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.

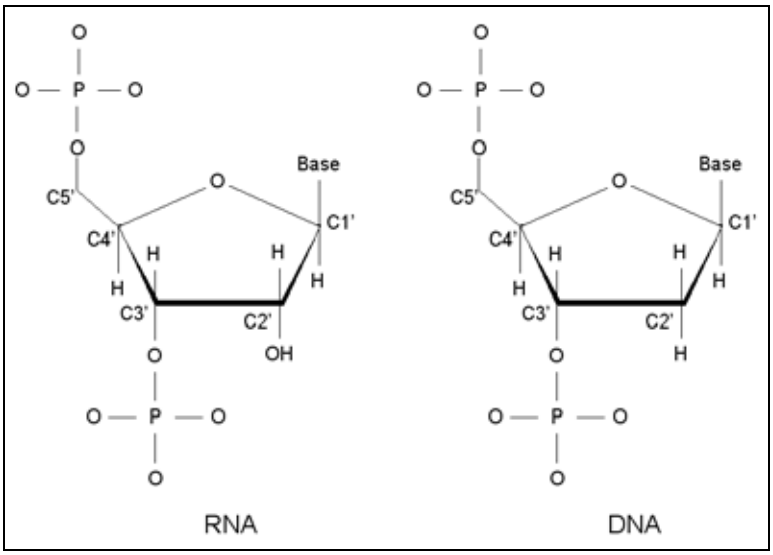

Figure 1-1. Organization of repeating units in RNA and DNA respectively.

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.

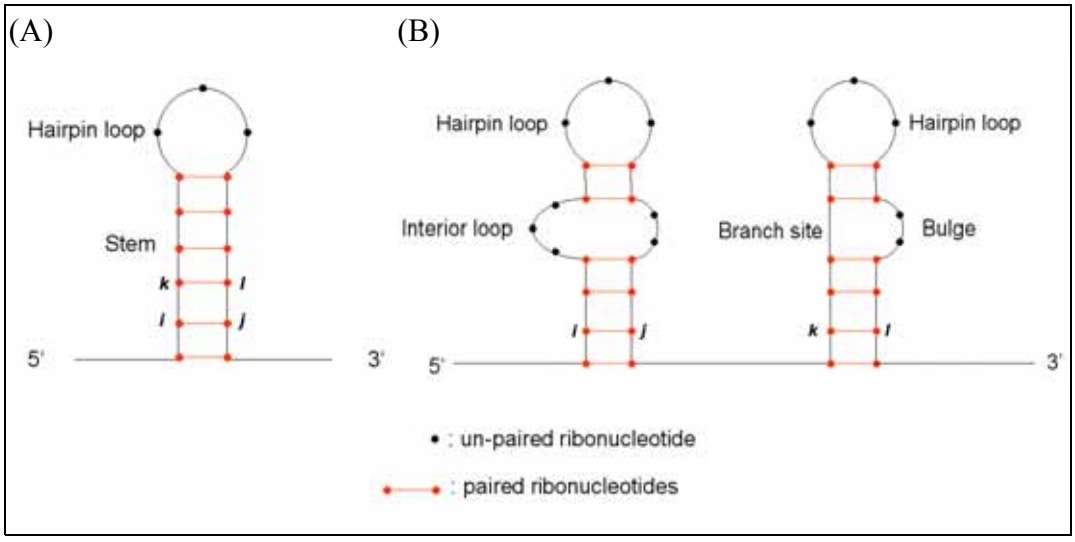

Figure 1-2. Elements of RNA secondary structures

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).

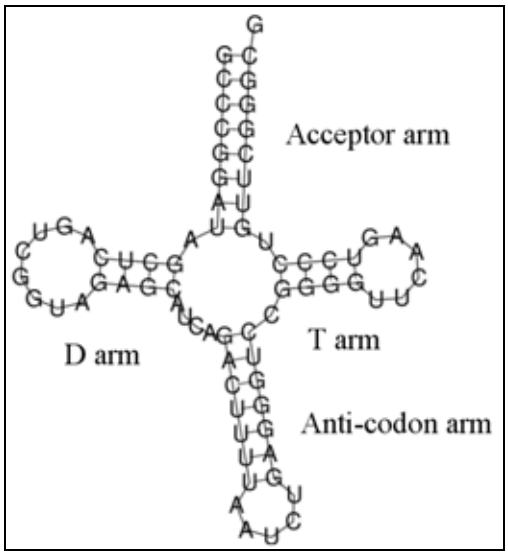

Figure 1-3. The cloverleaf-like secondary structure of a tRNA

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)

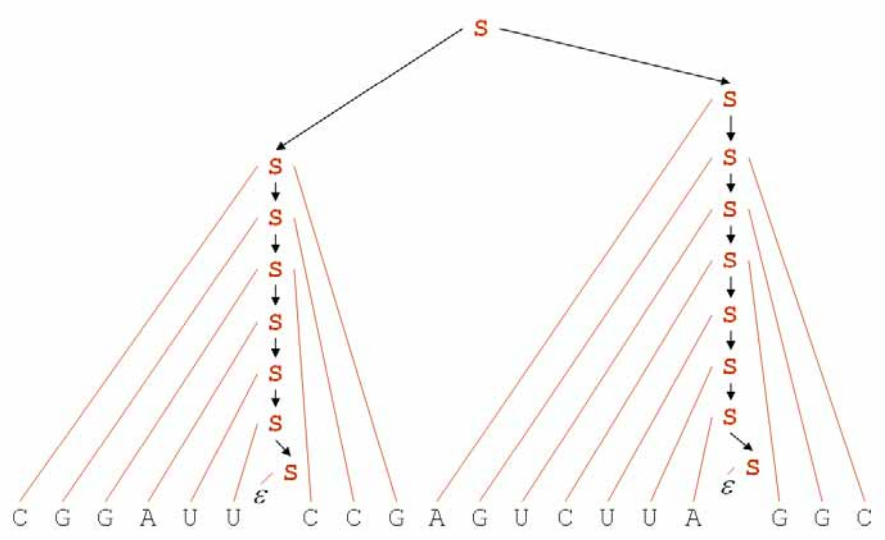

Figure 1-5 Two representations of the pairwise correlations in an RNA molecule with two non-interlaced hairpins

(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:

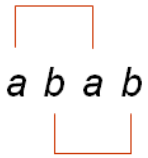

Figure 1-6. A crossing interaction that may be found in RNA tertiary structures

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.