

# UNEARTHING THE BURIED TREASURES - COMPUTATIONAL IDENTIFICATION AND ANALYSIS OF NONCODING RNAs

Byung-Jun Yoon, *Student Member, IEEE*, and P. P. Vaidyanathan, *Fellow, IEEE* <sup>\*†</sup>

Submitted to the IEEE Signal Processing Magazine - *Special Issue on Signal Processing Methods in Genomics and Proteomics*.

## 1 Introduction

The central dogma of molecular biology states that the genetic information flows from DNA to RNA to protein. This dogma has exerted a substantial influence on our understanding of the genetic activities in the cells. Under this influence, the prevailing assumption until the recent past was that genes are basically repositories for protein coding information, and proteins are responsible for most of the important biological functions in all cells. In the meanwhile, the importance of RNAs has remained rather obscure, and the RNA was mainly viewed as a passive intermediary that bridges the gap between DNA and protein. Except for classic examples such as *tRNAs* (transfer RNAs) and *rRNAs* (ribosomal RNAs), functional noncoding RNAs were considered to be rare.

However, this view has experienced a dramatic change during the last decade, as systematic screening of various genomes identified myriads of **noncoding RNAs (ncRNAs)**, which are RNA molecules that function without being translated into proteins [11, 40]. It has been realized that many ncRNAs play important roles in various biological processes. As RNAs can interact with other RNAs and DNAs in a sequence-specific manner, they are especially useful in tasks that require highly specific nucleotide recognition [11]. Good examples are the *miRNAs* (microRNAs) that regulate gene expression by targeting *mRNAs* (messenger RNAs) [4, 20], and the *siRNAs* (small interfering RNAs) that take part in the *RNAi* (RNA interference) pathways for gene silencing [29, 30]. Recent developments show that ncRNAs are extensively involved in many gene regulatory mechanisms [14, 17].

The roles of ncRNAs known to this day are truly diverse. These include transcription and translation control, chromosome replication, RNA processing and modification, and protein degradation and translocation [40], just to name a few. These days, it is even claimed that ncRNAs dominate the genomic output of the higher organisms such as mammals, and it is being suggested that the greater portion of their genome (which does not encode proteins) is dedicated to the control and regulation of cell development [27]. As more and more evidences pile up, greater attention is paid to ncRNAs, which have been neglected for a long time. Researchers began to realize that the vast majority of the genome that was regarded as “junk”, mainly because it was not well understood, may indeed hold

---

<sup>\*</sup>Both authors are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA.

<sup>†</sup>Work supported in parts by the NSF grant CCF-0428326 and the Microsoft Research Graduate Fellowship.

the key for the best kept secrets in life, such as the mechanism of alternative splicing, the control of epigenetic variations and so forth [27]. The complete range and extent of the role of ncRNAs are not so obvious at this point, but it is certain that a comprehensive understanding of cellular processes is not possible without understanding the functions of ncRNAs [47].

## 1.1 Finding ncRNAs

Although several systematic searches for ncRNAs in recent years have unveiled a large number of novel ncRNAs, it is believed that there are still numerous ncRNAs that are waiting to be discovered [11, 27, 40]. Typical estimates of the number of ncRNAs in the human genome are in the order of tens of thousands [27, 48], but the present genome annotation on ncRNAs is too incomplete to derive a more accurate estimate. Given the vast amount of genomic data that is currently available, it is practically impossible to identify all ncRNAs solely by experimental means. In order to expedite the annotation process, we desperately need the help of computational methods that can be used for identifying novel ncRNAs.

In this paper, we give a tutorial review of the various methods that can be used in the computational identification and analysis of ncRNAs. Most of all, we focus on statistical models that can be utilized for building probabilistic representations of RNA families. We review the main characteristics of these models and show how they can be used to identify new ncRNA genes, which are portions of DNA that give rise to ncRNA transcripts. The main emphasis of the discussion lies on methods for finding new members (or homologues) of known ncRNA families, but we also briefly mention about recent developments in techniques for finding novel ncRNAs at the end of the paper.

## 2 RNA Secondary Structure

Let us first consider the general characteristics of RNAs. The RNA is a nucleic acid that consists of a string of nucleotides (or bases), A, C, G and U, where uracil (U) is chemically similar to thymine (T) in the DNA. Different from DNAs, which exist in a double-stranded form, an RNA is generally a single-stranded molecule. The nucleotides A/U and C/G in an RNA molecule can form hydrogen bonded base-pairs, which are typically called **complementary base-pairs**<sup>1</sup>. If there exist complementary parts in a given RNA, these parts can form consecutive base-pairs, making the RNA fold onto itself. This complementary base-pairing determines the three-dimensional structure of the RNA to a considerable extent, and the two-dimensional structure resulting from the base-pairing is referred as the **RNA secondary structure**.

Fig. 1 shows two examples of RNA secondary structures. We can see that both RNAs display characteristic secondary structures after folding. As indicated in Fig. 1 (a), the consecutive base-pairs that are stacked onto each other after folding is called a **stem**, and the sequence of unpaired bases bounded by base-pairs is called a **loop**. The secondary structure of the RNA in Fig. 1 (a) consists of two **stem-loops** (or **hairpins**). In many cases, the base-pairings occur in a nested manner, where no

---

<sup>1</sup>Sometimes, the bases *G* and *U* can also form pairs.

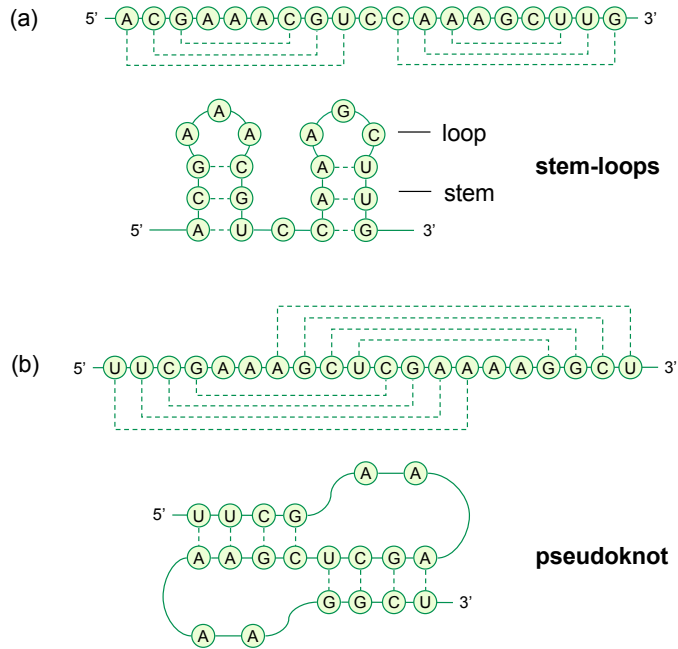


Figure 1: Two examples of RNAs with secondary structures. The primary sequence of each RNA is shown along with its structure after folding. The dashed lines indicate interactions between bases. (a) RNA with two stem-loops. (b) RNA with a pseudoknot.

interactions between bases cross each other. To be more precise, consider a base-pair between locations  $i$  and  $j$  ( $i < j$ ), and another base-pair between locations  $k$  and  $l$  ( $k < l$ ). We say that these two base-pairs are nested if they satisfy  $i < k < l < j$  or  $k < i < j < l$ . Secondary structures with crossing interactions, where there exist base-pairs at  $(i, j)$  and  $(k, l)$  that satisfy  $i < k < j < l$  or  $k < i < l < j$ , are called **pseudoknots**. One such example is shown in Fig. 1 (b). Although RNA pseudoknots are observed less frequently than secondary structures with only nested base-pairs, there are still many RNAs that are known to contain functionally important pseudoknots [42].

RNA secondary structures are known to play crucial roles in carrying out the functions of many ncRNAs. An intriguing example can be observed in **riboswitches**, which are regulatory RNA elements that have been recently found [26, 44]. Riboswitches are highly structured RNA domains that are found in the noncoding regions of various mRNAs. They make structural changes upon binding specific metabolites, thereby regulating the expression of the corresponding genes. Two common mechanisms of riboswitches in bacteria are illustrated in Fig. 2. The first mechanism works by *translation control* as shown in Fig. 2 (a). In the presence of the effector metabolite, the riboswitch changes its conformation by binding it. This structural change sequesters the *ribosome-binding site* (RBS), which prevents the ribosome from binding to the mRNA. The second mechanism is based on *transcription control*. In this case, the riboswitch forms a terminator stem upon binding the metabolite. This causes a premature termination of transcription, preventing the synthesis of the full-size mRNA. Riboswitches play pivotal roles in regulating several metabolic pathways, and they are prevalent in bacteria [26, 44]. Recent results show that similar metabolite-binding RNA domains are also present in eukaryotes (organisms

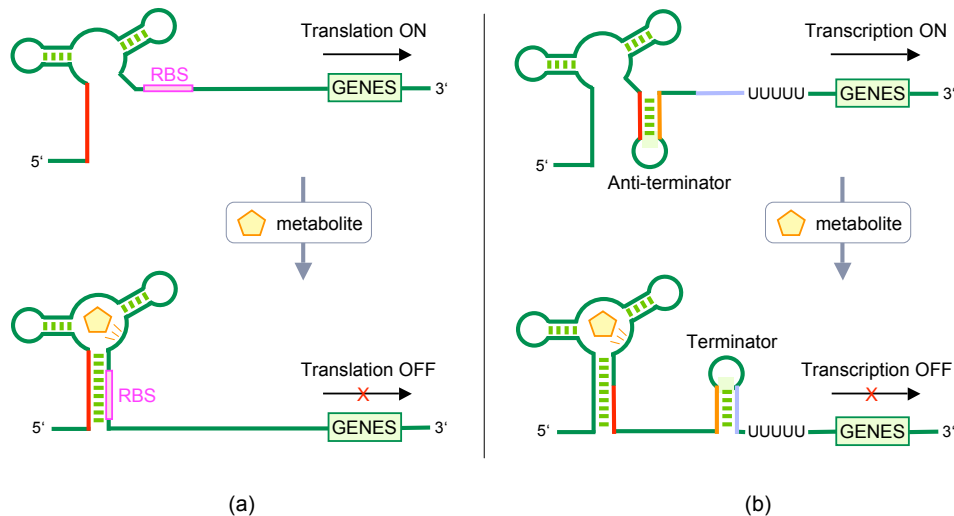


Figure 2: Two common mechanisms of riboswitches in bacteria. (a) **Translation control.** In the presence of the effector metabolite, the riboswitch changes its structure and sequesters the ribosome-binding site (RBS). This inhibits the translation initiation, thereby down-regulating the gene. (b) **Transcription control.** Upon binding the metabolite, the riboswitch forms a terminator stem, which prevents the generation of the full-size mRNA.

with cell nucleus) such as plants and fungi, although their gene-control mechanisms may be different from those in bacteria [41].

As we can see in this example, the structure of an RNA molecule is closely related to its function. For this reason, predicting the secondary structure of an RNA molecule based on its primary sequence has been of interest to many researchers. Since the RNA secondary structure is essentially governed by the base-pairing of nucleotides, many computational methods have been proposed for finding the “optimal base-pairing” of an RNA in an efficient manner. Such algorithms are typically called **RNA folding algorithms** [32, 37, 49, 54]. A good introduction to these algorithms can be found in [13].

### 3 Searching for Homologous RNAs

In biology, we say that two (or more) sequences are *homologous* if they are similar because of shared ancestry [5, 9]. Similar to protein-coding genes, ncRNA sequences can also be grouped into families of homologous sequences [18]. Sequences that belong to the same family often share a number of common statistical characteristics, although the reverse is not necessarily true. Given a new sequence, we can take advantage of these family-specific characteristics to determine whether it belongs to a specific sequence family. Its membership in a certain family can often be used to infer the function of the sequence.

In fact, many computational methods for biological sequence analysis make use of the above idea in one way or another [9], especially those used for gene identification. Suppose we have a set of related sequences that belong to the same family (e.g. tRNAs). Based on these sequences, we can extract the common features of the sequence family, and use them to search the database in order to find new

sequences (novel tRNAs) that share these features. Such computational screening may identify new members of a known sequence family, in a fast and efficient manner. This approach is typically called **homology search** (or **similarity search**).

### 3.1 Sequence-Based Homology Search

Most of the search methods that have been used for finding homologous protein-coding genes have been based on *sequence similarity*. Popular search algorithms such as **BLAST (Basic Local Alignment Search Tool)** [1] and **FASTA** [33] use known members in a sequence family to look for high-scoring local alignments in the target database. Another approach picks up common “patterns” or “motifs” in a set of related sequences and searches the database for regions that match these patterns. One example of such an approach is the **PROSITE database** [3], which has compiled biologically significant patterns of protein families. A more general approach would be to build a probabilistic representation of an entire sequence family and employ it in the search. One of the most popular models for constructing such a representation is the **profile-HMM (profile hidden Markov model)** [9, 22], which is an HMM with a linear structure that repetitively use a set of three states (match, insert, delete). As profile-HMMs can effectively describe distinct symbol probabilities at different locations and easily deal with additional insertions and deletions at any location, they have been widely used in several applications such as protein-coding gene-identification [23] and sequence alignment [9].

### 3.2 Statistical Model for RNA Sequences

The sequence-based methods described in the previous section (BLAST, FASTA, PROSITE, profile-HMM) are very useful for identifying homologous DNAs and proteins, but they often behave poorly when applied to RNA homology search. The main reason is the following. Many functional ncRNAs preserve their secondary structures more than they preserve their primary sequences [9]. Sometimes, these base-paired structures are still preserved among related RNAs, even when their similarity in the primary sequence level can be hardly recognized. Therefore, when evaluating the similarity between two RNA molecules, it is important to take both their primary sequences and their secondary structures into consideration.

As observed by Eddy in [12], this combined scoring scheme is much more effective in comparing (and also aligning) RNA sequences, and it can greatly enhance the discriminative power of an RNA homology search. This can be clearly seen from the example illustrated in Fig. 3. In this example, we have a query sequence that has a stem-loop structure. Let us perform ungapped pairwise alignments between the query sequence and each of the RNAs shown in Fig. 3 (b) and Fig. 3 (c). Both RNA-1 and RNA-2 differ from the query sequence RNA-0 at four locations. As the four mismatches (or “base substitutions”) in both alignments are identical, the primary sequence alignment score for RNA-1 and RNA-0 will be exactly the same as the alignment score for RNA-2 and RNA-0. However, we can see in Fig. 3 (b) and Fig. 3 (c) that RNA-1 preserves the secondary structure of the original query sequence, while RNA-2 does not. Apparently, RNA-1 is a better match to the query RNA-0, and therefore we should give it a higher score than RNA-2.

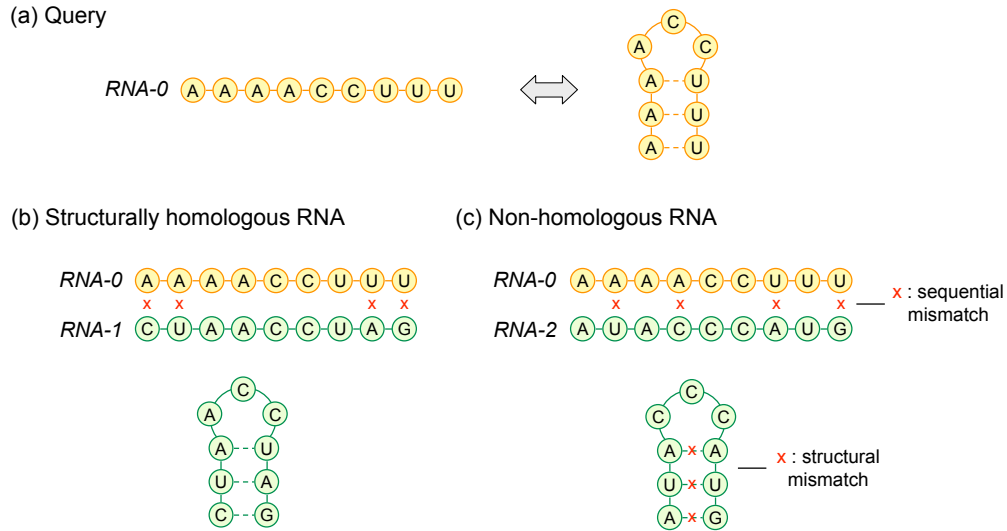


Figure 3: Ungapped alignment between two RNA sequences. (a) An RNA with a stem-loop structure is used as the query sequence. (b) A structurally homologous RNA that has also a stem-loop structure. (c) A structurally non-homologous RNA that does not fold to a stem-loop structure.

As this example shows, when computing a similarity measure between RNAs, it is important to consider their resemblance in the structural level as well as in the sequence level. Now, the question is how to combine the contributions from the sequence similarity and the structural similarity in a reasonable way. To answer this question, let us examine the effect of a conserved RNA secondary structure on its primary sequence. RNA sequences often undergo **compensatory mutations** in order to preserve their secondary structures. For a given base-pair in an RNA molecule, if the base in one side is changed to another base, the base in the other side is also changed such that the base-pair is still maintained. As a result, we can observe strong correlation between the two base positions in homologous RNAs as illustrated in Fig. 4. From this point of view, we can understand base-pairing in an RNA secondary structure in terms of pairwise correlations between distant bases in the primary sequence of the RNA. This shows that in order to model RNAs with conserved secondary structures, we need a statistical model which can describe such pairwise correlations. However, most statistical models that have been used for analyzing DNAs and proteins (including profile-HMMs) do not have

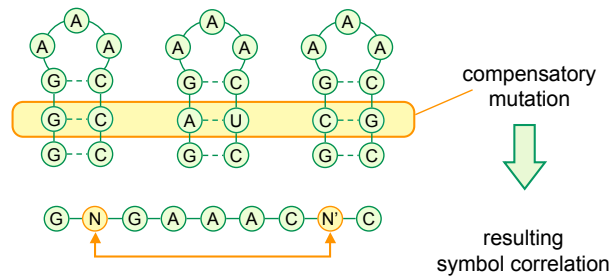


Figure 4: Compensatory mutations give rise to strong pairwise correlations in the primary sequence of an RNA.

### [Box-I] TRANSFORMATIONAL GRAMMARS

In computational linguistics, a **transformational grammar** is defined as a set of rules that can be used to describe (or generate) a set of symbol sequences over a given alphabet [6]. A transformational grammar can be characterized by the following components: **terminal symbols**, **nonterminal symbols**, and **production rules**. Terminal symbols are observable symbols that appear in the final symbol sequence, and nonterminal symbols are abstract symbols that are used to define the production rules. A production rule is defined as  $\alpha \rightarrow \beta$ , where  $\alpha$  and  $\beta$  are strings of terminal and/or nonterminal symbols, and it describes how a given string can be transformed into another string. We can generate various symbol sequences by applying these production rules repetitively. Chomsky categorized transformational grammars into four classes [6]. These are the **regular grammars**, **context-free grammars**, **context-sensitive grammars** and **unrestricted grammars**, in the order of decreasing restrictions on the production rules. These four classes comprise the so-called **Chomsky hierarchy of transformational grammars**. For further details on this topic, refer to texts on formal language theory such as [19].

the descriptive power to deal with such complex base correlations.

RNA sequences with secondary structures can be viewed as a kind of biological **palindromes**. Palindromes are symmetric sequences that read the same forwards and backwards, such as “I prefer pi”, “step on no pets”, and so on. Similarly, the base-pairing in an RNA secondary structure gives rise to symmetric (or *reverse complementary*, to be more precise) regions in its primary sequence that are analogous to palindromes. According to the **Chomsky hierarchy of transformational grammars** [6] (see Box-I for a brief introduction to transformational grammars), HMMs can be viewed as **stochastic regular grammars**. Regular grammars are the simplest among the four classes in the hierarchy, and it is known that they are inherently incapable of describing a palindromic language. It is of course possible that a regular grammar generates a palindrome as part of its language, but the point is that it is not capable of generating *only* such palindromes. Therefore, regular grammars cannot effectively discriminate palindromic sequences from non-palindromic ones, making them unsuitable for constructing RNA profiles.

In order to represent complex correlations that are frequently observed in ncRNA sequences, we need more complex models with larger descriptive power than the regular grammars. In the following sections, we review two statistical models - stochastic context-free grammars and profile context-sensitive HMMs - that are capable of describing such correlations. These models can be effectively used for building representations of RNA sequence families and performing RNA homology search.

## 4 Stochastic Context-Free Grammars and Covariance Models

Regular grammars allow only left-emissions of symbols, generating sequences left-to-right. However, context-free grammars (CFGs) incorporate additional production rules that allow pairwise-emissions, where one symbol is emitted to the left and the other symbol is emitted to the right. Thanks to these additional rules, CFGs become capable of describing sequences with nested correlations.

By using CFGs, we can easily write grammars that model RNA secondary structures. For example,

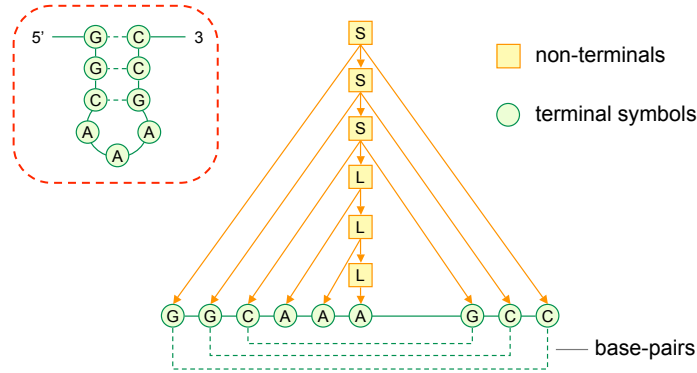


Figure 5: The parse-tree shows how the RNA sequence “GGCAAAGCC” can be generated from the given context-free grammar. It gives a graphical representation of the generation process ‘ $S \rightarrow gSc \rightarrow ggLgcc \rightarrow \dots \rightarrow ggcaaagcc$ ’.

the following grammar<sup>2</sup> can generate a RNA stem-loop with any number of base-pairs and a variable length loop. (The notation ‘|’ means ‘or’.)

$$\begin{aligned}
 S &\longrightarrow aSu \mid uSa \mid cSg \mid gSc \mid aLu \mid uLa \mid cLg \mid gLc \\
 L &\longrightarrow aL \mid cL \mid gL \mid uL \mid a \mid c \mid g \mid u
 \end{aligned}$$

The generation of a symbol string by a CFG can be conveniently expressed using a tree-structured graph, called a **parse-tree**. An example of such a parse-tree is given in Fig. 5, which shows how the RNA sequence “GGCAAAGCC” can be generated from the above grammar. It gives a graphical representation of the process

$$S \rightarrow gSc \rightarrow ggLgcc \rightarrow ggcaLgcc \rightarrow ggcaaLgcc \rightarrow ggcaaagcc,$$

which shows how the production rules are applied.

In fact, a large class of RNA secondary structures can be effectively modeled using CFGs, making them an attractive choice for constructing probabilistic profiles of RNA families. For this reason, there have been several attempts to use **stochastic context-free grammars (SCFGs)** in RNA sequence analysis [10, 39]. For example, the **covariance model (CMs)** [10] is the SCFG-analogue of the profile-HMM, which is suitable for modeling consensus RNA sequences from multiple sequence alignments. As a profile-HMM is obtained by using a set of three states (match, insert, delete) for each position in the multiple alignment and interconnecting them, a CM is obtained by constructing a tree-like directed-graph of states by repetitively using the basic building blocks called *CM nodes*. Each node can be viewed as a “super-state” that consists of one or more states, where the number of states depends on the type of the node. A typical CM has the following kinds of nodes: **S** (start of a new tree), **P** (pairwise-emission), **L** (left-emission), **R** (right-emission), **B** (bifurcation), **E** (end node). Each of these nodes can deal with a match, a deletion, and additional insertions at the given location using a combination of match, insert, and delete states.

<sup>2</sup>Note that the bases  $A, C, G, U$  - which are terminal symbols in this case - are written in lower case letters to differentiate them from other nonterminal symbols.

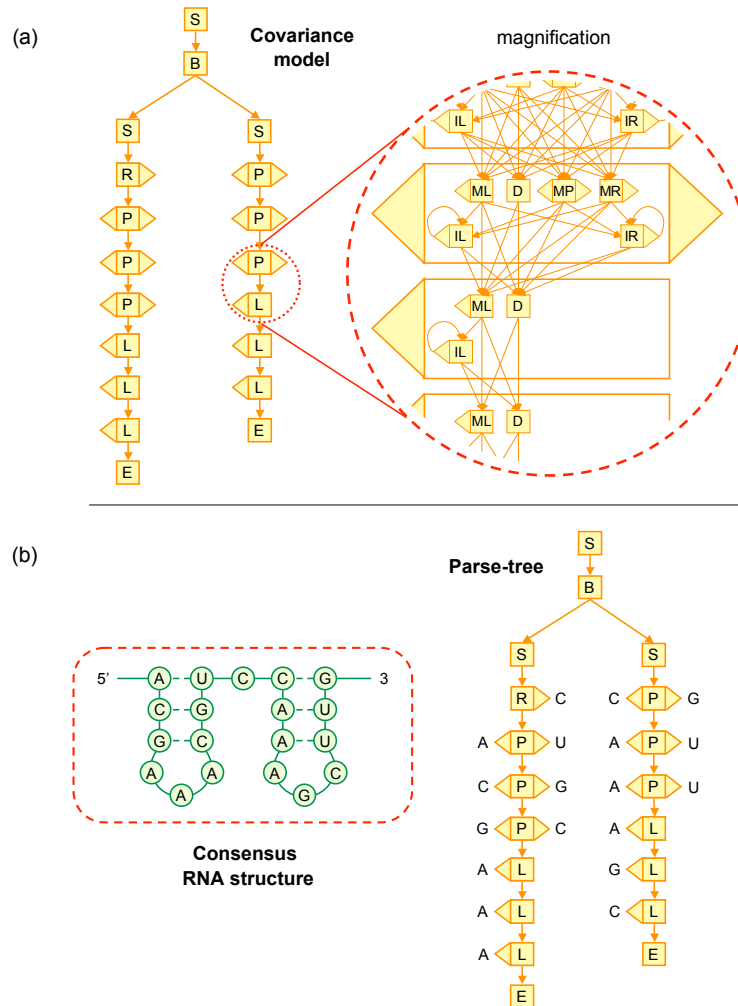


Figure 6: (a) A covariance model that represents the consensus RNA structure shown in Fig. 1 (a). Each node consists of a number states. For example, the P-node has six states and the L-node consists of three states, as shown in the magnified figure on the right. (b) The RNA sequence shown inside the box is aligned to the CM, showing a parse-tree.

Constructing a CM based on a multiple alignment is rather straightforward. We first predict the consensus secondary structure by finding the base-pairs through identifying the covarying columns in the alignment. Once the consensus structure is found, we build the corresponding consensus RNA structure tree, which looks similar to a parse-tree of the consensus RNA sequence. Then we replace each node in the constructed tree by one of the CM nodes (S, P, L, R, B, E) to obtain the final model.

An example of such a CM is shown in Fig. 6 (a), where the model is constructed from the consensus RNA structure illustrated in Fig. 1 (a). As we can see in the magnified figure on the right, each CM node consists of several states. For example, the P-node consists of six states - MP (match-pair), ML (match-left), MR (match-right), D (delete), IL (insert-left), IR (insert-right) - and the L-node consists of three states. The box in Fig. 6 (b) shows an example of an RNA sequence folded to its consensus secondary structure. This RNA sequence can be aligned to the given CM, resulting in a parse-tree shown on the right of Fig. 6 (b). Similarly, we can align every sequence in the multiple alignment to the CM

and count the emission and transition events at each CM state to estimate the emission and transition probabilities. We can simply use the relative frequencies, or use these frequencies as the initial seed and run an EM (expectation-maximization) algorithm called the **inside-outside algorithm** [24] to optimize the model parameters. Further details on CMs can be found in [9].

CMs obtained in this manner can be used for finding homologous RNAs in a database. When CMs were first proposed, they were applied to the prediction of tRNA genes [10] and achieved an impressive 99.8% overall sensitivity<sup>3</sup> at a relatively low false positive rate of  $< 0.002$  per Mb (megabase; 1 million nucleotides) [25]. For finding the best alignment between the CM and an RNA sequence, they used a variant of the **Cocke-Younger-Kasami (CYK) algorithm** [19, 24] which is the SCFG-analogue of the Viterbi algorithm. One major problem of a CM-based search is the slow scanning speed due to the high computational complexity of the CYK algorithm. The time-complexity of a general CYK algorithm is  $O(L^3M^3)$ , where  $M$  is the number states and  $L$  is the length of the target sequence. For more restricted SCFGs such as CMs, the complexity decreases to  $O(L^3M)$  [9], but it is still much slower than the Viterbi algorithm. For this reason, it is sometimes advantageous to use a hybrid approach to speed up the search. A later version of the tRNA-prediction algorithm called the *TRNASCAN-SE* [25] combines other prediction algorithms with the CM-based approach, where the simpler algorithms are used as pre-filters. This hybrid method has a comparable sensitivity (99.5%) and a much lower false positive rate ( $< 0.00007$ ), while running nearly 1,500 times faster than the original program that is fully based on a CM [25].

There exists also a BLAST-like search tool that uses only a single RNA sequence and its secondary structure to look for homologues [21]. It is shown to outperform programs that use only the primary sequence information, but its computational cost is too high to be used in practice, unless a clustered computing environment is available.

## 5 Profile Context-Sensitive HMMs

Although the SCFG-based models can be used for modeling various RNAs, their descriptive power is limited to nested correlations, hence they are not capable of dealing with RNA pseudoknots. As we can see from the example shown in Fig. 1 (b), RNA pseudoknots have crossing dependencies between bases, and in order to model such dependencies we have to resort to more complex models such as the **context-sensitive grammars (CSGs)**. However, parsing a general CSG is an NP-complete problem [16], hence computationally intractable. For this reason, several different subclasses of CSGs have been proposed [28, 34], which have the descriptive power of modeling most RNA pseudoknots and computationally tractable at the same time. The grammar proposed by Rivas and Eddy [34] incorporates several symbol rearranging rules to obtain crossing interactions in the final symbol sequence, and the method proposed by Matsui et al. [28] uses *tree adjoining grammars (TAGs)* for modeling pseudoknots. Both models can deal with large classes of correlations that include most of the known pseudoknots,

---

<sup>3</sup>Two metrics called **sensitivity (SN)** and **specificity (SP)** are frequently used to evaluate the performance of a gene finder. They are defined as  $SN = TP/(TP+FN)$  and  $SP = TP/(TP+FP)$ , where TP is the number of true-positives, FN is the number of false-negatives, and FP is the number of false-positives.

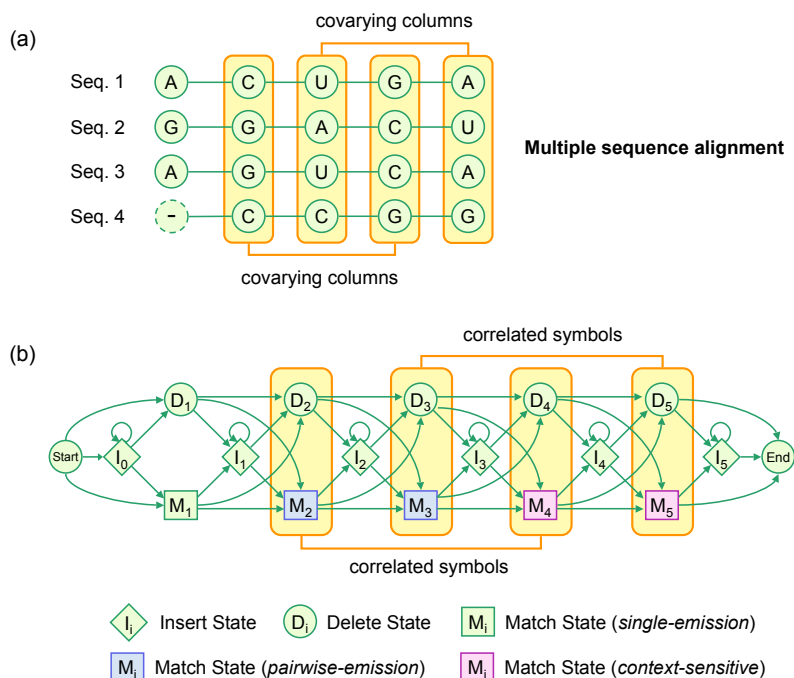


Figure 7: Constructing a profile-csHMM. (a) We first form a multiple sequence alignment of related RNAs. The base-pairs in the consensus secondary structure can be predicted from the covarying columns. (b) The corresponding profile-csHMM is constructed from the multiple alignment.

but neither model can represent all of them.

Instead of using CSGs, we can use **context-sensitive HMMs (csHMMs)** that have been recently proposed [50, 51] (see Box-II). The csHMMs are extensions of traditional HMMs that are capable of modeling *any* kind of pairwise correlations between distant symbols (including crossing correlations). **Profile context-sensitive HMMs (profile-csHMMs)** [53], which are specifically structured csHMMs with a repetitive structure, can be especially useful in modeling RNA profiles. The basic structure of a profile-csHMM is quite similar to that of a profile-HMM. It repetitively uses a set of match, insert, and delete states to model each position in the multiple alignment. However, unlike profile-HMMs, there can be three different kinds of match states depending on the type of correlation at the base position that is being modeled. If the base position is not involved in base-pairing, we use a single-emission state for the match state at the given position. For two positions that form a base-pair, we use a pairwise-emission state in the front and the corresponding context-sensitive state in the rear position, to model the correlation between these positions. Note that additional bases that are inserted to the alignment do not have an explicit correlation with others, hence single-emission states are used for insert states. Delete states are non-emitting states as in the traditional profile-HMMs. Fig. 7 shows a simple example that demonstrates how a profile-csHMM can be constructed from an RNA multiple sequence alignment. We can see in Fig. 7 (a) that the first position is not correlated to any other position, hence the match state  $M_1$  uses a single-emission state. The second position and the fourth positions are correlated, so we use a pairwise-emission state at  $M_2$  and the corresponding context-sensitive state at  $M_4$ . Similarly, a pairwise-emission state is used at  $M_3$  with the corresponding context-sensitive state

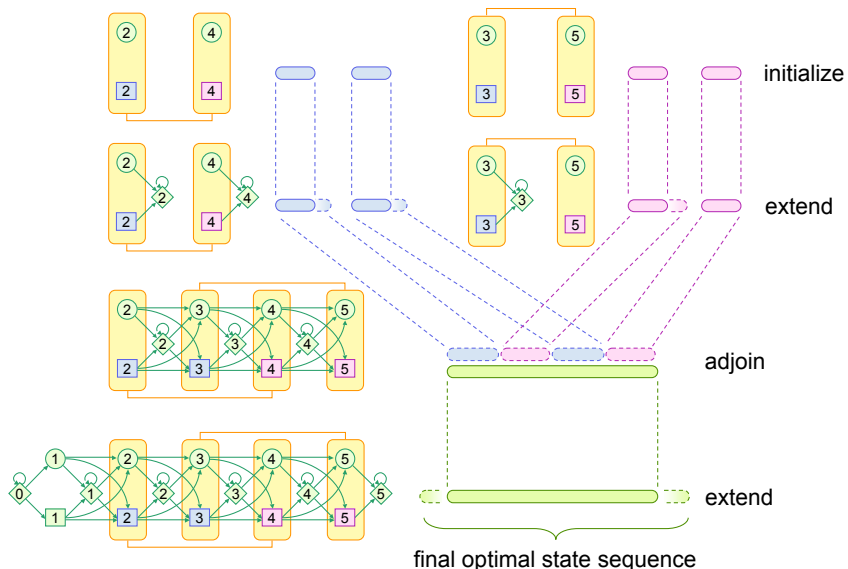


Figure 8: Illustration of the sequential component adjoining (SCA) algorithm. The optimal state sequence of longer subsequences can be found by extending and adjoining those of shorter subsequences.

at  $M_5$ . Once the model has been obtained, we can estimate its model parameters in a similar way as we estimate the parameters of a CM.

As we can see from the example illustrated in Fig. 7, the profile-csHMM provides a simple and intuitive method for constructing an RNA profile from a multiple sequence alignment. Moreover, it can represent *any* kind of pairwise dependencies between distant symbols, hence capable of dealing with all kinds of RNA pseudoknots. However, in order to use profile-csHMMs in practical applications, we need an efficient algorithm for finding the best alignment between the model and an observation sequence in a systematic way.<sup>4</sup> In fact, we can use the **sequential component adjoining (SCA) algorithm** [53], which can be viewed as a generalization of the Viterbi algorithm and the CYK algorithm. The basic philosophy underlying the SCA algorithm is similar to that of other dynamic programming algorithms; it first finds the optimal alignment for short subsequences, and uses this information to find the optimal alignment of longer subsequences. By iterating this process, we can ultimately find the global optimal alignment. Nevertheless, there are two main differences between the SCA algorithm and other algorithms such as the Viterbi algorithm and the CYK algorithm. In the first place, instead of using a fixed number of indices to designate the intermediate subsequences, the SCA algorithm uses a set of variable number of closed intervals to signify a subsequence. This significantly increases the number of ways in which the intermediate subsequences can be defined and extended. In the second place, the SCA algorithm extends and adjoins the optimal alignments of shorter subsequences according to a model-specific order. Note that in the Viterbi algorithm, the optimal subsequences were extended from left to right, and in the CYK algorithm they were extended from the inside to the outward direction. However, in the SCA algorithm, we define this extension/adjoining order in a

<sup>4</sup>This is equivalent to finding the optimal state sequence.

Model	Representable correlations			Representable sequences	Computational complexity
	Linear	Nested	Crossing		
Profile-HMM	O	X	X	coding-genes, proteins	$O(LM^2)$
Covariance Model	O	O	X	RNAs (no pseudoknots)	$O(L^3M)$
Profile-csHMM	O	O	O	RNAs (including pseudoknots)	variable

Table 1: Comparison between statistical models.

model-dependent manner such that all the correlations in the profile-csHMM are taken care of. Fig. 8 illustrates one possible way of obtaining the final optimal state sequence of a given observation sequence, based on the profile-csHMM shown in Fig. 7 (b). The overall computational complexity of the SCA algorithm depends on the specific correlation structure of the profile-csHMM. For sequential(linear) correlations (as in traditional HMMs), the complexity will be the same as that of the Viterbi algorithm, and for nested correlations (as in SCFGs), it will be identical to the complexity of the CYK algorithm. Table 1 compares the profile-csHMM with other statistical models that have been discussed so far.

The profile-csHMM is a relatively recent development that can provide an effective framework for constructing profiles for RNA families (including RNA pseudoknots) and building computational RNA analysis tools [53]. It opens up a lot of interesting theoretical issues as well as many possible applications in RNA analysis, including the prediction and alignment of RNA pseudoknots.

## 6 Beyond Homology Search: Identifying Novel ncRNAs

RNA homology search based on CMs or profile-csHMMs can be highly useful for predicting homologous ncRNA genes in genome sequences. However, these models are family-specific and they can be used only for searching homologues of known RNAs. Building a general purpose gene finder for predicting novel ncRNA genes is a much more challenging task.

Until now, various signal processing techniques have been applied to the prediction of protein-coding genes, which include DFT [2, 43], digital filters [45, 46], hidden Markov models (HMMs) [22] and many others. Among them, HMM-based methods have been especially successful. State-of-the-art gene finders (primarily based on HMMs) boast high prediction ratios that are far above 90%, achieving nearly perfect prediction results in simple organisms such as bacteria and yeast. However, these methods are not suitable for predicting ncRNA genes due to the following reasons. First of all, many ncRNAs lack the various statistical cues that have been used for identifying protein-coding genes. Unlike coding genes, their primary sequences do not display strong composition bias with strength comparable to the codon bias<sup>5</sup> in protein-coding genes [12]. They do not have open reading frames (ORFs)<sup>6</sup> that were effectively used in coding-gene finders [31]. Moreover, many ncRNAs are consider-

<sup>5</sup>A codon is a tri-nucleotide unit that codes for a single amino acid. Nonuniform usage of codons can give rise to a strong period-3 property in a DNA sequence.

<sup>6</sup>An ORF is any sequence of DNA that can potentially encode a protein. It starts with a start codon and ends with a stop codon [5]. Usually, the existence of a long ORF is a reasonable indication of the presence of a protein-coding gene.

**[Box-II] CONTEXT-SENSITIVE HMM**

The **context-sensitive HMM (csHMM)** can be viewed as an extension of the traditional HMM, where some states have variable emission and transition probabilities that depend on the “context” [50]. Such context-dependency can be quite effective in modeling certain types of correlations, and similar extensions have been previously proposed for different applications. (For example, see [15] for a related model that was used in image compression.) The csHMM has three different classes of hidden states, namely, **single-emission states  $S_n$** , **pairwise-emission states  $P_n$** , and **context-sensitive states  $C_n$** . Single-emission states  $S_n$  are identical to the regular states in traditional HMMs. Pairwise-emission states  $P_n$  are similar to single-emission states except that the symbols emitted at  $P_n$  are stored in the associated auxiliary memory  $Z_n$ , which can be a stack or a queue. Context-sensitive states  $C_n$  are fundamentally different from the others, in the sense that their probabilities are not fixed, but depend on the context. When we enter  $C_n$ , it first accesses the memory  $Z_n$  and retrieves a symbol  $x$ . (Note that this symbol was previously emitted at the corresponding pairwise-emission state  $P_n$ .) Once the symbol is retrieved, the emission probabilities of  $C_n$  are adjusted according to the value of  $x$ . For example, we can adjust the probabilities so that  $C_n$  emits the same symbol  $x$  with high probability (possibly, with probability one). The transition probabilities at  $C_n$  are also variable and they depend on whether the memory  $Z_n$  is empty or not. This context-sensitive property increases the descriptive power of the HMM significantly, and the csHMMs are capable of modeling various pairwise symbol correlations including crossing correlations.

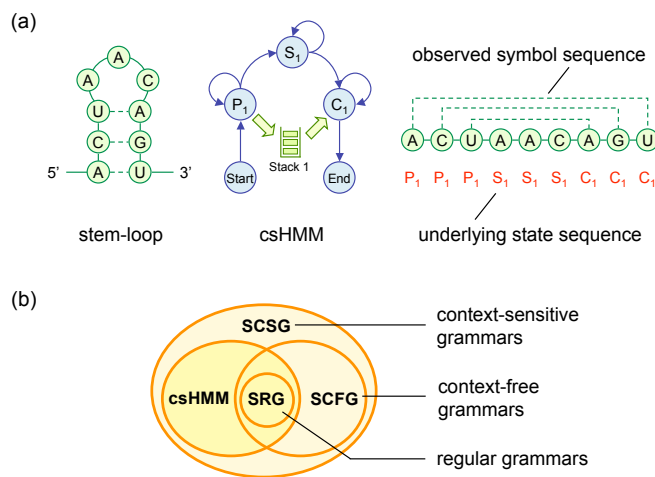


Figure 9: (a) An example of a simple csHMM that models a stem-loop. (b) The Venn-diagram shows the location of the csHMM in the Chomsky hierarchy.

The example in Fig. 9 (a) shows a simple csHMM that can model stem-loops. The single-emission state  $S_1$  generates the loop part, where the bases are not correlated to others. The states  $P_1$  and  $C_1$  together generate the stem part. Firstly, the bases generated by  $P_1$  are pushed onto the stack. Secondly, when we enter  $C_1$ , it pops the base on the top of the stack and the emission probabilities of  $C_1$  are adjusted such that it emits the complementary base. The transition probabilities of  $C_1$  are set so that it makes self-transitions until the stack becomes empty. In this way, we can always generate sequences with stem-loops. The Venn-diagram in Fig. 9 (b) shows where the csHMM is located in the Chomsky hierarchy. As we can see, the csHMM fully includes stochastic regular grammars (SRGs), and it is a proper subset of stochastic context-sensitive grammars (SCSGs). The csHMM has a significant overlap with stochastic context-free grammars (SCFGs), but neither of them fully contains the other. However, csHMMs are capable of modeling various crossing dependencies, which cannot be done using SCFGs. An in-depth introduction to csHMMs can be found in [52].

ably shorter than coding-genes, where a typical ncRNA has less than a few hundred nucleotides [18]. (An extreme example is the miRNA which has only about 21-25 nucleotides, in general [38].) This makes it difficult to judge whether the statistical property inside the ncRNA genes is different from that of the rest in a statistically meaningful manner.

Although traditional protein-coding gene finders cannot be directly used for identifying novel ncRNA genes, we can utilize the native characteristics of RNAs for building ncRNA gene finders. For example, as many ncRNAs have well-conserved secondary structures, we can exploit this property for finding ncRNA genes. However, an RNA sequence can have a large number of thermodynamically plausible secondary structures that have no biological significance [13]. In fact, it has been realized that the existence of a plausible secondary structure is not a sufficient evidence for detecting ncRNAs [35]. What is more important is whether the given secondary structure is preserved across different species, which can serve as a compelling evidence of its biological significance. For this reason, most ncRNA gene-prediction algorithms take advantage of multiple sequence data for finding novel ncRNAs [7, 8, 36, 47].

A common strategy of many general purpose ncRNA gene finders - such as **QRNA** [35], **ddbRNA** [8], **MSARI** [7], and **RNAz** [47] - can be summarized as follows [31]. They first look for regions in genome sequences that are conserved across different species, and form a multiple sequence alignment between these regions. Based on the alignment, they investigate whether there exists a common secondary structure that is preserved in all sequences. This information is used to decide whether these regions correspond to a functional ncRNA or not. Some of these algorithms have been used for screening the genomes of several organisms, and the detection results indicate that the aforementioned strategy is indeed quite effective. For example, RNAz - which is the current state-of-the-art algorithm for predicting novel ncRNAs - achieves an average sensitivity of 84.17% at 96.42% specificity, and 75.27% sensitivity at 98.93% specificity [47]. Recently, RNAz has been used to perform a comparative screening of several vertebrate genomes, and it predicted more than 30,000 putative ncRNA genes in the human genome [48]. Among them, almost a thousand ncRNA genes were conserved in all four vertebrate genomes included in the screening, which strongly suggests that these ncRNAs are biologically functional.

Despite the initial success of these ncRNA gene finders, there is yet a large room for improvement. In fact, the average prediction ratios of the existing algorithms are not as high as one might hope, and they still do not work well for certain classes of RNAs.<sup>7</sup> However, the performance of ncRNA gene finders has been improving at a fast pace, and it is clear that computational gene finders will play important roles in unveiling more and more novel ncRNAs in the future.

## 7 Conclusions

Unlike protein-coding genes, ncRNA genes have remained unnoticed until relatively recently. Compared to the annotation of protein-coding genes, which is nearly complete in many genomes that have

---

<sup>7</sup>For example, the sensitivity of RNAz for U70 snoRNAs (small nucleolar RNAs) is below 62%, and for tmRNAs (transfer-messenger RNA) it is below 25% [48].

been sequenced so far, the annotation of ncRNA genes have just begun. At present, it is even difficult to give a reliable estimate of the total number of ncRNAs in a genome. Given the enormous amount of genomic data, which is still increasing nowadays, we cannot stress strongly enough the importance of computational methods in finding ncRNA genes and analyzing them. Interestingly enough, many methods that are widely used in RNA sequence analysis have been already extensively used in the signal-processing community for a long time. For example, SCFGs that are frequently used for constructing RNA-profiles were originally used in speech recognition and natural language processing [24]. Moreover, profile-HMMs and profile-csHMMs are variants of traditional HMMs that have been also extensively used in speech and audio processing. The emerging field of computational RNA sequence analysis poses plenty of interesting questions to researchers across diverse areas, and we believe that the signal processing community can make a meaningful contribution to the advancement of this field.

## Acknowledgments

We would like to thank the anonymous reviewers for their insightful remarks and valuable suggestions, which were very helpful in improving the paper.

## References

- [1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, "Basic local alignment search tool", *Journal of Molecular Biology*, vol. 215, pp. 403-410, 1990.
- [2] D. Anastassiou, "Genomic signal processing", *IEEE Signal Processing Magazine*, pp. 8-20, July 2001.
- [3] A. Bairoch and P. Bucher, "PROSITE: recent developments", *Nucleic Acids Research*, vol. 22, pp. 3583-3589, 1994.
- [4] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function", *Cell*, vol. 116, pp. 281-297, 2004.
- [5] T. A. Brown, *Genomes*, John Wiley & Sons Inc., NY, USA, 2002.
- [6] N. Chomsky, "On certain formal properties of grammars", *Information and Control*, vol. 2, pp. 137-167, 1959.
- [7] A. Coventri, D. J. Kleitman, and B. Berger, "MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure", *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 12102-12107, 2004.
- [8] D. di Bernardo, T. Down, and T. Hubbard, "ddbRNA: detection of conserved secondary structures in multiple alignments", *Bioinformatics*, vol. 19, pp. 1606-1611, 2003.
- [9] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, Cambridge Univ. Press, Cambridge, UK, 1998.
- [10] S. R. Eddy and R. Durbin, "RNA sequence analysis using covariance models", *Nucleic Acids Research*, vol. 22, pp. 2079-2088, 1994.
- [11] S. R. Eddy, "Non-coding RNA genes and the modern RNA world", *Nature Reviews Genetics*, vol. 2, pp. 919-929, 2001.
- [12] S. R. Eddy, "Computational genomics of noncoding RNA genes", *Cell*, vol. 109, pp. 137-40, 2002.

- [13] S. R. Eddy, "How do RNA folding algorithms work?", *Nature Biotechnology*, vol. 22, pp. 1457-1458, 2004.
- [14] V. A. Erdmann, M. Z. Barciszewska, M. Szymanski, A. Hochberg, N. de Groot, and J. Barciszewski, "The non-coding RNAs as riboregulators", *Nucleic Acids Research*, vol. 29, pp. 189-193, 2001.
- [15] S. Forchhammer and J. Rissanen, "Partially hidden Markov models", *IEEE Transactions on Information Theory*, vol. 42, pp. 1253-1256, 1996.
- [16] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, New York, 1979.
- [17] S. Gottesman, "Stealth regulation: biological circuits with small RNA switches", *Genes & Development*, vol. 16, pp. 2829-2842, 2002.
- [18] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna and S. R. Eddy, "Rfam: an RNA family database", *Nucleic Acids Res.*, vol. 31, pp. 439-441, 2003.
- [19] M. A. Harrison, *Introduction to formal language theory*, Addison-Wesley, 1978.
- [20] L. He and J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation", *Nature Reviews Genetics*, vol. 5, pp. 522-531, 2004.
- [21] R. J. Klein and S. R. Eddy, "RSEARCH: Finding homologs of single structured RNA sequences", *BMC Bioinformatics*, vol. 4, 44, 2003.
- [22] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov models in computational biology: applications to protein modeling", *Journal of Molecular Biology*, vol. 235, pp. 1501-1531, 1994.
- [23] A. Krogh, I. S. Mian, and D. Haussler, "A hidden Markov model that finds genes in E. coli DNA", *Nucleic Acids Research*, vol. 22, pp. 4768-4778, 1994.
- [24] K. Lari and S. J. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm", *Computer Speech and Language*, vol. 4, pp. 35-56, 1990.
- [25] T. M. Lowe and S. R. Eddy, "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence", *Nucleic Acids Res.*, vol. 25, pp. 955-964, 1997.
- [26] M. Mandal and R. R. Breaker, "Gene regulation by riboswitches", *Nature Reviews Mol. Cell Bio.*, vol. 5, pp. 451-463, 2004.
- [27] J. S. Mattick, "Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms", *BioEssays*, vol. 25, pp. 930-939, 2003.
- [28] H. Matsui, K. Sato, and Y. Sakakibara, "Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures", *Bioinformatics*, vol. 21, pp. 2611-2617, 2005.
- [29] M. A. Matzke and J. A. Birchler, "RNAi-mediated pathways in the nucleus", *Nature Reviews Genetics*, vol. 6, 2005.
- [30] M. T. McManus and P. A. Sharp, "Gene silencing in mammals by small interfering RNAs", *Nature Reviews Genetics*, vol. 3, 2002.
- [31] V. Moulton, "Tracking down noncoding RNAs", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 2269-2270, 2005.
- [32] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman, "Algorithms for loop mathings", *SIAM Journal of Applied Mathematics*, vol. 35, pp. 68-82, 1978.
- [33] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison", *Proc. Natl. Acad. Sci. USA*, vol. 4, pp. 2444-2448, 1988.
- [34] E. Rivas and S. R. Eddy, "The language of RNA: a formal grammar that includes pseudoknots", *Bioinformatics*, vol. 16, pp. 334-340, 2000.

- [35] E. Rivas and S. R. Eddy, "Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs", *Bioinformatics*, vol. 16, pp. 583-605, 2000.
- [36] E. Rivas and S. R. Eddy, "Noncoding RNA gene detection using comparative sequence analysis", *BMC Bioinformatics*, vol. 2, 8, 2001.
- [37] J. Ruan, G. D. Stormo, and W. Zhang, "An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots", *Bioinformatics*, vol. 20, 2004.
- [38] G. Ruvkun, "Glimpses of a tiny RNA world", *Science*, vol. 294, pp. 797-799, 2001.
- [39] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R. C. Underwood and D. Haussler, "Stochastic context-free grammars for tRNA modeling", *Nucleic Acids Res.*, vol. 22, pp. 5112-5120, 1994.
- [40] S. Gisela, "An expanding universe of noncoding RNAs", *Science*, vol. 296, pp. 1260-1263, 2002.
- [41] N. Sudarsan, J. E. Barrick, and R. R. Breaker, "Metabolite-binding RNA domains are present in the genes of eukaryotes", *RNA*, vol. 9, pp. 644-647, 2003.
- [42] E. B. ten Dam, C. W. A. Pleij, and D. Draper, "Structural and functional aspects of RNA pseudoknots", *Biochemistry*, vol. 31, pp. 11665-11676, 1992.
- [43] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences", *Computer Applications in the Biosciences*, vol. 13, no. 3, pp. 263-270, 1997.
- [44] B. J. Tucker and R. R. Breaker, "Riboswitches as versatile gene control elements", *Current Opinion in Structural Biology*, vol. 15, pp. 342-348, 2005.
- [45] P. P. Vaidyanathan and B.-J. Yoon, "The role of signal-processing concepts in genomics and proteomics", *Journal of the Franklin Institute*, vol. 341, pp 111-135, 2003.
- [46] P. P. Vaidyanathan, "Genomics and proteomics: a signal processor's tour", *IEEE Circuits and Systems Magazine*, vol. 4, no.4, pp. 6-29, 2005.
- [47] S. Washietl, I. L. Hofacker, and P. F. Stadler, "Fast and reliable prediction of noncoding RNAs", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 2454-2459, 2005.
- [48] S. Washietl, I. L. Hofacker, M. Lukasser, A. Hüttenhofer and P. F. Stadler, "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome", *Nature Biotechnology*, vol. 23, pp. 1383-1390, 2005.
- [49] C. Witwer, I. L. Hofacker, and P. F. Stadler, "Prediction of consensus RNA secondary structures including pseudoknots", *IEEE Trans. Comp. Biology and Bioinformatics*, vol. 1, pp. 66-77, 2004.
- [50] B.-J. Yoon and P. P. Vaidyanathan, "HMM with auxiliary memory: A new tool for modeling RNA secondary structures", *Proc. 38th Asilomar Conference on Signals, Systems, and Computers*, Monterey, CA, Nov. 2004.
- [51] B.-J. Yoon and P. P. Vaidyanathan, "An overview of the role of context-sensitive HMMs in the prediction of ncRNA genes", *Proc. IEEE Workshop on Statistical Signal Processing*, Bordeaux, France, July 2005.
- [52] B.-J. Yoon and P. P. Vaidyanathan, "Context-sensitive hidden Markov models for modeling long-range dependencies in symbol sequences", *IEEE Transactions on Signal Processing*, to appear.
- [53] B.-J. Yoon and P. P. Vaidyanathan, "Profile context-sensitive HMMs for probabilistic modeling of sequences with complex correlations", *Proc. 31st IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, May 2006.
- [54] M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information", *Nucleic Acids Research*, vol. 9, pp. 133-148, 1981.