RNA SECONDARY STRUCTURE PREDICTION USING CONTEXT-SENSITIVE HIDDEN MARKOV MODELS

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

It has been believed for decades, that proteins are responsible for most of the genetically important functions in all cells. Due to this reason, most of the research in molecular biology was focused on identifying genes that encode proteins, and their roles in the genetic network. Recent studies indicate that non-coding RNAs play important roles in various processes. Such ncRNA genes cannot be effectively identified using traditional gene-finders that aim at protein-coding genes. Many ncRNAs conserve their secondary structures as well as their primary sequences, which have to be taken into account when looking for ncRNA genes. In this paper, we propose a new method based on context-sensitive HMMs, which can be used for predicting RNA secondary structure. It is demonstrated that the proposed model can predict the secondary structure very accurately, at a low computational cost.

1. INTRODUCTION

For many decades, it has been generally accepted that the genetic information flows from DNA to RNA to protein. This has been the central dogma of biology, and most of the research has been focused on identifying genes that encode proteins. Proteins have been believed to perform most of the important functions in all cells, which range from structural and catalytic functions to genetic regulations. In the meantime, RNAs have been mainly viewed as an intermediary between DNAs and proteins, except for several infrastructural RNAs, such as the tRNAs and rRNAs that are used in the protein-coding machinery. Therefore, "genes" were almost synonymously used for protein-coding regions, and the vast majority in the genome that does not encode proteins has been regarded as "junk" that is practically information-less.

Recently, startling observations have been made regarding non-coding RNAs (ncRNAs) by numerous groups of biologists [1, 2, 3]. A number of evidences have been found, which show that the importance of ncRNAs have been underestimated. In fact, recent results indicate that ncRNAs that have evaded our eyes for a long time, constitute the majority of the genomic programming in the higher organisms [1]. It has been found that ncRNAs have important roles in various processes. They affect transcription and the chromosome structure, take part in RNA processing and modification, regulate mRNA stability and translation, and also affect protein stability and transport [2]. Until now, surprisingly many functions of ncRNAs have been found, but there are still countless ncRNAs whose functions are unknown. Moreover, although active research has unveiled many ncRNAs that were not previously known to us, the ncRNAs that have been identified until now are still considered to be the tip of an iceberg [4].

Given the huge amount of genomic data that is available these days, it is nearly impossible to identify all these ncRNAs by experimental means. During the last decade, computational sequence analysis has become very popular, and it has expedited the process of annotating protein-coding genes. A number of gene-finders have been proposed that were based on a variety of approaches, including hidden Markov models (HMM) [5, 6, 7], neural networks [8], digital filters [9], and so forth. A number of methods, especially those based on HMMs, have been quite successful in identifying protein-coding regions. However, none of the traditional methods can be directly applied to predicting ncRNA genes. Many interesting RNAs conserve their secondary structures more than they conserve their primary sequences [7, 10]. In most organisms, ncRNA genes do not display strong sequence composition biases [10], which is the reason why the traditional approaches that are mainly based on base composition statistics simply fail. Therefore, in order to predict ncRNA genes effectively, we have to consider the primary sequence and the secondary structure of an RNA at the same time.

Until now, several RNA gene-finders have been built using stochastic context-free grammars (SCFGs) [11, 12]. In this paper, we propose a new method that can be used for modeling and predicting RNA secondary structures. The proposed method is based on *context-sensitive HMMs* [13] and it has certain advantages over the SCFG-based one. It can provide an efficient framework for building gene-finders that can search for ncRNA genes.

2. RNA SECONDARY STRUCTURE

RNA is a nucleic acid that consists of a sequence of nucleotides A, C, G and U[1]. The nucleotide A forms a hydrogen bonded pair with U, and C forms a pair with G, which are called *complementary base-pairs*. RNA is generally a single-stranded molecule, and it typically folds onto itself to form consecutive 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*. An example of a simple RNA secondary structure is shown in Fig. 1. This kind of structure is called a *stem-loop* or a *hairpin*, and it is frequently observed in various RNAs. As can be seen in this example, there are pairwise interactions 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.

0-7803-8665-5/04/$20.00 ©2004 IEEE

[1]U is chemically similar to T in DNA.

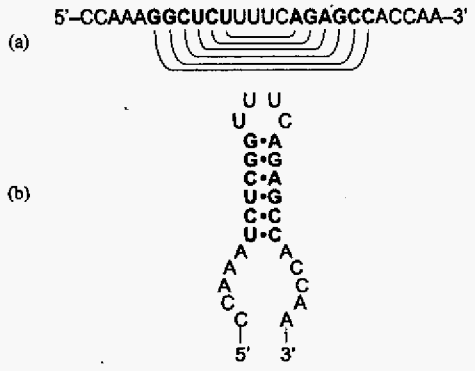

Fig. 1. The 3' end of a histone mRNA. (a) Primary sequence before folding. The lines indicate interactions between base-pairs. (b) Consensus secondary structure.

Such RNAs with nested pairwise interactions are in principle similar to palindromes, which are sequences that read the same forwards and backwards. The palindrome language, which consists of all possible palindromes, is a classic example of languages that cannot be represented using the so-called *regular grammars* in the Chomsky hierarchy of transformational grammars [14]. Hidden Markov models can be viewed as stochastic regular grammars, which indicates that HMMs cannot describe palindrome languages. It is of course possible that regular grammars - hence, HMMs - generate palindromes as part of their language. However, they are not capable of generating *only* such palindromes, thus not able to effectively differentiate palindromic sequences from non-palindromic ones. In order to describe such languages, we have to use higher-order grammars such as the *stochastic context-free grammars* (SCFGs) [7, 15].

Until now, several SCFG-based methods have been introduced [11, 12] that can identify tRNAs with high accuracy. One major disadvantage of the SCFG-based approaches is their high computational cost, which renders the prediction of large RNAs infeasible. Instead of using SCFGs, we can use *context-sensitive HMMs* that have been recently introduced [13]. The context-sensitive HMM is capable of modeling pairwise interactions between distant bases in RNAs. It has the advantage that the states in the model directly correspond to the base locations in an RNA sequence. Moreover, it can be easily incorporated into existing gene-finders that are based on HMMs. In addition to this, context-sensitive HMMs have efficient algorithms for finding the optimal alignment [16] and computing the probability [17] of a given observation sequence. The computational complexity of these algorithms is $O(L^2M^3)$ for sequences with a single nested structure, which is smaller than $O(L^3M^3)$ of the algorithms for general SCFGs [15], where L is the length of the sequence and M is the number of different states (non-terminals).

3. CONTEXT-SENSITIVE HIDDEN MARKOV MODELS

In this section, we briefly describe the concept of context-sensitive HMMs. The context-sensitive HMM can be viewed as an extension of the traditional HMM, where some of the states are equipped with auxiliary memory. Symbols that are emitted at certain states are stored in the memory, and they serve as the *context* that affects the emission and transition probabilities of the model. There are three different classes of states, which are the *single-emission state* S_n, the *pairwise-emission state* P_n and the *context-sensitive state* C_n. P_n and C_n always exist in pairs. For example, if there are two pairwise-emission states P_1 and P_2 in the model, then the model should also have two context-sensitive states C_1 and C_2. Each pair (P_n, C_n) is associated with an auxiliary memory, such as a stack or a queue. Figure 2 shows examples of the triplets of P_n, C_n and the n-th memory element.

The single-emission state S_n is identical to the states in traditional HMMs. As we enter S_n, a symbol is emitted according to its emission probabilities. After the emission, a transition is made to the next state, following the specified transition probabilities. The pairwise-emission state P_n is similar to the single-emission state S_n, except that the emitted symbol is stored in the associated memory. The data stored in the memory affects the emission probabilities and the transition probabilities of C_n, in the future. Once the emitted symbol is stored, it makes a transition to the next state according to the transition probabilities associated with P_n. The context-sensitive state C_n is quite different from the others, in the sense that its emission probabilities and the transition probabilities are not fixed. In fact, these probabilities are affected by the *context*, or the data stored in the associated memory element. This is the reason why C_n is called a context-sensitive state. When we enter C_n, the associated memory is accessed and a symbol x is retrieved. Once the symbol is retrieved, 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). Another distinctive character of the context-sensitive state C_n is that the transition to the state C_n is not always allowed. Let us consider the case when some state attempts to make a transition to C_n. Before making the transition, the auxiliary memory that is associated with C_n is examined first. If the memory is empty, the transition is not allowed, and it is forced to make a transition to another state. This is done by setting the transition probability to C_n to zero and adjusting the remaining probabilities in a corresponding manner. This restriction is necessary to maintain the same number of P_n and C_n in a state sequence. Let $\mathbf{s} = s_1 s_2 \dots s_L$ be a feasible state sequence of an observed symbol string $\mathbf{x} = x_1 x_2 \dots x_L$. Then the number of occurrence of P_n in $\mathbf{s}$ is kept the same as the number of occurrence of C_n in $\mathbf{s}$ by the previous restriction. This is reasonable, since if there are more C_n states than there are P_n states,

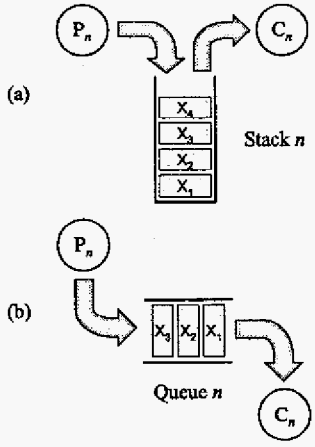

Fig. 2. Examples of the triplets of P_n, C_n, and the associated memory element. (a) When using a stack. (b) When using a queue.

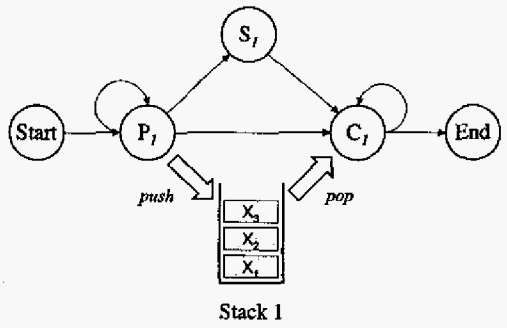

Fig. 3. An example of a context-sensitive HMM that generates only palindromes.

the emission probabilities of the context-sensitive state C_n cannot be properly determined. On the other hand, if there are more P_n states than C_n states, the symbols that were emitted by surplus P_n states do not affect the probabilities at all, hence they may be simply replaced by single-emission states.

Based on the described context-sensitive HMM, we can easily construct a simple model that generates only palindromes. For example, we may use the structure shown in Fig. 3. In this model, the auxiliary memory that is associated with the pair (P_1, C_1) is a stack. Initially, the model begins at the pairwise-emission state P_1. It makes several self-transitions to generate a number of symbols, which are pushed onto the stack. At some point, it moves to the context-sensitive state C_1. Once we enter the context-sensitive state C_1, the emission probabilities and the transition probabilities of C_1 are adjusted, such that the state always emits the symbol on the top of the stack and makes self-transitions until the stack becomes empty. In this way, C_1 emits the same symbols as were emitted by P_1, but in the reverse order. If we denote the number of symbols that were emitted by P_1 as N, the generated string will always be a palindrome of the form $x_1 \cdots x_N x_N \cdots x_1$ or $x_1 \cdots x_N x_{N+1} x_N \cdots x_1$.

Since the probabilities in the context-sensitive HMM change according to the context of the system, those algorithms that were used in traditional HMMs, such as the Viterbi's algorithm for finding the optimal state sequence and the forward algorithm for computing the probability of an observation sequence, cannot be used. However, there exist efficient algorithms for finding the optimal alignment [16] and computing the probability of observed symbol strings [17] for context-sensitive HMMs. These algorithms can be used for models with a single nested structure, in which case they are faster than the algorithms[2] for general SCFGs [15].

4. PREDICTING RNA SECONDARY STRUCTURE

Context-sensitive HMMs can effectively describe the pairwise interactions between bases that are remote from each other, while taking the dependencies between adjacent bases into account. This makes the context-sensitive HMM ideal for modeling RNA sequences with conserved secondary structures. In order to demonstrate this, let us construct a model for predicting the secondary structure of a short RNA with a conserved hairpin structure. One interesting RNA family that conserves such a structure is the histone mRNA. Typically, metazoan mRNAs end in the so-called poly(A) tail. However, the replication-dependent histone mRNA is an exception, which terminates with a conserved 26-nucleotide that contains a 16-nucleotide stem-loop [18]. Note that this region is not translated into a protein. The 3' end of the histone mRNA plays a critical role for binding to the stem-loop binding protein (SLBP), and it is known to affect the nucleocytoplasmic transport of the mRNA and the cytoplasmic regulation [18]. The consensus secondary structure of the 3' end of histone mRNAs is depicted in Fig. 1 (b).

We represent this consensus structure using a simple context-sensitive HMM as shown in Fig. 4. In this model, S_1 corresponds to the 5' flanking region and S_3 corresponds to the 3' flanking region. The pair (P_1, C_1) generates the stem part of the secondary structure and S_2 generates the loop. After constructing this model, we computed the model parameters based on a number of RNA sequences whose secondary structures were already known. We used the 65 seed alignments in the Rfam database [19] that consist of hand-curated alignments of known members of the RNA family. Given these alignments, each RNA sequence can be readily aligned to the context-sensitive HMM in Fig. 4. Now that we know the state sequence of each RNA in the training set, we can count the number of emission and transition events at respective states. The observed counts can be used to estimate the emission probabilities as well as the transition probabilities of the model. In order to make the prediction performance of the model more robust against small errors in the sequences, we may add *pseudo-counts* to the observed counts. Essentially, this corresponds to using mean posterior estimation by incorporating a prior from the Dirichlet distributions [7].

Once we have trained the context-sensitive HMM, we used the same model for predicting the secondary structure of other RNA sequences. For testing purpose, we first collected 500 sequences from the Rfam database of histone 3' UTR stem-loop, and removed sequences that had ambiguous bases. We also removed sequences that were included in the training set. The final test set contained 461 RNA sequences. For each sequence in the training set, we used the dynamic programming algorithm in [16] to find the optimal state sequence. The predicted optimal alignment is compared with the true alignment of the test sequence to evaluate the performance of the model. In order to assess the prediction accuracy, we first counted the number of true-positives (TP), the number of false-positives (FP) and the number of false-negatives (FN). True-positives are defined as the base-pairs that are correctly

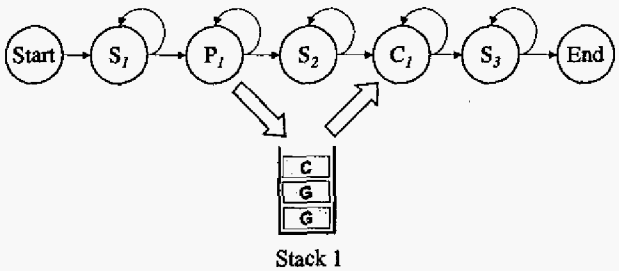

Fig. 4. Context-sensitive HMM that models the 3' end of histone3 mRNA. This model can generate sequences that fold onto themselves to form a stemp-loop.

[2]CYK algorithm can be used for finding the optimal alignment and the inside algorithm can be used for computing the probability of an observed sequence (the scoring problem).

Fig. 5. Example of a typical false-positive (shown in dotted line).

predicted. False-positives are the predicted base-pairs that do not form pairs in the trusted alignment. Finally, false-negatives are the base-pairs in the trusted alignment that could not be predicted. From these values, the sensitivity (SN) and the positive predictive value (PPV) of them system were computed as follows

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}.$$

Considering the simplicity of the model that we have used, the prediction accuracy was surprisingly high. The sensitivity and the specificity of the constructed prediction system were

$$\mathrm{SN} = 0.9912, \quad \mathrm{PPV} = 0.9599.$$

It can be seen from above, that the number of false-negatives were negligible, resulting in a sensitivity that is close to unity. Most of the false-positives occurred inside the loop, where the predictor formed an additional base-pair between U and A as shown in Fig. 5. The dotted line that connects the base U and A in the loop indicates the incorrect base-pair. This kind of errors can be easily prevented by imposing a restriction on the loop size, such that it has at least four nucleotides. This improves the specificity to $\mathrm{PPV} = 0.9934$. The model in Fig. 4 can be improved further, if we also account for bulges in the stem by introducing more states. We may also build a gene-finder based on a similar model, that can search for unknown histone 3' stem-loops in DNA sequences that have not been annotated yet.

5. CONCLUDING REMARKS

The context-sensitive HMM is an efficient tool that can reflect pairwise interactions between distant symbols effectively. As demonstrated in section 4, they are suitable for modeling and predicting RNAs that conserve secondary structures as well as primary sequences. By building a good model that closely represents the consensus structure of the RNA family that is of our interest, we can predict the secondary structure of an unannotated RNA sequence with high accuracy. The context-sensitive HMM has efficient algorithms for finding the optimal alignment and scoring observed symbol strings, which make it possible to build practical systems based on the model. Context-sensitive HMMs can also be used for modeling pseudoknots, where the pairwise interactions between bases are allowed to cross each other. Pseudoknots are found in many interesting RNAs, and identifying these pseudoknots is important in several applications, such as the three-dimensional structure prediction. One interesting problem would be how to modify the alignment and scoring algorithms in [16, 17] to account for such pseudoknots.

6. ACKNOWLEDGMENTS

This work was supported in part by the ONR grant N00014-99-1-1002, USA.

7. REFERENCES

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

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

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

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] S. R. Eddy, "Computational genomics of noncoding RNA genes", *Cell*, vol. 109, pp. 137-40, 2002.

[11] 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.

[12] 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.

[13] Byung-Jun Yoon and P. P. Vaidyanathan, "HMM with auxiliary memory: a new tool for modeling RNA secondary structures", Proc. 28th Asilomar Conference on Signals, Systems, and Computers, Monterey, CA, Nov. 2004.

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

[15] 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.

[16] Byung-Jun Yoon and P. P. Vaidyanathan, "Optimal alignment algorithm for context-sensitive hidden Markov models", in submission.

[17] Byung-Jun Yoon and P. P. Vaidyanathan, "Dynamic programming algorithm for scoring context-sensitive HMMs", in submission.

[18] Z. Dominski and W. F. Marzluff, "Formation of the 3' end of histone mRNA", *Gene*, vol. 239, pp. 1-14, 1999.

[19] 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.