AN OVERVIEW OF THE ROLE OF CONTEXT-SENSITIVE HMMS IN THE PREDICTION OF NCRNA GENES

Byung-Jun Yoon and P. P. Vaidyanathan

Dept. of Electrical Engineering
California Institute of Technology, Pasadena, CA 91125, USA
E-mail: bjyoon@caltech.edu, ppvnath@systems.caltech.edu

ABSTRACT

Non-coding RNAs (ncRNA) are RNA molecules that function in the cells without being translated into proteins. In recent years, much evidence has been found that ncRNAs play a crucial role in various biological processes. As a result, there has been an increasing interest in the prediction of ncRNA genes. Due to the conserved secondary structure in ncRNAs, there exist pairwise dependencies between distant bases. These dependencies cannot be effectively modeled using traditional HMMs, and we need a more complex model such as the context-sensitive HMM (csHMM). In this paper, we overview the role of csHMMs in the RNA secondary structure analysis and the prediction of ncRNA genes. It is demonstrated that the context-sensitive HMMs can serve as an efficient framework for these purposes.

1. INTRODUCTION

The central dogma of molecular biology states that the genetic information in the cells flow from DNA to RNA to protein. The hereditary information stored in the DNA is transcribed into a messenger RNA (mRNA), which is then translated into a protein. In the early stage of genomics research, most researchers have concentrated on prokaryotes, which are organisms (mostly unicellular) whose cells do not have a nucleus [1]. In prokaryotes, it has been found that the genome consists mostly of protein-coding genes. Most part of their DNA is used for encoding proteins, and these proteins control most of the genetic information in the cell. Based on these results, it has been assumed that this must be also true in eukaryotes. Eukaryotes are organisms with cell nuclei, and they can be either unicellular or multicellular [1]. All the complex organisms such as humans, fruit flies, rice plants are eukaryotes. For a long time, it has been believed that proteins must be responsible for all important functions in the cell machinery, which include not only structural and catalytic functions but also regulatory functions. In the meanwhile, RNA has been simply viewed as a passive intermediary that interconnects DNA and protein, with the exception of several infrastructural RNAs, such as the transfer RNA (tRNA) and the ribosomal RNA (rRNA).

However, the sequencing of eukaryotic genomes has revealed that most part of their genomes is not used for encoding proteins. For example, analysis of the human genome has revealed that less than 1.5% of the entire DNA sequence is translated into proteins [3]. Although only a small portion of the human genome encodes proteins, estimates suggest that around $97-98\%$ of the transcriptional output of the genome is non-coding RNA (RNA transcripts that are not translated into proteins) [4]. This inevitably raises the question whether the human genome is full of useless transcriptions.

In recent years, startling observations have been made by many researchers regarding these non-coding RNAs (ncRNAs) [5, 6, 7]. Numerous evidences have been found which showed that ncRNAs play important roles in various cellular processes. For example, it has been found that ncRNAs affect transcription and the chromosome structure, regulate mRNA stability and translation, affect protein stability and transport, and are also involved in RNA processing and modification [5]. Moreover, it is even suggested that the ncRNAs constitute the majority of genomic programming in the complex organisms [2, 4]. Systematic research on ncRNAs has found surprisingly diverse functions of ncRNAs, but there exist numerous ncRNAs whose precise functions are not known yet. Moreover, the ncRNAs that have been identified till now are still considered to be only a small fraction of the existing ncRNAs [7].

In this paper, we consider the problem of computationally identifying ncRNA genes in a DNA sequence that has not been annotated yet. Until now, many protein-coding gene finders have been proposed, where methods based on hidden Markov models (HMMs) have been especially successful [8, 9]. However, these traditional gene finders cannot give satisfactory results in ncRNA gene prediction, and we need a more complex model such as the context-sensitive HMM (csHMM) [10] for this purpose. In the following discussion, we elaborate on the role of context-sensitive HMMs in the computational analysis of RNA sequences and propose an efficient framework for building non-coding RNA gene finders. The organization of the paper is as follows. In Sec. 2, we consider traditional protein-coding gene finders and discuss the difficulties in applying them to ncRNA gene prediction. Sec. 3 reviews the characteristics of RNAs and RNA secondary structures, and Sec. 4 briefly overviews the concept of context-sensitive HMMs that can be used for representing RNAs with conserved secondary structures. In Sec. 5, we present the csHMM database search algorithm, and a ncRNA gene-finder that can identify IREs (iron response elements) is considered in Sec. 6. The paper is concluded in Sec. 7.

2. GENE PREDICTION

Thanks to the success of genome sequencing projects, we are currently experiencing an explosion in genomic data. Given the huge amount of data that is available to us these days, it is practically impossible to identify all ncRNA genes solely by experimental means. In order to expedite the search process, we need computational tools for predicting the probable locations of these genes. In fact, many gene finders have been proposed during the last

Work supported in parts by the NSF grant CCF-0428326 and the Microsoft Research Graduate Fellowship.

0-7803-9404-6/05/$20.00 ©2005 IEEE 1378

decade, whose target was mostly protein-coding genes. These gene-finders have been built based on various approaches, which include hidden Markov models (HMMs) [8, 9], discrete Fourier transform [12, 13], digital filters [14, 15], neural networks [16], and so forth. A number of methods, especially those based on HMMs, have been quite successful in identifying protein-coding regions. Unfortunately, none of these traditional methods can be directly used for identifying ncRNA genes due to the following reasons. In many organisms, ncRNA genes do not display strong sequence composition bias [11]. In addition to this, ncRNA genes are considerably shorter than protein-coding genes, which makes it difficult to judge whether the statistical properties inside the genes are different from the properties of the rest in a statistically meaningful manner. To make matters worse, ncRNA genes do not have *codons*, hence start codons and stop codons that have been conveniently used for locating coding genes cannot be utilized any more. Instead, many functional RNAs are known to conserve their secondary structures more than they conserve their primary sequences [17]. Therefore, traditional gene-finders that are mainly based on base composition statistics cannot give satisfactory results in predicting ncRNA genes. In order to build ncRNA gene-finders, we have to consider both the primary sequence and the secondary structure of the RNA that is of our interest.

3. RNA SECONDARY STRUCTURE

The RNA is a nucleic acid that consists of a string of nucleotides A, C, G and U. These symbols stand for adenine, cytosine, guanine, and uracil, respectively. Uracil (U) is chemically similar to thymine (T) in the DNA. The nucleotide A can form a hydrogen-bonded pair with U, and C can form a pair with G. These hydrogen-bonded base-pairs are called *complementary base-pairs* or *Watson-Crick pairs*. The RNA is generally a single-stranded molecule, and it typically folds onto itself to form base-pairs that are stacked onto each other, which is called a *stem*. The structure that results from these base-pairs is called the *RNA secondary structure*. Fig. 1 shows an example of a simple RNA secondary structure. This kind of structure is called a *stem-loop* (or a *hairpin*), and it is frequently observed in various RNAs. As we can see in Fig. 1 (a), there exist pairwise correlations between bases that are distant from each other. Most of the pairwise interactions in RNAs occur in a nested fashion, where the interactions do not cross each other. When there exist crossing interactions, they are called *pseudoknots*.

RNA sequences with nested pairwise correlations are in principle similar to *palindromes*, which are sequences that are symmetric around the center. The palindrome language, which consists of all possible palindromes, is a classic example of a language that cannot be modeled using the so-called *regular grammars* in the Chomsky hierarchy of transformational grammars [18]. It is known that HMMs, which have been widely used for building protein-coding gene finders, can be viewed as *stochastic regular grammars*, hence they cannot be used for representing such complex dependencies [11]. It is of course possible to construct a HMM that generates also palindromes, but the point is that we cannot restrict the model to generate *only* palindromes. Therefore, conventional HMMs cannot "effectively" differentiate palindromes from non-palindromes, which makes it not suitable for building ncRNA gene finders.

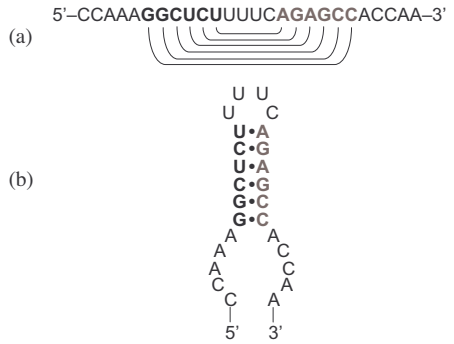


Fig. 1. The 3' UTR (untranslated region) of a histone mRNA. (a) Primary sequence of the RNA. The lines indicate the correlations between distant bases. (b) Consensus secondary structure.

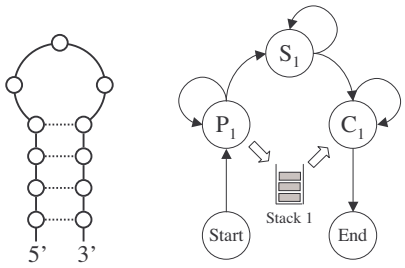
In order to represent complex pairwise correlations that are observed in the RNA sequences with conserved secondary structures, we need more complex models that have larger descriptive power. One possibility is to use the so-called *stochastic context-free grammars* (SCFGs). SCFGs have been used by several ncRNA gene finders and RNA analysis tools, with satisfactory results [19, 20]. Instead of using the SCFG, we may use the *context-sensitive HMMs* (csHMM) that have been proposed recently [10]. Context-sensitive HMMs have several advantages over SCFGs [10], and they have been applied to RNA secondary structure prediction, where they showed promising results [21]. In the following section, we briefly overview the concept of csHMM and demonstrate how they can be used for modeling RNA sequences with various secondary structures.

4. CONTEXT-SENSITIVE HMM

The context-sensitive HMM is an extension of the traditional HMM, where some states are equipped with auxiliary memory [10]. Symbols that are emitted at specific states are stored in the associated memory, and this data serves as the *context* of the model, which affects the emission and transition probabilities of certain future states. There are three distinct types of states, namely, the single-emission state S_n, the pairwise-emission state P_n, and the context-sensitive state C_n. The single-emission state S_n is identical to the state in a regular HMM. The pairwise-emission state P_n is identical to the single-emission state except that it stores the emitted symbols in the auxiliary memory. This data is used to adjust the probabilities of the corresponding context-sensitive state C_n, in the future. Therefore, the model has to be constructed such that P_n always comes before C_n (they need not be adjacent to each other), since the probabilities at C_n cannot be properly decided without the context. When we enter C_n, we first retrieve a symbol x from the associated memory. Note that x is one of the symbols that were previously emitted at the corresponding pairwise-emission state P_n. The emission probabilities of C_n are adjusted according to the value of x. For example, we may adjust the probabilities such that C_n emits the same symbol x with high probability (possibly, with probability one). Table 1 summarizes the properties of the three classes of states in the csHMM.

	Memory Access	Probabilities
S_n	N/A	fixed
P_n	stores the emitted symbol	fixed
C_n	retrieves the stored symbol	context-dependent

Table 1. Properties of the three classes of states S_n, P_n, and C_n.



(a) (b)

Fig. 2. (a) A stem-loop. Each node represents a base in the RNA, and the dotted lines indicate the correlations between bases that form base-pairs. (b) A csHMM that can generate sequences with a stem-loop structure

By using context-sensitive HMMs, we can easily construct a model that can represent RNA sequences with a specific secondary structure. Fig. 2 illustrates an example of a csHMM that can represent RNA sequences with a stem-loop structure. In this model, P_1 and C_1 are associated with a stack, and they work together to generate the stem part. P_1 stores the emitted symbols (bases) in the stack, and as we enter C_1, we retrieve the bases that were previously emitted by P_1. The emission probabilities of C_1 are adjusted such that it emits the complementary bases of the retrieved ones. The single-emission state S_1 is used to generate the loop part in the structure. In a similar manner, we can also model other RNA sequences with various RNA secondary structures. As shown in this example, the csHMM can provide an effective framework for modeling RNA secondary structures and building RNA analysis tools.

As the context-sensitive HMMs are capable of dealing with complex correlations that do not satisfy the Markov assumption, algorithms such as the Viterbi algorithm and the forward algorithm that have been used with traditional HMMs cannot be used any more. In order to use csHMMs in practical applications, we need new algorithms that can take care of the pairwise correlations in the symbol sequence. Recently, algorithms have been proposed for finding the optimal state sequence (alignment problem) [22] and computing the probability (scoring problem) [23] of an observation sequence based on the given csHMM. These algorithms can be used for analyzing sequences such as the RNA sequences with a stem-loop structure, which have single nested correlations. An example of a sequence with single nested correlations is shown in Fig. 3 (a). Multiple nested correlations illustrated in Fig. 3 (b) and crossing correlations that are depicted in Fig. 3 (c) are not considered in this case.

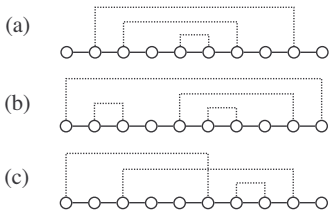


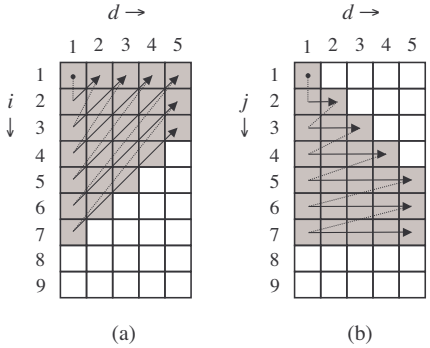
Fig. 3. Examples of sequences with different correlations. (a) Single nested correlations. (b) Multiple nested correlations. (c) Crossing correlations.

5. DATABASE SEARCH ALGORITHM

In order to build a ncRNA gene finder based on csHMMs, we first have to construct a csHMM that closely represents the target gene. Once we have a good model for the ncRNA gene, we have to search the database to find the regions that match the given csHMM reasonably well. For this purpose, we need an efficient database search algorithm that can be used with context-sensitive HMMs. In this section, we propose a csHMM database search algorithm, which is a variant of the optimal alignment algorithm proposed in [22].

Let us first define the variables that are needed to describe the algorithm. Let $\mathbf{x} = x_1x_2...x_L$ be the observed symbol sequence, where L is the length of the observation. The underlying state sequence (path) is denoted as $\mathbf{s} = s_1s_2 \dots s_L$. We assume that there are M states in the context-sensitive HMM. M_1 is the number of state-pairs (P_n, C_n) and M_2 is defined as the number of single-emission states, hence we have $M = 2M_1 + M_2$. It is assumed that all pairwise interactions between P_n and C_n occur in a nested manner (with a single nested structure) and do not cross each other. For notational convenience, we also define the following sets $\mathcal{P} = \{P_1, \dots, P_{M_1}\}$, $\mathcal{C} = \{C_1, \dots, C_{M_1}\}$ and $\mathcal{S} = \{S_1, \dots, S_{M_2}\}$. We denote the transition probability from state v to w as $t(v, w)$. The emission probability of a symbol x at a state $v \in \mathcal{S}$ or $v \in \mathcal{P}$ is defined as $e(x|v)$. Since the emission probabilities at a context-sensitive state $v \in \mathcal{C}$ depends on the symbol x_p that was previously emitted at the corresponding pairwise-emission state, we denote the emission probability at $v \in \mathcal{C}$ as $e(x|v, x_p)$.

In the alignment algorithm in [22], the variable $\gamma(i, j, v, w)$ is defined as the log-probability of the optimal path among all sub-paths $s_i \cdots s_j$ with $s_i = v$ and $s_j = w$, where it is assumed that all pairwise-emission states P_n are paired with the corresponding context-sensitive states C_n inside the sub-path. In many cases, we can limit the maximum length of the ncRNA gene, which reduces the overall computational complexity of the algorithm significantly. Let us define $d = j - i + 1$ to be the length of the sub-sequence, where we restrict it to be $d \leq D$ for some D. Based on this setting, we may define either $\gamma(i, d, v, w)$ or $\gamma(j, d, v, w)$, in a similar manner. Using one of these variables instead of $\gamma(i, j, v, w)$ can minimize the memory requirement as well. Depending on which variable we use, there exist two different schemes for computing these variables iteratively. This is illustrated in Fig.4. In the following algorithm, we use the vari-



(a) (b)

Fig. 4. Two different update schemes. (a) When using the variable $\gamma(i, d, v, w)$. (b) When using the variable $\gamma(j, d, v, w)$.

able $\gamma(j, d, v, w)$, hence adopting the update scheme in Fig. 4 (b), which is similar to the scheme elaborated in [11]. Now, the database search algorithm can be defined as follows.

Database Search Algorithm

For $j = 1, \ldots, L, d = 1, \ldots, \min(D, j)$ and $v = 1, \ldots, M, w = 1, \ldots, M$.

(i) $d = 1$ $(v = w)$

$$\gamma(j, d, v, v) \quad = \quad \begin{cases} \log e(x_i|v) & v \notin \mathcal{P}, \mathcal{C} \\ -\infty & \text{otherwise} \end{cases}$$

(ii) $d = 1$ $(v \neq w)$

$$\gamma(j, d, v, w) = -\infty$$

(iii) $v = P_n, w = C_m (n \neq m)$, or $v \in \mathcal{C}$, or $w \in \mathcal{P}$

$$\gamma(j, d, v, w) = -\infty$$

(iv) $v = P_n, w = C_n, d = 2$

$$\gamma(j, d, v, w) = \log e(x_{j-1}|v) + \log t(v, w) + \log e(x_j|w, x_{j-1})$$

(v) $v = P_n, w = C_n, d > 2$

$$\gamma(j, d, v, w) \quad = \quad \max_{u_1, u_2} \Big[\log e(x_{j-d+1}|v) + \log t(v, u_1) + \gamma(j-1, d-2, u_1, u_2) + \log t(u_2, w) + \log e(x_j|w, x_{j-d+1}) \Big]$$

(vi) $v \in \mathcal{P}, w \notin \mathcal{C}$

$$\gamma(j, d, v, w) \quad = \quad \max_{u} \Big[\gamma(j-1, d-1, v, u) + \log t(u, w) + \log e(x_j|w) \Big]$$

(vii) $v \notin \mathcal{P}, w \in \mathcal{C}$

$$\gamma(j, d, v, w) \quad = \quad \max_{u} \Big[\log e(x_{j-d+1}|v) + \log t(v, u) + \gamma(j, d-1, u, w) \Big]$$

(viii) $v \notin \mathcal{P}, w \notin \mathcal{C}$

In this case, the variables $\gamma(j, d, v, w)$ can be updated using either the update formula (vi) or (vii). □

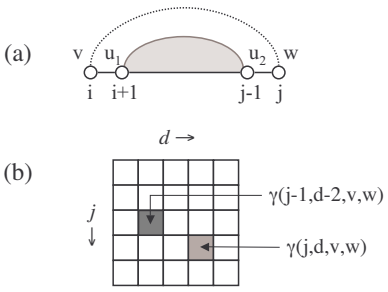


Fig. 5. Illustration of the step (v).

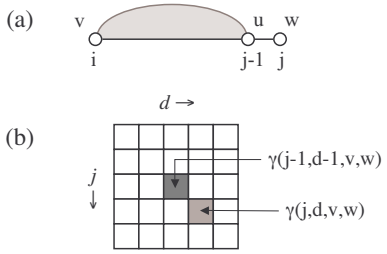


Fig. 6. Illustration of the step (vi).

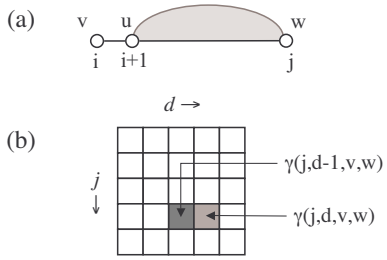


Fig. 7. Illustration of the step (vii).

As we can see from above, the log-probability $\gamma(j, d, v, w)$ is computed in an iterative manner, starting from a shorter sequence and extending it progressively. Whenever there exist pairwise correlations between symbols, the emission of these symbols are considered at the same time. As mentioned earlier, when computing $\gamma(j, d, v, w)$, we consider only those paths, where all the P_n states are paired with the corresponding C_n states inside the path. Therefore, the log-probability $\gamma(j, d, v, w)$ is set to $-\infty$, whenever P_n and C_n do not form pairs. For example, in case (iii), when the leftmost state $s_i (= s_{j-d+1}) \in \mathcal{C}$, it cannot be paired with the corresponding pairwise-emission state, since there are no more states to the left of s_i. This is also true when the rightmost state is $s_j \in \mathcal{P}$.

Now, let us consider the case when $v = P_n$ and $w = C_n$. When $d = 2$, we can simply compute $\gamma(j, d, v, w)$ as in (iv), by considering the emission of $x_i (= x_{j-d+1})$ and x_j together. When $d > 2$, we can compute $\gamma(j, d, v, w)$ as follows. Since s_i has to form a pair with s_j as shown in Fig. 5 (a) by the dotted line, the pairwise-emission states and the corresponding context-sensitive states inside $s_{i+1} \ldots s_{j-1}$ have to exist in pairs. This is indicated by the shaded region in Fig. 5 (a). As the log-probability of the optimal path for $s_{i+1} \ldots s_{j-1}$ is already stored in $\gamma(j-1, d-2, u_1, u_2)$, we can compute $\gamma(j, d, v, w)$ by extending $\gamma(j-1, d-2, u_1, u_2)$ as shown in (v).

Fig. 6 illustrates the case when $v \in \mathcal{P}$ and $w \notin \mathcal{C}$. Since there can be no interaction between s_j and any other state $s_k (i \leq k \leq j-1)$, all the pairwise-emission states and the context-sensitive states inside $s_i \ldots s_{j-1}$ should exist in pairs. Therefore $\gamma(j, d, v, w)$ can be computed by extending $\gamma(j-1, d-1, v, u)$ to the right by one symbol, as described in step (vi) of the algorithm. Similarly, when $v \notin \mathcal{P}$ and $w \in \mathcal{C}$ as in Fig. 7, we can compute $\gamma(j, d, v, w)$ based on $\gamma(j, d-1, u, w)$ as described in (vii).

Careful examination of the search algorithm shows that its computational complexity is

$$O(LDM_1M^2) + O(LDM_2^2M), \tag{1}$$

which grows linearly with the length L of the entire sequence, or the size of the database. If we do not limit the maximum length to be D, the complexity will be $O(L^2 M_1 M^2) + O(L^2 M_2^2 M)$, which is not a linear function of L any more.

6. PREDICTION OF IRON RESPONSE ELEMENTS

In order to demonstrate the idea, we constructed a csHMM that can be used for finding the regions in the given DNA sequence, which are transcribed into *iron response elements* (IREs). IREs are found in the 5' or 3' UTRs of various messenger RNAs. It is known that the *iron regulatory proteins* (IRPs) bind to the IREs in order to control the iron metabolism inside the cell [24]. The IRE has a well-conserved hairpin structure that has either an interior loop or a bulge. The consensus secondary structures of the IREs are shown in Fig. 8. As shown in this figure, certain bases are especially well-conserved in the IREs. For example, the IREs have a loop that consists of six bases, which has the pattern "CAGWGH". In this pattern, W can be either A or U (T), and H can be either A, C or U (T). There can be non-canonical base-pairs in the stem such as the GU/UG pairs, and the lower stem can be of variable length.

We used the context-sensitive HMM in Fig. 9 to represent the IREs with conserved secondary structures. The lower stem is modeled using the pairwise-emission state P_1 and the context-sensitive state C_1. This state-pair is associated with a stack as shown in Fig. 9. Similarly, the upper stem is modeled by the state-pair (P_2, C_2), which uses a separate stack. Note that (P_1, C_1) and (P_2, C_2) are capable of representing stems of variable lengths. The loop and the bulge are modeled using single-emission states, since the bases in these parts do not form pairs with other bases. Each single-emission state S_n uses a different set of emission probabilities, in order to specify which base is conserved at each location.

Based on this context-sensitive HMM, we used the database search algorithm elaborated in Sec. 5 to find IREs in several DNA sequences. We chose four DNA sequences in the human genome that are known to contain functional IREs. These sequences have been previously used for testing the performance of *PatSearch*, a pattern matching program that can find functional elements with various patterns in DNA or protein sequences [25]. We ran the

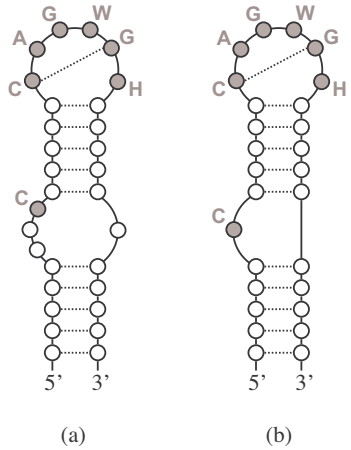


Fig. 8. The consensus secondary structures of the iron response elements (IREs).

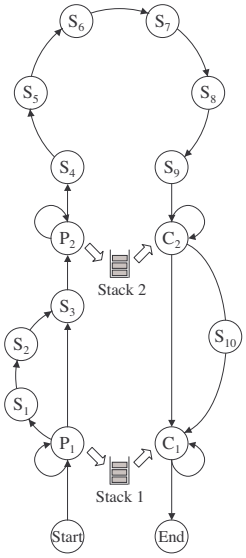


Fig. 9. A csHMM that represents the IREs.

search algorithm for finding high-scoring regions in the database. When there were overlaps between several high-scoring regions, only the one with the highest score was stored as a match. The search results are summarized in Table 2. The first column in in Table 2 shows the EMBL accession number and the second column shows the UTRdb ID of each DNA sequence [26]. The search results of the csHMM-based IRE finder are shown in the third column. The fourth column contains the prediction results of the PatSearch program. As summarized in Table 2, the csHMM-based IRE finder was able to find all the IREs in the given DNA sequences, and there were no false predictions. Interestingly enough, the csHMM-based approach was able to identify the start position and the end position of the IREs more precisely than the PatSearch. For example, in the second DNA sequence (accession number: Y09188), the proposed method predicted that there exists an IRE between 6 and 31, which matches the data in the UTRdb [26] and the Rfam database [27] exactly. Similarly, in the third sequence (accession number: D28463), the csHMM-based method predicted the location of the IRE to be between 32 and 59, which matches the data in the Rfam database. In the fourth sequence (accession number: J04755), the predicted location of the csHMM-based method was identical to the location stored in the Rfam database. In [25], it is reported that the PatSearch software has predicted the location of the IRE to be between 34 and 56, which is completely different from the true location. Since the sequence between 34 and 56 does not match the typical pattern of the IREs, the wrong position reported in [25] is probably a simple typo.

EMBL AC	UTRdb ID	csHMM	PatSearch
X60364	5HSA001988	13-35	13-35
Y09188	5HSA003829	6-31	8-30
D28463	5HSA003858	32-59	35-57
J04755	5HSA013930	951-978	34-56

Table 2. The database search result for finding IREs.

7. CONCLUDING REMARKS

In this paper, we considered the role of context-sensitive HMMs in building non-coding RNA gene finders. The csHMM is an extension of the traditional HMM, where some states are equipped with auxiliary memory. Emissions made at certain states are stored in the associated memory, and this serves as the context of the model, which affects the emission and transition probabilities of specific future states. This increases the descriptive power of the model tremendously, and as a result, csHMMs are capable of representing complex correlations between distant symbols, which are not possible using regular HMMs. The context-senstive HMMs can effectively model RNA sequences with various secondary structures, hence they can provide an efficient framework for building RNA analysis tools, especially, ncRNA gene finders. We introduced a database search algorithm that can be used with csHMMs, whose computational complexity increases only linearly with the size of the database that is to be searched. We demonstrated that the csHMM-based gene finder could achieve satisfactory prediction results, which makes the use of csHMMs for building ncRNA gene finders look very promising. Future work includes building a more flexible ncRNA gene-finders based on the proposed framework and extending the proposed algorithm for pseudoknots.

8. REFERENCES

[1] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Ra?, K. Roberts, and P. Walter, *Essential cell biology*, Garland Publishing Inc., New York, 1998.

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

[3] E. S. Lander, "Initial sequencing and analysis of the human genome", *Nature*, vol. 409, pp. 860-921, 2001.

[4] J. S. Mattick, "Non-coding RNAs: the architects of eukaryotic complexity", EMBO Reports, vol. 2, pp. 986-991, 2001.

[5] S. Gisela, "An expanding universe of noncoding RNAs", *Science*, vol. 296, pp. 1260-1263, 2002.

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

[7] G. Ruvkun, "Glimpses of a tiny RNA world", *Science*, vol. 294, pp. 797-799, 2001.

[8] A. Krogh, I. Saira Mian, D. Haussler, "A hidden Markov model that finds genes in E. coli DNA", *Nucleic Acids Res.*, vol. 22, pp. 4768-4778, 1994.

[9] S. L. Salzberg, A. L. Delcher, S. Kasif, O. White, "Microbial gene identification using interpolated Markov models", *Nucleic Acids Res.*, vol. 26, pp. 544-548, 1998.

[10] Byung-Jun 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.

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

[12] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences", CABIOS, vol. 13, no. 3, pp. 263-270, 1997.

[13] D. Anastassiou, "Genomic signal processing", *IEEE Signal Processing Magazine*, pp. 8-20, July 2001.

[14] P. P. Vaidyanathan and Byung-Jun Yoon, "Gene and exon prediction using allpass-based filters", *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics*, Raleigh, NC, Oct. 2002.

[15] P. P. Vaidyanathan and Byung-Jun Yoon, "The role of signal-processing concepts in genomics and proteomics", *Journal of the Franklin Institute*, vol. 341, pp 111-135, 2003.

[16] Y. Cai and C. Chen, "Artificial neural network method for discriminating coding regions of eukaryotic genes", *Comput. Appl. Biosci.*, vol. 11, pp. 497-501, 1995.

[17] S. R. Eddy, "Computational genomics of noncoding RNA genes", Cell 109 (2) (2002) 137-40.

[18] N. Chomsky, "On certain formal properties of grammars", *Information and Control*, vol. 2, pp. 137-167, 1959.

[19] 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.

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

[21] Byung-Jun Yoon and P. P. Vaidyanathan, "RNA secondary structure prediction using context-sensitive hidden Markov models", *Proc. International Workshop on Biomedical Circuits and Systems (BioCAS)*, Singapore, Dec. 2004.

[22] Byung-Jun Yoon and P. P. Vaidyanathan, "Optimal alignment algorithm for context-sensitive hidden Markov models", *Proc. 30th International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, Mar. 2005.

[23] Byung-Jun Yoon and P. P. Vaidyanathan, "Scoring algorithm for context-sensitive HMMs with application to RNA secondary structure analysis", *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics*, Newport, RI, May 2005.

[24] M. W. Hentze and L. C. Kuhn, "Molecular control of vertebrate iron metabolism: mRNA-basedregulatory circuits operated by iron, nitric oxide, and oxidativestress", *Proc. Natl. Acad. Sci.*, vol. 93, pp. 8175-8182, 1996.

[25] G. Pesole, S. Liuni, and M. D'Souza, "PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance", *Bioinformatics*, vol. 16, no. 5, pp. 439-450, 2000.

[26] G. Pesole, S. Liuni, G. Grillo, F. Licciulli, F. Mignone, and C. Saccone, "UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs.", *Nucleic Acids Res.*, vol. 30, pp. 335-340, 2002.

[27] 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.