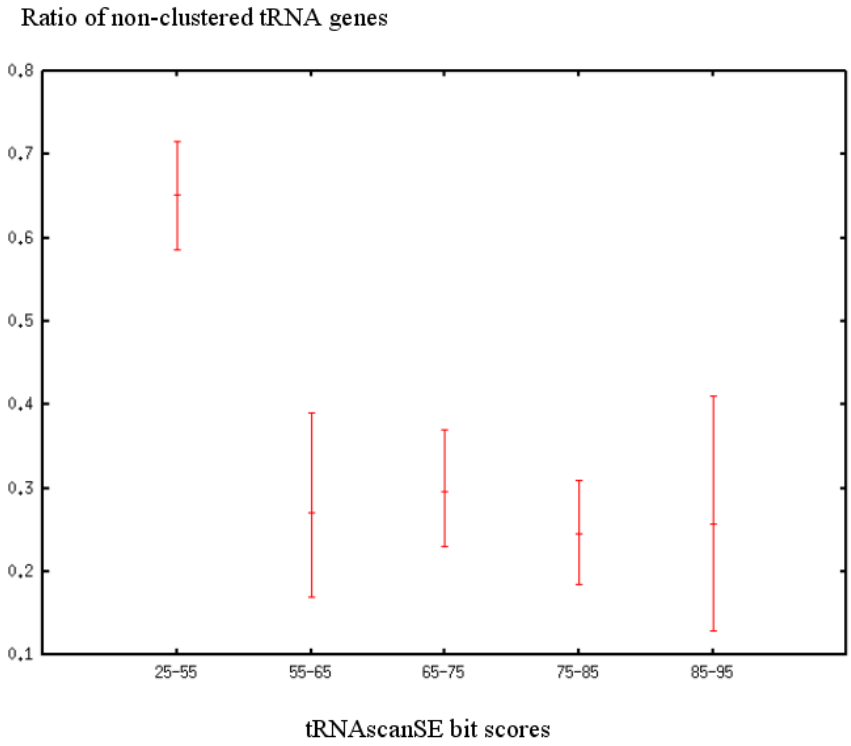


Figure 3-9. The human low-scoring tRNA genes are enriched with non-clustered ones

Each red bar is the 95% confidence interval for each bin. The confidence intervals shown here were estimated as described in subsection 3.2.1.3.

3.2.2.2. The mammalian clustered tRNA genes can cover 46 necessary anticodons

In this subsection, the functionality of non-clustered tRNA genes is explored indirectly on the basis of the need for their roles in protein translation. If clustered tRNA genes are shown not to include all the anticodons required for protein translation, this will be evidence that non-clustered tRNA genes are necessarily functional. Conversely, if clustered tRNA genes provide all the required anticodons, non-clustered tRNA genes may not be necessarily required for protein translation.

These results indicate that clustered tRNA genes in the human genome can cover all 46 types of tRNAs and exactly satisfy the wobble rules (Table 3-2, compare “yeast” and “clustered” ones). Although, the human non-clustered tRNA genes can also cover 46 types of tRNAs, there are several cases that violate the wobble rules (Table 3-2, compare “yeast” and

"non-clustered" ones). Besides, in the mouse clustered tRNA genes, additional anticodons were found (Table 3-3). These results suggest that the clustered tRNA genes in mammalian genomes may be sufficient to provide the necessary types of tRNAs for translating proteins.

tRNA types	Yeast	Human				
		All	Clustered, dist < 10kb	Clustered, dist < 6kb	Non-clustered, dist < 10kb	Non-clustered, dist < 6kb
Ala	3	3	3	3	3	3
Arg	5	5	5	5	5	5
Asn	1	2	1	1	2	2
Asp	1	1	1	1	1	1
Cys	1	1	1	1	1	1
Gln	2	2	2	2	2	2
Glu	2	2	2	2	2	2
Gly	3	3	3	3	2	2
His	1	1	1	1	1	1
Ile	2	2	2	2	2	2
Leu	5	5	5	5	5	5
Lys	2	2	2	2	2	2
Met*	2	2	2	2	2	2
Phe	1	1	1	1	1	1
Pro	3	3	3	3	1	1
Ser	4	4	4	4	4	4
Thr	3	3	3	3	3	3
Trp	1	1	1	1	1	1
Tyr	1	2	1	1	2	2
Val	3	3	3	3	3	3
Total	45	47	45	45	45	45

Table 3-2. Comparison between types of anticodons of yeast and the human tRNAs

Each number indicates the distinct types of tRNA anticodons corresponding to a particular amino acid. For example, there are 2 distinct types of anticodons found in the yeast tRNA genes corresponding to the tRNAs carrying isoleucine (Ile). Each red box is used to indicate that for a particular amino acid, the number of corresponding anticodon types that can be found in a category (clustered, non-clustered, *etc.*) of human tRNA genes is different from that of the anticodon types found in yeast tRNA genes.

"*" means that there are two types of tRNAs with exactly the same anticodon for Met_i and Met_m respectively.

tRNA types	Yeast	Mouse			
		All	Clustered, dist < 1 mb	Clustered, dist < 10 kb	Clustered, dist < 6 kb
Ala	3	4	4	3	3
Arg	5	6	5	5	5
Asn	1	2	1	1	1
Asp	1	2	1	1	1
Cys	1	2	2	1	1
Gln	2	2	2	2	2
Glu	2	2	2	2	2
Gly	3	4	4	4	4
His	1	2	2	1	1
Ile	2	3	3	2	1
Leu	5	6	5	5	5
Lys	2	2	2	2	2
Met*	2	2	2	2	2
Phe	1	2	2	1	1
Pro	3	4	3	3	3
Ser	4	6	4	4	4
Thr	3	4	3	3	3
Trp	1	1	1	1	1
Tyr	1	2	1	1	1
Val	3	4	4	4	4
total	46	62	53	48	47

Table 3-3. Comparison between types of anticodons of yeast and mouse tRNAs

The color-coding convention used in this table follows that of Table 3-2.

The anticodon types of non-clustered mouse tRNA genes were not listed. The types of anticodons that can be found in non-clustered mouse tRNA genes exceed the essential types of anticodons (the column "yeast"). It is difficult to determine which of them may not be the anticodons of *bona fide* mouse tRNA genes. The purpose of this table is thus to demonstrate that clustered mouse tRNA genes can cover the anticodons essential for protein translation.

"*"means that there are two types of tRNAs with exactly one anticodons for Met_i and Met_m respectively.

3.2.3. Discussion

3.2.3.1. Clustering may be a useful criterion for filtering out tRNA pseudogenes

Three threads of evidence imply that maybe the clustered tRNA genes in the mammalian genomes are functionally more important than the non-clustered tRNA genes are. First, the human low-scoring tRNA genes, which are more likely to be pseudogenes, are significantly enriched with non-clustered tRNA genes. Second, the finding that clustered tRNA genes should be sufficient for protein translation implies that non-clustered tRNA genes may not necessarily be required for protein translation. Third, ~56% of human clustered tRNA genes are human-mouse synteny-conserved, while only ~40% of human non-clustered tRNA genes are human-mouse synteny-conserved (for details see section 2.2 and Figure 2-7).

3.3. Summary

In the first part of this chapter (section 3.1), I explored the tendency of the synteny-non-conserved tRNA genes retrieved from Rfam to be pseudogenes. Results relevant to genome-wide ncRNA finding include that:

- ~65% of human synteny-non-conserved tRNA genes retrieved from Rfam are nuclear mitochondrial tRNA sequences (numt-tRNAs).
- Evidence suggests that these numt-tRNAs are currently non-functional in the human genome. The observed patterns of mutation are weakly suggestive of a mechanism of pseudogenisation that involves promoter inactivation.
- Once numt-tRNAs were disregarded, it was apparent that many of the remaining low-scoring synteny-non-conserved tRNA genes might not necessarily be pseudogenes.

In the second part of this chapter (section 3.2), I explored the functionality of human

non-clustered tRNA genes. The main results are that:

- Low-scoring tRNA genes are enriched with non-clustered tRNA genes.
- Mammalian clustered tRNA genes can provide sufficient types of tRNAs to cover all the anticodons required for protein translation. This is consistent with non-clusters tRNA genes not needing to be functional, but does not demonstrate that they are non-functional.

With respect to the functionality of synteny-non-conserved ncRNAs in mammalian genomes, there are two hypotheses. In the following, I summarize the pieces of evidence for or against each of these:

1. Hypothesis: synteny-non-conserved ncRNA genes are pseudogenes.
   - Evidence against this hypothesis:
     - The majority (71.9%) of human nuclear tRNA derived low-scoring and synteny-non-conserved (Rfam) tRNA sequences still preserve their internal promoters to a certain extent (see subsection 3.1.2.4.2.). They may not be functional tRNA genes but may be transcribable.
     - Some synteny-non-conserved and non-clustered (tRNAscanSE) tRNA gene loci are also high-scoring, suggesting that these loci may not necessarily be pseudogenes (see the high-scoring bins in Figure 3-9).
   - Conclusion:
     - Evidence is weak, but is suggestive that synteny-non-conserved ncRNAs are a mixture of functional ncRNAs and pseudogenes.

2. Hypothesis: non-clustered tRNA genes are pseudogenes.
   - Evidence for this hypothesis:

- The set of low-scoring tRNA genes in the human genome is significantly enriched with non-clustered tRNA genes (see subsection 3.2.2.1. and Figure 3-9).
- Clustered tRNA genes can cover 46 types of anticodons required for protein translation, implying that non-clustered tRNA genes may be functionally less important for translation (see subsection 3.2.2.2.).
- ~56% of human clustered tRNA genes are human-mouse synteny-conserved, while only ~40% of human non-clustered tRNA genes are human-mouse synteny-conserved (see section 2.2 and Figure 2-7).

◆ Evidence against this hypothesis:

- Some non-clustered tRNA genes are high-scoring as well as synteny-conserved in mammalian genomes (see subsection 2.2.2.7.), not suggesting that they are pseudogenes.

◆ Conclusion:

- Evidence is weak, but suggestive that non-clustered tRNAs may be more likely to be pseudogenes.

In conclusion, evidence weakly supports that synteny-non-conserved ncRNAs are a mixture of functional ncRNAs and pseudogenes. Besides, non-clustered tRNA genes may be more likely to be pseudogenes.

Chapter 4. Modelling functional elements associated with ncRNAs

So far in thesis, the main focus has been on discussing issues related to applying comparative-genomics based approaches for genome-wide ncRNA finding. This is due to the fact that till now these approaches have been believed to be one of the most promising ncRNA finding strategies. With the evidence presented in the previous chapters, this belief has therefore been challenged, due to the finding of insufficient covariations, the existence of numerous synteny-non-conserved and potentially functional ncRNAs, *etc*. There is another related limitation of alignment approaches to this general problem: if a set of functionally related ncRNAs are mainly constrained at the structural level, their sequences may become very divergent at the primary-sequence level, making alignment very difficult (Torarinsson et al. 2006).

Accordingly, it is appropriate to consider what approaches might be viable for genome-wide ncRNA finding which do not rely on comparative genomics. One possible strategy is to apply machine learning techniques which can, given a set of unaligned functional ncRNAs, generate models of functional elements implicated in either the transcription or functioning of ncRNAs. Such models can then be used to scan the genomes in order to find novel ncRNAs.

From this chapter, I consider the computational modelling of two types of functional elements that may be associated with ncRNAs:

- the transcription start sites (TSSs) of ncRNAs
- the functional elements/sites that are associated with RNA motifs in RNA transcripts

In the first part of this chapter, I introduce the computational approaches that may be used

to find the transcription regulatory regions, including enhancers/silencers and transcription start sites (TSSs). I start with a brief introduction of transcription regulatory regions, as well as the basics of available motif models and relevant machine learning techniques that have been used to discover motifs. Then I introduce an existing system, Eponine, which was designed to generate predictive models of functional sites, such as TSSs, in genomes.

In the second part of this chapter, I consider the direct detection of RNA motifs in genomes. I explore the possibility of applying available computational approaches for identifying RNA structural motifs in genomes. I also introduce a new model I have created for the purpose of discovering the functional sites which are associated with RNA structural motifs.

4.1. Computational detection of transcription regulatory regions

Access to and recognition of transcription units by transcription machinery are two critical steps in the generation of functional transcripts of all genes, including both protein-coding and ncRNA genes. The essential components involved in transcription initiation include RNA polymerases, transcription factors (TFs), DNA templates, and transcription regulatory elements on genomic DNA sequences. The regulatory elements that are on the same chromosome as the respective transcription units are also called *cis*-regulatory elements. Based on the distance from the genes they regulate, *cis*-regulatory elements can be further categorized into promoters, which are in close proximity to transcription start sites (TSSs), and enhancers/silencers, which can be at great distance from TSSs. A regulatory element may consist of multiple transcription factor binding sites (TFBSs) that can specifically interact with different TFs. A set of TFBSs for a particular TF may share unique sequence patterns, which are generally short and degenerate.

For each gene, the interaction of its promoter with a specific type of RNA polymerase and with a set of TFs determines the exact transcription start point. Different RNA polymerases together with specific sets of TFs favour different promoter sequences. In eukaryotes, there are three different types of RNA polymerases for transcribing genes into RNA molecules. RNA polymerase I only transcribes tandemly repeated ribosomal RNA genes (except 5S rRNA genes). RNA polymerase III transcribes tRNA genes, 5S rRNA genes, and some small nuclear RNA genes. RNA polymerase II transcribes all protein-coding genes. There is evidence indicating that RNA polymerase II is also responsible for transcribing many structural ncRNA and mRNA-like ncRNA genes (Lee et al. 2004). Genes that are transcribed by RNA polymerase I are referred to as pol I genes, and so forth. Modelling promoters of pol II or pol III genes is therefore potentially useful for ncRNA finding. In fact, the internal promoters of tRNA genes have been used as an important signal for tRNA finding in eukaryotic genomes (Fichant and Burks 1991; Pavesi et al. 1994; Lowe and Eddy 1997).

Enhancers/silencers are another type of transcription regulatory element. Their function may be independent of their orientations and distances relative to respective transcription start sites (For review see Khoury and Gruss 1983). Interaction of enhancers/silencers with transcription factors can alter the transcription efficiency of associated transcription units. One important regulatory mechanism of enhancers is inducing chromatin remodelling in eukaryotic cells (For reviews see Vignali et al. 2000; Berger 2002). The genomic DNA of eukaryotes is packaged with histone and non-histone proteins into compact chromatin. To allow transcription to be initiated, the structure of compact chromatin must be remodelled in order to allow efficient access by RNA polymerases. In particular, a class of complex enhancers, locus control regions (LCRs), may consist of multiple regions for initiating chromatin remodelling (For review see Dean 2006). While an enhancer can regulate transcription of only one gene, LCRs can be effective on a cluster of genes. For example, an LCR in mammalian genomes is

suggested to regulate the temporal expression of the beta-globin locus, which consists of at least four genes (For review see Li et al. 2002).

Many computational methods have been developed in order to address the problems relevant to finding transcription regulatory regions in genomes. For instance, many motif finders have been developed to detect over-represented motifs. However, the over-represented motifs so discovered may not directly be useful for discriminating functional sites in genomes. One reason is that the individual interaction between a TF and its TFBS is rarely sufficient to trigger a particular regulatory mechanism. For instance, in eukaryotes, the transcription initiation may be associated with multiple TFBSs (for review see Sandelin et al. 2007). Consequently, for the purpose of finding particular functional sites in genomes, I consider the systems which can model the association of multiple TFBSs with particular functional sites.

In the following two subsections, I introduce the approaches for finding motifs and functional sites. In the first subsection (4.1.1.), existing computational approaches for discovering over-represented motifs are briefly introduced. Although these approaches were not directly used in the work presented in this thesis, this introduction provides essential knowledge for using methods that can perform selective classification of functional sites in the genomes. In the second subsection (4.1.2.), I introduce the computational approaches that can be used to model particular functional sites, such as TSSs and TTSs in genomes. The approaches described and developed here are applied in chapters 5 and 6.

4.1.1. Computational detection of over-represented motifs

Computational detection of over-represented motifs in a set of related sequences can be helpful when studying the regulatory mechanisms of gene expression. Although determination of the functional TFBSs for a TF in genomes can currently only be achieved by experiment, many computational systems have been designed for the purpose of finding over-represented

patterns in a set of sequences containing genes known to be regulated by a particular TF. If over-represented motifs can distinguish sequences with genes with similar functions from background genomic sequences, these features can be suspected to be candidate regulatory elements, possibly TFBSs of the same TF(s).

Over the past decades, many computational approaches have been developed in order to find the over-represented motifs among a set of related sequences. There are two main issues in discovering motifs: 1) the type of model used to represent motifs; 2) the approach used to learn the parameters of the motif model. In the following of this section, these two issues are discussed.

4.1.1.1. Motif models

The first step towards modelling transcription regulatory regions is using a formulation to describe a set of TFBSs for a particular TF. There are at least two types of motif models that have been used for this purpose: *consensus* based models, and *profile* based models.

4.1.1.1.1. Consensus based models

A consensus is a string of simple symbols for describing the most probable nucleotide at each position of TFBSs. A consensus model is suitable for describing a set of TFBSs that are completely identical. Consensus based models have also been extended to incorporate ambiguous symbols. One strategy is to use the IUPAC-IUB alphabet (Nomenclature Committee of the International Union of Biochemistry 1986) to code the ambiguous symbols (Tompa 1999). For example, if both A and G are observed at a particular position of a set of TFBSs, “R” (purine) is thus used to represent this position; if all four types of nucleotides are observed, then “N” is used.

The significance of a consensus can be evaluated by several different scoring schemes. One widely used scoring scheme is the *z*-score, which measures how unlikely a consensus

with certain occurrences in a given set of sequences is found given a background distribution (Tompa 1999). In brief, the *z*-score is the number of standard deviations of the observed frequency of a consensus from its expected frequency. The expected frequency of a consensus can be calculated by counting the number of occurrence in a set of random sequences, which can be generated using a high-order Markov chain modelling the background distribution (Sinha and Tompa 2002).

4.1.1.1.2. Profile based models

One problem with the consensus based motif model is its insufficiency for describing the differential preference toward different symbols at a particular position of a motif. A more flexible, and possibly more powerful, motif model is a *profile* based model, which can describe the alignment of a set of functionally related TFBSs. A widely used profile based model for representing motifs is a position frequency matrix (PFM) (also as position specific frequency matrix, PSFM) (for review see Wasserman and Sandelin 2004), which is a type of product-multinomial model. A PFM consists of a series of columns. Each column of a PFM is a multinomial distribution over all possible symbols of the alphabet used in each position of a motif. By using a PFM, each position of a sequence motif is treated independently, although this assumption may be biologically imprecise as shown in some analyses of protein-DNA interactions (Barash et al. 2003).

The probability of emitting a particular sequence pattern that starts at the i^{th} position of a sequence x from a PFM can be evaluated by:

$$M(x,i) = \prod_{l=1}^{|M|} P_l(x(i+l-1)) \qquad [4\text{-}1]$$

$|M|$ is the number of columns of the PFM. P_l returns the probability of a particular symbol emitted by the l^{th} column of the model. $x(i + l - 1)$ is the symbol at the $(i + l - 1)^{th}$ position of x. For modelling TFBSs, the possible symbols for each column consist of adenine (A), guanine

(G), cytosine (C), and thymine (T). A PFM can be displayed in the form of sequence logos (Schneider and Stephens 1990). A sequence logo for a PFM contains of a series of columns of stacked symbols, where the height of each symbol is proportional to its information content at each position. In the rest of this thesis, sequence logos are used to represent the primary-sequence motifs.

One advantage of using PFMs to describe motifs is that it is very easy to connect a motif model to statistical information theory. The statistical significance of a motif can be assessed by calculating the information content of a PFM. The information content at the l^{th} position of a site is:

$$I(l) = \sum_b P_{l,b} \log_2 \frac{P_{l,b}}{P_b} \qquad [4\text{-}2]$$

, where b refers to each of the possible bases; $P_{l,b}$ is the probability of base b at the l^{th} position; P_b is the frequency of base b in the background sequences (*e.g.* non-site sequences in the genomes). This formulation is equivalent to the relative entropy and the Kullback-Leibler distance, between the foreground motif model and the background sequence model (for review see Stormo 2000). Usually the base composition in the background sequence model is assumed to be independent and identically distributed (i.i.d.). One simple approach is to assume that each base in the background is equally probable and thus P_b is 0.25 for each base.

In order to search for a particular pattern in a given sequence, a PFM value is usually converted into a sum of a series of log-likelihood ratios with respect to a background sequence model B:

$$W(x,i) = \sum_{l=1}^{|M|} \log_2 \frac{P_l(x(i+l-1))}{B(x(i+l-1))} \qquad [4\text{-}3]$$

This conversion gives a position specific scoring matrix (PSSM), which is also called a position weight matrix (PWM) (for review see Wasserman and Sandelin 2004). Given a

sequence region, a PWM can be used to evaluate the log-likelihood ratio between the foreground motif model and the background sequence model. A higher log-likelihood ratio can be interpreted as that the foreground model is more likely to generate a given sequence pattern than is the background model. The PWM scores have been shown to be proportional to the binding energy contribution of the bases (Berg and von Hippel 1987; Stormo 2000). A PWM can be used to scan for candidate TFBSs in a long sequence. For finding TFBSs in a sequence of length N, all $N - |M| + 1$ sub-sequences of length $|M|$ must be enumerated and scored.

4.1.1.2. Algorithms for discovering motifs

In an *in silico* motif finding problem, the positions, patterns, and lengths of over-represented motifs in a set of related sequences may be initially unknown. Motif finding algorithms must be capable of optimizing these parameters given a set of sequences. In order to simplify the motif finding problem, existing motif finding algorithms usually require a user-defined motif length. Consequently, the parameters that need to be learned are the motif patterns, and their respective positions in individual sequences. Based on the models used, motif finding methods can be classified into *consensus* based and *profile* based methods, which are briefly introduced in the following, respectively.

4.1.1.2.1. Consensus based methods

Consensus based motif finding methods discover over-represented motifs by exhaustive enumeration of a set of motifs (Tompa 1999; Marsan and Sagot 2000; Pavesi et al. 2001). These methods usually use the following two steps to discover over-represented motifs:

- Enumerate all possible *m*-mer substrings in the given set of sequences.
- Score and rank the *m*-mer substrings by using some statistical measures, such as the *z*-score.

Consensus based methods can be very fast, if a suitable indexing structure, such as the suffix tree (Marsan and Sagot 2000), is used for organizing the sequences. While some evidence suggested that consensus based motif finding methods may suffer from high false positive rates (Osada et al. 2004), a recent survey reveals that these methods can have a performance comparable to that of profile-based methods (Tompa et al. 2005). However, there are considerations in using consensus based methods. Firstly, generating one consensus optimal for predicting new sites is not straightforward. Similar substrings must be clustered into fewer groups in a post-processing stage (Marsan and Sagot 2000). Secondly, for computational efficiency, some consensus based methods such as YMF (Sinha and Tompa 2000) and Weeder (Pavesi et al. 2001) restrict the number of mismatches allowed in a pattern. When several positions in a set of TFBSs with respect to a TF are weakly constrained, as in the cases of eukaryotes, consensus based methods may not work well (Pavesi et al. 2001).

4.1.1.2.2. Profile based methods

Profile based motif finding methods discover over-represented motifs by selecting oligonucleotides from the set of input sequences and then aligning them to generate profiles. These methods generally consist of two components:

- A likelihood function which can evaluate how likely a particular motif is to be over-represented given a set of sequences.
- An optimization procedure which can maximize the likelihood function.

A basic form of the likelihood functions used in many profile-based motif finding systems (for review see Stormo 2000) is the information content of a motif, as the formulation presented in [4-2]. The positions of a motif in individual sequences are referred to as the missing data. An important task of the optimization procedure is to search for the solution of missing data which may maximize the likelihood function. Two of the most widely used

optimization algorithms are the Expectation Maximization (EM) (Lawrence and Reilly 1990; Bailey and Elkan 1994) and Gibbs Sampling (Lawrence et al. 1993).

EM algorithm

The EM algorithm is a general approach for maximizing a likelihood function with missing data. The EM algorithm iterates between two steps: in the first step, the expected values of the missing data are estimated, conditioned on the proposed model parameters; in the second step, given the expected values of the missing data, the new model parameters that can maximize the log likelihood function are chosen. The first step is the expectation step (E-step) and the second step is the maximization step (M-step). These two steps are iterated until a convergence criterion is satisfied.

There have been many extensions to the original EM based motif finding algorithm (Lawrence and Reilly 1990). For instance, the MEME (multiple expectation maximization for motif elicitation) algorithm is designed to model motifs with zero-or-one occurrences per sequence (ZOOPS) (Bailey and Elkan 1994), although the original EM motif finding algorithms were designed to find one occurrence per sequence. Another significant improvement to EM made in the MEME algorithm is its capability to detect multiple motifs within a single run.

Gibbs Sampling

In mathematics and physics, Gibbs Sampling is a sampling algorithm that is used to explore the joint probability of two or more random variables. It is a special case of the Metropolis-Hastings algorithm, which is a type of Markov chain Monte Carlo algorithm. A Gibbs Sampling approach for motif finding also consists of an iteration of two steps: predictive update step and sampling step (Lawrence et al. 1993), which correspond to the E-step and the M step of an EM algorithm respectively. However, unlike the deterministic

process used in EM to find the missing data (*i.e.* the start sites of a motif in individual sequences), a stochastic process is adopted in the Gibbs Sampling motif finding algorithm (Lawrence et al. 1993). At the predictive update step of Gibbs Sampling, a sequence *z* is chosen and the other sequences are used to derive the model parameters, given the current site positions. At the sampling step, the probability of generating the site in each position of sequence *z* can thus be estimated conditioned on the current motif model. The new site position in sequence *z* is sampled with the probability distribution of the site positions.

Several improvements have been made to enhance the capability of the original Gibbs Sampling based motif finders (for review see Pavesi et al. 2004). The capabilities of the enhanced Gibbs Sampling motif finders include finding multiple motifs simultaneously (Thompson et al. 2003), modelling two-block motifs (GuhaThakurta and Stormo 2001; Liu et al. 2001), *etc*.

4.1.1.3. Considerations when using motif finding methods

Although many motif finding algorithms have been developed, computational detection of functional motifs in real genomes remains a challenging problem. Several independent surveys indicated that, in the context of genome-wide TFBS finding, the performance of available motif finding algorithms is far from being satisfactory (Hu et al. 2005; Tompa et al. 2005). An important finding is that most of the existing motif finding systems are not very effective in discriminating functional sites, particularly when complex genomes, such as the human and mouse genomes, are investigated.

Several possible reasons to the poor performance of existing motif-finding approaches have been proposed:

- The optimization procedure may get stuck in local optima.
- The background model used in many methods may be too simple to reflect the true

background in complex genomes.

- The architecture of functional sites may not be properly modelled as a single motif. For instance, TSSs may associate with two or more TFBSs.

A number of improvements have been made in order to address these issues (for review see Pavesi et al. 2004; MacIsaac and Fraenkel 2006). In the following subsection, I introduce methods that may be more suitable for prediction of functional sites in complex genomes.

4.1.2. Computational detection of functional sites

In transcription, TSSs are determined by the binding of multiple TFs to a set of TFBSs in close proximity to TSSs (for review see Fickett and Hatzigeorgiou 1997). For example, the transcription initiation of mammalian tRNA genes by RNA polymerase III is regulated by the binding of TFs to the A and B boxes (Hsieh et al. 1999) (Figure 4-1), which are within certain distances downstream of TSSs (Pavesi et al. 1994).

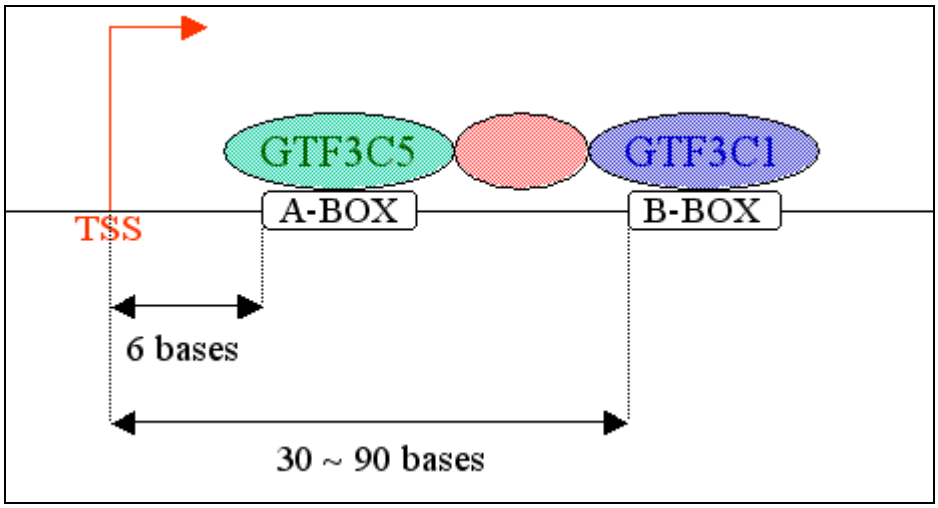


Figure 4-1. The transcription initiation of mammalian tRNA genes is regulated by A and B boxes

One computational approach for TSS finding is to model the promoters of genes, since promoters are in close proximity to TSSs. Although a number of TSS finding systems based

on promoter modelling have been developed, most of them are specifically designed for finding the TSSs of protein-coding genes (for review see Fickett and Hatzigeorgiou 1997). For the purpose of finding the TSSs of ncRNAs, a system that can be used to learn new models given a new set of training sequences is of interest.

A possible approach to model TSSs is using Hidden markov models (HMMs). Complex HMMs, which recruit various states for modelling multiple signals associated with splicing and translation, have been used for finding eukaryotic protein-coding genes (Burge and Karlin 1997). Presumably *ad hoc* designed HMMs should be able to model complex regulatory elements by adequately connecting the states of relevant TFBSs. However, there are some concerns for applying HMMs to TSS modelling. First, it is generally difficult to guess a suitable HMM topology for any types of regulatory elements. Second, the parameter tuning of complex HMMs may easily be trapped in a local optimum (Durbin et al. 1998).

Over the past few years, several new systems have been developed to model regulatory modules which may consist of multiple TFBSs (Wasserman and Fickett 1998; GuhaThakurta and Stormo 2001; Bailey and Noble 2003; Zhou and Wong 2004; Aerts et al. 2005). Motif finding systems that use regulatory module models may potentially be applicable to finding promoters. However, for the purpose of predicting TSSs, there are concerns with these systems. First, the distance constraints between motifs in a module are generally un-modelled, or merely modelled by using a linear gap penalty (Bailey and Noble 2003), which appears to be unsuitable for describing the distance range between TFBSs, as observed in the tRNA gene promoters (Figure 4-1). Second, these module finding systems may report just an approximate area for regulatory modules, but not an actually functional site, which is not what we would expect from a TSS prediction algorithm.

Here, for the purpose of modelling the TSSs of ncRNA genes in the mammalian genomes, I chose to use an available system, Eponine (Down and Hubbard 2002), which was originally

designed to model the TSSs of mammalian protein-coding genes. One feature of Eponine is that it has been designed to perform predictions of functional sites in genomes. Eponine has been demonstrated to be effective in discriminating TSSs (Down and Hubbard 2002) and transcription termination sites (TTSs) (Ramadass 2004) in mammalian genomes. In the following subsection (4.1.2.1.), I introduce the basics of the original Eponine implementation.

4.1.2.1. Modelling functional sites using Eponine

4.1.2.1.1. The Eponine Anchored Sequence Model

The Eponine Anchored Sequence Model (EAS) is a classification model that is aimed to be applied to individual points within a large genome, *i.e.* exact reference positions on the genome sequence, such as the base pair at which transcription starts (TSS). An essential component of the EAS model is a positioned constraint (PC), which consists of:

- A position weight matrix (PWM) which models a signal that may contribute to the classification of a particular functional site.
- A discrete probability distribution to describe the position of a PWM relative to the reference site.

In the EAS model, the score of a PC can be calculated as:

$$\phi(x,a) = \frac{\log(\sum_{i=-\infty}^{+\infty} P(i)\cdot W(x,i+a)\ )}{|W|} \qquad [4\text{-}4]$$

where x is a DNA sequence; a is a pre-defined reference site for each sequence x; $P(i)$ is a discrete probability distribution for modelling the distance of a motif from the reference site (*i.e.* TSS, TTS, *etc.*); $W(x,i+a)$ is the PWM score for offset i relative to the reference site a. $P(i)$ is usually in the form of a discrete Gaussian distribution. It should be noted is that, the PWM used in the Eponine models is actually a probability frequency matrix (PFM, see [4-1]) normalized with background base compositions. The difference between the PWM used in

Eponine and the general form of PWM (see [4-3]) is that, the latter is equivalent to the logarithm of the former. For simplicity, the term PWM is still used in describing the Eponine models, in order to be consistent with the terminology used in the papers relevant to Eponine (Down and Hubbard 2002; Down et al. 2006).

A particular point about the this scoring function is that, this function may allow, not only a strong motif with a very sharp position distribution relative to a particular reference site, but also short motifs with very broad distributions. This is caused by the summation of the position-constrained PWM scores across a region on a sequence. This design may be advantageous to the situation where there are general compositional biases toward some particular oligonucleotides, as what we have observed in the case of CpG overrepresentation in eukaryotic promoters. However, it should be noted that, by using such a scoring function, the EAS model is not designed to find optimal motifs that are over-represented in a set of sequences. Therefore, the EAS model is specifically designed to discriminate functional sites in the genomic context, *i.e.* the individual points within a large genome.

It should be noted that the final score of each PC for each sequence must be normalized by $|W|$, the number of columns in each PWM. At first glance this normalization seems to be unnecessary; however, it is critical for learning the EAS models. The reason is that, in optimizing the parameters of the EAS models, the widths of PWMs are not a pre-defined and fixed value. The learning system of Eponine learns a set of optimal PWMs from a pool of candidate PWMs of varied widths. If a PWM score is not normalized, a PWM with more columns may be preferred. Similar normalization strategies has been used by some of the motif finding systems where the lengths of motifs are not pre-defined, such as the Gibbs Motif Sampler (Lawrence et al. 1993).

Learning the EAS models

The EAS model is so built by taking the weighted sum of a number of PC scores. This complex model is equivalent to the generalized linear model (GLM) (McCullagh and Nelder 1983), where each PC in this complex model is equivalent to a basis function in GLMs.

The general formulation of a GLM can be expressed as:

$$\eta(x) = \sum_m \beta_m \phi_m(x) + C \qquad [4\text{-}5]$$

The term, *x*, represents a sequence. ϕ is a set of basis functions. β is a set of weights associated with individual basis functions. "C" is the constant. For binary classifications (*e.g.* classifying sequences into positive and negative ones), one logistic function,

$$\sigma(\eta) = \frac{1}{1+e^{-\eta}} \qquad [4\text{-}6]$$

can be used to transform the raw output of GLMs to fit a sigmoid curve. Thus, the output of this transformation can be used to decide whether an input *x* belongs to a particular class.

For training an EAS model, the parameters that need to be learned include PCs, and the weights that associate with PCs. Each PC consists of a PWM and an associated probability position distribution, which also need to be learned. At the initial stage of training, the parameters of PWMs and associated position distributions should be largely unknown. A trainer should be able to recruit informative PWMs and discard non-informative ones. The Eponine trainer uses a combined strategy consisting of the relevance vector machine (RVM) algorithm (Tipping 1999) and a Monte Carlo sampling process:

- A number of random PWMs of certain widths, and random Gaussian position distributions, are initialized.
- Use the RVM algorithm to estimate the weights of PCs and thus prune

non-informative PCs.

- Recruit new PCs by using a Monte Carlo sampling process to adjust the widths and weights of PWMs, as well as the parameters (*i.e.* mean and width) that decide the shape of Gaussian position distributions.

The RVM algorithm is the core algorithm for learning informative PCs. Since the RVM is so important for training the EAS model for classification, it is discussed in the following.

The Relevance Vector Machine

The RVM is a Bayesian approach to learn parameters of GLMs (Tipping 1999). It can take a set of basis functions, corresponding to PCs in the EAS model, and then use a "pruning prior" to discard the basis functions that do not contribute significantly to a particular classification problem.

In general, the Bayesian way for estimating parameters for classification can be written as:

$$P(\beta \mid X, T) = \frac{P(T \mid X, \beta) \times P(\beta)}{P(T \mid X)} \qquad [4\text{-}7]$$

$P(\beta \mid X, T)$ is the posterior probability of a model with parameter set β, given paired input and target data, X and T, where $X = (x_1, x_2, \ldots, x_N)$, represents the N input points (*i.e.* sequences in this thesis), and $T = (t_1, t_2, \ldots, t_N)$, represents respective targets (or responses). $P(T \mid X, \beta)$ is the likelihood of the model given the data. $P(\beta)$ is the prior probability of β and $P(T \mid X)$ is the normalization constant. For binary classifications where $t_n = [0, 1]$, the likelihood can be calculated by:

$$P(T \mid X, \beta) = \prod_{n=1}^{N} \sigma(\eta_n)^{t_n} (1 - \sigma(\eta_n))^{1-t_n} \qquad [4\text{-}8]$$

, where η_n is the predicted output (of a GLM) for an input x_n.

When there is no prior knowledge of the model parameters (*e.g.* β_m's in [9]), a non-informative prior can be used. A non-informative prior can be a uniform distribution or a very broad exponential-family distribution. However, choosing an informative prior may enable the learning of a sparse model, which contains only a few basis functions. An advantage of training a sparse model is reducing the chance of overfitting to data. To achieve sparsity, the RVM framework uses an automatic relevance determination (ARD) Gaussian prior over each weight (Tipping 1999):

$$P(\beta_m \mid \alpha_m) = G(\beta_m \mid 0, \alpha_m^{-1}) \qquad [4\text{-}9]$$

, where the hyperparameter, α_m, is the inverse variance of each mean-zero Gaussian distribution. This choice of prior implies that there is a strong preference that many β_m's are close to zero. After optimizing parameter β and hyperparameter α, basis functions that are not informative for classification can be decided. If α_m is extremely large, the variance of the respective Gaussian distribution will be very small and the distribution, $P(\beta_m \mid \alpha_m)$, will peak at 0. A zero weight means that the associated basis function is non-informative and could be dropped.

For optimizing GLMs, the RVM algorithm has been shown to achieve a better sparsity than do other relevant algorithms (Tipping 1999). Thus, by using the RVM algorithm, the Eponine trainer is capable of exploring a large parameter space in order to select a set of PCs which can optimize the EAS model for classification. (Down and Hubbard 2002).

4.1.2.1.2. The Eponine Windowed Sequence model (EWS)

Using the EAS model for functional sites requires a set of positive training sequences, where reference points must be labelled in these sequences. TSSs and TTSs are extremely fortunate cases because lots of experimental evidence is available to indicate relatively definable regions for these sites. However, for other cases where the existence of common

regulatory elements in a set of functionally related sequences is only suspected, it is difficult to adequately label training sequences with reference sites and thus the EAS strategy is not expected to work properly. An alternative is the Eponine windowed sequence (EWS) model, which is more suitable for modelling common motifs whose locations in individual sequences are varied or unknown.

The basic formulation of basis functions used in the EWS model is:

$$\phi(x) = Z \times \underset{s=u}{\overset{v}{optimal}}\left(\prod_{k=1}^{K}\left(\sum_{i=-\infty}^{\infty} P_k(i) \cdot W_k(x, s+i)^{\frac{1}{|W_k|}}\right)^{\frac{1}{K}}\right) \qquad [4\text{-}10]$$

and

$$Z = \frac{1}{|u| - |v| + 1} \qquad [4\text{-}11]$$

where the interval [u, v] is the u^{th} position to the v^{th} position that are accessible by the basis function ϕ, on sequence x; P_k is the discrete probability distribution of the distance between the k^{th} PWM (W_k) and the first PWM (W_1). This complex basis function is called the convolved sensors basis function (CSBF) in the EWS models.

A CSBF may contain more than one position constrained PWM. The reason for normalizing CSBFs with $1/k$ is similar to the use of $\frac{1}{|W|}$ for normalizing the PWMs in the EAS models (see subsection 4.1.2.1.1.), because currently the number of PWMs in a CSBF is not fixed. Otherwise, without a normalization factor, a CSBF with more PWMs may be preferred by the Eponine trainer. The normalization factors, $1/k$ and $\frac{1}{|W|}$, are modifications to the original Eponine implementation (Down 2002; Down and Hubbard 2004).

In order to explain how the score calculation in [4-10] is performed, I use a CSBF consisting three position-constrained PWMs as an example (Figure 4-2). Given a sequence x,

the score on the first position is calculated by multiplying the three scores given by position-constrained PWMs 1 ~ 3. Although in the plot there is just a single fixed point for each position-constrained PWM (Figure 4-2, upper-left), it should be noted that the score for each position-constrained PWM is a summation over a position distribution *P*. The final score of a CSBF given sequence *x* is the optimal one in all the scores on the interval [*u*, *v*].

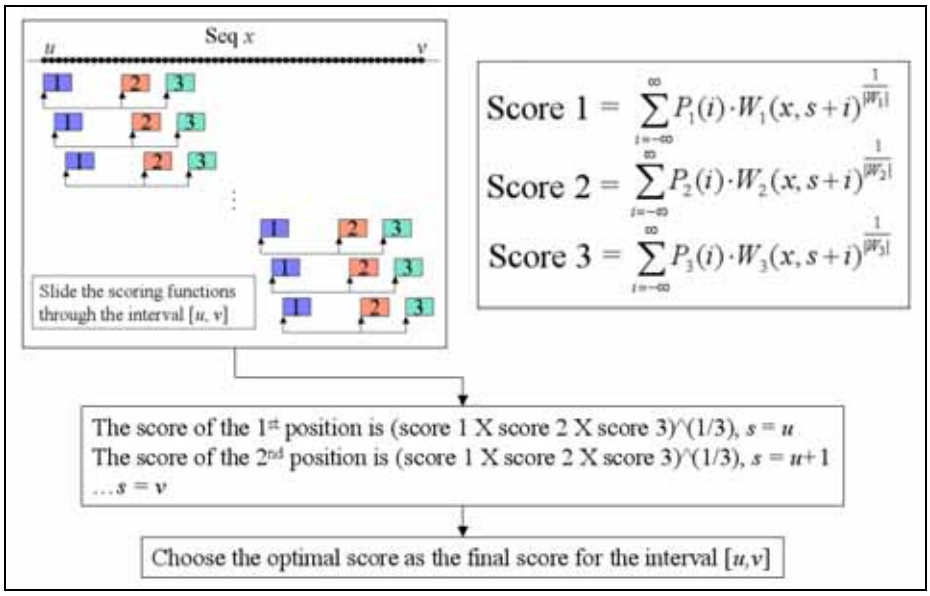


Figure 4-2. How to calculate the score of a CSBF consisting of three PWMs and associated position distributions

Learning the EWS models

For training the EWS model, two types of parameters must be learned: 1) the probability distribution of positions and 2) PWMs. For distributions of positions, the training process is very similar to that for training the EAS models (see subsection 4.1.2.1.1.), except that the reference site is replaced with one of the position constrained PWMs in each CSBF. The Monte Carlo sampling process is used to optimize the choice of CSBFs. A new member PWM is randomly sampled from the pool of CSBFs, and then the so generated new CSBFs, will be re-weighted and pruned by using a RVM strategy. Through iterating the Monte Carlo sampling process and the pruning process using the RVM, an EWS model consisting of a set

of CSBFs could be learned.

4.2. Modelling local RNA motifs

In the previous parts of this thesis, ncRNA classifiers and the modelling of transcription regulatory elements of ncRNAs have been discussed. Due to the particular types of signals that are used in these methods, there are certain limitations on the scopes of their applications. Firstly, existing comparative algorithms may overlook the RNA structural motifs spanning only a region in a transcript. Secondly, when modelling transcription regulatory elements, any RNA motifs implicated in the regulation of ncRNA expression are essentially ignored.

The transcripts of ncRNA genes are not the only RNAs that may contain RNA structural motifs. Evidence suggests that local RNA structures may be implicated in the regulation of protein translation (for review see Kozak 2005). Besides, single-stranded regions in transcripts can also be part of functional motifs (for review see Mattaj 1993). The local RNA motifs discussed here are considered as a composite of primary-sequence patterns and local RNA structures, where different parts of a composite motif may be separated by unstructured and/or functionally unimportant regions of variable length.

One type of computational approach for identifying local RNA motifs is to search for the consensus RNA motifs in a group of functionally related transcripts. Existing algorithms for finding consensus RNA motifs in transcripts can be generalized into three major categories: variants of the Sankoff's algorithm, variants of stochastic context-free grammars (SCFGs), and variants of genetic algorithms. In the following subsection (4.2.1.), I briefly introduce existing algorithms for finding local RNA motifs, and the considerations in using these algorithms.

As previously discussed (see subsection 4.1.2.), computational modelling of functional sites requires algorithms that can combine the contribution from multiple TFs. A similar

approach is required to combine the contributions of local RNA motifs to generate a predictive model. In an attempt to address this, I developed a new RNA motif extension to the Eponine modelling system. The addition of this new extension allows the modelling of functional sites as a composite of primary-sequence and secondary-structure motifs from a set of unaligned functionally related sequences. This is described in subsection 4.2.2.

4.2.1. Available methods for finding consensus RNA motifs in sequences

4.2.1.1. The Sankoff's algorithm and variants

Given a set of sequences, Sankoff's algorithm can generate optimal primary-sequence alignment and secondary-structure minimum free energy (MFE, see subsection 1.3.1) simultaneously (Sankoff 1985). However, the time complexity is $O(N^{3K})$ and the space complexity is $O(N^{2K})$, where N is the sequence length and K is the number of sequences. It is therefore not practical to apply Sankoff's algorithm to finding consensus RNA motifs in a set of sequences. Variants of Sankoff's algorithm have thus been created in order to find consensus RNA motifs in an acceptable time. Two modifications have been adopted by different implementations in order to accelerate the search process. Firstly, only local hairpins are considered by inhibiting branching configuration. A branching configuration is the partition of one sequence into two structural regions in the base-pair dependent energy rule (Nussinov and Jacobson 1980). Inhibiting branching configuration is equivalent to taking out $W(i,k-1)$ from [1.2] of subsection 1.3.1.1. , reducing the time complexity from $O(N^6)$ to $O(N^4)$ for pairwise alignments.

The second modification for accelerating Sankoff's algorithm is to use progressive alignment methods. The strategy of progressive alignment methods is to find the best pairwise alignments first, and then other alignments or single sequences can be consecutively added to

existing alignments. In the primary form of progressive alignment methods, once a group of sequences have been aligned, their relations cannot be altered at later steps. The procedure of combining alignments terminates when all sequences have been aligned. The time complexity can be $O(L^4N^4)$, where L is the average sequence length; N is the number of sequences (Gorodkin et al. 2001).

Progressive alignment methods can efficiently generate acceptable multiple sequence alignments; however, these methods are greedy and alignments can be trapped in a local optimum. The reason for this is that the best pairwise alignments do not necessarily contain optimal motifs shared by all sequences, and globally optimal motifs may be only sub-optimal when comparing two sequences. When finding primary-sequence motifs, additional approaches can be used to improve multiple sequence alignments. Related techniques include iterative refinement methods, simulated annealing, Gibbs sampling, *etc* (For reviews see Durbin et al. 1998). Nonetheless, no variants of Sankoff's algorithm use these approaches and the primary form of progressive alignment methods is still the most common strategy used by variants of Sankoff's algorithms.

4.2.1.2. The stochastic context-free grammars (SCFGs)

Just as in the prediction of RNA secondary structures, statistical models, such as SCFGs (see subsection 1.3.3) and McCaskill's sampling algorithm (McCaskill 1990), can replace MFE for finding the consensus RNA motifs among sequences. PMcomp/PMmulti (Hofacker et al. 2004) uses McCaskill's sampling algorithm to do pairwise/multiple structural alignments. Its time complexity and space complexity is as high as $O(N^6)$ and $O(N^4)$ respectively for pairwise alignments. The computational complexity of PMcomp/PMmulti is not less than that of Sankoff's algorithm. For multiple structural alignments, it also uses progressive alignment methods in order to restrict computational complexity. For pure SCFGs-based algorithms that can do *ab initio* structural alignments, the computational complexity is at least as high as for

the original Sankoff's algorithm. In order to reduce complexity, variants of SCFGs (Knudsen and Hein 1999; Knudsen and Hein 2003) take alignments that are generated by popular multiple-sequence-alignment programs, such as ClustalW, and then refine alignments using SCFGs. One problem with this approach is that the quality of initial multiple sequence alignments nearly determines the performance of variants of SCFGs. If the initial alignments were trapped in a local optimum in terms of RNA motifs, it seems unlikely that further refinement at the structural level could give optimal answers (Knudsen and Hein 1999). In addition, perfectly identical RNA secondary structures, which may not be always practical for modelling RNA motifs in genomes, are sometimes assumed (Knudsen and Hein 2003).

4.2.1.3. Genetic-algorithm based approaches

Unlike the current implementations of variants of Sankoff's algorithm or variants of SCFGs, GA-based approaches are less easily trapped in a local optimum. Although GA-based approaches are not guaranteed to find the optimal solution, they can be very good in predicting RNA structures (Chen et al. 2000; Taneda 2005). One problem with the current GA-based approaches is that primary-sequence motifs are not generally considered as part of RNA motifs; few GA-based approaches have been designed to find both types of motifs simultaneously.

4.2.1.4. Uncategorized RNA-motif finding approaches

There are other types of consensus RNA-motif finding algorithms that cannot easily be classified into the above categories. One type of algorithms is to take folded sequences and then align the predicted RNA structures. These programs do not predict RNA structures by themselves. Instead, the structure of each sequence may be taken from the prediction made by MFE-based RNA secondary-structure prediction algorithms, such as Mfold (Zuker 1989) and RNAfold of the Vienna package (Hofacker 2003). MARNA (Siebert and Backofen 2005), RNAForester (Hochsmann et al. 2004), and RNADistance (Hofacker 2003) are three examples.

For instance, from the predicted RNA structures for sequences, MARNA identifies seeds of both primary-sequence and RNA structural motifs and then feeds these motifs to T-Coffee (Notredame et al. 2000). One concern with such algorithms is that their performance can be influenced by the accuracy of the optimal global structures predicted. Besides, these algorithms may be vulnerable to the cases where the consensus RNA motifs between a set of sequences is quite different from the optimal structures for individual sequences.

Another type of algorithms, such as RNAalifold (Hofacker et al. 2002) and MSARI (Coventry et al. 2004), are designed to find consensus RNA motifs in primary-sequence alignments that are generated by using popular multiple sequence alignment programs, such as ClustalW. These algorithms take compensatory mutations as the evidence for supporting the existence of a global RNA motif (Coventry et al. 2004; Washietl et al. 2005). One concern with these algorithms is that, they depend on the primary-sequence alignments, which may, under certain circumstances, be incapable of revealing the consensus RNA structures between sequences. Their performance should be sensitive to the sequence identities between given sequences, although the required identities were not clearly defined in their original papers.

Consequently, currently available algorithms are not practical enough for modelling regulatory RNA motifs in genomes, since there are so many considerations and restrictions in using them. Given a set of functionally related regions in transcripts, there should be an algorithm that can model both common primary-sequence and structural motifs efficiently. The resulting model should be potentially applicable to genome-wide regulatory RNA motif finding. Therefore, I extended Eponine to include local RNA structural motifs in order to create an ncRNA modelling tool, which can be applied to finding RNA-motif associated functional sites in genomes.

4.2.2. Extending Eponine to include RNA structural motifs

Both the EAS and EWS models of the Eponine package (see subsection 4.1.2.1.) are useful for modelling primary-sequence motifs and the relations of motifs to other reference sites. Similarly the Eponine RNA-motif extension should model both RNA structural motifs and the relations of structural motifs to other sites. RNA motifs should be considered as yet another type of motifs that are in sequences, except that RNA motifs possess structural features, including stems and loops. In brief, the Eponine RNA-motif extension aims at modelling the regulatory RNA motifs that are constituted by specific arrangement of both primary-sequence motifs and structural motifs, with appropriate scoring scheme.

Primary-sequence motifs are modelled by PWMs in the EAS and EWS models. Similarly, a formal description of structural features must be chosen in order to extend both the Eponine models to include structural motifs. One possibility for modelling individual hairpins is to use Covariance Models (CMs), which are SCFG-based RNA profiles. However, for several reasons, I decided that CMs may not be adequate for extending Eponine models. Firstly, training an Eponine RNA-motif model that consists of CMs can be very time-consuming, because numerous CMs can be temporarily generated in the training process and each must be assessed and updated. The time complexity of evaluating each CM is at least $O(L^3)$, where L is the length of each candidate region for a particular hairpin (Durbin et al. 1998). Secondly, it is difficult to adapt the scores of CMs on sequences for EAS and EWS models. Distributions of the CM scores may vary greatly across different types of RNA motifs. There is no obvious solution for combining the CM scores and the PWM scores in order to model primary-sequence and structural motifs simultaneously.

Another question for modelling RNA motifs is how to properly address variations of structural features. Although variations in hairpins are commonly believed to be disastrous for

some structural RNA genes, evidence indicates that a certain degree of variation exists in RNA structural motifs of similar functions. An example is the transcription termination signals of bacterial genes, where the sizes of stems can vary from 5 to 30 base pairs and the lengths of loops vary from 3 to 9 bases (de Hoon et al. 2005).

Using existing probabilistic models cannot properly address dimensional variations of RNA structural motifs. For instance, standard CMs using general topologies can tolerate small size variations of hairpins, but they cannot model these variations explicitly. To explicitly model such variations, CMs need additional techniques, such as duration modelling. Duration modelling is a technique used for addressing the length distribution explicitly (Durbin et al. 1998). However, if such techniques are used, the computational complexity will be much higher. In addition, other structural features, such as folding energies of hairpins, may still need to be modelled by other yet unmentioned techniques.

Therefore, in developing the RNA motif extension of Eponine, I decided to use a local RNA structural model which is not based the classic probabilistic model of RNA structures, such as CMs. The new model should be able to model a variety of features of local RNA hairpins. There are two steps in training the models: firstly, candidate hairpins for each sequence should be first located; and secondly, the Eponine trainer learns a model describing the structural features of the consensus RNA motifs of these sequences. In the following two subsections, I introduce the implementation of the Eponine RNA-motif extension, including the approaches to locate local hairpins (subsection 4.2.2.1.) and the way structural features are modelled (subsection 4.2.2.2.).

4.2.2.1. Locating local hairpins

The RNA motifs, which the Eponine RNA-motif extension is designed to model, are specific arrangements of a set of single-stranded and double-stranded regions in sequences. Consequently, predicting and evaluating RNA secondary structures of given sequences is

necessary. It is reasonable to assume that any position in each sequence can be the start point of a hairpin structure. Proposed RNA motif models should evaluate all hairpins that may start at each position of each sequence.

Predicting hairpins that may be functionally important is not straightforward. Firstly, optimal structures can be predicted only for regions of restricted length, but not for the full-length region of long sequences. The time complexity for predicting optimal structures by using either MFE or SCFGs is proportional to the cubic sequence length. Given any fragment of genomic sequence, one practical strategy for finding candidate functional motifs is to chop the original sequence into consecutively windowed regions and then predict hairpins for individual regions. Although this approach may sacrifice some hairpins that span a region larger than the window size, stable hairpins within windowed regions can still be predicted. It is also reasonable to infer that long-range interactions in large hairpins should depend on stable hairpins within windowed regions. By evaluating hairpins in windowed regions, trained models can be applied to genome-wide RNA motifs finding; all regions in each sequence can be consecutively evaluated by sliding the windows through all positions. Similar strategies have been used by other algorithms for genome-wide ncRNA finding (Rivas and Eddy 2001; di Bernardo et al. 2003). The time complexity of folding windowed RNA secondary structures for multiple sequences is thus $O(LNM^3)$, where L is number of sequences; N is the average number of windows per sequence; M is the length of windowed regions.

Secondly, predicting the sub-optimal hairpins for each sequence seems necessary. Evidence suggests that optimal structures do not necessarily represent the functional forms of various regulatory RNA motifs. In addition, RNA folding may alter in response to certain conditions, such as the binding of ligands, increases in di-ionic strength in solution, interaction with RNA binding proteins, post-transcription modifications, *etc*. For finding consensus RNA motifs among sequences, only optimal folding for each sequence may not be sufficient.

Exhaustively enumerating all possible hairpins that may fold in each sequence is computationally expensive and impractical. There are at least two simpler approaches for predicting sub-optimal hairpins for each sequence. The first approach is to collect the optimal hairpin for each position of each windowed region (Figure 4-3, algorithm A). For each position *i* within a windowed region, the optimal hairpin, which is conditioned on that position *i* must pair with another position *j*, is saved, where $i < j <$ window size. By scanning sliding windows for each sequence, optimal hairpins that start at individual positions in each sequence are collected. These site-specific optimal hairpins are not necessarily the components of globally optimal structures. This approach is similar to Zuker's suboptimal folding algorithm, and to the inside and outside directions of the CYK algorithm (Durbin et al. 1998). The consideration of this approach is time complexity. In addition to the time complexity $O(N^3)$ for calculating the energy matrix in using Zuker's MFE algorithm, additional time complexity, $O(\text{window size}^2)$, is required in order to trace respective optimal hairpins for all possible paired positions in each windowed region.

By contrast, the second approach for collecting sub-optimal hairpins for each sequence is much simpler. Only the optimal structure for each windowed region is predicted (Figure 4-3, algorithm B). From the optimal structure, individual hairpins are extracted, and then saved with their respective start positions. By scanning sliding windows for each sequence, a series of optimal hairpins that start at distinct positions in each sequence are collected. Just like the situation of the first approach, these site-specific optimal hairpins are not necessarily the components of optimal global folding. The second approach can be much faster than the first one, because much less folding space is explored (Figure 4-3).

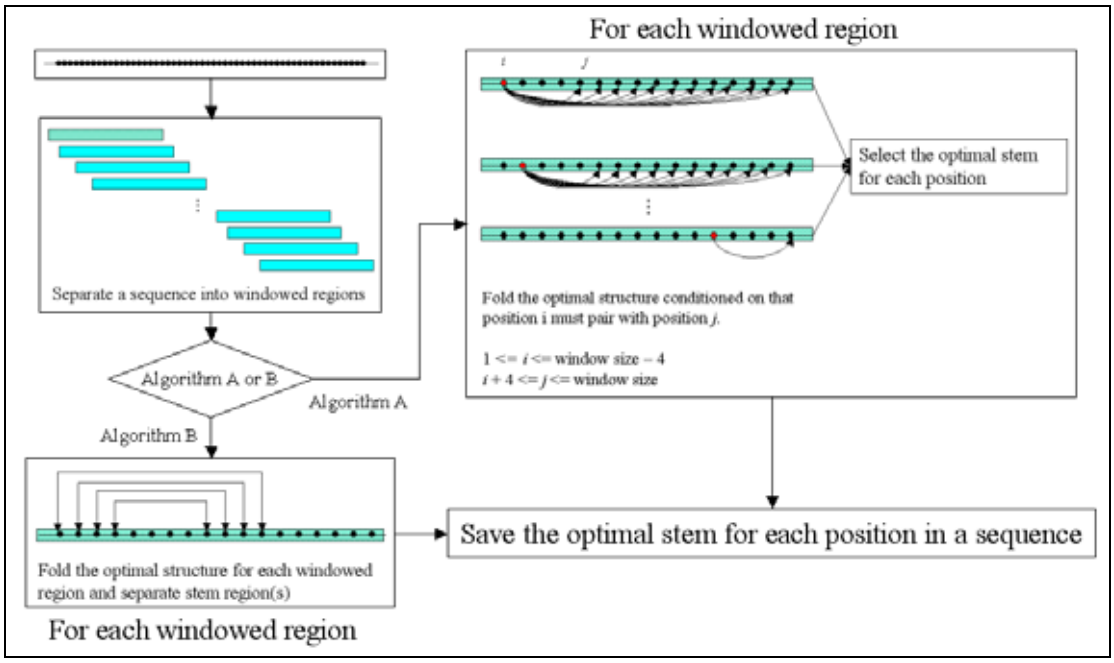


Figure 4-3. Two modes (algorithm A: the stringent mode and algorithm B: the fast mode) for finding local hairpins for windowed regions

In order to compare the performance of different approaches for predicting RNA structural motifs, human tRNAs of exactly the same length, 72 bases, were used as the test data set. Windows of different sizes were also tried to investigate possible effects. The targets for this evaluation included D arm, anticodon arm, and T arm (Figure 1-3), of 168 human tRNAs. The implementation for predicting RNA structures follows Zuker's MFOLD algorithm and uses the same parameters (Zuker 1989). The result reveals that the first approach (Algorithm A, Table 4-1) is better than the second one (Algorithm B, Table 4-1); however, it also suggests that the second approach is still useful, if the results of the second approach are compared to the predictions made by RNAfold (default, RNAfold, Table 4-1) (Hofacker et al. 1994-2006) with default parameters.

Algorithm A

	D arm	Anticodon arm	T arm
Window size: 50	112	150	132
Window size: 100	112	150	131

Algorithm B

	D arm	Anticodon arm	T arm
Window size: 50	80	146	131
Window size: 100	64	142	131

RNAfold

	D arm	Anticodon arm	T arm
default	35	28	58

Table 4-1. Performance of different algorithms for three hairpins of 168 human tRNAs

Algorithm A: The stringent mode. Individual hairpins are extracted from all optimal structures conditioned on that the i^{th} base should pair with the j^{th} base in each windowed regions, where $i < j <$ window size.

Algorithm B: The fast mode. Individual hairpins are extracted from the optimal structure for each windowed region.

Values in cells are the numbers of correct predictions (made by different algorithms) for respective arms. For D arm, the criteria of correct prediction is existence of a hairpin at 9^{th} or 10^{th} position, with stem size 3 ~ 4 base pairs and loop sizes 7 ~ 10 bases. For anticodon arm, the correct prediction should be at 26^{th} or 27^{th} position, with stem size 4 ~ 5 base pairs and loop size 7 ~ 9 bases. For T arm, the correct prediction should be at 48^{th} or 49^{th} position with stem size 4 ~ 5 base pairs and loop size 7 ~ 9 bases. The performance of RNAfold is assessed by using its default parameters.

In addition to the successful identification of three distinct hairpins of tRNAs, both Algorithms A and B predict extra hairpins. The biological significance of these extra hairpins is not clear. It is possible that these secondary structures could never fold in real tRNAs because they are relatively unstable compared to the optimal structures of individual tRNAs. By using the Eponine learning scheme, this redundancy should not be a serious problem, because only stable hairpins that can be consistently found in individual sequences are useful for distinguishing positive training sequences from negative training sequences. In the following text, algorithms A and B are referred to as the stringent model and the fast mode, respectively, of the Eponine RNA-motif extension.

4.2.2.2. Modelling structural features with probability distributions

Having evaluated the capability of the module responsible for locating local hairpins in sequences, consideration is now given to applying the Eponine training framework to model RNA motifs. One important issue is about designing a scoring scheme of the secondary structures in sequences.

Before moving further to discuss the scoring of complex RNA motifs composed of many hairpins, the scoring of a simple hairpin is first considered. In an oversimplified hairpin (Figure 1-2, A), there is only one single-stranded region (hairpin loop), and one non-interrupted double-stranded region (stem). Numerical parameters, which can potentially be applied to distinguishing one simple hairpin from the other, include dimensions of hairpins, free energy of the local region, free energy of the stem region, *etc*. Dimensions of each hairpin include loop size and stem size. If functions of RNA structural motifs depend on adequate combinations of individual features, then it seems reasonable to draw an analogy between primary-sequence motifs and features of RNA hairpins. Each feature of a hairpin seems analogous to each column of a PWM.

Each column of PWMs is a discrete distribution over all possible symbols in the used alphabet; similarly, each feature of hairpins can be modelled with a probability distribution. The mean of each distribution is the most frequently found value for one particular feature. For example, because the most frequently found stem size for *rho*-independent transcription termination signals is 9 (de Hoon et al. 2005), the mode of the corresponding discrete probability distribution should be 9. The deviation of each distribution can represent the degree of variations, such as different stem sizes that are observed in *rho*-independent transcription termination signals.

The probability of emitting a sequence x that harbours an RNA structural motif (RM) is:

$$RM(x,i) = (\prod_{r=1}^{R} P_r(F_r(x,i)))^{\frac{1}{R}} \quad [4\text{-}12]$$

, where R is the number of features that are used to model each hairpin; P_r is the proposed probability distribution of the r^{th} feature of a particular RNA structural motif; the model of this structural motif is $P = (P_1, P_2, P_3 \ldots, P_R)$; F_r is the function that returns the numerical value of the r^{th} feature of a hairpin, which folds at the i^{th} position of sequence x. $\frac{1}{R}$ is used to normalize the score of each hairpin. It seems this normalization is unnecessary; however, it is very important for modelling primary-sequence and structural motifs simultaneously. For each primary-sequence motif, the PWM score is the normalized joint probability of individual positions. For generating a scoring scheme that can sensibly combine scores from both PWM scores and RM scores, a similar normalization that is applied to PWM scores should also be applied to hairpin scores.

Compatibility between RM scores and PWM scores is one of most critical issues in developing the Eponine RNA-motif extension. If the extension uses an inappropriate scoring scheme that may make the order of magnitude of RM scores significantly different from that of PWM scores, the trained models may be biased to contain only RMs or only PWMs. Before the use of normalized RM scores, empirical rules have been used in order to make non-normalized RM scores compatible with PWM scores. For example, by comparing distributions of the scores of PWMs and non-normalized RMs, some multiplication factors were derived for transforming RM scores. However, the optimal value of the multiplication factor may change greatly under different conditions, especially when more than two different structural features are used to model RNA structural motifs.

By using joint probability of structural features to score each hairpin, many structural features can be modelled explicitly. By contrast, some features, such as stability of a particular

hairpin, cannot be modelled explicitly by using CMs. In addition, with normalized RM scores, distinct features can be treated as individual columns of a PWM. Theoretically, it is possible for the Eponine trainer to randomly choose distinct features to learn an optimal sensor for an RNA structural motif, just as the addition and subtraction of columns in learning the optimal PWM for modelling a primary-sequence motif (for details see subsection 4.1.2.1.1.).

Currently, the probability distribution for modelling each structural feature is a discrete Gaussian distribution; however, it should be noted that a discrete Gaussian distribution may not be the best one for describing all the distributions of stem size, loop size, local energy, *etc*. If there is a strong peak in the distribution of structural features, the width (deviation) of a Gaussian distribution should be assigned a small value, such that the there are light tails in this distribution. However, in cases where the distribution of features is flat within a certain range, the width of the Gaussian distribution must be a large value in order to simulate the flatness in local regions.

4.2.2.3. Applying RM scores to the EAS and EWS models

With the RM scoring scheme created in the previous section, the Eponine RNA-motif extension is able to model RNA motifs that are composed of primary-sequence patterns and secondary-structure motifs. In the following, the way the RM scoring scheme is adapted into the existing Eponine sequence models is introduced.

4.2.2.3.1. Using RM scores in the EAS model – the Eponine Anchored RNA-motif model

The formulation of basis functions for the EAS model is:

$$\phi(x) = \frac{\log(\sum_{i=-\infty}^{+\infty} P(i) \cdot W'(x, i+a) \ )}{|W'|} \qquad [4\text{-}13]$$

For modelling structural motifs:

$$W'(x, i+a) = \exp(RM(x, i+a)) \quad [4\text{-}14]$$

and

$$|W'| = 1 \quad [4\text{-}15]$$

The operation "exp" is used for avoiding the exceptional situations where the returned value from a *RM* is 0. This situation may occur when there are no significant RNA motifs starting at a particular position in a sequence. $|W'|$ is assigned with 1, because the normalization has been performed in the calculating the value of each *RM* (see [4-12]). Apart from that, for modelling primary-sequence motifs, W' is simply replaced with *W*. Such an extension to the Eponine EAS model is referred to as the Eponine Anchored RNA-motif model (the EAR model)

The new Eponine trainer uses a Monte Carlo sampling process for learning an optimal set of positioned RMs: 1) the mean and width of distributions are assigned randomly; 2) new RMs are generated by sampling features from all hairpins predicted in all training sequences; 3) new RMs can also be generated by adjusting the mean or the width of randomly chosen distributions of structural features in existing RMs. After positioned RMs are updated, the Eponine trainer uses the RVM to re-estimate their respective weights, which correspond to weights of basis functions in GLMs.

4.2.2.3.2. Using RM scores in the EWS model – the Eponine Windowed RNA-motif model

The formulation of basis functions for the EWS model is:

$$\phi(x) = Z \times \underset{s=u}{\overset{v}{optimal}}\left(\prod_{k=1}^{K}\left(\sum_{i=-\infty}^{\infty} P_k(i) \cdot W'_k(x, s+i)^{\frac{1}{|W'_k|}}\right)^{\frac{1}{K}}\right) \quad [4\text{-}16]$$

For modelling structural motifs, W' is substituted with *RM*. For modelling

primary-sequence motifs, W' is substituted with W. Such an extension to the Eponine EWS model is referred to as the Eponine Windowed RNA-motif model (the EWR model).

The Eponine trainer uses a Monte Carlo sampling process, which is similar to the optimization of RMs for the EAS models, to optimize the parameters of RMs for the EWS models.

Consequently, by using the scoring scheme designed to simultaneously model RNA structural and primary-sequence motifs, Eponine is now capable of modelling the consensus RNA motifs in a set of anchored or unanchored sequences.

4.3. Summary

In this chapter, I introduced methods for motif finding and functional site finding in preparation for modelling regulatory regions that may be implicated in the transcription of ncRNAs. For the purpose of finding functional sites, such as TSSs and TTSs, in complex genomes, there are three main requirements:

- Modelling an association of multiple motifs to describe functional sites.
- Modelling the distribution of individual motifs with respect to a particular functional site location.
- High selectivity in classification of functional sites in a large genome.

At the time of preparation of this thesis, Eponine appears to be one system that takes all these issues into consideration. Therefore, in the next chapter, the Eponine sequence models are applied to the modelling of the TSSs of mammalian RNA polymerase III genes.

In addition, I developed a new RNA-motif extension to the Eponine sequence models.

This new extension is particularly designed for finding the consensus RNA motifs in a set of sequences. The unique features of this new tool include that:

- It is an alignment-independent method.
- The models so trained may consist of primary-sequence patterns and secondary-structure motifs, which may give insights to the functional regions in a set of sequences.
- It is a local RNA-motif modelling tool, which means that a global conservation of RNA secondary structures in the set of sequences under investigation is not required.
- It may still work if not all the sequences under investigation fold into the same RNA motifs.
- The models so trained may potentially be useful for discriminating in genomes the functional sites associated with RNA motifs.

Chapter 6 is dedicated to the evaluation of the capability of the new RNA-motif modelling tool. The potential applications of the Eponine RNA-motif extension in genome-wide ncRNA finding will also be explored.

Chapter 5. Modelling the transcription regulatory elements of mammalian RNA polymerase III genes

Most existing ncRNA finding algorithms are designed to find structural ncRNAs. These algorithms can be regarded as being structure-dependent, because they use the potential of a particular genomic region to fold into high-order RNA structures as a signal of the existence of ncRNAs. However, structure-dependent ncRNA finding algorithms will fail to predict non-structured ncRNAs, whose functions do not depend on folding into high-order structures. In addition, a non-transcribable genomic region may be misclassified as an ncRNA locus simply because a region of structure-formation potential is predicted by structure-dependent algorithms. Therefore, 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 this chapter, the possibility of using a type of structure-independent genome-wide ncRNA finding approach is explored, based on the modelling of the transcription regulatory elements.

Transcription regulatory elements have been used as a signal for finding particular classes of ncRNAs, such as tRNAs (Fichant and Burks 1991; Pavesi et al. 1994; Lowe and Eddy 1997). However, the identification of transcription regulatory elements is currently used as a screening step, not as a determination step, in genome-wide ncRNA finding. If transcription regulatory element methods are used alone for genome-wide ncRNA finding, the false-positive rate can be very high. For instance, eufindtRNA, which is an internal-promoter finding program, predicts over 1,300 candidate loci for tRNAs on human chromosome 1 (in the NCBI 35 assembly), but only less than ~10% (120) of them may be functional tRNAs based on evaluation using structure-folding potentials.

It is essentially unknown why the methods designed to predict the transcription

regulatory elements of ncRNAs appear to suffer from high false positive rates. Some possible explanations are as follows. Firstly, it is possible that existing promoter models were not built specifically for finding mammalian tRNA genes. The specificity of these tools may have been sacrificed, to a certain extent, in order to make them sensitive enough for finding tRNA genes in multiple organisms. Secondly, internal promoters may be just part of the signal required for determining the transcription specificity of tRNA genes in mammalian genomes. It is possible that other non-promoter transcription regulatory elements, such as enhancers/silencers and LCRs, may play a role in the specific initiation of tRNA transcription. Thirdly, some of the non-tRNA loci which appear to contain the internal-promoter-like patterns might correspond to novel non-tRNA ncRNA genes.

Consequently, the specific aims of this chapter include:

- Learning 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.
- Finding evidence to support the functionality of the predicted non-tRNA pol III genes.

The Eponine system, described in chapter 4, appears to be suitable for these purposes. Eponine models have previously been used to predict functional sites, such as transcription start sites (TSSs) and transcription termination sites (TTSs), in complex genomes. Given a set of training sequences, the Eponine trainer can simultaneously learn the important signals, in the form of PWMs, and the "architectural" relationship (*i.e.* the distance distribution) of PWMs to a particular type of functional sites (for a detailed discussion see section 4.1, chapter 4). Eponine is one of the few systems that have been applied to learning a model capable of selectively predicting the TSSs of protein coding genes in mammalian genomes (Down and

Hubbard 2002). Given Eponine's success in modelling RNA polymerase II (pol II) TSSs, one interesting question is whether Eponine models are useful for predicting the ncRNAs transcribed by pol III in mammalian genomes. Therefore, in this chapter, the Eponine system was taken as a quick approach for modelling the transcription regulatory regions of mammalian pol III genes.

In this chapter, the Eponine Anchored Sequence (EAS) model (see section 4.1, chapter 4) was tried for creating a new model for discriminating pol III genes in the mammalian genomes.

5.1. Modelling the transcription start sites of mammalian pol III type II genes

In this section, the Eponine Anchor Sequence (EAS) model was used to model the transcription start sites (TSSs) of pol III genes. A suitable training set should consist of the genes that contain promoters with similar architectures, because the EAS model is not designed for managing a heterogeneous set of functional sites that are each associated with distinct combinations of transcription factor binding sites (TFBSs). For that reason, a brief introduction to the types of promoter architectures of eukaryotic pol III genes is given in the following.

There are three distinct types of promoter architecture that have been found in eukaryotic pol III genes, where each type of promoter is associated with a unique combination of distinct TFBSs (see Table 5-1) (for review see Paule and White 2000). The promoters of type I and type II genes are intragenic. Type I (*e.g.* 5S rRNAs) and type II (*e.g.* tRNAs) genes share an "A box" (sometimes also known as the "A block"), which is the binding site of TFIIIC. A "C box" (sometimes also as the "C block"), which is the binding site of TFIIIA, is unique to type I genes. A "B box" (sometimes also as the "B block"), which is the binding site of TFIIIB, is

unique to type II genes. Although "A boxes" for tRNAs and 5S rRNAs can be exchanged, the distances to their respective TSSs vary: it seems that the distance for tRNA genes is 10 bases, while the distance for 5S rRNAs is 50 bases. Although there are no TATA boxes for mammalian type I and II genes, the transcription factors (TFs) that interact with intragenic TFBSs seem to guide TATA-Box Binding Protein (TBP) to the upstream regions of type I and II genes and TBP can recruit pol III to the correct transcription start sites. On the other hand, promoters of type III genes are 5' to the TSS in the upstream region. Unique TFBSs of type III genes are the TATA box, the proximal sequence element (PSE), and the distal sequence element (DSE).

Type	Genes	Core TFs	TFBFs
Type I	5S rRNAs, *etc.*	TFIIIA, TFIIIC, TFIIIB, TBP, polIII	A box and C box (Intragenic regions)
Type II	tRNAs, VARNAs, 7SL, *etc.*	TFIIIC, TFIIIB, TBP, pol III	A box and B box (Intragenic regions)
Type III	U6 snRNAs, 7SK, *etc.*	TFIIIC1, TFIIIB, TBP, SNAPc, pol III	PSE, TATA box, DSE (Upstream regions)

Table 5-1. The TFs and the TFBSs associated with three distinct types of eukaryotic pol III genes

Given these distinct architectures, when creating a model that may discriminate tRNA genes as well as other pol III genes, the sources of training sequences needs to be limited to those of pol III type II genes. In the set of pol III type II genes, VARNA1 genes can be another source of training sequences, in addition to tRNAs. To date, more than 40 VARNA1 genes have been found. Although there are other pol III type II genes such as 7SL, these genes are not as numerous as VARNA1 genes. VARNA1s are encoded in adenoviruses (Weinmann et al. 1974) and they are transcribed by the mammalian RNA pol III machinery. Hence, VARNA1 genes can be considered as mammalian pol III type II genes, because there is evidence that the promoters of VARNA1 genes are similar to these of mammalian tRNA genes (Cannon et al. 1986; Wu et al. 1987).

Thus, in this section (5.1), VARNA1s and tRNAs were used as training sequences to generate an Eponine EAS model for pol III type II TSSs.

5.1.1. Materials and methods

5.1.1.1. Training and test data sets

For the purpose of creating an EAS model, one set of positive sequences and one set of negative sequences are required.

The human tRNA genes and adenovirus VARNA1 genes were used as the positive sequences. The set of mouse tRNA genes predicted by tRNAscanSE were not included because the set might contain a large number of pseudogenes (Mouse Genome Sequencing Consortium 2002). A set of negative sequences were recruited by taking random samples from the human genome. The preparation of these sequences for training and testing is described in the following subsections (subsections 5.1.1.1.1. , 5.1.1.1.2. , and 5.1.1.1.3.).

5.1.1.1.1. Preparation of human tRNA sequences

In order to avoid over fitting of a learned model to training data, validation is necessary. One type of validation is to evaluate the performance of trained models on test data that is independent of the training set. If the performance of a trained model is significantly worse than on the training data, this may indicate that this model has been over fitted to the training data.

Therefore, the recruited tRNA genes were partitioned into two groups, one for training and the other for testing. Due to the high redundancy in the set of human tRNA genes, proper partitioning became an issue. For instance, there can be as many as 20 nearly identical copies for a particular anti-codon type of tRNA genes. When using a random sampling process, it is unlikely that all the highly similar tRNA genes would be grouped into a single set. Here, I took advantage of the forty tRNA-gene subgroups already prepared in section 2.2, chapter 2,

where these subgroups were generated according to the anti-codon types and pairwise sequence identities of tRNA genes (for details see materials and methods of section 2.2, chapter 2). These forty subgroups were re-merged into two groups, group 1 and 2, based on the pairwise identities between the consensus sequences of the subgroups. The grouping process was carried out in a progressive manner, where the two groups with the highest consensus identity were merged first, and then the groups with the next highest identity were successively merged.

Group 1 and group 2 consisted of 200 and 167 human tRNA genes, respectively. The inter-dependence between the training set and the test set was further assessed by comparing the inter-group and intra-group sequence identities. Each sequence was used to search for its most similar sequences in the same group and in the other group, respectively. The results reveal that there is a clear sequence-identity difference between these two groups, since all the intra-group best pairwise identities were greater than 83% and all the inter-group best pairwise identities were smaller than 78% (Figure 5-1). The results suggest that the tRNA genes in group 1 are distinct from the tRNA genes in group 2. The tRNA genes in group 1 (group-1 tRNA genes) were used for training and the tRNA genes in group 2 (group-2 tRNA genes) were used for testing.

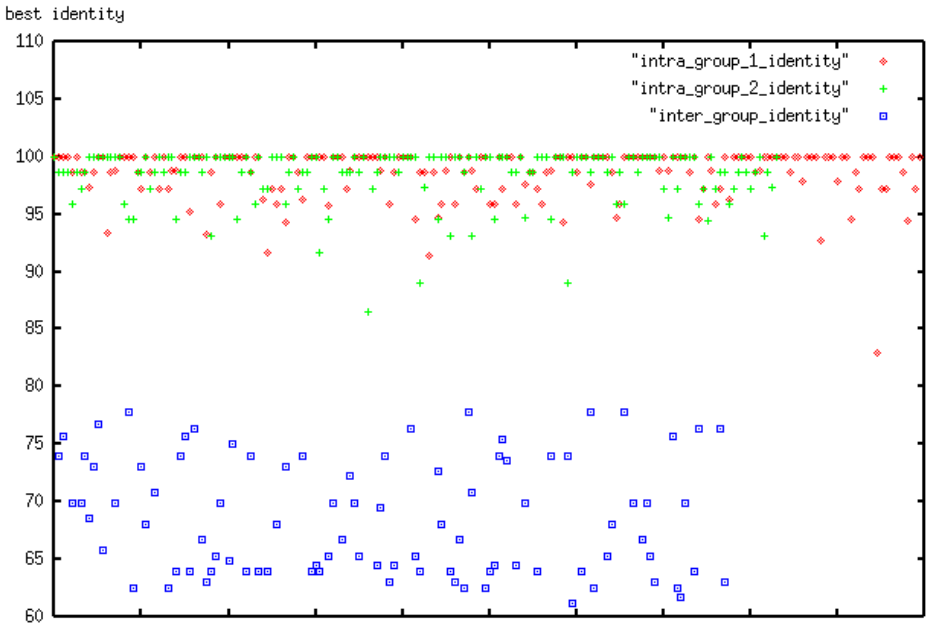


Figure 5-1. Separation of the sequence identity distributions between intra-group and inter-group sequences of tRNA genes.

When preparing the tRNA sequences for training and test, the first base of the cloverleaf-like structure of each recruited tRNA gene was used as the anchoring point. 100 bases upstream and 150 bases downstream with respect to the anchoring point in each human tRNA gene were retrieved. The purpose of including the upstream and downstream flanking regions of the recruited tRNA genes in training sequences is to explore if there are motifs other than the A box and B box that can be used to model the TSSs of pol III type II genes.

5.1.1.1.2. Preparation of VARNA1 sequences

VARNA1 genes were used as another source of sequences for building a pol III type II TSS model. Forty-three regions containing VARNA1 genes were retrieved from GenBank by using the keyword "VARNA1". VARNA1 sequences were extracted from these regions by using the locations indicated in the GenBank annotation. By using TGICL (TIGR 2002-2003), VARNA1 genes were clustered into 5 subgroups (for the detailed procedure for the sequence clustering, see section 2.2, chapter 2). The 5 subgroups were further merged into two independent groups. Group 1 and group 2 consisted of 9 and 32 VARNA1 genes, respectively.

An assessment on the sequence independence, as mentioned in preparing the human tRNA genes for training, was also performed here. The results show that all the intra-group best pairwise identities were greater than 95%; all the inter-group best pairwise identities were smaller than 86%. The results suggest that the VARNA1 genes in group 1 are distinct from the VARNA1 genes in group 2.

Group-1 VARNA1 genes together with group-1 tRNA genes were used for training (Table 5-2, Training). Group-2 VARNA1 genes and group-2 tRNA genes were used for testing (Table 5-2, Testing). The 32 genes used for testing actually correspond to 9 distinct ones, because many of them have exactly the same sequences. Likewise, the 9 genes used for training correspond to 7 distinct ones.

	Training	Testing
Human tRNA genes	200 (group 1)	167 (group 2)
Adenovirus VARNA1 genes	9 (group 1)	32 (group 2)
Subtotal	209	199

Table 5-2. The training and test data sets for creating an EAS model for pol III type II TSSs

When preparing the VARNA1 sequences for training and test, the first base of each gene was used as the anchoring point; 100 bases upstream and 150 bases downstream with respect to the anchoring point in each VARNA1 gene were retrieved. The purpose of including flanking sequences for training is the same as described to prepare tRNA sequences for training in the previous subsection (see subsection 5.1.1.1.1.).

5.1.1.1.3. Preparation of negative sequences

Two sets of ten thousand random sequences were sampled from the human genome as negative training and test sequences, respectively. These random sequences were 250 bases in length.

5.1.1.2. Evaluation of the performance of EAS models against the test data set

When evaluating the accuracy of trained EAS models against the test data set prepared as described in 5.1.1.1. , the 100th base of each test sequence was taken as the anchoring point. A true positive was determined, if any region within 5 bases away from the anchoring point of a positive test sequence was predicted as a hit. A false positive was determined, if any region within 5 bases away from the anchoring point of a negative test sequence was predicted as a hit.

5.1.1.3. Presentation of the performances of different models

The performances of all trained models will be presented in the form of coverage-accuracy (C-A) plots. Coverage (sensitivity) is the proportion of true positive sequences that can be correctly predicted; accuracy (positive predictive value) is the proportion of true positive sequences in the set of predicted sequences. For example, with a specific threshold, if 150 out of 199 positive test sequences are successfully predicted and 5 out of 10000 negative test sequences are incorrectly classified as the pol III type II genes, the accuracy is 96.8% (150/(150+5)) and the coverage is 75.4% (150/199).

The C-A plot can be considered as an alternative presentation of Receiver Operating Characteristic (ROC) curves, except that the size of negative test sequences is not considered in the former plot. Plotting these characteristics is especially useful when comparing the performances of two competing models when using an extremely large negative data set, such as random sequences from the human genome. Suppose that there are two models, where model X predicts 150 false positives from 10,000 negative test sequences, while model Y predicts 100 false positives. Both models can predict 150 true positives from 200 positive test sequences. The false positive rates are 1.5% and 1% respectively. In contrast, the accuracies for these models are 50% (150/(150+150)) and 60% (150/(150+100)), respectively, and thus the difference between their performances can be easily seen in a C-A plot. Consequently, for

evaluating the performances of methods that are designed for finding functional sites in large and complex genomes, C-A plots are more suitable than the classic ROC curves.

5.1.1.4. Evaluation of the performance of EAS models against real genomic sequences

The performance of EAS pol III type II TSS models was also evaluated against human chromosomes 11 and 13. The human genome assembly used in this evaluation was NCBI 35. These sequences were retrieved from the Ensembl ftp site (ftp://ftp.ensembl.org/pub/).

When using EAS pol III type II TSS models to scan a chromosome, each position can be the start of a putative pol III type II gene. Consecutive hits would be clustered together if all of their scores were higher than a particular threshold. Such hits were regarded as a single record of prediction.

5.1.1.5. Determining overlapped genomic hits predicted by using different methods

An EAS pol III type II TTS model predicts the transcription start sites in genomes. By contrast, existing tRNA gene finding algorithms, such as eufindtRNA and tRNAscanSE, predict a range, namely the start and end positions for each putative tRNA gene. To determine the overlapped hits predicted using different methods, the following approach was used. If a tRNAscanSE (or eufindtRNA) predicted hit was within 100 bases downstream of an EAS pol III type II TTS model predicted site, the two hits predicted by different methods were considered to represent the same gene.

5.1.2. Results

5.1.2.1. Naïve training by using default parameters

Using the training sequences prepared as described in 5.1.1. , an Eponine Anchored Sequence (EAS) model for the mammalian pol III type II promoters was trained. Figure 5-2 is a schematic presentation of the constraint distributions relative to the anchoring point as

indicated by the blue triangle. The anchoring point in this figure corresponds to the transcription start site of pol III type II genes. The relative width of the position distributions for each hairpin is shown by the width drawn. The sequence under each constraint is motif consensus sequence. The sequence logos of the motifs in this model were presented in Figure 5-3. In the remaining part of this thesis, other Eponine models will be presented using this convention.

There were several problems with this model. Firstly, the model was unable to distinguish *bona fide* tRNA genes from random sequences (data not shown). Secondly, both the patterns of A box and B box were much shorter than what have been suggested by experimental approaches (DeFranco et al. 1980; Galli et al. 1981). Further investigation revealed that between VARNA1s and the human tRNAs, the 8th to 22nd positions, which are supposedly the "A box", are very different.

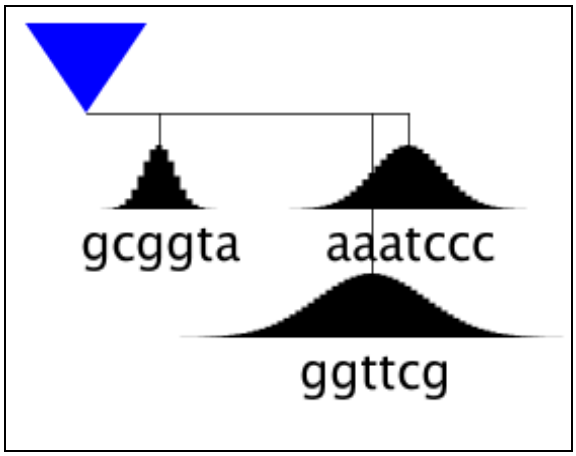


Figure 5-2. An EAS model for pol III type II promoters (naïve training)

Weight: 6.52, position: 6, width: 7.36 Weight: 11.28, position: 53, width: 3.18

Weight: 6.91, position: 68, width: 5.99

Figure 5-3. The sequence logos of position-constrained motif matrices of the naïve EAS model (Figure 5-2) for pol III type II promoters

The value of "weight" for each motif corresponds to the weight associated with each basis function in the GLM of an EAS model. The value of "position" for each motif corresponds to the mean of the discrete Gaussian distribution used to model the position of a motif relative to the reference site. The value of "width" corresponds to the width of the discrete Gaussian distribution (for other details about these parameters see subsection 4.1.2.1.1)

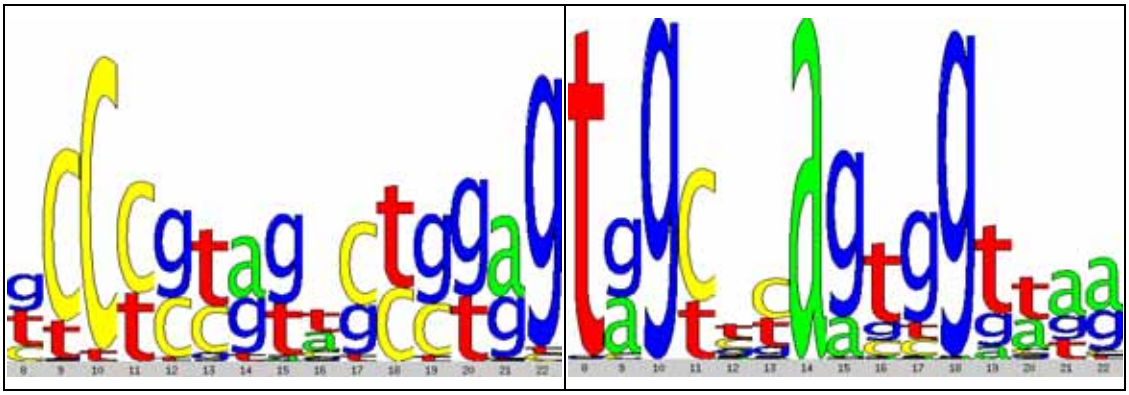


Figure 5-4. Comparison between the sequence logos of the 8th-22nd positions of VARNA1s (left) and tRNAs (right)

5.1.2.2. Optimizing the anchoring points

From the results presented above, VARNA1s, which are viral genes rather than real mammalian genes, seem to be unsuitable for training pol III type II promoter models. However, on investigation it was found that the poor training was probably due to the incorrect assignment of the anchoring points for the recruited sequences. The first base of the cloverleaf-like structure of tRNAs, is in fact not the transcription start site. The real transcription start sites of mammalian tRNAs are at the 5' regions upstream of the first base of cloverleaf-like structures. After transcription, the 5' dangling sequences of the raw tRNA transcripts must be cut off by RNase P (for review see Gopalan et al. 2002). On the other hand, transcription start sites of VARNA1s are generally used as the first bases for VARNA1 genes in the GenBank annotation.

After adjusting the anchoring points of the recruited sequences, manual alignments reveal that respective "A boxes" of VARNA1s and the human tRNAs are quite similar (Figure 5-5). These results show that when inconsistent anchoring points are provided, the Eponine trainer for the EAS models can be incapable of optimizing the PWMs.

Figure 5-5. Comparison between sequence logos of the presumable internal promoter regions of VARNA1s (left) and tRNAs (right) (after adjusting anchoring points of VARNA1s)

5.1.2.3. The EAS pol III type II promoter model

Using the sequences with correct anchoring points, a new EAS pol III type II promoter model was trained. This model is called "model 1" in the remainder of section 5.1. This model appears to be quite complex (Figure 5-6). There are five distinct motifs at the 6th, 19th, 43rd, 52nd, and 53rd positions. Respective weights for these motifs in the generalized linear models are 4.76, 8.34, 4.37, 9.01, and 12.58.

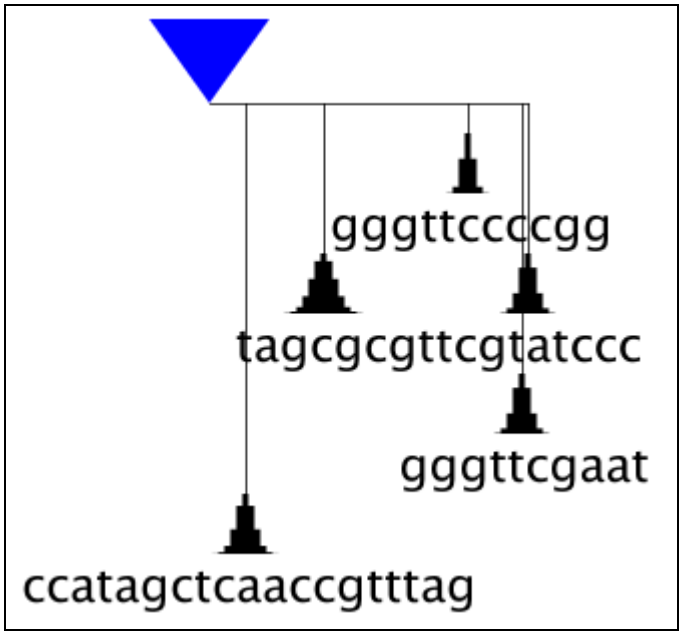


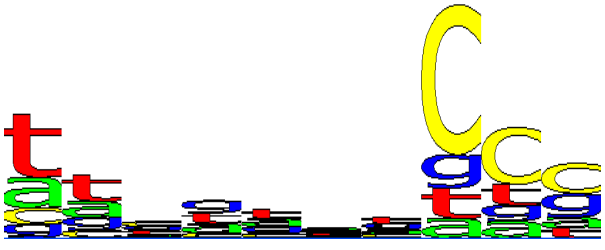
Figure 5-6. An EAS pol III type II promoter model (after adjusting the anchoring points of VARNA1s) (model 1)

Weight: 12.57, position: 6, width: 1.45

Weight: 4.37, position: 19, width: 1.87

Weight: 4.76, position: 43, width: 0.87

Weight: 9. position: 52, width: 1.30



Weight: 8.34, position: 53, width: 1.30

Figure 5-7. The sequence logos of position-constrained motif matrices of model 1 (Figure 5-6)

The annotation used in this figure follows the convention of Figure 5-3

The motifs in the new model fit the current knowledge about transcription regulation of mammalian pol III type II genes. The motifs that start at 6th and 19th positions correspond to the 5' and 3' parts of the "A box" respectively. The motifs that start at 43rd, 52nd, and 53rd positions, which are similar to one another, correspond to the "B box". The three positions represent discrete preferred sites of the "B box" in mammalian tRNA genes. The variation in the location of the "B box" is consistent with the previous reports which indicated the flexibility in distance between the "A box" and the "B box" in eukaryotic tRNA genes (Camier et al. 1990; Pavesi et al. 1994).

5.1.2.3.1. The performance of model 1 – using the recruited test sequences

The performance of model 1 was initially assessed against 199 positive test sequences recruited as described in 5.1.1.1.1. and 5.1.1.1.2. , and a set of 10,000 negative test sequences prepared as described in 5.1.1.1.3. The results reveal that model 1 can achieve 100% accuracy at 70% coverage on this data set (Figure 5-8, model 1). The high accuracy suggests that model 1 may have a low false positive rate. Besides, at this accuracy and coverage, ~50% distinct VARNA1 sequences in the test data set were successfully predicted. These results suggest that model 1 can potentially be applicable to genome-wide pol III type II gene finding. The performance of model 1 is further evaluated using real genomic sequences in the following subsection (5.1.2.3.2.).

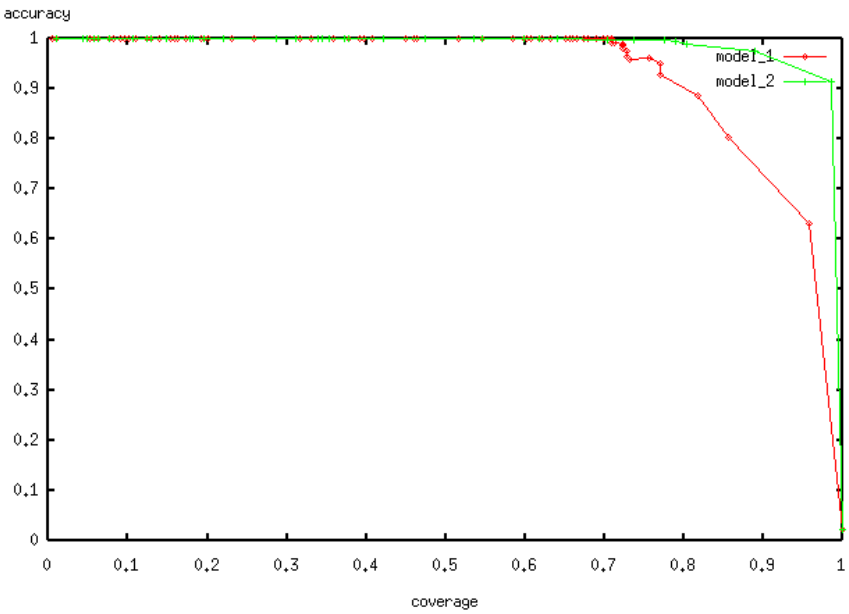


Figure 5-8. C-A plots of model 1 and model 2 on the test data set

5.1.2.3.2. The performance of model 1 – using human chromosomes 11 and 13

In order to assess the performance of model 1 in the context of real genomic sequences, this model was used to scan human chromosomes 11 and 13. In this subsection, a threshold corresponding to 100% accuracy and 66% coverage assessed against the test data set was chosen (Figure 5-8, model 1). It was found that the sizes of clustered hits were generally within the range of 1 to 3 bases, and none of them were longer than 5 bases (for definition of clustered hits see subsection 5.1.1.4.). This suggests that model 1 can detect pol III type II TSSs with good positional accuracy.

To compare the predictions made by using different methods, overlapped hits were determined as described in subsection 5.1.1.5. The methods discussed here include tRNAscanSE, eufindtRNA, and model 1. The predictions made by eufindtRNA were also compared here because eufindtRNA is a pure pol III type II promoter finding algorithm, not considering the structure-formation potential in a candidate region. In brief, eufindtRNA can be considered as an algorithm based on pure motif models. By contrast, tRNAscanSE is a hierarchical system which filters initial predictions made by other algorithms (*e.g.*

eufindtRNA, *etc*.), using structure-formation potential (for more details about how tRNAscanSE works see subsection 2.1.1.1. , chapter 2).

The results reveal that, for discriminating tRNA genes in the human genome, the performance of this model is comparable to existing algorithms (Figure 5-9 and Figure 5-10). Notably, the TSSs predicted by using model 1 and eufindtRNA frequently overlapped with MIRs. MIRs are mammalian interspersed repeats (Smit and Riggs 1995), which are tRNA-derived short interspersed repetitive elements (SINEs). The expected lengths of MIRs are ~260 bases. If the 300 bases upstream and downstream of the first base of each prediction were checked, as many as ~66% and ~51% of the TSSs predicted by model 1 on human chromosomes 11 and 13 respectively overlapped with MIRs (Table 5-3, model 1). Besides, ~57% and ~46% of the TSSs predicted by eufindtRNA on human chromosomes 11 and 13 respectively overlapped with MIRs (Table 5-3, eufindtRNA). In addition, 90.1% (20/22) and 100% (10/10) of the predictions made concurrently by both methods overlapped with MIRs (Figure 5-9 and Figure 5-10).

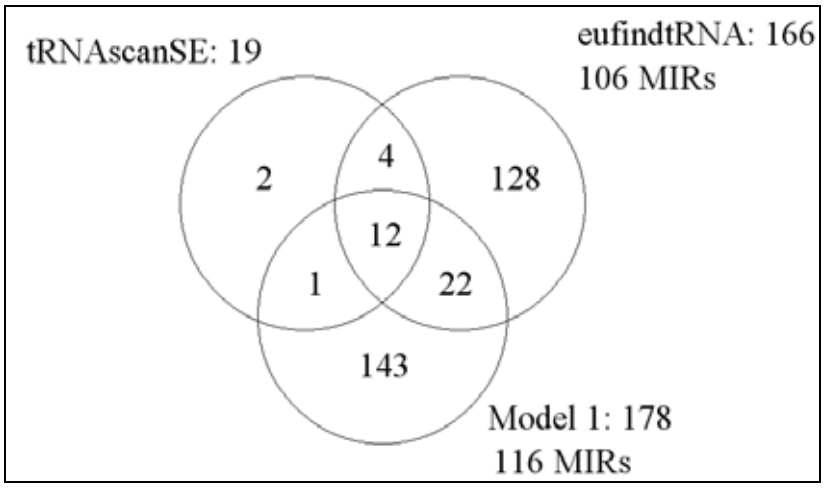


Figure 5-9. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 1) for human chromosome 11

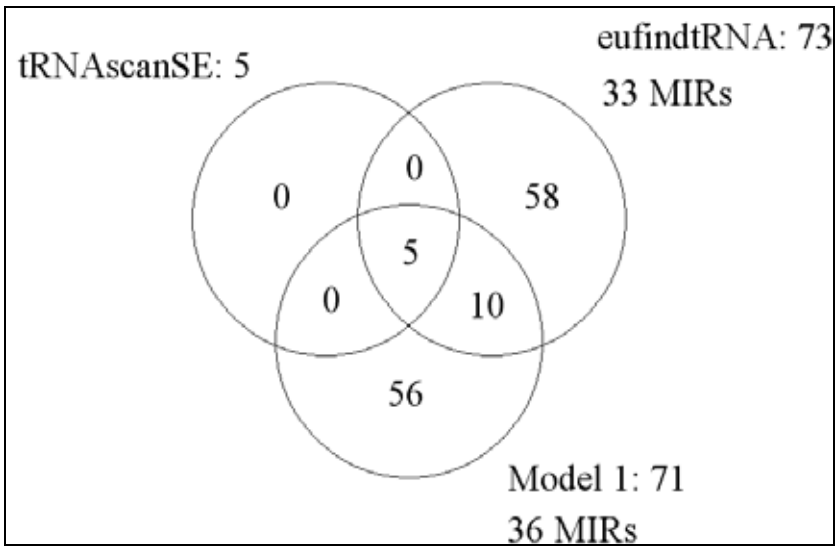


Figure 5-10. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 1) for human chromosome 13

	Human chromosome 11	Human chromosome 13
EufindtRNA	63.9% (106/166)	45.2% (33/73)
Model 1	65.2% (116/178)	50.7% (36/71)

Table 5-3. Ratios of MIRs in different predictions for pol III type II genes on human chromosomes 11 and 13

MIRs – functional transcripts or pseudogenes?

It was surprising that more than half the pol III type II TSSs predicted by both model 1 and eufindtRNA are MIRs. Since less than 6% and 3% of the sequences on human chromosomes 11 and 13 respectively are MIRs, there is obviously an enrichment of MIRs in the sets of predicted TSSs.

In order to explore whether these predicted TSSs correspond to functional transcription units, two approaches were taken. Firstly, the MIRs predicted by both model 1 and eufindtRNA were used as negative sequences for training a revised EAS pol III type II TSS

model (see subsection 5.1.2.4.). If MIRs are pseudogenes, their promoters should have been at least partially degraded and thus including MIRs in negative training sequences may improve the specificity of the Eponine pol III type II TSS model. Secondly, the conservation of these MIRs in human-mouse syntenic regions was examined (see subsection 5.1.2.5.). If some MIRs are synteny-conserved, they are more likely to be functional elements.

5.1.2.4. Model 2 – using MIRs as the negative training sequences

The MIRs that were detected by both model 1 and eufindtRNA on human chromosomes 11 and 13 were added into the set of negative training sequences. The trained model (Figure 5-11) appears to be more complex than the model trained using random human genomic sequences as the only source of negative training sequences (Figure 5-6) however maintains the motifs of model 1. This new model is referred to as model 2. There are six distinct motifs at position 5, 15, 18, 18, 21, and 53. While the final motif in model 2 corresponds to the "B box", the "A box" is now represented by five motifs and there are overlaps between motifs. The performance of model 2 is slightly better than model 1 (Figure 5-8), since its accuracy is higher than model 1 when coverage is 90% ~ 100%.

5.1.2.4.1. The performance of model 2 – using human chromosomes 11 and 13

In order to compare the performance of model 2 with that of model 1 in the context of real genomic sequences, model 2 was also used to scan human chromosomes 11 and 13. In this subsection, a threshold corresponding to 100% accuracy and 55% coverage evaluated against the test data set was chosen when using model 2. Given this threshold, the number of predictions made by model 2 on human chromosomes 11 and 13 was comparable to that previously made by using model 1 (see the denominators in Table 5-4). Besides, model 2 had good positional accuracy, similar to that of model 1 (for the positional accuracy of model 1 see subsection 5.1.2.3.2.).

Using model 2 to scan human chromosomes 11 and 13, far fewer of the TSSs predicted

overlapped with MIRs than when using the previous model (model 1) (Table 5-4). Only ~16% and 10% of predictions on human chromosomes 11 (Figure 5-13) and 13 (Figure 5-14) respectively overlapped with MIRs. Besides, no MIRs on human chromosomes 11 and 13 were predicted concurrently by eufindtRNA and model 2. However, one problem with model 2 is that, the prediction coverage of tRNA genes on human chromosome 13 is decreased from 100% to 60% (Figure 5-14) and on human chromosome 11 is decreased from 68% to 63%. The result suggests that it is difficult to train a pol III type II TSS model that can completely avoid predicting TSSs which appear to be only associated with MIR elements.

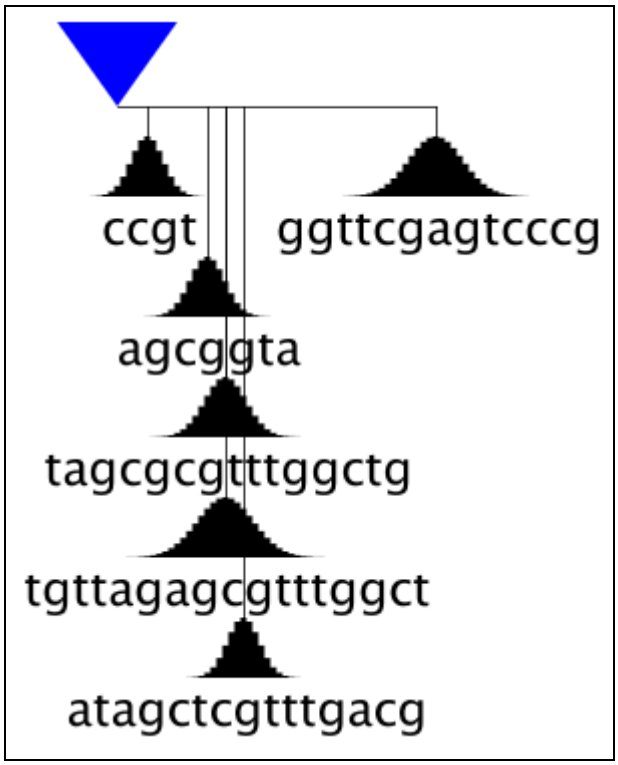
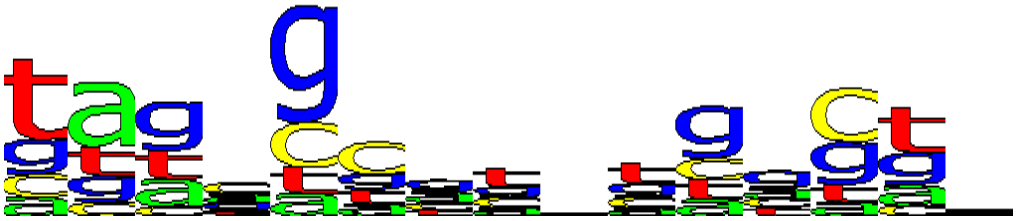


Figure 5-11. An EAS pol III type II model (using MIRs as negative training sequences) (model 2)

Weight: 4.18, position: 5, width: 2.66

Weight: 5.33, position: 15, width: 2.89



Weight: 8.11, position: 18, width: 3.44

Weight: 4.45, position: 18, width: 4.64

Weight: 18.85, position: 21, width: 2.66

Weight: 11.80, position: 53, width: 4.50

Figure 5-12. The sequence logos of position-constrained motif matrices of model 2 (Figure 5-11)

The annotation used in this figure follows the convention of Figure 5-3.

	Human chromosome 11	Human chromosome 13
Model 1	65.2% (116/178)	50.7% (36/71)
Model 2	16.0% (25/156)	10% (9/90)

Table 5-4. Ratios of MIRs in the predictions made models 1 and 2 for pol III type II genes on human chromosomes 11 and 13

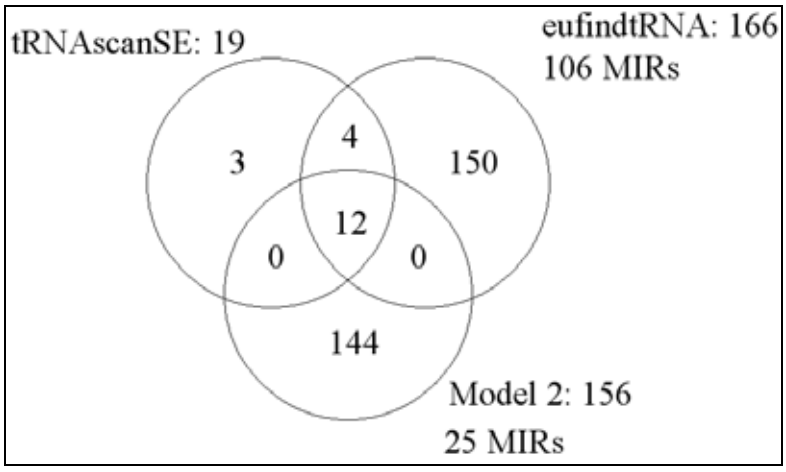


Figure 5-13. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 2) for human chromosome 11

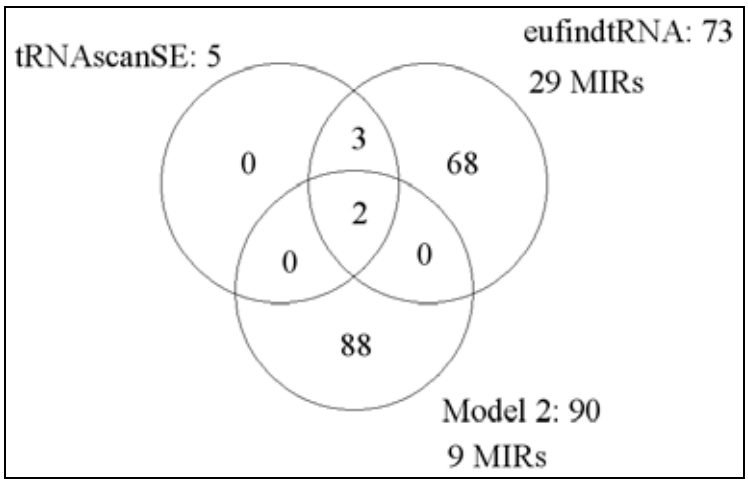


Figure 5-14. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 2) for human chromosome 13

One interpretation of these results is that modelling TSSs alone, *i.e.* without considering the structure-formation potentials, is insufficient to distinguish functional pol III type II genes from inactive MIRs. However another interpretation is that the predictions are correct and that this finding implies that some MIRs are still being actively transcribed. There is evidence to suggest that transcripts of repetitive elements may not be completely non-functional. For example, mouse B2 RNAs, which are the transcripts of a class of tRNA-derived SINEs, can specifically bind RNA polymerase II holozymes to repress transcript synthesis in response to heat shock (Allen et al. 2004; Espinoza et al. 2004). The EAS pol III type II models also predict TSSs which are not associated with tRNAs or MIRs. While some of these predictions may be false positives, it is also possible that some correspond to novel functional genes.

Consequently, in the following subsection (5.1.2.5.), I explore the functionality of the predicted TSSs that do not correspond to tRNA genes. These sites may include MIRs as well as non-MIR elements. The synteny-conservation of these regions was taken as an indicator of their functionality. If regions near the predicted TSSs are conserved in the human-mouse syntenic regions, this supports the idea of them being functional transcripts.

5.1.2.5. Investigating the human-mouse synteny-conservation of the predicted pol III type II TSSs

The human-mouse synteny-conservation of the pol III type II TSSs predicted by model 1 and eufindtRNA were examined. The method used here followed the same procedures as described in section 2.1, chapter 2. The results reveal that only a few of the predicted TSSs on human chromosomes 11 and 13 are synteny-conserved (Table 5-5). Most of those that were synteny-conserved were found in the intronic regions of protein-coding genes (Table 5-6). In general, the identities between the human and mouse synteny-conserved signals are lower than 80%, except that on human chromosome 13 one pair of human-mouse synteny-conserved signals predicted by model 1 has 95% identity. Does this case represent a novel pol III type II gene? It is difficult to make this conclusion because the high identity may be evolutionarily constrained by the function of the protein-coding genes, but not necessarily by the function of any pol III type II genes. In addition, most of the alignments of the other synteny-conserved predictions in Table 5-6 contain many indels.

Therefore, the conclusion is that synteny-conservation provides no clear evidence to support the functionality of the predicted pol III type II TSSs not associate with tRNA genes.

	Methods	Non-tRNA predictions	Non-tRNA predictions in syntenic regions in the mouse genome
Human chromosome 11	Model 1	165	5 [1]
	EufindtRNA	150	5 [2]
	Model 1 and eufindtRNA	22	0
Human chromosome 13	Model 1	66	2
	EufindtRNA	68	0
	Model 1 and eufindtRNA	10	0

Table 5-5. The synteny conservation of the non-tRNA pol III type II signals on human chromosomes 11 and 13

[1]: there are 3 MIRs in these 5 synteny-conserved signals. [2]: all the 5 synteny-conserved signals are MIRs.

	Methods	Synteny-conserved signals	Overlapping with known genes	
			Protein-coding regions	Unknown
Human chromosome 11	Model 1	5	5 (introns)	0
	EufindtRNA	5	3 (introns)	2
Human chromosome 13	Model 1	2	1 (exon)	1
	EufindtRNA	0	0	0

Table 5-6. Distributions of the synteny-conserved pol III type II promoter signals in intronic and exonic regions

"Unknown" means that there are no genes annotated in the regions predicted to be pol III type II genes

5.1.3. Discussion

I attempted to model pol III TSSs using the Eponine system because of its success when applied to the similar problem of modelling RNA polymerase II (pol II) TSSs (Down and Hubbard 2002). However, the results from modelling of the TSSs of mammalian pol III type II genes have been less clear. Firstly, creating a general pol III TSS model proved impractical due to the substantially different promoter subgroups, so it was decided to concentrate efforts on modelling the largest pol III type II subgroup. It was possible to train models that could be used to scan entire human chromosomes predicting the TSSs of majority of known pol III type II genes (tRNAs) while making relatively few other predictions. However the proportion of other predictions was much higher than when Eponine was used to predict TSSs for pol II genes (Down and Hubbard 2002). Numerous TSSs predicted by using the EAS pol III type II model overlapped with MIR repetitive elements. A similar phenomenon was also observed when the tRNA-gene finder, eufindtRNA, which primarily identifies the internal promoters, was used. The biological significance of these MIRs that may have good pol III type II promoters is unknown. No evidence can be found to support the suggestion that these MIRs might generate functional transcripts.

There are a number of possible ways of explaining these results including the following:

- If we assume the majority of predictions that do not match known pol III type II genes are false positives, maybe this indicates that the Eponine system is not sufficient to model pol III type II TSSs completely. One possibility might be that the Monte Carlo method used in the Eponine trainer was unable to learn optimal PWMs representing the internal promoters of mammalian pol III type II genes with the datasets used here, which were smaller than used for pol II training.

- Alternatively, it might be that the internal promoters are insufficient for regulating the transcription of mammalian pol III type II genes, making apparently valid pol III type II predictions non functional. Other non-promoter regulatory regions, such as locus control regions (LCRs) and enhancers/silencers, might be necessary for the transcription regulation of mammalian pol III type II genes. The observation that tRNA genes tend to exist in clusters might fit with some additional regulatory process.

With respect to the first possibility, further exploration of promoter modelling using other motif-finding approaches to predict pol III type II TSSs could be considered as future work. Since the original goal of the first part of this chapter was to test Eponine as a quick approach for modelling the TSSs of mammalian pol III type II genes, a comprehensive assessment of the performances of other approaches for modelling and discovering the TSSs is beyond the scope of this chapter.

With respect to the second possibility I explored if it is possible to detect any evidence for non-promoter transcription regulatory regions associated with mammalian tRNA gene clusters. However, the initial attempt to look for signals in regions around these tRNA gene clusters (as described in section 2.2, chapter 2) was inconclusive (data not shown) and thus future work is needed.

5.2. Summary

In this chapter, an attempt was made to model the transcription regulatory regions of mammalian tRNA genes. In the first part of this chapter, the transcription start site of mammalian pol III type II genes, including tRNA genes and VARNA1 genes, was modelled by using the Eponine Anchor Sequence (EAS) model. Important findings are as follows:

- The performance of the EAS pol III type II TSS models is comparable to that of existing methods, such as eufindtRNA, for identifying the TSSs of tRNA genes.
- Both the EAS pol III type II TSS models and the internal-promoter based tRNA gene finder may predict many repetitive elements, MIRs.
- By using MIRs as the negative training sequences, the performance of the new EAS pol III type II model cannot be further improved.

One future work is to try other motif-finding approaches to predict pol III type II TSSs. Another future work is to search for non-promoter regions regulating transcription of pol III type II genes that are clustered in mammalian genomes.

Chapter 6. Finding RNA motifs in genomes

In chapter 4 of this thesis, a new RNA-motif modelling tool based on the functional-site modelling tool -- Eponine was created. This new tool is particularly designed for modelling functional sites that may be associated with local RNA motifs. In addition, the models so trained should be capable of discriminating ncRNAs in genomes. Unlike other comparative algorithms that can be used for genome-wide ncRNA finding, this tool is not dependent on sequence alignments. Thus this tool may potentially provide an alternative approach for genome-wide ncRNA finding.

In this chapter, I assessed the capability of the Eponine RNA-motif extension. Two types of capabilities are of interest:

- The capability of the Eponine RNA-motif extension to find the consensus RNA motifs, consisting of both primary-sequence and secondary-structure motifs, in a set of transcripts
- The capability of the models so learned to discriminate a particular type of ncRNAs in genomes

Three types of different ncRNAs with distinct structural features were used to perform the capability assessment. The modelling of the mammalian tRNAs is discussed in subsection 6.1.1. The modelling of the *rho*-independent transcription terminators of bacteria is discussed in subsection 6.1.2. The modelling of the pseudoknots in the 3' untranslated regions (UTR) of viral genes is discussed in subsection 6.1.3.

6.1. Using the Eponine RNA-motif extension

6.1.1. Modelling RNA-motifs of mammalian tRNAs

The set of mammalian tRNAs was chosen as the starting case for assessing the capability of the Eponine RNA-motif extension, since the consensus clover-leaf secondary structure features of tRNAs have been studied for decades. tRNAs are also widely used as a data set for evaluating the performances of RNA secondary-structure prediction programs and ncRNA classifiers.

In this subsection, further assessment is made of the performances of the stringent and the fast modes of the Eponine RNA-motif extension (for definitions of the stringent mode and the fast mode, see Figure 4-3 and subsection 4.2.2.1.). It was shown that when identifying the canonical secondary structures of tRNAs, the stringent mode was better than the fast mode (see Table 4-1). An issue which was not investigated is the effect of using different structure-scanning modes on performance in the context of discriminating ncRNAs in genomes. If the models trained using the fast mode do not perform significantly worse than the models trained using the stringent mode, maybe the fast mode could be sufficient for the purpose of discriminating ncRNAs in genomes.

Consequently, there are two purposes of this subsection. Firstly, the performances of pure structural-motif models trained using the stringent mode and the fast mode, respectively, are compared. Secondly, I demonstrate that the Eponine RNA-motif extension can be used to train a discrimination model consisting of both primary-sequence patterns and RNA secondary-structure motifs.

6.1.1.1. Materials and methods

6.1.1.1.1. Recruiting the genomic sequences for training and testing

The sets of human tRNA genes created in section 5.1, chapter 5, were used for assessing the capabilities of the Eponine RNA-motif extension. The human tRNAs of group 1 were used for training models, and the tRNAs of group 2 were used for testing the performances of these trained models (Table 6-1, positive sequences). In order to realize the effect of using genomic sequences on modelling consensus RNA motifs, the flanking regions of human tRNA genes were included. The first base of the cloverleaf-like structure of each tRNA was used as the anchoring point; 100 bases upstream and 150 bases downstream with respect to the anchoring point in each human tRNA gene were retrieved. Two thousand random sequences and ten thousand random sequences were sampled from the human genome as negative training sequences and negative test sequences, respectively (Table 6-1, negative sequences). The human genome assembly used for random sampling was NCBI 35. These sequences were retrieved from the Ensembl ftp site (ftp://ftp.ensembl.org/pub/). These random sequences were 250 bases in length.

	Positive sequences	Negative sequences
Training data	200 genomic sequences of human tRNAs (group 1)	2000 random sequences from the human genome
Test data	167 genomic sequences of human tRNAs (group 2)	10,000 random sequences from the human genome

Table 6-1. The training and test data sets for modelling the human tRNAs

6.1.1.1.2. Determination of the performance of EAR models against the test data set

The training sequences described in the previous subsection were used to train the Eponine Anchored RNA-motif models (the EAR models, see subsection 4.2.2.3.1, chapter 4). When evaluating the performance of trained models, the 100th base of each test sequence was taken as the anchoring point. A true positive was determined if any region within 5 bases away

from the anchoring point of a positive sequence was predicted as a hit. A false positive was determined if any region within 5 bases away from the anchoring point of a negative sequence was predicted as a hit.

6.1.1.1.3. Setting the parameters of the Eponine RNA-motif extension

The size of windowed regions for predicting the local RNA structural motifs was set to 50 bases when running the Eponine RNA-motif extension. As a result, only the base pairs within each windowed region of 50 bases would be considered in the trained models. The windows were limited to 50 bases in this subsection for several reasons. Firstly, finding a consensus global RNA structure in a set of sequences is not the objective of designing the Eponine RNA-motif extension. It is instead designed to use consensus local RNA motifs for discriminating a particular type of ncRNAs in genomes. Secondly, one purpose of this subsection is to compare the performances of different RNA-motif scanning modes, *i.e.* the stringent mode and the fast mode (for the details of these two modes, see section 4.2, chapter 4). If evidence strongly suggests that long-range canonical base pairs are essential for discriminating a particular type of ncRNAs, the size of windowed regions can certainly be increased at the cost of computational time.

6.1.1.2. Results

6.1.1.2.1. Pure secondary-structure models of human tRNAs

By using the stringent mode, an EAR model consisting of eight hairpins was trained (Table 6-2 and Figure 6-1 A). While it might seem that too many hairpins were found, the eight hairpins can be grouped into five distinctly positioned hairpins, namely, hairpins that start at 10^{th}, 15^{th}, 27^{th}, 49^{th}, and 59^{th} positions respectively in tRNA molecules. Among these predicted consensus hairpins, hairpins that start at 10^{th}, 27^{th}, and 49^{th} positions clearly correspond to three well-known hairpins, D arm, anticodon arm, and T arm, respectively in tRNAs. The hairpin that starts at 59^{th} position can be viewed as a shifted T arm, because some

tRNA genes contain intronic sequences and the distance between the first base of cloverleaf-like structure and T arm is therefore longer than that in the tRNAs without introns.

Weight	Position	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
2.05	10	0.48	10	1.2	3	0.7
2.13	10	0.41	8	0.5	4	0.2
1.83	15	0.33	6	2.6	3	0.3
2.51	26	1.07	9	0.0	4	0.2
2.32	27	1.96	7	0.1	5	0.5
2.08	49	1.00	7	1.0	3	0.6
1.54	50	10.14	7	0.3	5	0.1
1.68	59	0.00	5	1.0	4	0.2

Table 6-2. The trained parameters of an anchored RNA structural model for mammalian tRNAs by using the stringent mode for locating local hairpins

The titles, "Weight", "Position", and "Width", are used as described in Figure 5-3. "Loop size" is the mean of the discrete Gaussian distribution used to model a loop region. "Stem size" is the mean of the discrete Gaussian distribution used to model a stem region.

A fast-mode EAR model consisting of ten hairpins was also trained (Table 6-3). Just as the hairpin groups in the stringent-mode EAR model, these ten hairpins can be categorized into four distinctly positioned hairpin groups, namely, hairpins that start at 3^{rd}, 10^{th}, 27^{th}, and 47^{th} positions respectively in tRNA molecules. The latter three correspond to three well-known hairpins, D arm, anticodon arm, and T arm respectively in tRNA molecules.

It seems that the model trained using the stringent mode for locating local hairpins is slightly simpler than the model trained by using the fast mode, although most likely this is caused by chance. In the current implementation of the Eponine RNA-motif extension, similar sub-models of individual hairpins are not merged and in different training runs the numbers of hairpins found may differ. In brief, the difference between the numbers of hairpins found by

two models does not suggest that one of the models may be better than the other one.

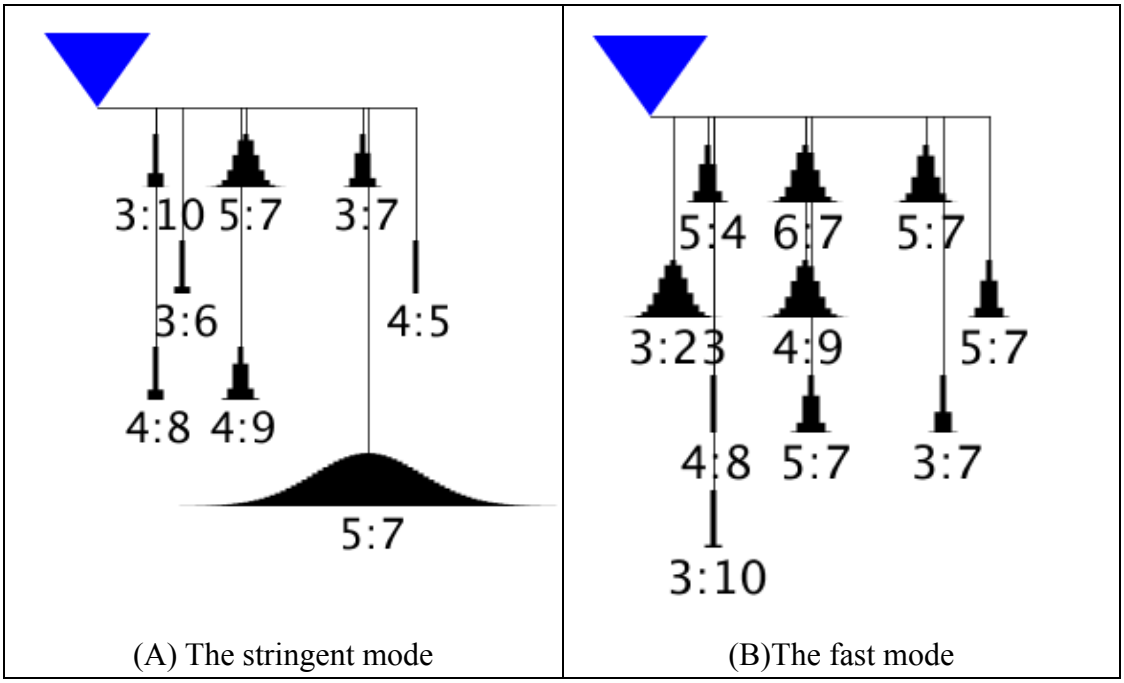
Weight	Position	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
1.97	3	2.27	23	0.7	3	0.1
2.69	9	1.05	4	0.2	5	0.3
2.78	10	0.08	8	0.7	4	0.1
2.34	10	0.23	10	0.8	3	0.5
1.51	26	1.83	7	0.1	6	1.1
1.35	26	2.06	9	0.0	4	0.1
1.82	27	1.00	7	0.7	5	0.1
2.89	47	1.52	7	0.0	5	0.2
1.73	50	0.59	7	2.9	3	0.0
1.38	58	1.00	7	1.2	5	0.0

Table 6-3. The trained parameters of an anchored RNA structural model for mammalian tRNAs by using the fast mode for locating local hairpins

The titles used in this table follow the convention of Figure 5-3 and Table 6-2.

Evaluating the performances of the fast mode and the stringent mode

By using the test data set recruited as described in 6.1.1.1.1. the performances of the models trained respectively using the fast mode and the stringent mode of the Eponine RNA-motif extension were evaluated. The results suggest that the performance of the fast mode can be as good as that of the stringent mode (Figure 6-4, fast mode and stringent mode). Although using the fast mode risks missing important hairpins, it can still be used for finding consensus RNA structural motifs in sequences when sufficient positive sequences are used for training. Since by using the fast mode the CPU time is about 40%-60% of the time taken by using the stringent mode, all models in the following were trained by using the fast mode, unless otherwise indicated.



(A) The stringent mode (B)The fast mode

Figure 6-1. Two Eponine anchored RNA structural models for mammalian tRNAs

The diagrams were prepared following the convention used in Figure 5-2, except that the motifs shown here are RNA structural motifs. The constraints drawn with two numbers under them correspond to RNA hairpins. These numbers are used to describe the dimension of a consensus hairpin. Each dimension consists of the stem size and the loop size that are separated by a colon. For example, in the right most hairpin in (A), 4:5 means that the size of this stem is 4 base pairs and the length of the loop is 5 bases.

6.1.1.2.2. A mixed primary-sequence and RNA secondary-structure model

Here, the capability of the Eponine RNA-motif extension to model both primary-sequence and RNA secondary-structure motifs was evaluated by using the human tRNAs recruited as described in 6.1.1.1.1. The results reveal that the EAR model is capable of finding both primary-sequence and RNA secondary-structure motifs of tRNAs (Figure 6-2). Such models that contain both primary-sequence and RNA structural motifs are referred to as mixed models in this thesis.

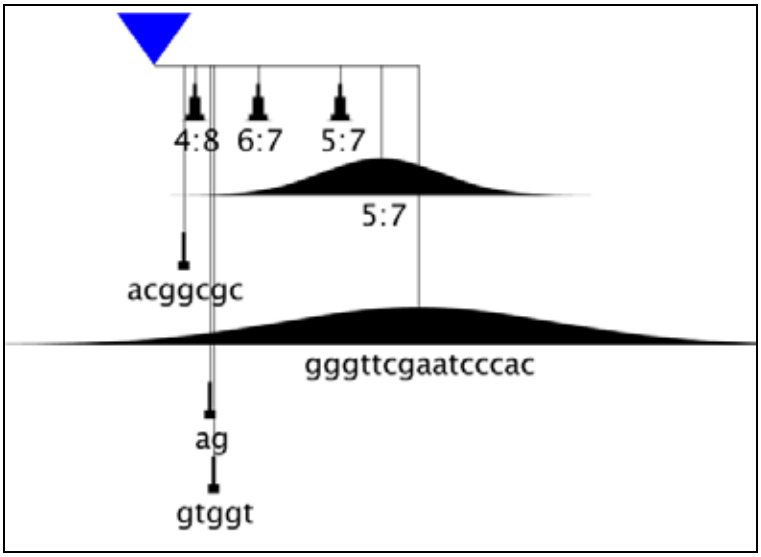


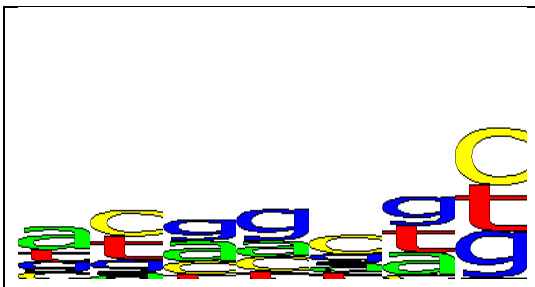
Figure 6-2. An Eponine anchored and mixed (primary-sequence and RNA structural) model

This figure is drawn following the convention used in Figure 6-1.

Weight	Position	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
5.06	8	0.45	Not available (a PWM of 7 columns)			
1.97	11	1.00	8	0.15	4	0.01
1.76	15	0.45	Not available (a PWM of 2 columns)			
4.15	16	0.45	Not available (a PWM of 5 columns)			
1.48	28	1.00	7	0.46	6	2.39
2.19	50	1.00	7	0.39	5	0.52
2.40	61	16.11	7	0.04	5	0.05
31.88	71	36.50	Not available (a PWM of 15 columns)			

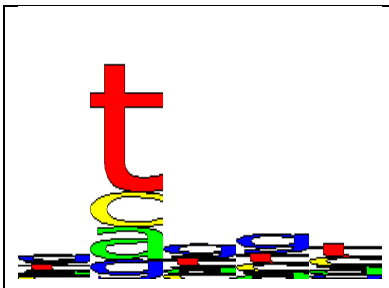
Table 6-4. The trained parameters of the EAS mixed model presented in Figure 6-2

The titles used in this table follow the convention of Figure 5-3 and Table 6-2

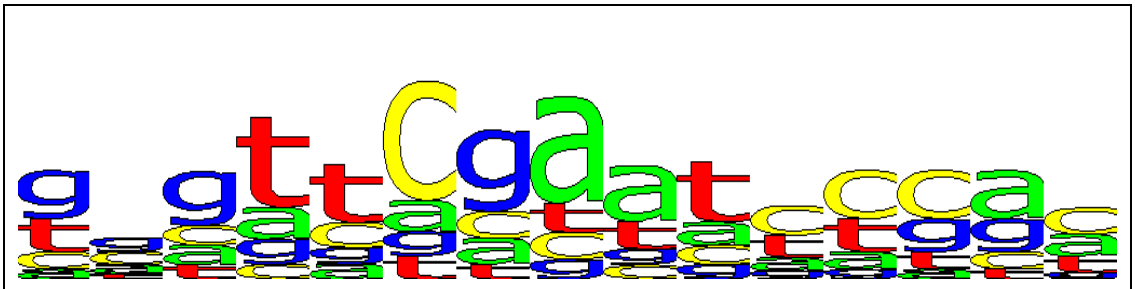


Position: 8

Position: 15



Position: 16



Position: 71

Figure 6-3. The sequence logos of position-constrained motif matrices in the Eponine EAS mixed model presented in Figure 6-2 and Table 6-4.

"Position" corresponds to the "Position" column in Table 6-4.

Evaluating the performances of the mixed model of human tRNAs

The capability of the trained mixed model to differentiate human tRNAs from random genomic sequences was also evaluated using the test data set recruited as described in 6.1.1.1.1. The results reveal that a mixed model ("mixed model, fast mode", Figure 6-4) can perform better than models consisting of only RNA structural motifs ("structure-only" models,

Figure 6-4). For discriminating tRNAs in the human genome, the false positive rate of the mixed model should be much lower than that of the models consisting of only RNA secondary-structure motifs (comparing the "structural-only" models with the mixed model, Figure 6-4).

For comparison, a pure primary-sequence model, which did not consist of RNA motifs, was trained taking the training data set as described in 6.1.1.1.1. The performance of this pure primary-sequence model was also evaluated using the test data set recruited as described in 6.1.1.1.1. However, in this evaluation, the accuracy of the mixed model for human tRNAs ("mixed model, fast mode", Figure 6-4) was not as good as this pure primary-sequence model ("pure primary-sequence model", Figure 6-4) when the coverage (sensitivity) was set to be higher than 90%. There were 10 false positives predicted by the mixed model, while only 2 false positives were found by using the pure primary-sequence model.

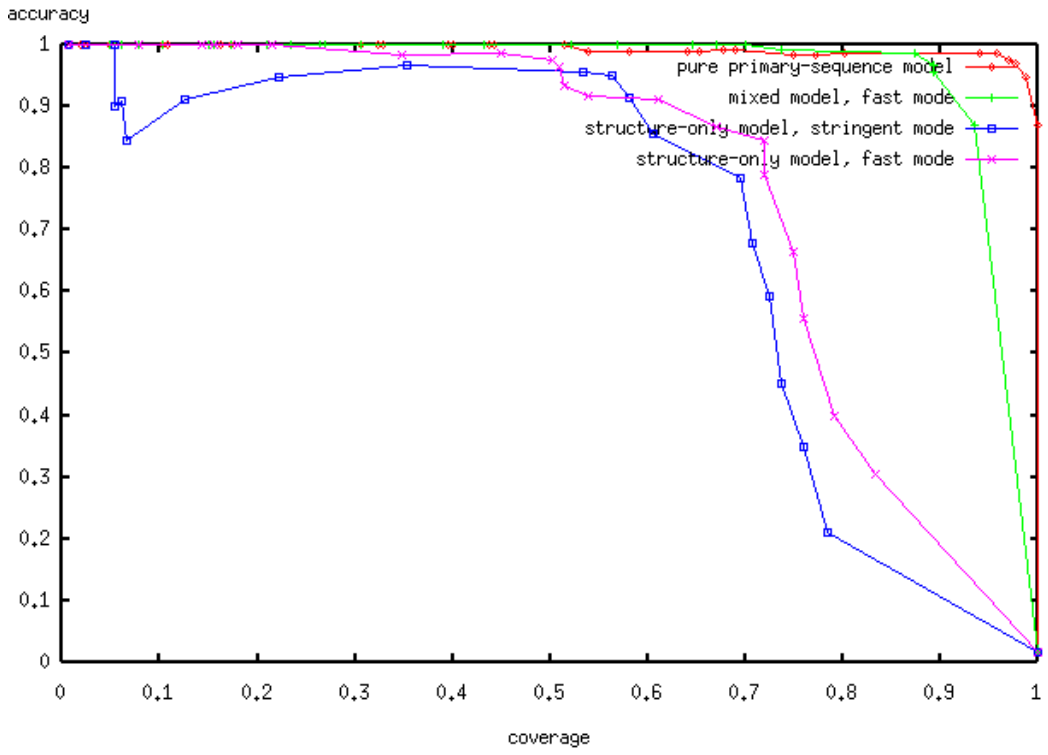


Figure 6-4. Comparison of performances among models trained by different modes for classifying human tRNA genes from random genomic sequences

6.1.1.3. The false positives predicted by using the mixed model

To explore why a mixed model discovered more false positives, the features of the 10 high-scoring false positives were examined in detail. The conservation of the internal promoter in each sequence, and the conservation of local RNA motifs corresponding to the D arm, anticodon arm, and T arm in the canonical tRNA clover-leaf like structures were evaluated.

The results reveal that most of the false positives predicted by the mixed model of human tRNAs contain only a subset of the motifs in the canonical tRNA structures (Table 6-5). In summary these false positives can be characterised as:

- A sequence with a strong internal promoter (as determined by eufindtRNA) can be identified as a tRNA.
- A sequence with a partial set of weak motifs, either in a combination of a weak internal promoter and a local RNA structural motif, or in a combination of two or more local RNA structural motifs, can be identified as a tRNA.
- Most of the false positives overlap with repetitive elements.

Serial ID	Internal promoters[1]	D arm	anticodon arm	T arm	Repeat
1	+	-	-	-	SINE/MIR
2	-	-	-	+[2]	LINE/L1
3	+	-	+	-	LINE/L1
4	-	-	+ (ss) (offset)	+	SINE/MIR
5	+	-	-	-	LTR/MaLR SINE/Alu
6	+	-	+ (ss)	+ (offset)	SINE/Alu
7	-	-	+ (ss) (offset)	+ (offset)	LINE/L1
8	-	-	+ (ss)	+ (ls)	LTR/MaLR SINE/Alu
9	+	-	-	+ (offset)	LINE/L1
10	+	-	-	+	(not available)

Table 6-5. The high-scoring false positives predicted by using the mixed model of human tRNAs

[1]: the internal promoters were determined by using eufindtRNA with a relaxed parameter set

[2]: there is an additional hairpin at the 3' side of the T arm. This additional hairpin also contributes to the final score.

(ss): a stem which is smaller than the corresponding canonical local RNA motif.

(ls): a stem which is longer than the corresponding canonical local RNA motif.

(offset): a hairpin is a few bases away from the best positions in the canonical tRNA structure.

(not available): not overlapping with repetitive elements

Due to the scoring scheme used in Eponine, these findings are not really surprising. Given a GLM-based RNA-motif model such as the mixed model of human tRNAs, the final score of a genomic locus is actually a transformed weighted sum of PWM scores and RM scores. Thus, a mixed model consisting of many local motifs may be apt to identify truncated ncRNAs and other ncRNA-derived sequences. In fact, such behaviour is not unique to the Eponine RNA-motif extension. A similar observation has been made in the development of tRNAscanSE (Lowe and Eddy 1997), where the tRNA covariance model was shown to discover some truncated tRNAs and tRNA-derived SINEs which could not be identified by using promoter-based methods (such as eufindtRNA), and hierarchical and rule-based systems (*e.g.* tRNAscanSE) for genome-wide tRNA finding.

6.1.2. Modelling *rho*-independent transcription termination

The modelling of human tRNA genes partially demonstrates the capability of the Eponine RNA-motif extension. Since many existing ncRNA-finding algorithms have also been shown to be capable of detecting the cloverleaf-like structures, the result of the modelling of human tRNAs only reveals that the Eponine RNA-motif extension has a function similar to other tools. Consequently, in this subsection, a more difficult case (for reasons see the discussion in the next two paragraphs), the *rho*-independent transcription terminators, was used to evaluate the capability of the Eponine RNA-motif extension.

The *rho*-independent transcription terminator, which consists of both primary-sequence and RNA structural motifs, is an important functional element for regulating the transcription termination of bacterial genes (Uptain and Chamberlin 1997). Unlike modelling tRNA genes, finding *rho*-independent transcription terminators is a topic that has received less investigation. Apparently, only *ad hoc* algorithms can find *rho*-independent transcription terminators in the bacterial genomes (d'Aubenton Carafa et al. 1990; Ermolaeva et al. 2000; Lesnik et al. 2001; de Hoon et al. 2005). Up to this point, no general-purpose RNA-motif finding algorithms have been used to find the consensus RNA motifs in these regions of transcription termination.

One reason that makes *rho*-independent termination signals an unpopular data set is that the boundaries of *rho*-independent termination signals are not so well defined as known ncRNA genes (such as tRNA genes). It is difficult to adequately align these regions. The identities of pairwise alignments of the regions around transcription termination sites are generally low. Fewer than 0.5% of pairwise alignments have identities greater than 60% (data not shown), if the alignments are generated by randomly choosing raw sequences that have been used by de Hoon *et al*. (de Hoon et al. 2005). Whether these low-identity alignments can reveal the structural relations among sequences cannot be confidently determined. However,

as has been discussed previously (see section 2.1, chapter 2, and section 4.2, chapter 4), most existing algorithms would not be expected to have good performance in finding structural signals in such data set.

Some *ad hoc* algorithms were claimed to have high specificity and high sensitivity in detecting *rho*-independent transcription terminators. However, there must be some doubt about the generality of such results given the training and optimisation processes used. Firstly, some models were actually tested with exactly the same sequences that have been used for training respective models (d'Aubenton Carafa et al. 1990; Lesnik et al. 2001; de Hoon et al. 2005). These models may be over fitted and unable to generalise to new data, something that has not been tested for because of the use of a non-independent test data set. Secondly, some algorithms discard all predictions in intragenic regions (Ermolaeva et al. 2000), even though the scores of these predictions exceed the computationally defined threshold. The eradication of this major source of false positives makes it impossible to properly estimate the accuracy and specificity of the predictions made by these algorithms.

6.1.2.1. Materials and methods

6.1.2.1.1. The data sets for training and testing the Eponine anchored RNA-motif model

In order to train and test the EAR models for *rho*-independent transcription terminators, 423 transcription terminators that have been used by de Hoon *et al*. (de Hoon et al. 2005) were divided into two data sets for training and testing respectively. Each sequence consists of 20 bases upstream and 50 bases downstream of the respective transcription termination site annotated by Hoon *et al*. (de Hoon et al. 2005).

Two sets of 2,000 negative sequences for training and testing models, respectively, were randomly taken from the *B. subtilis* genome (GenBank accession number: AL009126). These negative sequences were 70 bases in length.

6.1.2.1.2. Determination of the performance of EAR models against the test data set

When evaluating the performance of EAR models for *rho*-independent transcription terminators against the test data set, the 20th base of each sequence was taken as the anchoring point. A true positive was determined if any region within 5 bases away from the anchoring point of a positive sequence was predicted as a hit. A false positive was determined if any region within 5 bases away from the anchoring point of a negative sequence was predicted as a hit.

6.1.2.1.3. Scanning for rho-independent transcription terminators in genomes

When an EAR model for *rho*-independent transcription terminators was used to scan genomes, both strands of genomes were scanned. Each position in a genome can be the first base of a *rho*-independent transcription terminator. Consecutive hits would be clustered together if all of their scores were higher than a particular threshold and considered as a single prediction.

Determination of putative terminators of genes

For each gene, if a predicted *rho*-independent TTS on the same strand is within the range starting from 50 bases upstream of the stop codon, continuing till the 500 bases downstream of the stop codon, this TTS is considered as a putative terminator, unless if this TTS is within the coding region of the next gene. If there were more than one candidate hit for a particular gene, the one that was closer to the stop codon was used.

Determination of intragenic terminators

If an intragenic predicted hit is more than 50 bases from the stop codon of a gene, it is regarded as a true intragenic hit.

6.1.2.1.4. The data set for training and testing the Eponine Windowed RNA-motif model

To assess the capability of the Eponine Windowed RNA-motif model (the EWR model,

see subsection 4.2.2.3.2, chapter 4) to find consensus RNA motifs in a set of sequences where no reference points are known, a set of 423 *B. subtilis* genomic sequences that contain *rho*-independent transcription terminators was prepared. In order to make the assessment more challenging, the positions of *rho*-independent transcription terminators in respective sequences were randomly distributed between 1 and 100 (Figure 6-5). These sequences were randomly divided into a training set (212 sequences) and a test set (211 sequences). The negative sequences recruited for training and testing models were the same as described in subsection 6.1.2.1.1.

When evaluating the performance of EWR models for *rho*-independent transcription terminators, a true positive was determined if any position in a positive sequence was predicted as a hit. A false positive was determined if any position in a negative test sequence was predicted as a hit.

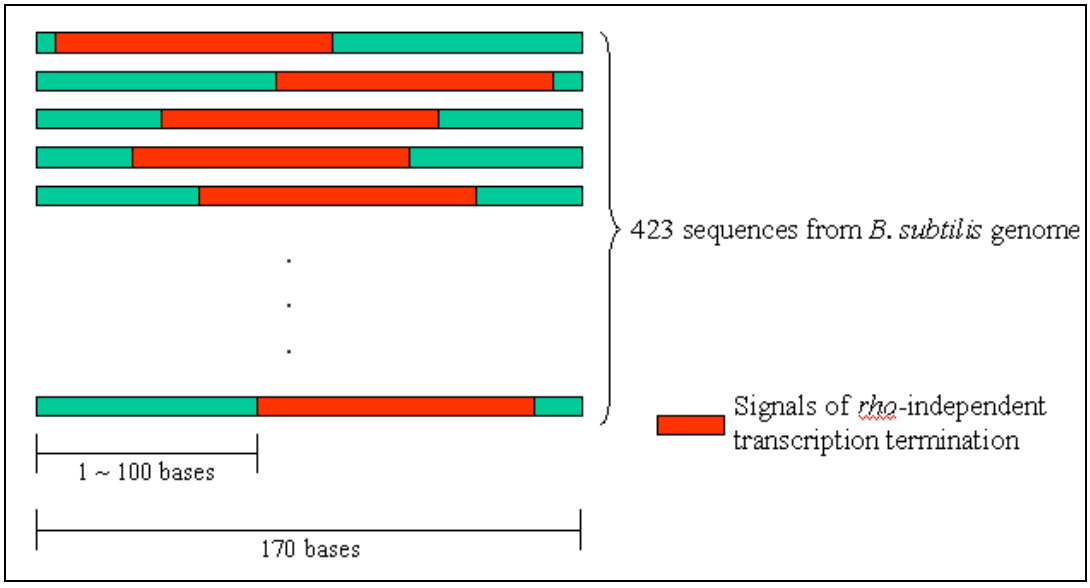


Figure 6-5. Preparation of a set of unanchored sequences that contain *rho*-independent transcription terminators at random positions

6.1.2.2. Results

6.1.2.2.1. The Eponine anchored RNA-motif model (EAR model)

The EAR mixed model for the *rho*-independent transcription terminators of *B. subtilis* consisted of five motifs (see Table 6-6 and Figure 6-6). This model is basically consistent with the current knowledge of the composition of the *rho*-independent terminators (For details see Lesnik et al. 2001), where the first two motifs (weights 0.85 and 5.30, Table 6-6) correspond to an A-region (adenosine-rich region); and a stable hairpin (weight 6.03, Table 6-6) is followed by a T-region (weight 13.62, Table 6-6) (thymidine-rich region in genome, corresponding to uridine-rich region in transcripts). An additional motif is at positive 5 (weight 4.17, Table 6-6). However, its importance is not clearly understood. Since it overlaps with the hairpin motif it may be capturing sequences preference within the hairpin of *rho*-independent transcription terminators. The Eponine sub-model for the hairpin of *rho*-independent transcription terminators is at position 5 (weight 6.03, Table 6-6); the stem size is 9 base pairs in length and the loop size is 12 bases in length. The standard deviation for the distribution of loop size is 16.5 bases, which is obviously larger than the mean loop size (12, Table 6-6). The heavy tail in the distribution of the loop size is consistent with the previous models of the *rho*-independent terminators of either *E. coli* or *B. subtilis* (d'Aubenton Carafa et al. 1990; Ermolaeva et al. 2000; Lesnik et al. 2001; de Hoon et al. 2005).

Weight	Position	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
0.85	-3	0.60	Not available (a PWM of 3 columns)			
5.30	1	0.63	Not available (a PWM of 5 columns)			
6.03	5	4.46	12	16.5	9	2.13
4.17	5	1.38	Not available (a PWM of 4 columns)			
13.62	29	17.96	Not available (a PWM of 7 columns)			

Table 6-6. The trained parameters of an EAR model for *bacillus rho*-independent transcription terminators

The titles used in this table follow the convention of Table 6-4.

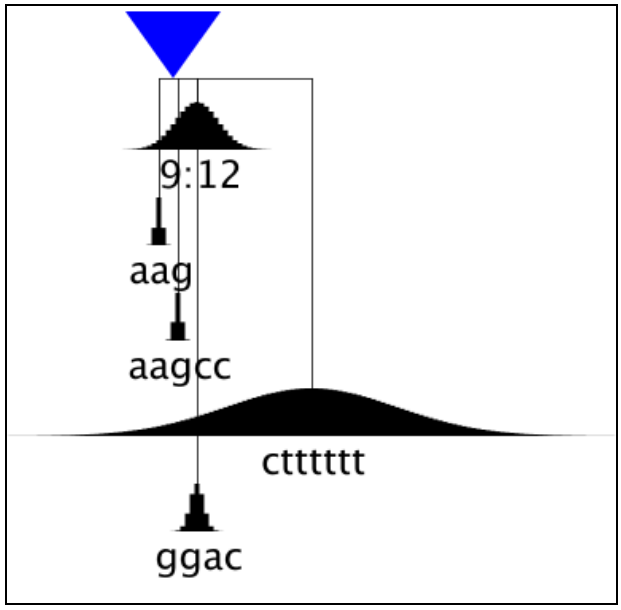
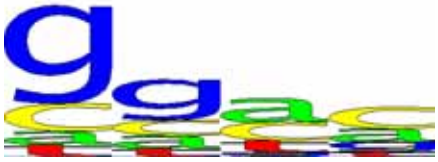


Figure 6-6. An EAR model for *rho*-independent transcription terminators

This figure is drawn following the convention used in Figure 6-1.

Position: -3

Position: 1



Position: 5

Position: 29

Figure 6-7. The sequence logos of the position-constrained motif matrices presented in Figure 6-6 and Table 6-6

"Position" corresponds to "Position" column in Table 6-6.

For comparison, a pure primary-sequence model, which did not consist of RNA motifs, was trained taking the training data set as described in 6.1.2.1.1. A structure-only model, which did not consist of primary-sequence motifs, was also trained using the same data set. C-A plots of different models for the *rho*-independent transcription terminators were calculated using the test data set of 211 positive sequences and 2000 negative sequences. The result reveals that the performance of the mixed model (see Table 6-6 and Figure 6-6) is better than that of the pure primary-sequence and structure-only models (Figure 6-8).

Discriminating the *rho*-independent transcription terminators in real bacterial genomes

In order to further assess the performances of the EAR mixed model and other algorithms, the sensitivities and specificities were estimated by using the result of scanning the full-length

genomic sequences of *B. subtilis* and *E. coli* K-12 (GenBank accession number: U00096) (Table 6-7). The predictions that overlap with experimentally verified *rho*-independent transcription terminators were counted as true positives. In order to avoid bias in the evaluation, only known terminators that were not used for training the respective algorithms/models were used to estimate sensitivities. Predictions in intragenic regions were taken as false positives for estimating false positive rates. Although some of the *rho*-independent transcription terminators may possibly reside in intragenic regions, the location distribution of true terminators should be greatly biased towards intergenic regions. While it is likely that some of the predictions that fall in intergenic regions are false positives, the ratio of intragenic predictions over all predictions provide at least an estimate of the false positive rate.

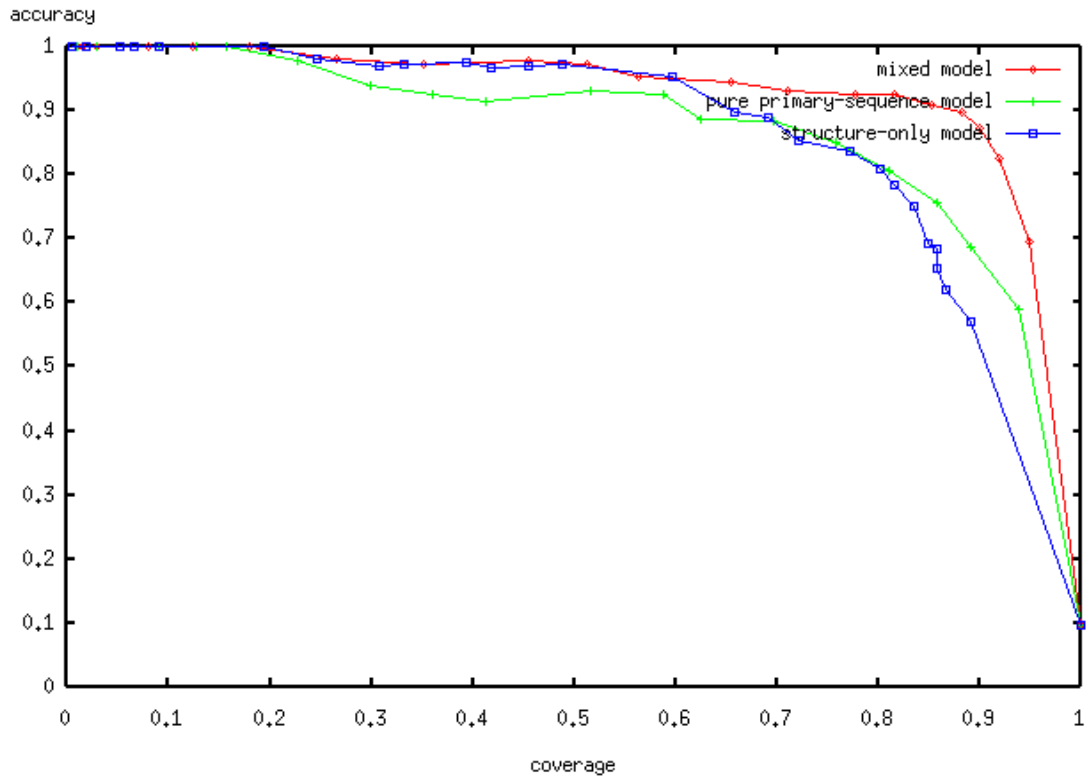


Figure 6-8. Comparison between the C-A plots of the mixed, the structure-only, and the primary-sequence-only models of *rho*-independent transcription terminators

(A) Performance for finding *rho*-independent transcription terminators in *B. subtilis*

Reference	The name of the software	Independent test data	Sensitivity	False positive rate	Intragenic hits
(Ermolaeva et al. 2000)	TransTerm	Yes[1]	86.2% (399/463)	NA[2]	NA[2]
(Lesnik et al. 2001)	RNAMotif	No	NA	NA	NA
(de Hoon et al. 2005)	NA	No	NA[3]	NA[3]	NA
This thesis, 2006	EAR mixed model	Yes	85.3% (180/211)	14% (766/5477)	766

(B) Performance for finding *rho*-independent transcription terminators in *E. coli*

Reference	Sensitivity	False positive rate	Intragenic hits
(Ermolaeva et al. 2000)	89%-98%	NA[2]	NA[2]
(Lesnik et al. 2001)	80%-100%	39% (2586/6635)	2586
(de Hoon et al. 2005)	67%	NA	NA
This thesis, 2006	81% (119/147)	16.6% (431/2604)	431

Table 6-7. Comparison of the performance of different algorithms in finding *rho*-independent transcription terminators in *B. subtilis*

(A) The performances of different algorithms for finding *rho*-independent transcription terminators in *B. subtilis*. (B) The performances of different algorithms for finding *rho*-independent transcription terminators in *E. coli*. Numbers in parentheses are the values that are used to estimate the sensitivities and the false positive rates for different algorithms. The sensitivities are the ratios of experimentally verified terminators that can be successfully predicted by different algorithms. The numbers of predictions that are in intragenic regions are taken as the numbers of false positives. The false positive rates are estimated by dividing the numbers of false positives with the numbers of all predictions. The statistics for TransTerm is estimated by using the results retrieved from http://www.cbcb.umd.edu/software/TransTerm/. The statistics for RNAMotif is retrieved directly from its original paper (Lesnik et al. 2001). The statistics for de Hoon et al.'s algorithm is taken directly from its original paper (de Hoon et al. 2005).

[1]: no negative sequences are used for estimating accuracy and specificity; only sensitivity is estimated by using positive sequences that are not used for training.

[2]: not available because intragenic hits are considered as background and invalidated in final output. For realizing the meaning of this table, see text for details.

[3]: not available because de Hoon *et al.*'s algorithm was trained by using *rho*-independent transcription terminators of *B. subtilis* as the positive training sequences.

NA: not available from respective papers and cannot be estimated by using results retrieved from related websites.

The results reveal that the EAR mixed model is competitive for predicting *rho*-independent transcription terminators in the bacterial genomes. Although the parameters of the EAR mixed model were trained using sequences from *B. subtilis*, this model can find *rho*-independent transcription terminators in *E. coli* with a reasonable sensitivity (81%, this thesis, Table 6-7 B) and a similar estimated false positive rate (16.6%).

In order to compare the EAR mixed model with other algorithms, each case is discussed separately because there are specific considerations associated with each algorithm. Firstly, the sensitivity, 81% (this thesis, Table 6-7 B), is obviously higher than the sensitivity (67%, de Hoon *et al*., Table 6-7 B) for finding *rho*-independent transcription terminators of *E. coli* by using de Hoon *et al*.'s algorithm. The latter was also trained by using sequences from *B. subtilis*. Although de Hoon *et al*.'s algorithm was claimed to have a specificity of 94% for finding *rho*-independent transcription terminators of *B. subtilis*, the high specificity was actually estimated by using only 567 non-terminating sequences (de Hoon et al. 2005), but not random intragenic regions in *B. subtilis*. In addition, the 567 negative sequences, which have been used for training the algorithm, are re-used for testing (de Hoon et al. 2005). The real specificity and false positive rates of de Hoon *et al*.'s algorithm should therefore be regarded as unknown.

Secondly, although the sensitivity (81%, this thesis, Table 6-7 B) of the EAR mixed model for predicting *rho*-independent transcription terminators of *E. coli* seems to be not as good as the sensitivity (80% ~ 100%, Table 6-7 B) of RNAMotif, the false positive rate of the EAR mixed model is estimated as only 14.7%, which is much lower than that (39%) of RNAMotif, calculated in a similar way. It should also be noted that the sensitivity of RNAMotif was estimated with exactly the same positive sequences that had been used for training. No predictions made for other bacterial genomes using RNAMotif can be found in original papers or on related websites.

Thirdly, the sensitivity (85.3%, this thesis, Table 6-7, A) of the EAR mixed model for finding terminators of *B. subtilis* was comparable to that (86.2%, Table 6-7, A) of TransTerm, even though it is impossible to estimate the false positive rates of TransTerm due to its peculiar way of estimating the confidence of predictions (Ermolaeva et al. 2000) (For details see discussions in the 5th paragraph in the introduction of this subsection, 6.1.2.).

Consequently, among the algorithms mentioned above, the EAR mixed model is the only *rho*-independent transcription terminator finding approach for which reasonably robust indicators of both sensitivity and specificity are available.

6.1.2.2.2. The Eponine windowed RNA-motif model (EWR model)

rho-independent transcription terminators should still be considered an easy case when evaluating ncRNA-finding algorithms, since there is a clearly definable reference point, namely the transcription termination site, in each sequence. When no obvious reference points are known, finding consensus RNA motifs is difficult for most available computational approaches. The Eponine windowed RNA motif model (EWR model) is specifically designed for such situations.

The results presented here (Figure 6-9) reveal that the EWR models are capable of finding key signals, corresponding to A-region (the motifs at offset 0 in sensors 1 and 2, Table 6-8), the stable hairpin (the motif at offset 26 in sensor 1, and the motif at offset 16 in sensor 2, Table 6-8), and T-region (the motif at offset 58 in sensor 1, and the motifs at offsets 42 and 79 in sensor 2, Table 6-8), for *rho*-independent transcription terminators in unanchored sequences (see subsection 6.1.2.1.4.). Although the performance of this EWR model (Figure 6-11) is not really comparable to the EAR mixed model, nearly 70% accuracy could be achieved when the coverage is 70%.

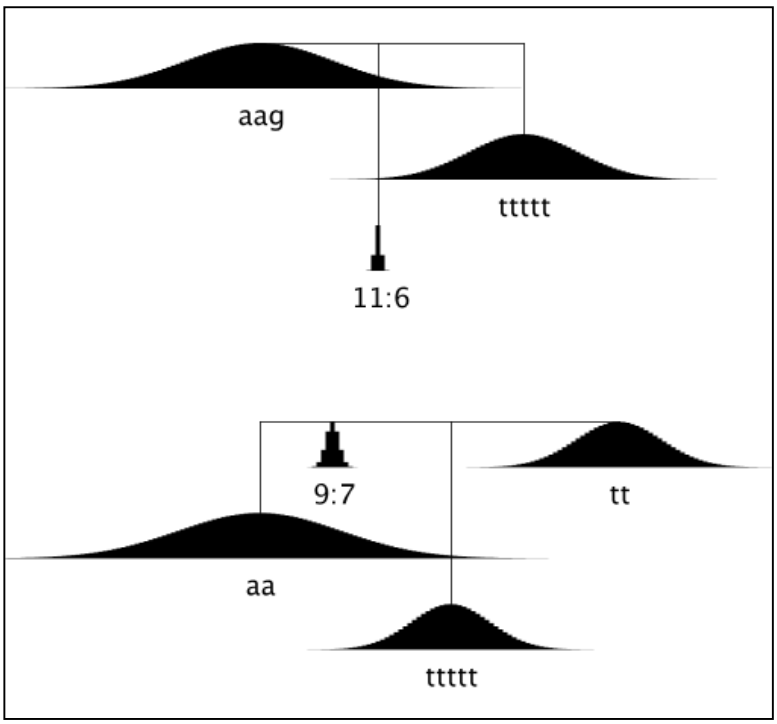


Figure 6-9. An EWR model for *rho*-independent transcriptional terminators

There are two convolved sensor basis functions (CSBFs, see subsection 4.1.2.1.2.) in the GLM of the EWR model for *rho*-independent transcription terminators. The upper one is referred to as sensor 1 and the lower one is referred to as sensor 2 in the following text.

Sensor 1:

Offset	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
0	16.25	Not available (a PWM of 3 columns)			
26	0.58	6	7.02	11	0.08
58	11.99	Not available (PWM, 5 columns)			

Sensor 2:

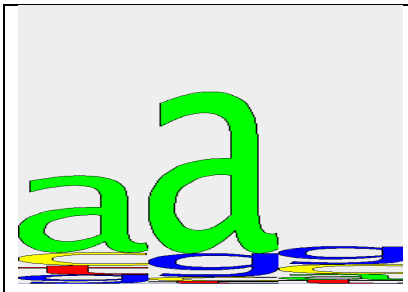
Offset	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
0	17.92	Not available (a PWM of 2 columns)			
16	1.39	7	8.69	9	0.13
42	8.62	Not available (a PWM of 5 columns)			
79	9.36	Not available (a PWM of 2 columns)			

Table 6-8. The trained parameters of an EWR model for *bacillus rho*-independent transcription terminators

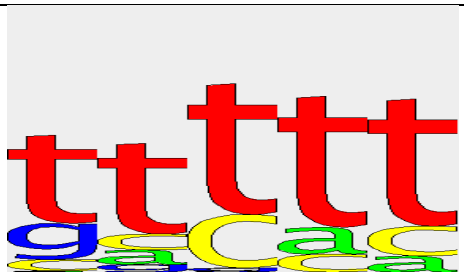
Sensor 1 is the convolved sensor basis function (CSBF) presented in the upper half of Figure 6-9 and sensor 2 is the CSBF presented in the lower half of Figure 6-9

"Offset" refers to the mean of the discrete Gaussian distribution used to model the distance between each motif and the first motif. Other titles follow the convention of Table 6-4.

Sensor 1:

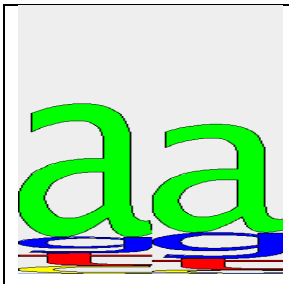


Offset: 0

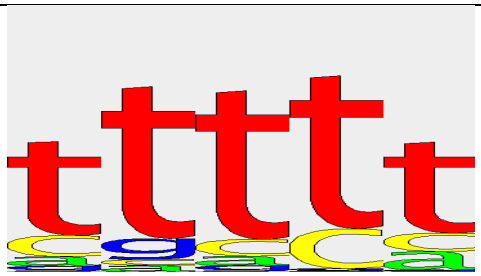


Offset: 58

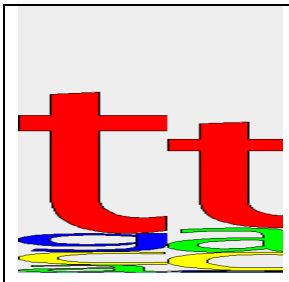
Sensor 2:



Offset: 0



Offset: 42



Offset: 79

Figure 6-10. The sequence logos of position-constrained motif matrices presented in Table 6-8 and Figure 6-9

"Offset" corresponds to "Offset" column in Table 6-8. Sensors 1 and 2 correspond to the sensors in Table 6-8 and Figure 6-9

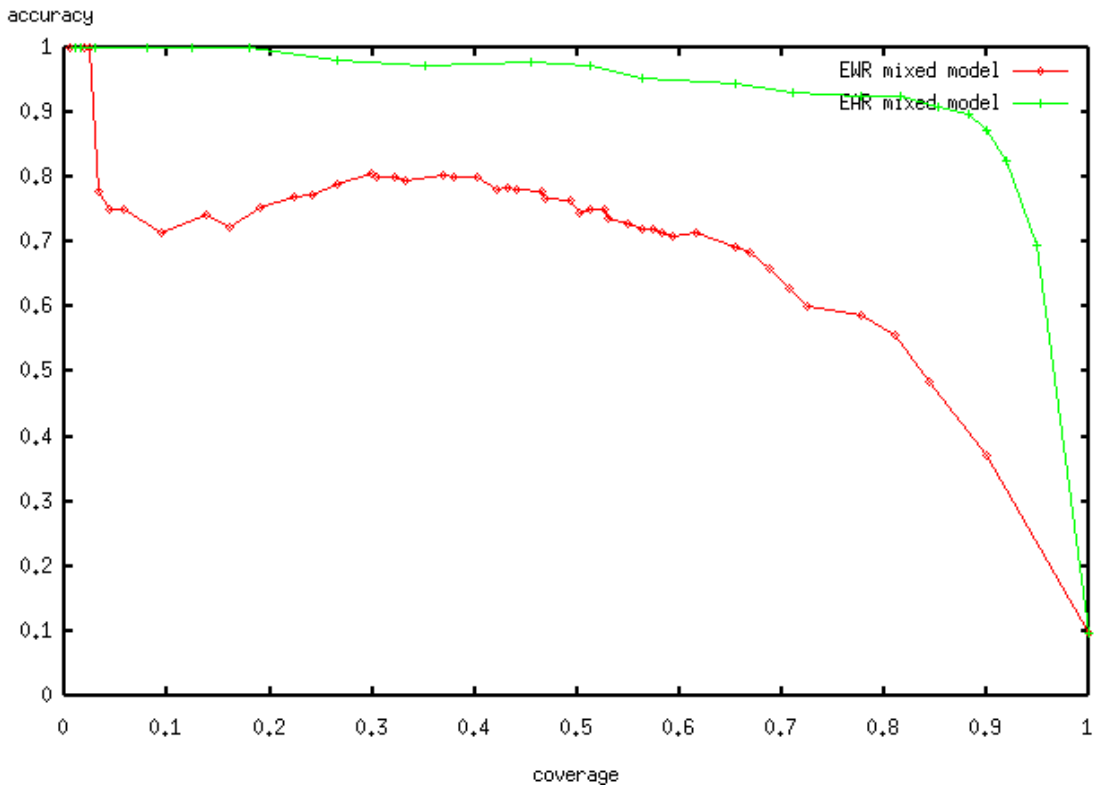


Figure 6-11. Comparison of the C-A plots of an EAR mixed model and an EWR model for *rho*-independent transcription terminators

6.1.2.3. Discussion

One obvious question about using the Eponine RNA extension to model *rho*-independent transcription terminators is the wide distribution of motif positions. For example, in the EAR mixed model (see subsection 6.1.2.2.1.), the width of the position distribution of the T-region is 17.96 (weight 13.62, Table 6-6). In the EWR model (see subsection 6.1.2.2.2.), there are also heavy tails for position distributions of both the A-region and the T-region (Figure 6-9). It seems that both of the EAR and the EWR models for *rho*-independent transcription terminators are inconsistent with the current view that the stable hairpin is immediately followed by the T-region. However, it should be noted that in the Eponine RNA-motif

extension, the first base of the respective hairpin is used as the position of each RNA structural motif. Consequently, in the EAR mixed model, the distances between the reference point (presumably the first base of the transcription termination signal) and the T-region in different sequences varies in response to the variations in the dimensions (loop size and stem size) of the stable hairpin in *rho*-independent transcription terminators. For similar reasons, it is not surprising that the wide position distributions of the T-region were also found in the EWR model of *rho*-independent transcription terminators. Consequently, the current implementation of the Eponine RNA-motif extension may not model ideally the proximity of motifs to their 5' adjacent structural motifs.

The inadequacy in modelling the exact relations between motifs and reference points separated by variable length structural motifs is a current weakness of the Eponine RNA-motif extension. For the purpose of modelling the relation between the hairpin and the T-region in the *rho*-independent transcription terminators, using the last base of the stem region as the location (reference point) for each structural motif might be helpful. However, switching the reference point for structural motifs is not expected to be a solution in all the situations, especially when the ncRNAs of unknown types are modelled as the most suitable reference points for a hairpin may vary from case to case. For example, in modelling the RNA motifs where the loop regions are responsible for the specific interaction with proteins, the most suitable anchoring point for hairpins could be the centre of the loop regions.

6.1.3. Modelling pseudoknots

Pseudoknots are seldom used for testing algorithms for finding consensus RNA motifs. Algorithms that were claimed to be capable of finding consensus pseudoknots in a set of sequences include GPRM (Hu 2002), ILM (Ruan et al. 2004), and comRNA (Ji et al. 2004). There are certain restrictions in using these algorithms. For example, GPRM and comRNA

cannot find primary-sequence motifs; users of GPRM must assign the expected number of hairpins in sequences; ILM requires pre-aligned sequences.

Although the Eponine RNA-motif extension is not specifically designed for finding consensus pseudoknots in sequences, it is not prohibited from finding consensus hairpins that overlap with each other, such as non-juxtaposed and non-nested stem regions in pseudoknots. In other words, the Eponine RNA-motif extension has the potential to find consensus pseudoknots in a set of sequences. The additional advantage of using a classification machine, such as the Eponine RNA-motif extension, is that the trained model may be applicable to finding new functionally related pseudoknots in genomes.

6.1.3.1. Materials and methods

To assess the capability of the Eponine RNA-motif extension for finding consensus pseudoknots, 18 sequences of 3' UTRs of genes of soil-borne rye mosaic viruses and soil-borne wheat mosaic viruses, which were also used by Hu (Hu 2002) for assessing GPRM, were recruited from the PseudoBase database (van Batenburg et al. 2001) as positive training sequences. Five hundred sequences of 40 bases in length were randomly sampled from the human genome and used as negative training sequences. The human genome assembly used for random sampling was NCBI 35. These sequences were retrieved from the Ensembl ftp site (ftp://ftp.ensembl.org/pub/).

These training sequences were used to train an EAR model as well as an EWR model. When the EAR model was used to model these pseudoknots, the first base of each sequence was used as the anchoring point.

6.1.3.2. Results

The resulting EWR model for the 3' UTRs of viral genes consisted of two consensus hairpins (Figure 6-12). The stem regions of these two hairpins were neither juxtaposed nor

nested. The distribution of the first base of the second hairpin peaks (offset: 5, hairpin ID 2, Table 6-9) at the end of the 5' stem of the first hairpin (stem size: 7, hairpin ID 1, Table 6-9). The most probable positions of the two hairpins were consistent with the configuration of the pseudoknots in these 3' UTRs of viral genes that were used for training. The result shows that the EWR models are capable of finding consensus pseudoknots in a set of sequences.

An EAR model for the pseudoknots in 3' UTR of viral genes was also trained. This EAR model also consisted of two hairpins (data not shown), which is consistent with the non-nested configuration of pseudoknots as shown in the EWR model.

Hairpin ID	Offset	Width of position distribution	Loop size	Width of loop size distribution	Stem size	Width of stem size distribution
1	0	2.7	4	8.8	7	0.8
2	5	2.7	9	4.1	4	0.2

Table 6-9. The trained parameters of an EWR model for pseudoknots in 3' UTRs of viral genes

The titles used in this table follow the convention of Table 6-8.

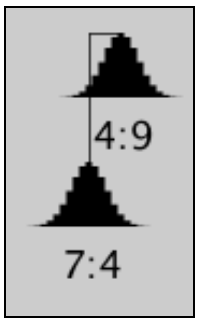


Figure 6-12. An EWR model for the 3' UTRs of viral genes

The notation used to describe RNA hairpins follows the convention of Figure 6-1.

6.2. Discussions

6.2.1. Considerations of using the Eponine RNA-motif extension

In order to train an Eponine RNA-motif model, a number of positive training sequences are required. For example, a set of ten sequences is insufficient for finding the pseudoknots in the 3' UTRs of viral genes with the current implementation and parameter settings of the Eponine RNA-motif extension. Training an Eponine RNA-motif model may require tens of positive sequences. In terms of finding functional RNA motifs, this requirement seems to be a weakness of the Eponine RNA-motif extension, compared to algorithms that can predict optimal RNA structures using only few sequences. Nonetheless, by using only a few sequences or even one sequence, available RNA-motif finding algorithms may also have difficulty in finding consensus structures in a set of unaligned sequences (Gardner and Giegerich 2004). Even though the algorithms that take pre-aligned sequences seem to have a good performance, none of them have been tested on alignments of real genomic sequences. Existing tests have generally been performed on alignments of well-trimmed sequences (Hofacker et al. 2002; Knudsen and Hein 2003; Coventry et al. 2004; Gardner and Giegerich 2004; Ruan et al. 2004). A similar situation is also true for the ncRNA classifying algorithms that utilise pre-aligned sequences (see also subsection 2.1.3.5. , chapter 2).

Another issue around using the Eponine RNA-motif extension is the computer time required for training a model. For example, it may take ~7 hours (24,108 seconds) and ~22 hours (79,661 seconds) to train an EAR mixed model and an EWR mixed model respectively for human tRNAs (Table 6-10). Within the trainer, predicting all local hairpins in each training sequence is not the most time-consuming step when using the Eponine RNA-motif extension. With the current implementation of the fast model of the Eponine RNA-motif extension, it takes less than 3 seconds by using an x86-64bit machine (3.2 Ghz Pentium IV EMT64, 64-bit

Linux) to predict local hairpins for a sequence of 250 bases in length. A significant proportion of time is actually spent using the Monte Carlo method to optimise parameters of PWMs and RNA motifs. For example, it is estimated that three-fourths of the CPU time used for training an EAR mixed model of tRNAs is spent in learning parameters of motifs, while only one-fourth of the CPU time (~6000/24108) is spent in predicting local RNA secondary structures (Table 6-10, tRNAs, EAR mixed model, CPU time).

	Training type	Sequence length	Number of positive sequences	Number of negative sequences	CPU time (x86-64bit) (seconds)
tRNAs	EAR mixed model	250	200	2000	24108.83
	EWR mixed model	250	200	2000	79661.45
Rho-independent transcription terminators	EAR mixed model	170	212	2000	15162.24
	EWR mixed model	170	212	2000	47300.76

Table 6-10. The execution time for training the EAR and the EWR models of tRNAs and rho-independent transcription terminators

"CPU time" is the CPU time of a 3.2 Ghz Pentium IV EMT64 machine which runs the 64-bit Linux OS.

When a trained model is applied to finding a particular type of RNA motifs in genomic sequences, most of the time will be spent on folding all windowed regions of genomic sequences. Using the Eponine RNA-motif models to scan the whole genome for searching RNA motifs can be very time-consuming. For example, using the EAR model to search for transcription termination terminators in the bacterial genomes took as long as one-week CPU time on an x86-64bit machine (3.2 Ghz Pentium IV EMT64, 64-bit Linux), scanning ~4-megabases x 2 (Table 6-11).

Organism	Genome length	CPU time (Pentium-4) (secs)
B. subtilis	4,214,630 x 2 strands	589755.91
E. coli	4,639,675 x 2 strands	638613.26

Table 6-11. The execution time for using the EAR model of *rho*-independent transcription terminators to scan the genomes of *B. subtilis* and *E. coli* respectively

6.2.2. Towards creating general EWR models of vertebrate ncRNAs

The scoring scheme of the Eponine RNA-motif extension is designed to allow a dynamic recruitment of relevant features. By using the Monte Carlo methods and the RVM strategy, theoretically the Eponine RNA-motif extension can determine the differential degrees of significance of various structural features for a particular hairpin and then choose the most relevant features for modelling it. In this project, however, this capability has not yet been evaluated. Lengths of stems and loops are currently the only features that have been recruited to model ncRNAs. It is possible that under certain circumstances, other features could significantly contribute to the model. While the hairpins of different classes of ncRNAs may vary in their stem and loop sizes, a recent report suggests that ncRNAs tend to have more stable structures than do random sequences (Clote et al. 2005). Although folding stability alone proved to be insufficient for identifying ncRNAs in genomes (Rivas and Eddy 2000), certain combinations of different structural features might be useful for genome-wide ncRNA finding.

One unfinished piece of work in this project is using the Eponine RNA-motif extension to create a general EWR model of vertebrate ncRNAs. There can be at least two approaches to fulfil this goal. Firstly, the EWR model can be used to find the consensus features of various classes of ncRNAs. In order to evaluate the performance of the trained model, a *k*-fold cross validation can be used. ncRNA classes can be divided into *k* groups and each group of ncRNAs is left out when training that particular model. The trained model could then be evaluated by using these ncRNAs. This process would be repeated until the *k* models had been evaluated.

Another possible approach for creating an EWR vertebrate-ncRNA-model is taking human-mouse syntenic alignments as the training sequences. The proposed approach can be,

not only a potential way to create a general ncRNA model, but also a useful strategy to look for undiscovered ncRNAs in mammalian genomes. The development of the Eponine RNA-motif extension provides a way to test hypotheses with regard to genome-wide ncRNA finding. The capability of this tool in genome-wide ncRNA finding is worthy of further exploration.

6.3. Summary

In this chapter, using three types of ncRNAs with distinct RNA structural motifs, I have demonstrated the capability of the Eponine RNA-motif extension to model the RNA motifs in transcripts. The applications of this extension include the following:

- When a particular type of functional sites is known for a set of sequences, Eponine anchored RNA-motif models can be used.
- When a functional site is suspected but the anchoring point in a set of transcripts is unknown, Eponine windowed RNA-motif models can be used.
- Eponine RNA models can be used for prediction, *i.e.* to search for novel sites of a particular type of ncRNAs in genomes.

There are some limitations of the tentative applications of the Eponine RNA-motif extension:

- The Eponine RNA-motif extension is designed to learn discrimination models consisting of local RNA motifs. This tool may not be capable of modelling the global consensus RNA secondary structure.
- For the purpose of discriminating novel functional sites in genomes, the trained model may be apt to find false positives that consist of only a subset of functional motifs.

There are some special issues that need to be taken into consideration in using the Eponine RNA-motif extension:

- A number of sequences are required for training the models.
- In training the models, significant amount of time may be spent in learning the parameters of PWMs and RNA motifs, due to the use of the Monte Carlo methods in optimization.

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.

Reference

Aerts, S., P. Van Loo, et al. (2005). "TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis." Nucleic Acids Res **33**(Web Server issue): W393-6.

Allen, T. A., S. Von Kaenel, et al. (2004). "The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock." Nat Struct Mol Biol **11**(9): 816-21.

Anderson, R. P. and J. R. Roth (1977). "Tandem genetic duplications in phage and bacteria." Annu Rev Microbiol **31**: 473-505.

Arney, K. L. (2003). "H19 and Igf2--enhancing the confusion?" Trends Genet **19**(1): 17-23.

Bafna, V. and S. Zhang (2004). "FastR: fast database search tool for non-coding RNA." Proc IEEE Comput Syst Bioinform Conf: 52-61.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bailey, T. L. and W. S. Noble (2003). "Searching for statistically significant regulatory modules." Bioinformatics **19 Suppl 2**: ii16-25.

Barash, Y., G. Elidan, et al. (2003). Modeling dependencies in protein-dna binding sites. In Proceedings of Seventh Annual International Conference on Computational Molecular Biology (RECOMB), ACM press, New York**:** 28-37.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-97.

Baskerville, S. and D. P. Bartel (2005). "Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes." Rna **11**(3): 241-7.

Beckmann, J. S., P. F. Johnson, et al. (1977). "Cloning of yeast transfer RNA genes in Escherichia coli." Science **196**(4286): 205-8.

Bentwich, I., A. Avniel, et al. (2005). "Identification of hundreds of conserved and nonconserved human microRNAs." Nat Genet **37**(7): 766-70.

Berg, O. G. and P. H. von Hippel (1987). "Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters." J Mol Biol **193**(4): 723-50.

Berger, S. L. (2002). "Histone modifications in transcriptional regulation." Curr Opin Genet Dev **12**(2): 142-8.

Birney, E., D. Andrews, et al. (2004). "Ensembl 2004." Nucleic Acids Res **32 Database issue**: D468-70.

Birney, E., D. Andrews, et al. (2006). "Ensembl 2006." Nucleic Acids Res **34**(Database issue): D556-61.

Borer, P. N., B. Dengler, et al. (1974). "Stability of ribonucleic acid double-stranded helices." J Mol Biol **86**(4): 843-53.

Brendel, V., G. H. Hamm, et al. (1986). "Terminators of transcription with RNA polymerase

from Escherichia coli: what they look like and how to find them." J Biomol Struct Dyn **3**(4): 705-23.

Brent, M. R. (2005). "Genome annotation past, present, and future: how to define an ORF at each locus." Genome Res **15**(12): 1777-86.

Burge, C. and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA." J Mol Biol **268**(1): 78-94.

Camier, S., R. E. Baker, et al. (1990). "On the flexible interaction of yeast factor tau with the bipartite promoter of tRNA genes." Nucleic Acids Res **18**(15): 4571-8.

Cannon, R. E., G. J. Wu, et al. (1986). "Functions of and interactions between the A and B blocks in adenovirus type 2-specific VARNA1 gene." Proc Natl Acad Sci U S A **83**(5): 1285-9.

Carninci, P., T. Kasukawa, et al. (2005). "The transcriptional landscape of the mammalian genome." Science **309**(5740): 1559-63.

Cavarelli, J., B. Rees, et al. (1993). "Yeast tRNA(Asp) recognition by its cognate class II aminoacyl-tRNA synthetase." Nature **362**(6416): 181-4.

Cech, T. R., A. J. Zaug, et al. (1981). "In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence." Cell **27**(3 Pt 2): 487-96.

Chang, Y. N., I. L. Pirtle, et al. (1986). "Nucleotide sequence and transcription of a human tRNA gene cluster with four genes." Gene **48**(1): 165-74.

Chen, J. H., S. Y. Le, et al. (2000). "Prediction of common secondary structures of RNAs: a genetic algorithm approach." Nucleic Acids Res **28**(4): 991-9.

Cheung, J., X. Estivill, et al. (2003). "Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence." Genome Biol **4**(4): R25.

Chiang, D., A. K. Joshi, et al. (2006). "Grammatical representations of macromolecular structure." J Comput Biol **13**(5): 1077-100.

Chiu, D. K. and T. Kolodziejczak (1991). "Inferring consensus structure from nucleic acid sequences." Comput Appl Biosci **7**(3): 347-52.

Chomsky, D. (1959). "On certain formal properties of grammars." Inform. Cont. **2**: 137-176.

Clamp, M., D. Andrews, et al. (2003). "Ensembl 2002: accommodating comparative genomics." Nucleic Acids Res **31**(1): 38-42.

Clote, P., F. Ferre, et al. (2005). "Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency." Rna **11**(5): 578-91.

Coventry, A., D. J. Kleitman, et al. (2004). "MSARI: multiple sequence alignments for statistical detection of RNA secondary structure." Proc Natl Acad Sci U S A **101**(33): 12102-7.

d'Aubenton Carafa, Y., E. Brody, et al. (1990). "Prediction of rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures." J

Mol Biol **216**(4): 835-58.

de Hoon, M. J., Y. Makita, et al. (2005). "Prediction of Transcriptional Terminators in Bacillus subtilis and Related Species." PLoS Comput Biol **1**(3): e25.

Dean, A. (2006). "On a chromosome far, far away: LCRs and gene expression." Trends Genet **22**(1): 38-45.

DeFranco, D., O. Schmidt, et al. (1980). "Two control regions for eukaryotic tRNA gene transcription." Proc Natl Acad Sci U S A **77**(6): 3365-8.

Dehal, P. and J. L. Boore (2005). "Two rounds of whole genome duplication in the ancestral vertebrate." PLoS Biol **3**(10): e314.

di Bernardo, D., T. Down, et al. (2003). "ddbRNA: detection of conserved secondary structures in multiple alignments." Bioinformatics **19**(13): 1606-11.

Dirks, R. M. and N. A. Pierce (2003). "A partition function algorithm for nucleic acid secondary structure including pseudoknots." J Comput Chem **24**(13): 1664-77.

Dittmar, K. A., J. M. Goodenbour, et al. (2006). "Tissue-Specific Differences in Human Transfer RNA Expression." PLoS Genet **2**(12): e221.

Domitrovich, A. M. and G. R. Kunkel (2003). "Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies." Nucleic Acids Res **31**(9): 2344-52.

Down, T., B. Leong, et al. (2006). "A machine learning strategy to identify candidate binding sites in human protein-coding sequence." BMC Bioinformatics **7**: 419.

Down, T. A. (2002). Computational localization of promoters and transcription start sites in mammalian genomes. The Wellcome Trust Sanger Institute. Hinxton, Cambridge, Cambridge.

Down, T. A. and T. J. Hubbard (2002). "Computational detection and location of transcription start sites in mammalian genomic DNA." Genome Res **12**(3): 458-61.

Down, T. A. and T. J. Hubbard (2004). "What can we learn from noncoding regions of similarity between genomes?" BMC Bioinformatics **5**: 131.

Durbin, R., S. R. Eddy, et al. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, Cambridge University Press.

Eddy, S. R. (2004). "How do RNA folding algorithms work?" Nat Biotechnol **22**(11): 1457-8.

Eddy, S. R. and R. Durbin (1994). "RNA sequence analysis using covariance models." Nucleic Acids Res **22**(11): 2079-88.

Ermolaeva, M. D., H. G. Khalak, et al. (2000). "Prediction of transcription terminators in bacterial genomes." J Mol Biol **301**(1): 27-33.

Espinoza, C. A., T. A. Allen, et al. (2004). "B2 RNA binds directly to RNA polymerase II to repress transcript synthesis." Nat Struct Mol Biol **11**(9): 822-9.

Farnham, P. J. and T. Platt (1981). "Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro." Nucleic Acids Res **9**(3):

563-77.

Fatica, A., M. Morlando, et al. (2000). "Yeast snoRNA accumulation relies on a cleavage-dependent/polyadenylation-independent 3'-processing apparatus." Embo J **19**(22): 6218-29.

Fichant, G. A. and C. Burks (1991). "Identifying potential tRNA genes in genomic DNA sequences." J Mol Biol **220**(3): 659-71.

Fickett, J. W. and A. G. Hatzigeorgiou (1997). "Eukaryotic promoter recognition." Genome Res **7**(9): 861-78.

Fields, D. S. and R. R. Gutell (1996). "An analysis of large rRNA sequences folded by a thermodynamic method." Fold Des **1**(6): 419-30.

Flamm, C., W. Fontana, et al. (2000). "RNA folding at elementary step resolution." Rna **6**(3): 325-38.

Fournier, M. J., W. L. Miller, et al. (1974). "Clustering of tRNA cistrons in Escherichia coli DNA." Biochem Biophys Res Commun **60**(3): 1148-54.

Frazer, K. A., L. Pachter, et al. (2004). "VISTA: computational tools for comparative genomics." Nucleic Acids Res **32**(Web Server issue): W273-9.

Galli, G., H. Hofstetter, et al. (1981). "Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements." Nature **294**(5842): 626-31.

Gardner, P. P. and R. Giegerich (2004). "A comprehensive comparison of comparative RNA structure prediction approaches." BMC Bioinformatics **5**(1): 140.

Gautheret, D. and A. Lambert (2001). "Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles." J Mol Biol **313**(5): 1003-11.

Giege, R., M. Sissler, et al. (1998). "Universal rules and idiosyncratic features in tRNA identity." Nucleic Acids Res **26**(22): 5017-35.

Giles, K. E., M. Caputi, et al. (2004). "Packaging and reverse transcription of snRNAs by retroviruses may generate pseudogenes." Rna **10**(2): 299-307.

Gish, W. (1996-2004). "WU-BLAST." http://blast.wustl.edu.

Gopalan, V., A. Vioque, et al. (2002). "RNase P: variations and uses." J Biol Chem **277**(9): 6759-62.

Gorodkin, J., S. L. Stricklin, et al. (2001). "Discovering common stem-loop motifs in unaligned RNA sequences." Nucleic Acids Res **29**(10): 2135-44.

Graur, D. and W.-H. Li (2000). Fundamentals of Molecular Evolution. Sunderland, Massachusetts, Sinauer Associates, Inc.

Gray, D. M. (1997). "Derivation of nearest-neighbor properties from data on nucleic acid oligomers. II. Thermodynamic parameters of DNA.RNA hybrids and DNA duplexes." Biopolymers **42**(7): 795-810.

Griffiths-Jones, S., A. Bateman, et al. (2003). "Rfam: an RNA family database." Nucleic Acids

Res **31**(1): 439-41.

Griffiths-Jones, S., R. J. Grocock, et al. (2006). "miRBase: microRNA sequences, targets and gene nomenclature." Nucleic Acids Res **34**(Database issue): D140-4.

Griffiths-Jones, S., S. Moxon, et al. (2005). "Rfam: annotating non-coding RNAs in complete genomes." Nucleic Acids Res **33 Database Issue**: D121-4.

Grosshans, H. and F. J. Slack (2002). "Micro-RNAs: small is plentiful." J Cell Biol **156**(1): 17-21.

Gu, X., Y. Wang, et al. (2002). "Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution." Nat Genet **31**(2): 205-9.

GuhaThakurta, D. and G. D. Stormo (2001). "Identifying target sites for cooperatively binding factors." Bioinformatics **17**(7): 608-21.

Gunnery, S., Y. Ma, et al. (1999). "Termination sequence requirements vary among genes transcribed by RNA polymerase III." J Mol Biol **286**(3): 745-57.

Gutell, R. R., N. Larsen, et al. (1994). "Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective." Microbiol Rev **58**(1): 10-26.

Gutell, R. R., A. Power, et al. (1992). "Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods." Nucleic Acids Res **20**(21): 5785-95.

Guthrie, C. and J. Abelson (1982). Organization and Expression of tRNA Genes in Saccharomyces cerevisiae. The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression. J. R. Broach. Woodbury, New York, Cold Spring Harbor Laboratory Press**:** 487-528.

Hallenberg, C., J. Nederby Nielsen, et al. (1994). "Characterization of 5S rRNA genes from mouse." Gene **142**(2): 291-5.

Havgaard, J. H., R. B. Lyngso, et al. (2005). "Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%." Bioinformatics **21**(9): 1815-24.

Hazkani-Covo, E., R. Sorek, et al. (2003). "Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications." J Mol Evol **56**(2): 169-74.

Helm, M. (2006). "Post-transcriptional nucleotide modification and alternative folding of RNA." Nucleic Acids Res **34**(2): 721-33.

Hermann, T. and E. Westhof (1999). "Non-Watson-Crick base pairs in RNA-protein recognition." Chem Biol **6**(12): R335-43.

Hillier, L. W., W. Miller, et al. (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature **432**(7018): 695-716.

Hochsmann, M., B. Voss, et al. (2004). "Pure multiple RNA secondary structure alignments: a

progressive profile approach." IEEE/ACM Trans Comput Biol Bioinform **1**(1): 53-62.

Hofacker, I. L. (2003). "Vienna RNA secondary structure server." Nucleic Acids Res **31**(13): 3429-31.

Hofacker, I. L. (2006). "RNAplot." http://www.tbi.univie.ac.at/~ivo/RNA/RNAplot.html.

Hofacker, I. L., S. H. Bernhart, et al. (2004). "Alignment of RNA base pairing probability matrices." Bioinformatics **20**(14): 2222-7.

Hofacker, I. L., M. Fekete, et al. (2002). "Secondary structure prediction for aligned RNA sequences." J Mol Biol **319**(5): 1059-66.

Hofacker, I. L., W. Fontana, et al. (1994-2006). "RNAfold." http://www.tbi.univie.ac.at/~ivo/RNA/RNAfold.html.

Hoffmann, A. A., C. M. Sgro, et al. (2004). "Chromosomal inversion polymorphisms and adaptation." Trends Ecol Evol **19**(9): 482-8.

Hsieh, Y. J., Z. Wang, et al. (1999). "Cloning and characterization of two evolutionarily conserved subunits (TFIIIC102 and TFIIIC63) of human TFIIIC and their involvement in functional interactions with TFIIIB and RNA polymerase III." Mol Cell Biol **19**(7): 4944-52.

Hu, J., B. Li, et al. (2005). "Limitations and potentials of current motif discovery algorithms." Nucleic Acids Res **33**(15): 4899-913.

Hu, Y. J. (2002). "Prediction of consensus structural motifs in a family of coregulated RNA sequences." Nucleic Acids Res **30**(17): 3886-93.

Huang, X. and A. Madan (1999). "CAP3: A DNA sequence assembly program." Genome Res **9**(9): 868-77.

Ingham, C. J., J. Dennis, et al. (1999). "Autogenous regulation of transcription termination factor Rho and the requirement for Nus factors in Bacillus subtilis." Mol Microbiol **31**(2): 651-63.

International Human Genome Sequencing Consortium (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

James, B. D., G. J. Olsen, et al. (1989). "Phylogenetic comparative analysis of RNA secondary structure." Methods Enzymol **180**: 227-39.

Ji, Y., X. Xu, et al. (2004). "A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences." Bioinformatics **20**(10): 1591-602.

Joyce, G. F. (2002). "The antiquity of RNA-based evolution." Nature **418**(6894): 214-21.

Kampa, D., J. Cheng, et al. (2004). "Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22." Genome Res **14**(3): 331-42.

Kanz, C., P. Aldebert, et al. (2005). "The EMBL Nucleotide Sequence Database." Nucleic Acids Res **33**(Database issue): D29-33.

Katayama, S., Y. Tomaru, et al. (2005). "Antisense transcription in the mammalian

transcriptome." Science **309**(5740): 1564-6.

Ke, A. and J. A. Doudna (2004). "Crystallization of RNA and RNA-protein complexes." Methods **34**(3): 408-14.

Khoury, G. and P. Gruss (1983). "Enhancer elements." Cell **33**(2): 313-4.

Kiss, T. and W. Filipowicz (1995). "Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns." Genes Dev **9**(11): 1411-24.

Klein, R. J. and S. R. Eddy (2003). "RSEARCH: finding homologs of single structured RNA sequences." BMC Bioinformatics **4**: 44.

Klosterman, P. S., D. K. Hendrix, et al. (2004). "Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns." Nucleic Acids Res **32**(8): 2342-52.

Knudsen, B. and J. Hein (1999). "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history." Bioinformatics **15**(6): 446-54.

Knudsen, B. and J. Hein (2003). "Pfold: RNA secondary structure prediction using stochastic context-free grammars." Nucleic Acids Res **31**(13): 3423-8.

Konings, D. A. and R. R. Gutell (1995). "A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs." Rna **1**(6): 559-74.

Kozak, M. (2005). "Regulation of translation via mRNA structure in prokaryotes and eukaryotes." Gene **361**: 13-37.

Kramerov, D. A. and N. S. Vassetzky (2005). "Short retroposons in eukaryotic genomes." Int Rev Cytol **247**: 165-221.

Kruger, K., P. J. Grabowski, et al. (1982). "Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena." Cell **31**(1): 147-57.

Lasser-Weiss, M., N. Bawnik, et al. (1981). "Isolation and characterization of cloned rat DNA fragment carrying tRNA genes." Nucleic Acids Res **9**(22): 5965-78.

Lawn, R. M., K. Schwartz, et al. (1997). "Convergent evolution of apolipoprotein(a) in primates and hedgehog." Proc Natl Acad Sci U S A **94**(22): 11992-7.

Lawrence, C. E., S. F. Altschul, et al. (1993). "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." Science **262**(5131): 208-14.

Lawrence, C. E. and A. A. Reilly (1990). "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences." Proteins **7**(1): 41-51.

Lee, Y., C. Ahn, et al. (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-9.

Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." Embo J **23**(20): 4051-60.

Leontis, N. B. and E. Westhof (2003). "Analysis of RNA motifs." Curr Opin Struct Biol **13**(3):

300-8.

Lesnik, E. A., R. Sampath, et al. (2001). "Prediction of rho-independent transcriptional terminators in Escherichia coli." Nucleic Acids Res **29**(17): 3583-94.

Li, Q., K. R. Peterson, et al. (2002). "Locus control regions." Blood **100**(9): 3077-86.

Lindgreen, S., P. P. Gardner, et al. (2006). "Measuring covariation in RNA alignments: physical realism improves information measures." Bioinformatics **22**(24): 2988-95.

Little, R. D. and D. C. Braaten (1989). "Genomic organization of human 5 S rDNA and sequence of one tandem repeat." Genomics **4**(3): 376-83.

Liu, X., D. L. Brutlag, et al. (2001). "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes." Pac Symp Biocomput: 127-38.

Lopez-Lastra, M., A. Rivas, et al. (2005). "Protein synthesis in eukaryotes: the growing biological relevance of cap-independent translation initiation." Biol Res **38**(2-3): 121-46.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-64.

Lukavsky, P. J. and J. D. Puglisi (2005). "Structure determination of large biological RNAs." Methods Enzymol **394**: 399-416.

Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-5.

Lyngso, R. B. and C. N. Pedersen (2000). "RNA pseudoknot prediction in energy-based models." J Comput Biol **7**(3-4): 409-27.

MacIsaac, K. D. and E. Fraenkel (2006). "Practical strategies for discovering regulatory DNA sequence motifs." PLoS Comput Biol **2**(4): e36.

Maestre, J., T. Tchenio, et al. (1995). "mRNA retroposition in human cells: processed pseudogene formation." Embo J **14**(24): 6333-8.

Mandal, M., B. Boese, et al. (2003). "Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria." Cell **113**(5): 577-86.

Marsan, L. and M. F. Sagot (2000). "Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification." J Comput Biol **7**(3-4): 345-62.

Matouk, I. J., N. DeGroot, et al. (2007). "The H19 non-coding RNA is essential for human tumor growth." PLoS ONE **2**(9): e845.

Matsui, H., K. Sato, et al. (2004). "Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures." Proc IEEE Comput Syst Bioinform Conf: 290-9.

Mattaj, I. W. (1993). "RNA recognition: a family matter?" Cell **73**(5): 837-40.

McCaskill, J. S. (1990). "The equilibrium partition function and base pair binding probabilities for RNA secondary structure." Biopolymers **29**(6-7): 1105-19.

McCullagh, P. and J. A. Nelder (1983). Generalized linear models, Chapman and Hall, London.

Morlando, M., P. Greco, et al. (2002). "Functional analysis of yeast snoRNA and snRNA 3'-end formation mediated by uncoupling of cleavage and polyadenylation." Mol Cell Biol **22**(5): 1379-89.

Mourier, T., A. J. Hansen, et al. (2001). "The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus." Mol Biol Evol **18**(9): 1833-7.

Mouse Genome Sequencing Consortium (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Mungall, A. J., S. A. Palmer, et al. (2003). "The DNA sequence and analysis of human chromosome 6." Nature **425**(6960): 805-11.

Nahvi, A., J. E. Barrick, et al. (2004). "Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes." Nucleic Acids Res **32**(1): 143-50.

Nakanishi, K. and O. Nureki (2005). "Recent progress of structural biology of tRNA processing and modification." Mol Cells **19**(2): 157-66.

Nam, J. W., K. R. Shin, et al. (2005). "Human microRNA prediction through a probabilistic co-learning model of sequence and structure." Nucleic Acids Res **33**(11): 3570-81.

Neidle, S. (2002). Nucleic acid structure and recognition, Oxford University Press, New York.

Nissen, P., J. Hansen, et al. (2000). "The structural basis of ribosome activity in peptide bond synthesis." Science **289**(5481): 920-30.

Nomenclature Committee of the International Union of Biochemistry, N.-I. (1986). "Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984." Proc Natl Acad Sci U S A **83**(1): 4-8.

Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol **302**(1): 205-17.

Nussinov, R. and A. B. Jacobson (1980). "Fast algorithm for predicting the secondary structure of single-stranded RNA." Proc Natl Acad Sci U S A **77**(11): 6903-13.

Ohler, U., S. Yekta, et al. (2004). "Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification." Rna **10**(9): 1309-22.

Ohshima, K., M. Hattori, et al. (2003). "Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates." Genome Biol **4**(11): R74.

Okazaki, Y., M. Furuno, et al. (2002). "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs." Nature **420**(6915): 563-73.

Olson, T., M. J. Fournier, et al. (1976). "Detection of a major conformational change in transfer ribonucleic acid by laser light scattering." J Mol Biol **102**(2): 193-203.

Onoa, B. and I. Tinoco, Jr. (2004). "RNA folding and unfolding." Curr Opin Struct Biol **14**(3): 374-9.

Osada, R., E. Zaslavsky, et al. (2004). "Comparative analysis of methods for representing and searching for transcription factor binding sites." Bioinformatics **20**(18): 3516-25.

Ota, T., Y. Suzuki, et al. (2004). "Complete sequencing and characterization of 21,243 full-length human cDNAs." Nat Genet **36**(1): 40-5.

Pasquinelli, A. E., B. J. Reinhart, et al. (2000). "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA." Nature **408**(6808): 86-9.

Passananti, C., B. Davies, et al. (1987). "Structure of an inverted duplication formed as a first step in a gene amplification event: implications for a model of gene amplification." Embo J **6**(6): 1697-703.

Paule, M. R. and R. J. White (2000). "Survey and summary: transcription by RNA polymerases I and III." Nucleic Acids Res **28**(6): 1283-98.

Pavesi, A., F. Conterio, et al. (1994). "Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions." Nucleic Acids Res **22**(7): 1247-56.

Pavesi, G., G. Mauri, et al. (2001). "An algorithm for finding signals of unknown length in DNA sequences." Bioinformatics **17 Suppl 1**: S207-14.

Pavesi, G., G. Mauri, et al. (2004). "In silico representation and discovery of transcription factor binding sites." Brief Bioinform **5**(3): 217-36.

Pedersen, J. S., G. Bejerano, et al. (2006). "Identification and classification of conserved RNA secondary structures in the human genome." PLoS Comput Biol **2**(4): e33.

Ramadass, A. S. (2004). Computational detection of gene regulatory signals in human genome sequence. The Wellcome Trust Sanger Institute. Hinxton, Cambridge, Cambridge.

Rangan, P. and S. A. Woodson (2003). "Structural requirement for Mg2+ binding in the group I intron core." J Mol Biol **329**(2): 229-38.

Ravasi, T., H. Suzuki, et al. (2006). "Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome." Genome Res **16**(1): 11-9.

Reams, A. B. and E. L. Neidle (2004). "Selection for gene clustering by tandem duplication." Annu Rev Microbiol **58**: 119-42.

Reddy, R., D. Henning, et al. (1987). "The capped U6 small nuclear RNA is transcribed by RNA polymerase III." J Biol Chem **262**(1): 75-81.

Ricchetti, M., F. Tekaia, et al. (2004). "Continued colonization of the human genome by mitochondrial DNA." PLoS Biol **2**(9): E273.

Riek, R., K. Pervushin, et al. (2000). "TROSY and CRINEPT: NMR with large molecular and supramolecular structures in solution." Trends Biochem Sci **25**(10): 462-8.

Rietveld, K., R. Van Poelgeest, et al. (1982). "The tRNA-like structure at the 3' terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA." Nucleic Acids Res **10**(6): 1929-46.

Ringner, M. and M. Krogh (2005). "Folding Free Energies of 5'-UTRs Impact Post-Transcriptional Regulation on a Genomic Scale in Yeast." PLoS Comput Biol **1**(7): e72.

Rivas, E. and S. R. Eddy (1999). "A dynamic programming algorithm for RNA structure prediction including pseudoknots." J Mol Biol **285**(5): 2053-68.

Rivas, E. and S. R. Eddy (2000). "The language of RNA: a formal grammar that includes pseudoknots." Bioinformatics **16**(4): 334-40.

Rivas, E. and S. R. Eddy (2000). "Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs." Bioinformatics **16**(7): 583-605.

Rivas, E. and S. R. Eddy (2001). "Noncoding RNA gene detection using comparative sequence analysis." BMC Bioinformatics **2**(1): 8.

Rivas, E., R. J. Klein, et al. (2001). "Computational identification of noncoding RNAs in E. coli by comparative genomics." Curr Biol **11**(17): 1369-73.

Rogic, S., A. K. Mackworth, et al. (2001). "Evaluation of gene-finding programs on mammalian sequences." Genome Res **11**(5): 817-32.

Romero, D. and R. Palacios (1997). "Gene amplification and genomic plasticity in prokaryotes." Annu Rev Genet **31**: 91-111.

Rould, M. A., J. J. Perona, et al. (1991). "Structural basis of anticodon loop recognition by glutaminyl-tRNA synthetase." Nature **352**(6332): 213-8.

Ruan, J., G. D. Stormo, et al. (2004). "An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots." Bioinformatics **20**(1): 58-66.

Sakakibara, Y., M. Brown, et al. (1994). "Stochastic context-free grammars for tRNA modeling." Nucleic Acids Res **22**(23): 5112-20.

Sandelin, A., P. Carninci, et al. (2007). "Mammalian RNA polymerase II core promoters: insights from genome-wide studies." Nat Rev Genet **8**(6): 424-36.

Sankoff, D. (1985). "Simultaneous solution of the RNA folding, alignment and protosequence problems." SIAM J. Appl. Math. **45**(5): 810-825.

Sankoff, D. and N. El-Mabrouk (2000). Genome Rearrangement. Current topics in computational biology. M. Zhang. Cambridge, MIT Press**:** 135-155.

Schmeing, T. M., A. C. Seila, et al. (2002). "A pre-translocational intermediate in protein synthesis observed in crystals of enzymatically active 50S subunits." Nat Struct Biol **9**(3): 225-30.

Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Res **18**(20): 6097-100.

Schwartz, S., W. J. Kent, et al. (2003). "Human-mouse alignments with BLASTZ." Genome Res **13**(1): 103-7.

Seitz, H., H. Royo, et al. (2004). "A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain." Genome Res **14**(9): 1741-8.

Shanab, G. M. and E. S. Maxwell (1991). "Proposed secondary structure of eukaryotic U14 snRNA." Nucleic Acids Res **19**(18): 4891-4.

Siebert, S. and R. Backofen (2005). "MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons." Bioinformatics **21**(16): 3352-9.

Siepel, A., G. Bejerano, et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res **15**(8): 1034-50.

Sinha, S. and M. Tompa (2000). "A statistical method for finding transcription factor binding sites." Proc Int Conf Intell Syst Mol Biol **8**: 344-54.

Sinha, S. and M. Tompa (2002). "Discovery of novel transcription factor binding sites by statistical overrepresentation." Nucleic Acids Res **30**(24): 5549-60.

Slack, F. J., M. Basson, et al. (2000). "The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor." Mol Cell **5**(4): 659-69.

Smit, A. F. (1999). "Interspersed repeats and other mementos of transposable elements in mammalian genomes." Curr Opin Genet Dev **9**(6): 657-63.

Smit, A. F. and A. D. Riggs (1995). "MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation." Nucleic Acids Res **23**(1): 98-102.

Smit, A. F. A. and P. Green (unpublished). "RepeatMasker." http://ftp.genome.washington.edu/RM/RepeatMasker.html.

Sprinzl, M., C. Horn, et al. (1998). "Compilation of tRNA sequences and sequences of tRNA genes." Nucleic Acids Res **26**(1): 148-53.

Sprinzl, M. and K. S. Vassilenko (2005). "Compilation of tRNA sequences and sequences of tRNA genes." Nucleic Acids Res **33**(Database issue): D139-40.

Staple, D. W. and S. E. Butcher (2005). "Pseudoknots: RNA structures with diverse functions." PLoS Biol **3**(6): e213.

Steege, D. A. (2000). "Emerging features of mRNA decay in bacteria." Rna **6**(8): 1079-90.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Storz, G. (2002). "An expanding universe of noncoding RNAs." Science **296**(5571): 1260-3.

Suzuki, H., K. Moriwaki, et al. (1994). "Sequences and evolutionary analysis of mouse 5S rDNAs." Mol Biol Evol **11**(4): 704-10.

Tamura, M., D. K. Hendrix, et al. (2004). "SCOR: Structural Classification of RNA, version 2.0." Nucleic Acids Res **32**(Database issue): D182-4.

Taneda, A. (2005). "Cofolga: a genetic algorithm for finding the common folding of two RNAs." Comput Biol Chem **29**(2): 111-9.

The ENCODE Project Consortium (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146):

799-816.

Thompson, W., E. C. Rouchka, et al. (2003). "Gibbs Recursive Sampler: finding transcription factor binding sites." Nucleic Acids Res **31**(13): 3580-5.

TIGR (2002-2003). "TIGR Gene Indices Clustering Tools (TGICL)."

Tipping, M. E. (1999). "The relevance vector machine." Advances in Neural Information Processing Systems 12 [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]: 652-658.

Tompa, M. (1999). "An exact method for finding short motifs in sequences, with application to the ribosome binding site problem." Proc Int Conf Intell Syst Mol Biol: 262-71.

Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." Nat Biotechnol **23**(1): 137-44.

Torarinsson, E., M. Sawera, et al. (2006). "Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure." Genome Res **16**(7): 885-9.

Torrents, D., M. Suyama, et al. (2003). "A genome-wide survey of human pseudogenes." Genome Res **13**(12): 2559-67.

Tourmen, Y., O. Baris, et al. (2002). "Structure and chromosomal distribution of human mitochondrial pseudogenes." Genomics **80**(1): 71-7.

Tsuzuki, T., H. Nomiyama, et al. (1983). "Presence of mitochondrial-DNA-like sequences in the human nuclear DNA." Gene **25**(2-3): 223-9.

Uemura, Y., A. Hasegawa, et al. (1999). "Tree adjoining grammars for RNA structure prediction." Theoretical Computer Science **210**(2): 277-303.

Uptain, S. M. and M. J. Chamberlin (1997). "Escherichia coli RNA polymerase terminates transcription efficiently at rho-independent terminators on single-stranded DNA templates." Proc Natl Acad Sci U S A **94**(25): 13548-53.

van Batenburg, F. H., A. P. Gultyaev, et al. (2001). "PseudoBase: structural information on RNA pseudoknots." Nucleic Acids Res **29**(1): 194-5.

Van de Peer, Y. (2004). "Computational approaches to unveiling ancient genome duplications." Nat Rev Genet **5**(10): 752-63.

Varani, G. and A. Pardi (1994). Structure of RNA. RNA-Protein Interactions. M. I.W. Oxford University, Oxford, IRL PRESS**:** 1-24.

Vignali, M., A. H. Hassan, et al. (2000). "ATP-dependent chromatin-remodeling complexes." Mol Cell Biol **20**(6): 1899-910.

Walter, A. E., D. H. Turner, et al. (1994). "Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding." Proc Natl Acad Sci U S A **91**(20): 9218-22.

Washietl, S., I. L. Hofacker, et al. (2005). "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome." Nat

Biotechnol **23**(11): 1383-90.

Washietl, S., I. L. Hofacker, et al. (2005). "Fast and reliable prediction of noncoding RNAs." Proc Natl Acad Sci U S A **102**(7): 2454-9.

Washietl, S., J. S. Pedersen, et al. (2007). "Structured RNAs in the ENCODE selected regions of the human genome." Genome Res **17**(6): 852-64.

Wasserman, W. W. and J. W. Fickett (1998). "Identification of regulatory regions which confer muscle-specific gene expression." J Mol Biol **278**(1): 167-81.

Wasserman, W. W. and A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements." Nat Rev Genet **5**(4): 276-87.

Weinmann, R., H. J. Raskas, et al. (1974). "Role of DNA-dependent RNA polymerases II and III in transcription of the adenovirus genome late in productive infection." Proc Natl Acad Sci U S A **71**(9): 3426-39.

Weischenfeldt, J., J. Lykke-Andersen, et al. (2005). "Messenger RNA surveillance: neutralizing natural nonsense." Curr Biol **15**(14): R559-62.

Westhof, E. and F. Michel (1994). Prediction and experimental investigation of RNA secondary and tertiary foldings. RNA-Protein Interactions. M. I.W. Oxford University, Oxford, IRL PRESS**:** 26-51.

Will, C. L. and R. Luhrmann (2001). "Spliceosomal UsnRNP biogenesis, structure and function." Curr Opin Cell Biol **13**(3): 290-301.

Wilson, J. H., J. S. Kim, et al. (1972). "Bacteriophage T4 transfer RNA. 3. Clustering of the genes for the T4 transfer RNA's." J Mol Biol **71**(3): 547-56.

Woischnik, M. and C. T. Moraes (2002). "Pattern of organization of human mitochondrial pseudogenes in the nuclear genome." Genome Res **12**(6): 885-93.

Wu, G. J., J. F. Railey, et al. (1987). "Defining the functional domains in the control region of the adenovirus type 2 specific VARNA1 gene." J Mol Biol **194**(3): 423-42.

Xue, C., F. Li, et al. (2005). "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine." BMC Bioinformatics **6**: 310.

Zhang, Z., P. M. Harrison, et al. (2003). "Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome." Genome Res **13**(12): 2541-58.

Zhou, Q. and W. H. Wong (2004). "CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling." Proc Natl Acad Sci U S A **101**(33): 12114-9.

Zuker, M. (1989). "On finding all suboptimal foldings of an RNA molecule." Science **244**(4900): 48-52.

Zuker, M. and P. Stiegler (1981). "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information." Nucleic Acids Res **9**(1): 133-48.

Appendix A . Tables related to the investigation of tRNA-gene order conservation in mammalian genomes

This appendix contains tables related to the investigation of tRNA-gene order conservation in the mammalian genomes (see chapter 2, section 2.2)

TGC	Ala1	GTG	His1	TGG	Pro1	GCT	Ser5
GGC	Ala2	ATG	His2	GGG	Pro2	ACT	Ser6
CGC	Ala3	TAT	Ile1	CGG	Pro3	TGT	Thr1
AGC	Ala4	GAT	Ile2	AGG	Pro4	GGT	Thr2
GCA	Cys1	AAT	Ile3	TTG	Gln1	CGT	Thr3
ACA	Cys2	TTT	Lys1	CTG	Gln2	AGT	Thr4
GTC	Asp1	CTT	Lys2	TCG	Arg1	TAC	Val1
ATC	Asp2	TAA	Leu1	GCG	Arg2	GAC	Val2
TTC	Glu1	CAA	Leu2	CCG	Arg3	CAC	Val3
CTC	Glu2	TAG	Leu3	ACG	Arg4	AAC	Val4
GAA	Phe1	GAG	Leu4	TCT	Arg5	CCA	Trp1
AAA	Phe2	CAG	Leu5	CCT	Arg6	GTA	Tyr1
TCC	Gly1	AAG	Leu6	TGA	Ser1	ATA	Tyr2
GCC	Gly2	CAT	Met1	GGA	Ser2	TTA	Ter1
CCC	Gly3	GTT	Asn1	CGA	Ser3	CTA	Ter2
ACC	Gly4	ATT	Asn2	AGA	Ser4	TCA	Sec1

Table A 1. Lookup table of anticodon types and the tRNA-gene symbols

cluster ID	chr	start	end
1.1.10	1	16,719,667	17,088,832
2.1.2	1	93,754,422	94,085,801
3.1.42	1	142,481,551	148,284,076
4.1.36	1	159,636,114	159,858,162
5.1.2	1	165,950,586	165,951,420
6.1.3	1	202,742,278	203,709,966
7.1.2	1	247,134,677	247,135,141
8.2.2	2	27,127,154	27,127,658
9.2.2	2	130,749,494	130,811,242
10.2.2	2	156,965,527	156,965,975
11.3.2	3	133,430,634	133,433,403
12.3.2	3	149,703,918	149,799,324
13.5.17	5	180,456,676	180,582,073
14.6.150	6	26,394,733	29,064,839
15.6.8	6	58,249,836	58,304,654
16.6.2	6	144,579,377	145,545,623
17.7.20	7	148,638,214	149,035,764
18.8.4	8	66,772,086	67,189,050
19.11.8	11	59,074,678	59,090,501

cluster ID	chr	start	end
20.11.2	11	75,624,205	75,624,588
21.12.2	12	97,421,412	97,422,232
22.12.5	12	123,972,254	123,990,536
23.13.2	13	40,532,874	40,928,132
24.14.14	14	20,147,335	20,222,086
25.15.3	15	43,278,096	43,280,712
26.15.2	15	76,939,959	77,824,124
27.16.17	16	3,140,676	3,359,885
28.16.2	16	22,114,533	22,216,043
29.16.2	16	55,891,364	55,891,975
30.16.5	16	69,369,615	70,017,969
31.17.18	17	7,963,198	8,071,107
32.17.2	17	19,352,086	19,704,837
33.17.8	17	34,161,560	35,527,152
34.17.3	17	70,541,596	70,542,875
35.18.2	18	41,553,749	41,923,341
36.19.2	19	1,334,361	1,334,635
37.19.2	19	4,675,082	4,675,719
38.X.3	X	3,766,418	3,843,344

Table A 2 The start and end coordinates of the tRNA gene clusters in the human genome (assembly NCBI 36)

Each cluster identifier (ID) is composed of three numbers separated by ".". The first number is a serial number. The second number (or X) is the chromosome on which a particular cluster resides. The third number is the number of tRNA gene loci in a particular cluster.

chr: chromosome

cluster ID	chr	start	end
1.1.3	1	73,971,393	74,985,840
2.1.2	1	107,331,203	107,332,257
3.1.2	1	134,861,508	134,861,945
4.1.2	1	167,478,309	167,479,017
5.1.26	1	172,870,617	173,506,186
6.2.2	2	56,997,464	56,997,850
7.2.2	2	118,738,191	118,747,667
8.2.3	2	122,066,935	122,069,480
9.3.2	3	3,109,391	3,135,216
10.3.5	3	19,820,110	20,371,715
11.3.2	3	51,446,283	51,447,407
12.3.30	3	96,396,659	97,766,935
13.4.2	4	56,953,853	57,727,180
14.4.3	4	131,397,335	132,386,642
15.4.2	4	149,499,077	150,476,012
16.5.2	5	31,164,664	31,165,168
17.5.5	5	125,693,626	125,698,919
18.5.2	5	142,649,903	142,755,501
19.6.52	6	47,908,583	48,294,102
20.6.2	6	86,211,030	86,369,597
21.7.2	7	28,081,759	28,502,820
22.7.3	7	98,690,607	99,418,054
23.7.2	7	120,626,628	120,708,747
24.8.2	8	97,592,760	97,593,215

cluster ID	chr	start	end
25.8.3	8	113,517,230	113,949,306
26.9.2	9	64,181,087	64,536,123
27.9.3	9	104,258,153	104,266,736
28.10.3	10	61,786,481	62,824,914
29.10.2	10	79,652,093	79,652,361
30.10.2	10	90,611,211	90,611,967
31.11.8	11	48,661,965	48,700,478
32.11.2	11	58,118,372	58,118,775
33.11.18	11	68,853,198	68,941,443
34.11.2	11	94,705,047	95,675,333
35.11.6	11	97,518,539	97,805,084
36.11.3	11	115,229,071	115,229,941
37.12.2	12	16,346,839	16,877,619
38.13.60	13	21,168,250	22,058,232
39.13.46	13	23,277,886	23,618,045
40.14.7	14	49,985,834	50,012,669
41.16.2	16	3,012,435	3,364,711
42.17.8	17	23,261,584	23,277,957
43.17.2	17	35,195,954	35,288,056
44.19.2	19	3,066,129	3,576,335
45.19.8	19	12,069,281	12,079,383
46.X.2	X	13,016,125	13,859,646
47.X.15	X	131,542,096	131,936,800
48.X.2	X	156,110,215	156,479,321

Table A 3. The start and end coordinates of the tRNA gene clusters in the mouse genome (assembly NCBI M36).

The convention used to assign the cluster ID to each cluster is the same as that used in Table A 2

human clusters (NCBI36)	mouse clusters (NCBI M36)	conservation type	quality of the human genome assembly	qualitiy of the mouse genome assembly
1.1.10	NA	synteny-non-conserved	CSN	FCS
2.1.2	coord: 3.122284970.122285054.-1	single-conserved	FCS	FCS
3.1.42	12.3.30	complicated	CSN	FCS
4.1.36	5.1.26	gapped	FCS	FCS
5.1.2	4.1.2	perfect	FCS	FCS
6.1.3	3.1.2	sub perfect type two	FCS	FCS
7.1.2	32.11.2	perfect	FCS	FCS
8.2.2	16.5.2	perfect	FCS	FCS
9.2.2	coord: 1.34379358.34379429.-1	single-conserved	FCS	FCS
10.2.2	6.2.2	perfect	FCS	FCS
11.3.2	27.9.3	sub perfect type two	FCS	FCS[1]
12.3.2	NA	synteny-non-conserved	FCS	WGS
13.5.17	31.11.8	gapped	FCS	FCS
14.6.150	38.13.60/39.13.46	gapped	FCS	FCS
15.6.8	NA	synteny-non-conserved	CSN	FCS
16.6.2	coord: 10.12612761.12612843.-1	single-conserved	FCS	FCS
17.7.20	19.6.52	gapped	FCS	FCS
18.8.4	10.3.5	sub perfect type two	FCS	FCS
19.11.8	45.19.8	sub perfect type one	FCS	FCS
20.11.2	22.7.3	sub perfect type two	FCS	FCS
21.12.2	30.10.2	perfect	FCS	FCS
22.12.5	17.5.5	sub perfect type one	FCS	FCS
23.13.2	NA	synteny-non-conserved	FCS	FCS
24.14.14	40.14.7	gapped	FCS	FCS
25.15.3	8.2.3	sub perfect type one	FCS	FCS
26.15.2	coord: 9.89924402.89924474.-1	single-conserved	FCS	FCS
27.16.17	42.17.8	gapped	FCS	FCS[2]
28.16.2	23.7.2	sub perfect type one	FCS	FCS
29.16.2	24.8.2	perfect	FCS	FCS
30.16.5	25.8.3	gapped	FCS	FCS[3]
31.17.18	33.11.18	sub perfect type one	FCS	FCS

human clusters (NCBI36)	mouse clusters (NCBI M36)	conservation type	quality of the human genome assembly	qualitiy of the mouse genome assembly
32.17.2	coord: 11.61224111.61224182.-1	single-conserved	FCS	FCS
33.17.8	35.11.6	gapped	FCS	FCS
34.17.3	36.11.3	perfect	FCS	FCS
35.18.2	NA	synteny-non-conserved	FCS	CSN
36.19.2	29.10.2	perfect	FCS	FCS
37.19.2	NA	synteny-non-conserved	FCS	FCS
38.X.3	NA	synteny-non-conserved	CSN	WGS

Table A 4. The synteny conservation of clustered human tRNA gene loci in the mouse genome

For columns 1 and 2 the cluster IDs are taken from Table A 1 and Table A 2 for human and mouse respectively.

NA: not available (when there is no corresponding cluster in the mouse genome).

coord: “coordinate” of a singlet tRNA gene locus in the mouse genome. This is used when the syntenic counterpart in the mouse genome is a singlet. The convention used here is chromosome:start:end:strand.

FCS: finished contig sequence; CSN: unfinished contig sequence (with gaps); WGS: whole genome shotgun sequence

[1]: mouse WGS between the 3' end tRNA gene and 3' boundary of the syntenic block

[2]: mouse WGS in the upstream region of the 5' end tRNA gene in this cluster

[3]: mouse WGS between the 5' end tRNA gene and 5' boundary of the syntenic block

singlet ID	coordinate (NCBI36)	coordinate (NCBIM36)	quality of the human genome assembly	quality of the mouse genome assembly
nc.1	1.55196130.55196202.-1	NA	FCS	FCS
nc.2	1.151910350.151910421.1	3.90561787.90561858.-1	FCS	FCS
nc.3	1.157378025.157378098.-1	1.175227004.175227077.1	FCS	FCS
nc.4	1.170424162.170424230.-1	NA	FCS	FCS
nc.5	1.178450899.178450971.-1	NA	FCS	FCS
nc.6	1.220704970.220705042.1	NA	FCS	CSN
nc.7	2.42891180.42891272.1	17.83770270.83770362.1	FCS	FCS
nc.8	2.70329627.70329697.-1	6.86369527.86369597.1	FCS	FCS
nc.9	2.74977554.74977622.1	NA	FCS	FCS
nc.10	2.117498979.117499050.-1	NA	FCS	FCS
nc.11	2.218818794.218818886.1	NA	FCS	FCS
nc.12	3.45705495.45705567.-1	9.123378123.123378195.-1	FCS	FCS
nc.13	3.126895867.126895938.-1	NA	FCS	FCS
nc.14	3.170972712.170972784.1	3.30792108.30792180.1	FCS	FCS
nc.15	3.185848789.185848859.-1	NA	FCS	FCS
nc.16	4.40603500.40603572.-1	NA	FCS	FCS
nc.17	4.124649455.124649526.-1	NA	FCS	FCS
nc.18	4.156604428.156604502.-1	NA	FCS	FCS
nc.19	5.26234296.26234368.-1	NA	FCS	FCS
nc.20	5.141754172.141754243.-1	NA	FCS	FCS
nc.21	5.159324619.159324696.-1	NA	FCS	FCS
nc.22	6.18944381.18944452.1	NA	FCS	WGS
nc.23	6.37395973.37396045.1	NA	FCS	FCS
nc.24	6.69971099.69971181.1	NA	FCS	FCS
nc.25	6.126143086.126143157.-1	10.30500556.30500627.1	FCS	FCS
nc.26	6.142620469.142620539.1	NA	FCS	FCS
nc.27	7.98905243.98905314.1	NA	FCS	FCS
nc.28	7.128210740.128210811.1	6.29338834.29338905.1	FCS	FCS
nc.29	7.138675986.138676058.1	6.38463539.38463611.1	FCS	FCS
nc.30	8.59667352.59667422.1	NA	FCS	FCS
nc.31	8.96351061.96351142.-1	4.10801211.10801292.1	FCS	FCS
nc.32	8.124238651.124238723.-1	15.57806066.57806138.-1	FCS	FCS
nc.33	9.5085085.5085156.1	NA	FCS	FCS
nc.34	9.14423938.14424009.-1	4.82090854.82090925.-1	FCS	FCS
nc.35	9.19393996.19394070.1	NA	FCS	FCS

singlet ID	coordinate (NCBI36)	coordinate (NCBIM36)	quality of the human genome assembly	quality of the mouse genome assembly
nc.36	9.76707810.76707881.-1	NA	FCS	FCS
nc.37	9.112000624.112000696.1	NA	FCS	FCS
nc.38	9.114656810.114656908.1	NA	FCS	FCS
nc.39	9.125695343.125695415.-1	NA	FCS	FCS
nc.40	9.130142176.130142266.-1	NA	FCS	FCS
nc.41	10.5935680.5935752.-1	NA	WGS	FCS
nc.42	10.22558444.22558517.-1	2.18504798.18504871.-1	WGS	FCS
nc.43	10.69194267.69194348.1	10.62824833.62824914.-1	FCS	FCS
nc.44	11.9253366.9253439.1	NA	FCS	FCS
nc.45	11.45246776.45246849.-1	NA	FCS	FCS
nc.46	11.50190455.50190526.-1	NA	FCS	FCS
nc.47	11.51216476.51216548.1	NA	FCS	FCS
nc.48	11.65872167.65872248.1	19.5038304.5038385.-1	FCS	FCS
nc.49	11.108541249.108541330.1	NA	FCS	FCS
nc.50	11.121935865.121935937.1	NA	FCS	FCS
nc.51	12.27734573.27734645.1	NA	FCS	FCS
nc.52	12.54870415.54870496.1	10.127861413.127861494.-1	FCS	FCS
nc.53	12.73137449.73137521.1	NA	FCS	FCS
nc.54	12.94953930.94954001.1	10.92882777.92882848.-1	FCS	FCS
nc.55	12.121426877.121426947.1	NA	FCS	FCS
nc.56	13.30146101.30146174.-1	5.149539350.149539423.-1	FCS	WGS
nc.57	13.44390062.44390133.-1	14.74886929.74887000.1	FCS	FCS
nc.58	13.93999905.93999977.-1	14.116971871.116971943.-1	FCS	FCS
nc.59	14.22468750.22468822.1	14.53471901.53471973.1	FCS	FCS
nc.60	14.31306567.31306637.-1	NA	FCS	FCS
nc.61	14.57776366.57776438.-1	12.71887725.71887797.-1	FCS	FCS
nc.62	14.72499432.72499503.1	NA	FCS	FCS
nc.63	14.88515195.88515267.1	NA	FCS	FCS
nc.64	14.101853182.101853255.1	12.111293145.111293218.1	FCS	FCS
nc.65	15.23878474.23878545.-1	7.58267184.58267255.1	FCS	FCS
nc.66	15.38673315.38673396.-1	2.118738191.118738272.-1	FCS	FCS
nc.67	15.63948454.63948525.-1	9.64536052.64536123.1	FCS	FCS
nc.68	15.87679308.87679380.1	7.79339932.79340004.1	FCS	FCS
nc.69	16.626737.626807.1	17.25602688.25602758.1	FCS	FCS
nc.70	16.14287251.14287322.1	16.13350901.13350972.1	FCS	FCS

singlet ID	coordinate (NCBI36)	coordinate (NCBIM36)	quality of the human genome assembly	quality of the mouse genome assembly
nc.71	16.72069717.72069789.-1	NA	FCS	FCS
nc.72	16.85975129.85975201.-1	8.124465281.124465353.-1	FCS	FCS
nc.73	17.15349410.15349483.1	NA	FCS	FCS
nc.74	17.26901213.26901284.1	11.79520845.79520916.1	FCS	FCS
nc.75	17.44624889.44624960.1	11.95675262.95675333.-1	FCS	FCS
nc.76	17.56218375.56218445.1	NA	CSN	FCS
nc.77	17.59957380.59957453.-1	NA	FCS	FCS
nc.78	17.63446475.63446547.-1	11.106828956.106829028.1	FCS	FCS
nc.79	17.78045886.78045957.-1	NA	CSN	FCS
nc.80	19.19713207.19713277.1	NA	FCS	WGS
nc.81	19.38359803.38359876.1	7.34943530.34943603.-1	FCS	FCS
nc.82	19.40758590.40758662.1	NA	FCS	FCS
nc.83	19.44594648.44594740.-1	7.28081759.28081853.1	FCS	FCS
nc.84	19.50673700.50673785.-1	7.18459766.18459851.1	FCS	FCS
nc.85	19.54729745.54729817.-1	NA	FCS	FCS
nc.86	19.57117208.57117280.-1	NA	FCS	WGS
nc.87	20.17803142.17803219.1	NA	FCS	FCS
nc.88	20.48385749.48385830.-1	NA	FCS	FCS
nc.89	21.14848387.14848457.1	NA	FCS	FCS
nc.90	21.17748978.17749048.-1	NA	FCS	FCS
nc.91	22.42877870.42877955.1	NA	FCS	FCS
nc.92	X.18602950.18603022.-1	X.156110215.156110287.1	FCS	FCS

Table A 5. The synteny conservation of non-clustered human tRNA gene loci (singlets) in the mouse genome

NA: not available (when there is no corresponding cluster in the mouse genome)

The assignment of a singlet ID follows the convention: “nc” (non-clustered). “serial number”.

The coordinates presented here follow the convention of that used in Table A 4.

Appendix B. The program sets written for this thesis

This appendix lists the main program sets that were particularly written for this thesis

Program set 1:

Table B 1. Functions of the program sets written for this thesis

Name: Search for synteny-conserved ncRNAs
Description of function:
Search for synteny-conserved ncRNAs in syntenic regions between two genomes, and determine the number of covariations between each pair of orthologous ncRNAs that are synteny-conserved. For each ncRNA locus in a particular genome, this program set can search for its corresponding syntenic blocks, which are defined by the unique best reciprocal homologue pairs (UBRHPs) that are determined by Ensembl, in other genome(s). For a particular ncRNA in one genome, its synteny-conserved counterpart is searched for in the corresponding syntenic region of the other genome initially using WUBLAST. This blast hit is then structurally aligned, using cmsearch (a program in the Infernal package) (Griffiths-Jones et al. 2003), to its consensus RNA structure, and the number of covariations between each pair of orthologous ncRNAs that are synteny-conserved are determined.

Program set 2

Name: Search and process synteny-conserved tRNA-gene Cluster
Description of function:
Search for synteny-conserved tRNA-gene clusters in the syntenic regions between two genomes, and examine the gene-order difference between two orthologous tRNA-gene clusters. For a tRNA-gene cluster in the human genome, this program set can search for its corresponding synteny-conserved clusters in other genomes in the syntenic regions defined by UBRHPs. A pair of orthologous tRNA-gene clusters are further analyzed by comparing the gene-order conservation between them.

Program set 3

Name: Align two ordered list of (tRNA-)gene symbols
Description of function:
Examine the gene-order conservation between two lists of (tRNA-)gene symbols. Using the dynamic programming library functions provided by biojava, this program set can align two lists of tRNA-gene symbols, which may be derived from a pair of syntenic regions from two genomes.

Program set 4

Name: RNA folding package
Description of function:
Predict the RNA secondary structure of a given sequence, and report the locations and sizes of stems and loops in this sequence. This program set provides an implementation of the Zuker's RNA secondary structure predicting algorithm. The thermodynamic parameters follow the ones used in (Zuker 1989). A set of adjunctive functions are provided in this program set, in order to facilitate the retrieval of local hairpins and the calculation of their thermodynamic stabilities.

Program set 5

Name: Eponine RNA motif extension, anchored
Description of function:
Prepare local hairpins and perform training of an Eponine anchored model which may consist of a set of RNA motifs. This program set provides a mechanism to extend Eponine anchored models to model RNA motifs. For each sequence recruited for training an Eponine anchored model, local RNA structures are predicted for each windowed region using Zuker's RNA secondary-structure predicting algorithm. Then SimpleStemLoopBasisSource uses the parameters of local hairpins as the basis to propose a new model. Other classes with the suffix BasisSource can optimize the parameters of a model using Monte Carlo sampling approaches. The parameters of an anchored model containing RNA motifs may consist of distributions of hairpin dimensions and/or stability and distance distributions between each motif and the anchored point of each sequence.

Program set 6

Name: Eponine RNA motif extension, unanchored
Description of function:
Prepare local hairpins and perform the learning of an Eponine unanchored model which may consist of a set of RNA motifs. This program set provides a mechanism to extend the Eponine unanchored models to model RNA motifs. For each sequence recruited for training an Eponine unanchored model, local RNA structures are predicted for each windowed region in this sequence using Zuker's algorithm. Then ConvolvedSensorsBasis uses the parameters of local hairpins as the basis to propose a new model. Other classes with the suffix BasisSource can optimize the parameters of a model by using Monte Carlo sampling approaches. The parameters of an unanchored model containing RNA motifs may consist of distributions of hairpin dimensions and/or stability and distance distributions between motifs.

Table B 2. Number of lines and file sizes of the program sets written for this thesis

Chap	Program set	Number of Lines	File size
2	Search-for-synteny-conserved-ncRNAs		
	● syntenic_proteins.pm	870	36k
	● protein_boundary.pm	320	9k
	● infernal.pm	192	5k
	● best_blast_hit.pm	103	6k
	● cmsearch_hit.pm	103	2k
	● paired_cmsearch_hit.pm	486	13k
	● other miscellaneous modules and scripts	1580	40k
2	Search-process-synteny-conserved-tRNACluster		
	● tRNAClusterDB.pm	130	3k
	● tRNASeqFasta.pm	321	3k
	● tRNAInfo.pm	112	2k
	● tRNAClusterDB_protein_boundary.pl	889	16k
	● other miscellaneous modules and scripts	274	8k
2, 3	Align two ordered list of (tRNA-)gene symbols		
	● AligntRNAName.java	511	15k
	● Other miscellaneous classes	209	7k
4, 6	RNA folding package		
	● Stem.java	32	1k
	● AbstractStem.java	52	2k
	● SimpleStem.java	204	5k
	● StemTools.java	93	3k
	● StrucTools.java	632	19k
	● StrucReport.java	183	4k
	● Pair.java	60	2k
	● Zuker.java	822	23k
4, 6	Eponine RNA motif extension, anchored		

	● AbstractStructureSampler.java	168	5k
	● SimpleStemLoopConstraint.java	793	23k
	● SimpleStemLoopBasisSource.java	348	9k
	● LocalEnergyDistBasisSource.java	59	2k
	● LocalEnergyOffsetBasisSource.java	59	2k
	● LoopSizeDistBasisSource.java	71	2k
	● LoopSizeOffsetBasisSource.java	71	2k
	● StemEnergyDistBasisSource.java	59	2k
	● StemEnergyOffsetBasisSource.java	59	2k
	● StemSizeDistBasisSource.java	69	2k
	● StemSizeOffsetBasisSource.java	71	2k
4, 6	Eponine RNA motif extension, unanchored		
	● AbstractStrucSampler.java	223	6k
	● ConvolvedSensorsBasis.java	798	24k
	● NewStruc1.java	380	10k
	● SampleLocalEnergyDist.java	65	2k
	● SampleLocalEnergyOffset.java	65	2k
	● SampleLoopSizeDist.java	71	2k
	● SampleLoopSizeOffset.java	76	2k
	● SampleStemEnergyDist.java	65	2k
	● SampleStemEnergyOffset.java	65	2k
	● SampleStemSizeDist.java	74	2k
	● SampleStemSizeOffset.java	75	2k