Effective Annotation of Noncoding RNA Families Using Profile Context-Sensitive HMMs

Byung-Jun Yoon
Dept. of Electrical & Computer Engineering
Texas A&M University
Collge Station, TX 77840-3128, USA
Email: bjyoon@tamu.edu

***Abstract*—Noncoding RNAs (ncRNAs) are RNA molecules that function without being translated into proteins. Systematic research on ncRNAs has shown that there exist many ncRNAs that are actively involved in various biological processes, playing key roles in controlling them. As the annotation of ncRNAs is still at an early stage, developing efficient computational tools for finding ncRNAs is of great importance. One effective way for finding new ncRNAs is to look for new RNAs that resemble the RNAs that have already been identified. Recently, a new model called the profile context-sensitive HMM (profile-csHMM) has been proposed, and it has been shown that they can provide a convenient framework for finding RNA homologues. In this paper, we give a brief review of profile-csHMMs and their application in RNA similarity search. We also introduce a number of recent advances related to profile-csHMMs and profile-csHMM based search.**

***Index Terms*—Profile context-sensitive HMM (profile-csHMM), noncoding RNA (ncRNA), RNA secondary structure, RNA similarity search, pseudoknot.**

I. Introduction

Noncoding RNAs (ncRNAs) are RNA molecules that function without being translated intro proteins. Systematic research on ncRNAs has revealed that there exist many ncRNAs that actively participate in various biological processes, playing crucial roles in them [3], [7], [10]. Although examples such as the transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) have been known for a long time, recent studies have revealed that the number and variety of ncRNAs as well as the extent of their roles are much larger than it was previously thought.

Unlike protein-coding genes, the annotation of ncRNAs is still at an early stage. Considering the enormous amount of sequence data that is available these days, we definitely need efficient computational methods that can expedite the annotation process. One effective way for identifying new ncRNAs is to search the database and look for RNAs that resemble known ncRNAs. Let us assume that we have a set of related RNAs that belong to the same functional family. Based on these RNAs, we can build a statistical model that closely represents them, and then use it to find new RNAs that look similar to them. This approach is typically referred as the *RNA similarity search* or *RNA homology search* [16].

As many ncRNAs have well-conserved secondary structures, it is important to take this structural information into account when looking for similar ncRNAs. Therefore, we need a statistical model that can represent various RNA secondary structures to perform an effective search. For this purpose, we can use the profile context-sensitive HMMs (profile-csHMMs) that have been recently proposed [15], [19]. It has been shown that profile-csHMMs can provide a convenient framework for performing RNA similarity search and building RNA sequence analysis tools [19].

In this paper, we provide a tutorial review of profile-csHMMs and their application in RNA similarity search. We also briefly cover a number of recent advances related to profile-csHMMs and profile-csHMM based search. The paper is organized as follows. In Sec. II, we give a brief review of RNA secondary structures and statistical models for representing and analyzing RNAs. In Sec. III, we review the concept of profile-csHMMs and show how they can be constructed from a multiple alignment of RNAs. In Sec. IV, we explain how profile-csHMMs can be used in RNA similarity search. Sec. V introduces two methods that can be employed to make profile-csHMM based searches faster. We conclude the paper in Sec. VI.

II. RNA Secondary Structure

RNA is a nucleic acid that consists of a string of four types of nucleotides. The nucleotides are typically denoted by A, C, G, and U, which stand for *adenine*, *cytosine*, *guanine*, and *uracil*, respectively. Uracil is chemically similar to *thymine* (T) in DNA, and T is replaced by U when a DNA is transcribed into an RNA. A-U and C-G can form hydrogen-bonded base-pairs, called the Watson-Crick base-pairs. Noncanonical base-pairs are also observed, where the most common noncanonical pair is the G-U wobble base-pair. Bases that can form a base-pair are typically said to be *complementary* to each other. Unlike DNA molecules that exist in double-stranded forms (called the DNA double helix), RNAs are generally single-stranded.

In many cases, the interactions between the complementary bases in a ncRNA make the molecule fold onto itself, forming a number of stacked base-pairs. This two-dimensional folding structure is typically called the *RNA secondary structure*, while the one-dimensional nucleotide sequence before folding is referred as the *primary sequence*. Fig. 1 shows two examples of RNA secondary structures. The RNA shown in Fig. 1(a) forms three base-pairs after folding. The stacked base-pairs

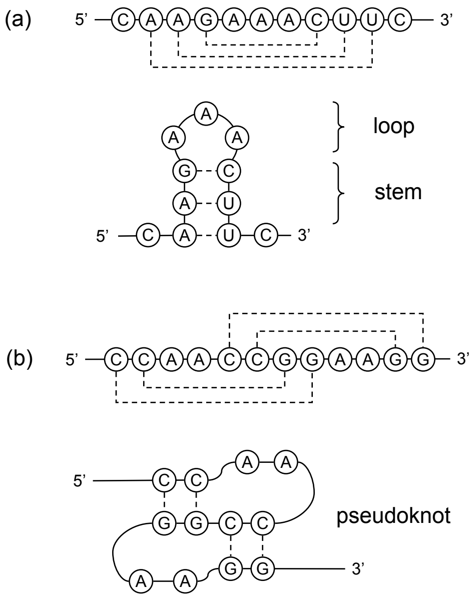

Fig. 1. Examples of RNA secondary structures. The dashed lines indicate the base-pairs formed between complementary bases. (a) RNA with a hairpin structure. (b) RNA with a pseudoknot.

are called a *stem* and the unpaired bases bounded by the base-pairs are called a *loop*. For this reason, the structure shown in Fig. 1(a) is called a *stem-loop* structure, or a *hairpin* structure, due to its shape. Long ncRNAs consist of multiple stem-loops resulting in a more complex structure. In most cases, the base-pairs in an RNA secondary structure occur in a nested manner. For convenience, let us denote a base-pair between positions i and j ($i < j$) as (i, j). We say that two base-pairs (i, j) and (m, n) are *nested*, if they satisfy either $i < m < n < j$ or $m < i < j < n$. Unlike the RNA in Fig. 1(a), where all the base-pairs occur in a nested manner, the RNA shown in Fig. 1(b) has crossing base-pairs. We say that two base-pairs (i, j) and (m, n) are *crossing*, if they satisfy either $i < m < j < n$ or $m < i < n < j$. RNAs with crossing base-pairs are called *pseudoknots*. Although RNA pseudoknots are not as common as RNAs with only nested base-pairs, there exist many ncRNAs that have pseudoknots [5], [11].

It has been observed that the secondary structure of a ncRNA often plays a crucial role in carrying out its function. As a result, RNAs in a same family have similar secondary structures as well similar primary sequences. For some RNA families, this structural similarity can be even higher that the sequence similarity. For this reason, it is important to consider both the structural similarity as well as the sequence similarity when performing an RNA similarity search. In fact, it has been observed that a *scoring scheme*[1] that can reasonably combine contributions from structural and sequence similarities can significantly enhance the discriminative power of the search compared to a sequence-based search [4], [16]. Therefore, in order to implement an effective scoring scheme for comparing RNAs, we need a statistical model that can faithfully represent the sequence and structure of an RNA at the same time. As we can see in Fig. 1, the secondary structure gives rise to correlations between distant bases in the primary sequence of the RNA. This implies that we need a statistical model that can describe such base correlations in order to handle RNAs.

Until now, a number of statistical models have been proposed that can be used for representing RNA secondary structures and implementing scoring schemes that combine sequence similarity and structural similarity [2], [6], [9]. However, these models can handle only a limited class of RNA secondary structures. For example, the stochastic context-free grammars[2] (SCFGs) [1] and the PHMMTSs (pair hidden Markov models on tree structures) [9] cannot handle RNAs with pseudoknots. This can be potentially a serious limitation, since there exist a number of RNAs with functionally important pseudoknots [5], [11]. PSTAGs (pair stochastic tree adjoining grammars) [6], which are a more recent development, can handle many known pseudoknots, but not all of them. Instead of using these models, we can use the profile-csHMMs [15], [19]. Unlike the previous models, profile-csHMMs can handle RNAs with *any* kind of base-pairs, including pseudoknots.

III. Profile Context-Sensitive HMM

The profile-csHMM is a specific kind of context-sensitive HMM (csHMM) [14] with a linear repetitive structure. Context-sensitive HMMs are extensions of conventional HMMs, where the emission and transition probabilities of certain future states depend on emissions that were previously made at corresponding past states. This context-sensitivity greatly increases the descriptive capability of the model, making csHMMs (and especially, profile-csHMMs) useful for representing RNAs with various secondary structures.

The structure of profile-csHMMs is similar to that of conventional profile-HMMs, where they repetitively use three kinds of states: *match states* M_k, *insert states* I_k, and *delete states* D_k. In the following, we show how profile-csHMMs can be constructed and how they work.

A. Constructing an ungapped model

Let us assume that we are given a multiple sequence alignment of RNAs that belong to the same family. Given such an alignment, we first construct an ungapped model that consists of match states alone, where M_k represents the k-th base in the consensus RNA sequence. Therefore, the number of match states in the profile-csHMM will be identical to the length of the consensus RNA that was used to construct the model. Each M_k has its own set of emission probabilities, which reflects the relative occurence of the four bases at the k-th position.

The main difference between a conventional profile-HMM and a profile-csHMM is that a profile-csHMM can use three

[1] A method for comparing sequences and computing a quantitative measure of their similarity.

[2] Covariance models (CMs) are a subclass of SCFGs with a special structure that is suitable for modeling RNAs [2]. The CM has been one of the most popular models in RNA sequence analysis.

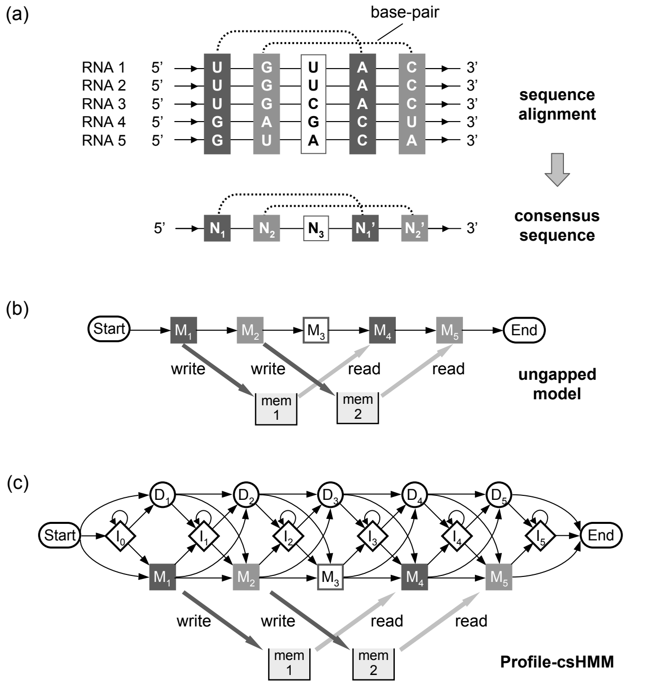

Fig. 2. Constructing a profile-csHMM from an RNA sequence alignment. (a) An RNA alignment with two crossing base-pairs. (b) An ungapped profile-csHMM corresponding to the consensus RNA sequence. (c) Full profile-csHMM that allows additional insertions and deletions in the consensus sequence.

different types of match states, namely, *single-emission match states*, *pairwise-emission match states*, and *context-sensitive match states*. To represent unpaired bases, we use single-emission match states. In order to represent a base-pair, we use a pairwise-emission match state and a corresponding context-sensitive match state to describe the correlation between the two bases.

As an example, let us consider the alignment in Fig. 2(a). We need five match states to represent the sequence, as there are five bases in the consensus RNA sequence. Now, we can see that the RNA in Fig. 2(a) has two crossing base-pairs. We first represent the base-pair between the first and the fourth bases by using a pairwise-emission state for M_1 and the corresponding context-sensitive state for M_4. Similarly, we use another pairwise-emission state for M_2 and the corresponding context-sensitive state for M_5. As the third base is unpaired, it is not correlated to any other base. Therefore, we simply use a single-emission state for M_3. By interconnecting the match states, we can obtain an *ungapped profile-csHMM* as shown in Fig. 2(b). The ungapped model serves as the "backbone" of the final profile-csHMM, and it can represent RNA sequences that match the consensus RNA sequence without any gap.

B. Insertions and deletions

Once the ungapped model has been constructed, we can add insert states I_k and delete states D_k to construct the full profile-csHMM. Firstly, the insert state I_k is used to model insertions between the positions k and $k+1$ in the original consensus RNA sequence. As an inserted base is not correlated with other bases, we use a single-emission state for I_k. Secondly, the delete state D_k is used to describe the deletion of the k-th symbol in the original RNA. As the delete states represent the bases that are missing, these states are *non-emitting states*, which are simply used as place-holders that interconnect other states. Fig. 2(c) shows the complete profile-csHMM that is constructed based on the RNA sequence alignment shown in Fig. 2(a).

C. Descriptive power of profile-csHMMs

As demonstrated in the previous example, profile-csHMMs provide a simple and intuitive way of representing RNA families with various secondary structures. Profile-csHMMs can represent *any* kind of base-pair correlations by properly arranging the pairwise-emission match states and the context-sensitive match states [15], [19]. As a result, profile-csHMMs can represent any kind of secondary structures, including those with pseudoknots. This generality is an important advantage over other models, such as SCFGs [1], PHMMTSs [9], and PSTAGs [6].

IV. Finding Similar RNAs

After constructing a profile-csHMM that represents the RNA family at hand, we can use this model to search the database for finding new RNAs that look similar to the given RNA family. Let us assume that we have a target RNA in a database, and we want to find out how much it resembles the reference RNA family. How can we compute the similarity score between the target and reference RNAs? One good solution would be to compute the maximum observation probability of the target RNA based on the profile-csHMM that represents the reference RNA family. Note that there can be many different *state sequences* (typically called *paths*) in the profile-csHMM that can generate the same symbol sequence, although different paths may have different probabilities. In order to compute the maximum probability of an observed symbol sequence (the target RNA), we have to find the optimal path that maximizes the observation probability. This can be viewed as finding the best alignment between the observed symbol sequence and the given model. For this reason, the problem of finding the optimal state sequence is typically referred as the *optimal alignment* problem.

Finding the optimal path by simply comparing all possible paths is practically infeasible, as the number of paths increases exponentially with the length of the observed sequence. Therefore, we need an efficient algorithm that can find the optimal path in a systematic manner. For traditional HMMs (and profile-HMMs), we can use the *Viterbi algorithm* [12], and for SCFGs, we can utilize the *CYK (Cocke-Younger-Kasami) algorithm* [1]. However, as profile-csHMMs can represent more complicated correlations than HMMs and SCFGs, we need

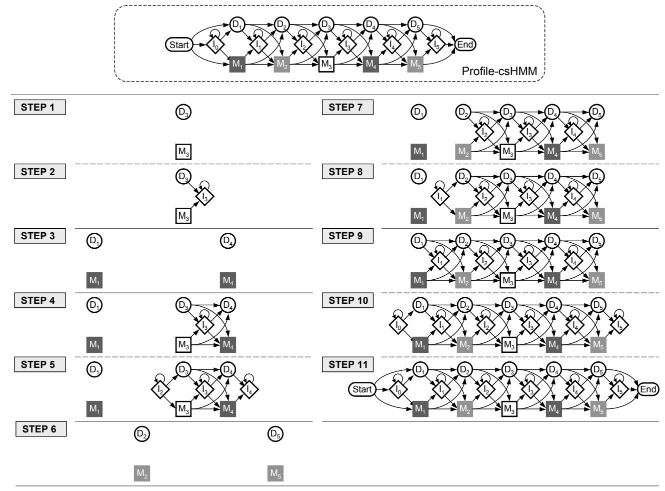

Fig. 3. Example of an adjoining order for the profile-csHMM in Fig. 2 (c).

a more general algorithm for solving the optimal alignment problem of profile-csHMMs.

A. Sequential component adjoining algorithm

Although the Viterbi algorithm and the CYK algorithm cannot be directly used with profile-csHMMs, we can generalize these algorithms to handle profile-csHMMs. Recently, an efficient dynamic programming algorithm called the *SCA (sequential component adjoining) algorithm* [15], [19] has been proposed for finding the optimal path in profile-csHMMs. The SCA algorithm makes the following important generalizations:

1) Instead of using a subsequence with a single interval, the SCA algorithm can define and use subsequences that consist of multiple nonoverlapping intervals.
2) Unlike the Viterbi algorithm that proceeds "left-to-right" or the CYK algorithm that proceeds "inside-to-outside", the SCA algorithm allows us to explicitly describe how the optimal subsequence of shorter subsequences can be extended and adjoined to find the optimal state sequence of longer subsequences.

The specific order that describes how we should proceed to find the final optimal state sequence is called the *adjoining order* [19], [20]. For example, Fig. 3 shows the adjoining order for the profile-csHMM shown in Fig. 2 (c). At each step of the adjoining process, we find the optimal state sequence that corresponds to the portion of the profile-csHMM at the given step, as illustrated in Fig. 2. By following these steps, we can ultimately find the optimal state sequence for the entire profile-csHMM. Detailed discussion on adjoining orders can be found in [19] and [20].

Unlike the Viterbi algorithm and the CYK algorithm, the SCA algorithm has a variable computational complexity that depends on the adjoining order [19]. As the adjoining order is specified based on the correlation structure of the profile-csHMM, the computational cost of the SCA algorithm ultimately depends on the consensus secondary structure of the reference RNA. Table I shows the computational complexity of the SCA algorithm for various RNA secondary structures. L is the length of the observed sequence (target RNA) and M is the number of states in the profile-csHMM. Note that M is proportional to the length of the reference RNA. As we can see in Table I, the computational cost increases as the structure becomes more complex. For typical RNA pseudoknots, the computational cost for finding the optimal path and computing the maximum observation probability is $O(L^4M)$. For RNA pseudoknots in the Rivas&Eddy class [8], the complexity can be as high as $O(L^6)$, and theoretically, the complexity can become even higher for RNAs with more complicated structures.

TABLE I
THE COMPUTATIONAL COMPLEXITY OF THE SCA ALGORITHM FOR VARIOUS RNA SECONDARY STRUCTURES.

RNA Secondary Structure	Computational Complexity
Hairpin Structure	$O(L^2M)$
tRNA Cloverleaf Structure	$O(L^3M)$
Typical pseudoknots	$O(L^4M)$
Rivas & Eddy class	$\leq\ O(L^6M)$

B. Optimizing the adjoining order

As mentioned in the previous section, the SCA algorithm has a variable computational complexity that depends on the adjoining order. However, the adjoining order is not unique, and for a given profile-csHMM, there typically exist many legitimate adjoining orders, where we can use any of these orders to find the unique optimal state sequence.[3] Although any properly defined adjoining order will find the optimal path, they may not necessarily have the same computational cost. Therefore, it is practically important to choose the "optimal" adjoining order that will minimize the computational cost for finding the optimal path. However, finding the optimal adjoining order can be quite difficult for profile-csHMMs that represent complex secondary structures, and we need an efficient method that can automatically find the optimal order for a given profile-csHMM. In fact, such an algorithm has been proposed in [20], and it can be used to find the best order that minimizes the computational cost of the SCA algorithm. It was shown that using the optimal adjoining order can improve the alignment speed up to 3.6 times for RNA pseudoknots [20].

C. Finding structural alignment of RNAs

Finding the optimal alignment between the reference RNA and the target RNA (whose structure is not yet known) gives us the probability of the target RNA, which can be used as a similarity score. Note that this score combines the contribution that comes from sequence similarity as well as the contribution from structural similarity. In addition to the similarity score, we can also predict the secondary structure of the target RNA based on this alignment [19]. Numerical experiments using several RNA families with pseudoknots obtained from the Rfam database [5] showed that profile-csHMMs could find highly accurate alignments at a relatively low computational cost. Table II summarizes the performance

[3] If the optimal path is not unique, different adjoining orders may give different results.

TABLE II
COMPARISON BETWEEN PROFILE-CSHMMS AND PSTAGS.

RNA Family	PROFILE-CSHMM			PSTAG		
	SN (%)	SP (%)	time (sec)	SN (%)	SP (%)	time (sec)
CORONA_PK3	**95.7**	**96.5**	0.68	94.6	95.5	37.2
HDV_RIBOZYME	**94.5**	95.3	0.58	94.1	**95.6**	207.5
TOMBUS_3_IV	**98.6**	**98.6**	0.42	97.4	97.4	270.9
FLAVI_PK3	**94.5**	**96.4**	1.87	-	-	-

of profile-csHMMs [19], [20] and that of PSTAGs [6] for comparison. These results have been obtained from cross-validation experiments as described in [20], using the optimal adjoining order.[4] The *sensitivity (SN)* and the *specificity (SP)* are computed by comparing the predicted base-pairs with the trusted base-pairs in the database, and they provide quantitative measures for evaluating the quality of the RNA alignments. Table II clearly shows that profile-csHMMs could find accurate alignments for all four RNA families that were used in the experiments. It should be noted that profile-csHMMs yielded good performance even for the FLAVI_PK3 family, which has a complex secondary structure that cannot be handled by PSTAGs. Furthermore, using additional heuristics for minimizing the computational cost [19], profile-csHMMs ran significantly faster than PSTAGs without sacrificing the prediction accuracy.

V. FAST RNA SEARCH USING PRESCREENING FILTERS

Although profile-csHMMs can find structural alignment of RNAs and compute their similarity scores reasonably fast, they are still too slow for scanning a large database, especially if the RNA of interest is long. In order to overcome this problem, two methods have been proposed to make profile-csHMM based searches faster [17], [18]. This was inspired by the *prescreening method* proposed for expediting CM-based searches [13]. The basic idea of the prescreening method is to first scan the database using a simple *prescreening filter*. The prescreening filter is a simple statistical model (such as a HMM) that can quickly filter out the regions that are dissimilar and pass only the similar regions to the next stage. We use a full profile-csHMM in the second stage to investigate these regions further. Fig. 4 illustrates the basic idea of the prescreening method. If the prescreening filter runs much faster than the profile-csHMM, yet capable of filtering out most of the dissimilar regions, the overall speed of the search can be significantly improved.

There are many possible ways for implementing the prescreening filter. In the following, we review two different prescreening methods proposed in [17] and [18].

A. Sequence-based prescreening

When carrying out an RNA similarity search, there will be many cases, where the target RNA is too different from the reference RNA in the sequence level that it cannot receive a high score even after considering the contribution that may come from structural similarity. As sequence similarity can be checked very easily using simple models, we do not have to use full profile-csHMMs for such RNAs. Based on this observation, a method has been proposed in [17] that uses conventional profile-HMMs as prescreening filters. It has been shown that the parameters of the profile-HMM prescreening filter can be chosen (based on the parameters of the profile-csHMM) such that there will be no loss in the prediction accuracy. Experimental results showed that the search speed could be improved up to eighty times using this method.

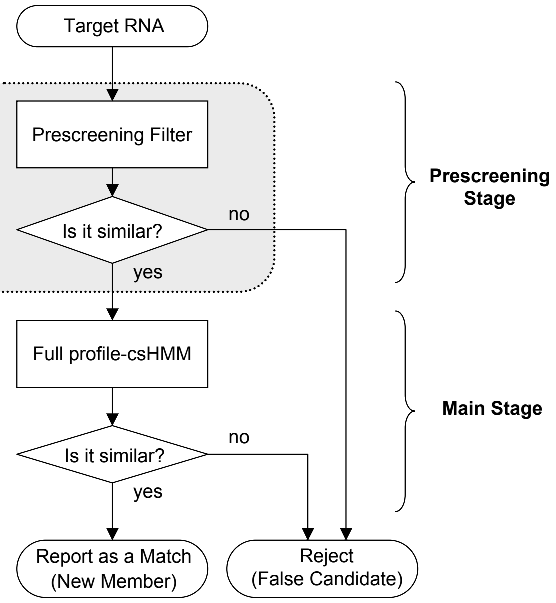

Fig. 4. Illustration of the prescreening method.

B. Structure-based prescreening

The main disadvantage of the sequence-based prescreening method is that it does not work well for RNA families whose members have low sequence similarity but reasonably high structural similarity. For such RNAs, the speed improvement using the method in [17] will be negligible. As these are the RNAs that we can benefit most by using a profile-csHMM based search, this can be indeed a serious disadvantage. In order to overcome this problem, a structure-based prescreening method has been proposed in [17]. This method constructs a matrix that reflects the consensus structure of the reference RNA, and uses this as a *matched filter* for finding RNAs with similar structures. This is illustrated in Fig. 5, where we can see how the method works. Fig. 5(a) shows the base-pairing matrix $\bar{\mathbf{P}}_r$ of the reference RNA, which reflects the secondary structure of the RNA. As $\bar{\mathbf{P}}_r$ is symmetric, we keep only the lower-triangular part to construct the matched filter matrix $\mathbf{S}$ (shown in Fig. 5(b)). Given a target RNA with an unknown structure, we compute its base-pairing matrix for all possible base-pairs. Then we keep only its upper-triangular part and denote it as $\bar{\mathbf{P}}_t$ (see Fig. 5(c)). Now, we can find out whether

[4] These numbers are obtained from [20]. Note that the CPU time for finding the structural alignment using profile-csHMMs does not include the time it took to estimate the "search depth" using conventional HMMs. As the Viterbi algorithm runs much faster than the SCA algorithm, the CPU time for estimating the search depth is small compared to the time for finding the alignment.

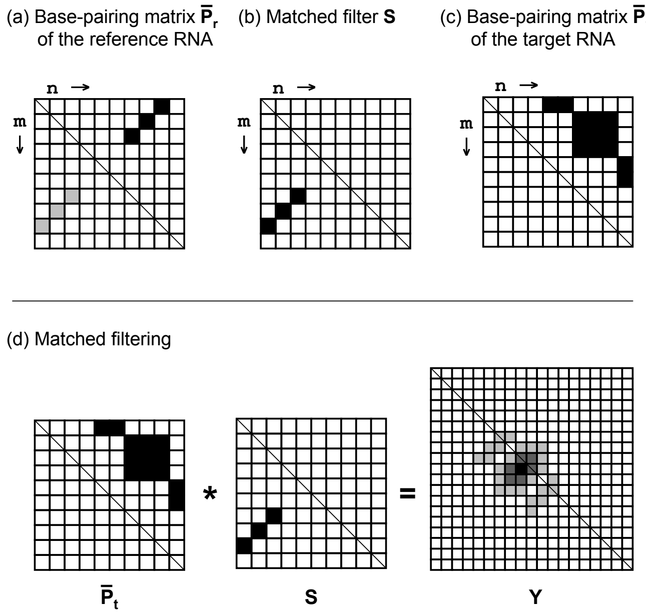
(a) Base-pairing matrix $\bar{\mathbf{P}}_r$ of the reference RNA
(b) Matched filter S
(c) Base-pairing matrix $\bar{\mathbf{P}}_t$ of the target RNA
n →
m ↓
(d) Matched filtering
*
=
$\bar{\mathbf{P}}_t$
S
Y

Fig. 5. Matched filtering based on stem patterns.

the target RNA has a similar structure as the reference RNA by finding the maximum overlap between the base-pairing matrices of the two RNAs. This can be done by computing the two-dimensional convolution

$$\mathbf{Y} = \bar{\mathbf{P}}_t * \mathbf{S},$$

and finding the largest element in $\mathbf{Y}$. Let λ be the value of this largest element. This λ gives us the maximum number of base-pairs that the two RNAs have in common. So, the larger λ we have, the closer will be the structures of the reference and the target RNAs. Thefore, we can use this λ as a "structural similarity score". As we can construct the matched filter matrix $\mathbf{S}$ for any kind of RNA secondary structure, this method can be used for any kind of RNAs, also including pseudoknots. Despite this generality, the computational complexity of this method is relatively low, as it simply involves a two-dimensional discrete convolution. In fact, the complexity will be only $O(L^2N)$, where L is the length of the target RNA and N is the number of base-pairs in the reference RNA.

In [18], it was shown that this method could improve the search speed around 30∼50 times for two RNA families with pseudoknots. This method can also be used in combination with other sequence-based prescreening methods. We expect that the overall search speed will be improved even further by combining the methods in [17] and [18].

VI. Concluding Remarks

In this paper, we reviewed the concept of profile-csHMMs and showed how they can be used in RNA similarity search. As shown throughout the paper, profile-csHMMs can conveniently represent RNAs with various secondary structures, and they provide a convenient framework for finding RNA homologues and building RNA sequence analysis tools. We also introduced a number of recent advances related to profile-csHMMs and profile-csHMM based search. Since the profile-csHMM is a recent invention, there exist many interesting theoretical and practical problems that need to be solved. For example, developing a more efficient prescreening method that can improve the search speed for general RNA families, the optimal parameterization of profile-csHMMs based on a small number of training sequences, and minimizing the memory requirement of the SCA algorithm are just a few examples of important topics for future research.

References

[1] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, Cambridge Univ. Press, Cambridge, UK, 1998.

[2] S. R. Eddy and R. Durbin, "RNA sequence analysis using covariance models", *Nucleic Acids Research*, vol. 22, 2079-2088, 1994.

[3] S. R. Eddy, "Non-coding RNA genes and the modern RNA world", *Nature Reviews Genetics*, vol. 2, pp. 919-929, 2001.

[4] S. R. Eddy, "Computational genomics of noncoding RNA genes", *Cell*, vol. 109, pp. 127-140, 2002.

[5] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: an RNA family database", *Nucleic Acids Research*, vol. 31, pp. 439-441, 2003.

[6] 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.

[7] J. S. Mattick, "Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms", *BioEssays*, vol. 25, pp. 930-939, 2003.

[8] E. Rivas and S. R. Eddy, "The language of RNA: a formal grammar that includes pseudoknots", *Bioinformatics*, vol. 16, pp. 334-340, 2000.

[9] Y. Sakakibara, "Pair hidden Markov models on tree structures", *Bioinformatics*, vol. 19, i232-i240, 2003.

[10] G. Storz, "An expanding universe of noncoding RNAs", *Science*, vol. 296, pp. 1260-1263, 2002.

[11] F. H. D. van Batenburg, A. P. Gultyaev, C. W. A. Pleij, J. Ng, and J. Oliehoek, "Pseudobase: a database with RNA pseudoknots", *Nucleic Acids Research*, vol. 28, pp. 201-204, 2000.

[12] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", *IEEE Transactions on Information Theory*, vol. IT-13, pp. 260-267, 1967.

[13] Z. Weinberg and W. L. Ruzzo, "Faster genome annotation of non-coding RNA families without loss of accuracy", *Proc. 8th Ann. Int. Conf. on Computational Molecular Biology (RECOMB)*, pp. 243-251, 2004.

[14] 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*, vol. 54, pp. 4169-4184, Nov. 2006.

[15] B.-J. Yoon and P. P. Vaidyanathan, "Profile context-sensitive HMMs for probabilistic modeling of sequences with complex correlations", *Proc. 31st International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, May 2006.

[16] B.-J. Yoon and P. P. Vaidyanathan, "Computational identification and analysis of noncoding RNAs - Unearthing the buried treasures in the genome", *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 64-74, Jan. 2007.

[17] B.-J. Yoon and P. P. Vaidyanathan, "Fast search of sequences with complex symbol correlations using profile context-sensitive HMMs and pre-screening filters", *Proc. 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, Apr. 2007.

[18] B.-J. Yoon and P. P. Vaidyanathan, "Fast Structural Similarity Search of Noncoding RNAs Based on Matched Filtering of Stem Patterns", *Proc. 41st Asilomar Conference on Signals, Systems, and Computers*, Monterey, CA, Nov. 2007.

[19] B.-J. Yoon and P. P. Vaidyanathan, "Structural alignment of RNAs using profile-csHMMs and its application to RNA homology Search: Overview and new results", *IEEE Trans. Automatic Control & IEEE Trans. Circuits and Systems: Part-I (Joint Special Issue on Systems Biology)*, accepted.

[20] B.-J. Yoon and P. P. Vaidyanathan, "Fast Structural Alignment of RNAs by Optimizing the Adjoining Order of Profile-csHMMs", *IEEE Journal of Selected Topics in Signal Processing*, submitted.