Scoring Algorithm for Context-Sensitive HMMs with Application to RNA Secondary Structure Analysis

Byung-Jun Yoon and P. P. Vaidyanathan
California Institute of Technology

***Abstract*— During the last decade, a number of evidences have been found that non-coding RNAs (ncRNA) are involved in various important processes. Many of these ncRNAs are known to conserve their secondary structure, which gives rise to complex dependencies between distant bases in the primary sequence. Therefore, we need more complex models than the traditional HMM in order to analyze ncRNAs. Recently, context-sensitive HMMs (csHMM) have been proposed for modeling RNA secondary structures. In this paper, we propose a dynamic programming algorithm that can be used for scoring csHMMs.**[1]

I. Introduction

It has been believed for a long time that proteins are responsible for most of the important biological functions in all cells. As a natural consequence, most of the research in genomics has been focused on protein-coding genes and their functions. In the meanwhile, RNA has been mainly viewed as a passive intermediary between DNA and protein, except for several infrastructural RNAs such as the tRNA and the rRNA.

However, a number of evidences have been found during the last decade, which showed that many non-coding RNAs (ncRNAs), which are RNAs that do not encode proteins, are involved in various biologically important processes [1], [2]. These days, it is even claimed that ncRNAs constitute the majority of genomic programming in the higher organisms such as mammals [3].

One interesting characteristic of many ncRNAs is that they conserve their secondary structure more than they conserve their primary sequence [4]. Figure 1 shows an example of an RNA with a simple secondary structure. As shown in Fig. 1, the single-stranded ncRNA can fold onto itself to form consecutive complimentary base-pairs, and the resulting structure is called the RNA secondary structure. Due to this structure, there exist pairwise dependencies between distant bases in the primary sequence. Such dependencies cannot be described using simple models including HMMs, and we need more complex models with greater descriptive power such as the *stochastic context-free grammar (SCFG)*. Until now, several methods have been proposed based on SCFGs for modeling RNA secondary structures [4], [5].

II. Context-Sensitive HMM

Recently, the concept of *context-sensitive HMM (csHMM)* has been proposed, which can be alternatively used for modeling RNA sequences with conserved secondary structures [6]. It is 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 the stored data serves as the context which affects the emission and transition probabilities of the model. In this way, the csHMM is capable of modeling sequences with complex dependencies. Examples of csHMMs that can generate sequences with various secondary structures can be found in [6].

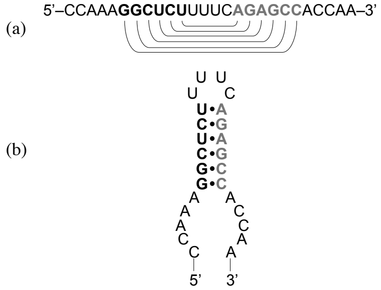

Fig. 1. (a) Primary sequence. Lines indicate pairwise dependencies between bases. (b) Secondary structure of the RNA.

Due to the context-sensitive characteristic of the csHMM, traditional algorithms that were used with regular HMMs cannot be used any more. In the following section, we describe a dynamic programming algorithm that can be used for computing the probability that the observed symbol string is generated by the given csHMM. This problem is traditionally called the ***scoring problem***, and the proposed algorithm can be viewed as the counterpart of the forward algorithm in regular HMMs.

III. The Scoring Algorithm

Let us first define the variables that are needed to describe the algorithm. We denote the observation sequence as $\mathbf{x} = x_1 x_2 ... x_L$, where L is the length of the sequence. s_n denotes the underlying state of the n-th symbol x_n. The csHMM is assumed to have M distinct states, which we denote by $1, 2, \ldots, M$. We assume that there are N pairs (P_n, C_n) of pairwise-emission states and context-sensitive states, where each pair is associated with a different stack. These $2N$ states are included in the set of M states $\{1, 2, \ldots, M\}$. It is assumed that all pairwise interactions between P_n states and C_n states are nested as shown in Fig. 2 (a) and they do not cross each other. In addition to this, we assume that multiple nested interactions as shown in Fig. 2 (b) are not allowed. For notational convenience, let us define the following sets $\mathcal{P} = \{P_1, \ldots, P_N\}$ and $\mathcal{C} = \{C_1, \ldots, C_N\}$.

[1] This work was supported in parts by the ONR grant N00014-99-1-1002 USA and the Microsoft Research fellowship.

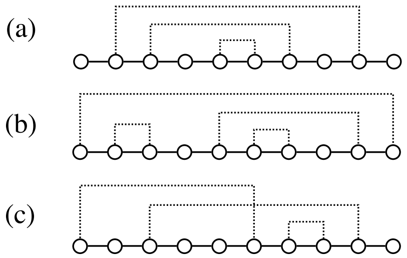

Fig. 2. (a) Single nested interactions. (b) Multiple nested interactions. (c) Crossing interactions.

The transition probability from state v to w is denoted as $t(v,w)$, and the probability that a symbol x will be emitted at a state v is denoted as $e(x|v)$. Now, let us define $\alpha(i,j,v,w)$ to be the probability of all paths $s_i \cdots s_j$ with $s_i = v$ and $s_j = w$, where every pairwise-emission state P_n in the path is paired with the corresponding context-sensitive state C_n. The variable $\alpha(i,j,v,w)$ will finally lead to the probability $P(\mathbf{x}|\Theta)$ that the observed symbol string $\mathbf{x}$ is generated by the given context-sensitive HMM with the set of parameters Θ. Now, the algorithm can be described as follows.

Initialization

For $i = 1, \ldots, L, v = 1, \ldots, M$.

$$\alpha(i,i,v,v) \;=\; \begin{cases} e(x_i|v) & v \notin \mathcal{P}, \mathcal{C} \\ 0 & \text{otherwise} \end{cases}$$

Iteration

For $i = 1, \ldots, L-1, j = i+1, \ldots, L$ and $v = 1, \ldots, M, w = 1, \ldots, M$.

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

$$\alpha(i,j,v,w) = 0$$

(ii) $v = P_n, w = C_n, j = i+1$

$$\alpha(i,j,v,w) = e(x_i|v)t(v,w)e(x_j|w)$$

(iii) $v = P_n, w = C_n, j \neq i+1$

$$\alpha(i,j,v,w) = \sum_{u_1,u_2} \Big[e(x_i|v)t(v,u_1) \times \alpha(i+1,j-1,u_1,u_2)t(u_2,w)e(x_j|w) \Big]$$

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

$$\alpha(i,j,v,w) = \sum_{u} \Big[\alpha(i,j-1,v,u)t(u,w)e(x_j|w) \Big]$$

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

$$\alpha(i,j,v,w) = \sum_{u} \Big[e(x_i|v)t(v,u)\alpha(i+1,j,u,w) \Big]$$

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

We can use either (iv) or (v).

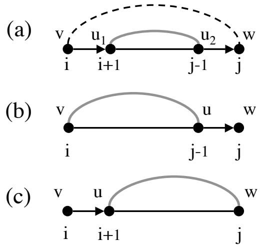

Fig. 3. Illustration of the iteration step of the scoring algorithm.

Termination

$$P(\mathbf{x}|\Theta) \;=\; \sum_{v,w} t(0,v)\alpha(1,L,v,w)t(w,0) \qquad \square$$

Note that $t(0,v)$ is the probability that the model will start at v, and $t(w,0)$ is the probability that the model will terminate after w. We can see that the probability $\alpha(i,j,v,w)$ is computed iteratively, starting from the inside to the outward direction. Every-time there is an interaction between s_i and s_j, they are considered together as shown in (ii) and (iii) of the iteration step. In this way, we can know which symbol was emitted at the pairwise-emission state, and therefore we can decide the probabilities of the corresponding context-sensitive state. This is illustrated in Fig. 3 (a). The dashed line denotes the interaction between $s_i = v$ and $s_j = w$. Since P_n and C_n exist in pairs inside $s_i \cdots s_j$, and as s_i is paired with s_j, all P_n and C_n states inside $s_{i+1} \cdots s_{j-1}$ must also exist in pairs. This is indicated by the grey line that connects the symbols at $i+1$ and $j-1$ in Fig. 3 (a). Therefore, the probability of $\alpha(i,j,v,w)$ can be computed as in (ii) of the iteration step. Similarly, Fig. 3 (b) and (c) respectively illustrate (iv) and (v) of the iteration step. It is not difficult to see that the overall computational complexity of the proposed algorithm is $O(L^2M^3)$.

IV. Conclusion

The context-sensitive HMM is an effective tool for modeling and analyzing RNAs with conserved secondary structures [6]. They have been applied for predicting the secondary structure of simple RNAs, and the results look very promising [7]. It would be interesting too build ncRNA gene-finders based on csHMMs, and it will be a topic for future research.

References

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

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

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

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

[5] 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.

[6] 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.

[7] 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.