FAST SEARCH OF SEQUENCES WITH COMPLEX SYMBOL CORRELATIONS USING PROFILE CONTEXT-SENSITIVE HMMS AND PRE-SCREENING FILTERS

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

Recently, profile context-sensitive HMMs (profile-csHMMs) have been proposed which are very effective in modeling the common patterns and motifs in related symbol sequences. Profile-csHMMs are capable of representing long-range correlations between distant symbols, even when these correlations are entangled in a complicated manner. This makes profile-csHMMs an useful tool in computational biology, especially in modeling noncoding RNAs (ncRNAs) and finding new ncRNA genes. However, a profile-csHMM based search is quite slow, hence not practical for searching a large database. In this paper, we propose a practical scheme for making the search speed significantly faster without any degradation in the prediction accuracy. The proposed method utilizes a pre-screening filter based on a profile-HMM, which filters out most sequences that will not be predicted as a match by the original profile-csHMM. Experimental results show that the proposed approach can make the search speed eighty times faster.

Index Terms— homology search, profile-csHMM, pseudoknot, noncoding RNA (ncRNA), context-sensitve HMM (csHMM).

1. INTRODUCTION

Modeling the common patterns and motifs in a set of related symbol sequences has been a problem of practical importance in various applications. The statistical model that reflects the key features of the given set can be used for finding similar sequences in a large database. This approach has been especially popular in computational biology, where it has been used for identifying new members in a known biological sequence family, such as protein-coding genes and noncoding RNAs (ncRNAs) [1]. This is typically called a *similarity search* or a *homology search*, and it has played a crucial role in the fast annotation of various genomes that have been obtained as a result of many genome sequencing projects.

A typical way of performing a similarity search is as follows. Firstly, we align the given symbol sequences based on their similarity. There exist various algorithms that can find a reasonably good multiple sequence alignment in an efficient manner [2]. The resulting alignment reveals the regions that are well conserved among different sequences, and we can also estimate the observation probabilities of different symbols at distinct positions. Based on the multiple alignment, we can build a statistical model that represents the "consensus sequence", or a "probabilistic profile", of the given sequence family. Once we have constructed the statistical model, it can be used to search a large database in order to find high-scoring regions, which are candidates for new members in the given sequence family.

Profile hidden Markov models (profile-HMMs), which are a subclass of HMMs with a linear repetitive structure, have been especially popular in building probabilistic profiles of protein sequences and protein-coding genes [1, 3]. Profile-HMMs are well-known for their efficiency in modeling the correlations between adjacent symbols, and they can be easily constructed from multiple sequence alignments. As a result, many protein-coding gene-finders are built on profile-HMMs and other variants of HMMs.

One problem of profile-HMMs is that their application is limited to sequences with a linear correlation structure. Correlations between distant symbols that are intertwined in a complicated way cannot be described using HMMs, as they do not satisfy the Markov property. For example, many ncRNAs have symbol correlations that appear in a nested manner, and sometimes, they even contain correlations that cross each other [6]. These sequences cannot be effectively represented by profile-HMMs. However, we can use the *profile context-sensitive HMMs* (profile-csHMMs) in such cases, which have been recently proposed [4]. Unlike most conventional models, including profile-HMMs and stochastic context-free grammars (SCFGs) that have been especially popular in computational biology [1], profile-csHMMs are capable of modeling *any* kind of pairwise symbol correlations. To the best of our knowledge, they are the first statistical model that can be practically used for representing and recognizing any kind of RNA pseudoknots.[1]

Unfortunately, the decoding algorithm for profile-csHMMs has a relatively high computational complexity due to the large descriptive power of the model [4]. This makes the profile-csHMM impractical for searching a large database, as the amount of time it takes for searching a huge database (e.g. the human genome has around three billion bases) can be prohibitively large.

In this paper, we propose a practical method for expediting a profile-csHMM based search. The proposed approach uses an efficient pre-screening filter based on a profile-HMM, which is constructed from the original profile-csHMM. This pre-screening filter will eliminate most sequences that will not be predicted as a "match" by the original profile-csHMM. Only a small fraction of sequences that passes this filter will be handed over to the profile-csHMM in the second stage. This can make the search significantly faster without sacrificing the prediction accuracy. The paper is organized as follows. We begin with a brief review of profile-csHMMs in the following section. Then we describe the proposed scheme in Sec. 3, and elaborate on how we should construct the pre-screening filter. Experimental results are shown in Sec. 4 to demonstrate the proposed idea.

Work supported in parts by the NSF grant CCF-0636799 and the Microsoft Research Graduate Fellowship.

[1] Pseudoknots are RNA sequences with crossing correlations [1].

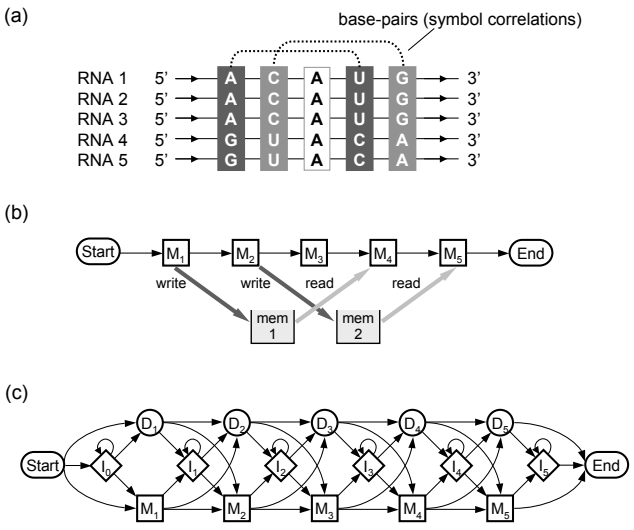

Fig. 1. An example of a profile-csHMM. (a) Multiple sequence alignment of five RNA sequences. Note that these RNAs have two base-pairs. (b) An ungapped profile-csHMM constructed from the alignment. (c) The final structure of the profile-csHMM that allows additional insertions and deletions at any location.

2. REVIEW OF PROFILE CONTEXT-SENSITIVE HMM

As the profile-HMMs are a subclass of HMMs with a linear repetitive structure, profile context-sensitive HMMs [4] are a subclass of context-sensitive HMMs (csHMMs), whose structure is similar to that of profile-HMMs. Context-sensitive HMMs are extensions of conventional HMMs, which have variable emission and transition probabilities that depend on the context [5]. This context-dependency increases the descriptive capability of the model significantly. Profile-csHMMs repetitively use three kinds of states, i.e., *match states* M_k, *insert states* I_k, and *delete states* D_k, to represent the distinct symbol emission probabilities at different locations and to model insertions and/or deletions at any position. In order to see how a profile-csHMM works, let us consider constructing a profile-csHMM from the sequence alignment shown in Fig. 1 (a). In this example, five RNAs are aligned to each other.[2] For simplicity, we assume that all RNAs have the same length and there is no gap in the alignment.

2.1. Constructing an ungapped profile-csHMM

As the average length of the RNAs is five, we first construct an *ungapped* profile-csHMM using five match states $M_1, M_2, \ldots, M_5$. This is illustrated in Fig. 1 (b). Each match state M_k represents the relative occurrences of distinct symbols at the k-th position. For example, as all the RNAs have an 'A' in the third position, we can adjust the emission probability of M_3 such that is emits 'A' with a high probability (or possibly, with probability one). One interesting thing that we can see in Fig. 1 (a) is that there exist pairwise correlations between non-adjacent symbols. For example, if there is an 'A' in the first position, it is followed by a 'U' in the fourth position, and if there is a 'G' in the first position, there will be a 'C' in the fourth position.[3] Such pairwise correlations between distant symbols are frequently observed in RNA sequences due to the so-called *RNA secondary structures* [1, 6]. In order to model the correlation between the first symbol and the fourth symbol, we use a *pairwise-emission state* for M_1 and a *context-sensitive state* for M_4. When we enter M_1, it emits a symbol according to the specified emission probabilities and stores the symbol in the auxiliary memory dedicated to the state-pair M_1 and M_4. Afterwards, when we enter the corresponding context-sensitive state M_4, it first reads the symbol x stored in the memory. The emission probabilities of M_4 are adjusted based on the value of x, such that it emits the complimentary symbol of x. Similarly, we use a pairwise-emission state for M_2 and a context-sensitive state for M_5 to model the correlation between the second and the fifth symbols. As the symbol in the third position is not correlated to any other symbol, we use a *single-emission state* for M_3.

2.2. Modeling insertions and deletions

In order to allow additional insertions and deletions in the original alignment, we add insert states I_k and delete states D_k to the ungapped model that has been obtained from the alignment. The insert state I_k is used to represent the case when a symbol is inserted between positions $k-1$ and k in the original alignment. We use a single-emission state for I_k, since inserted symbols are usually not correlated to any other symbol. Unlike the match states and the delete states, the insert states are allowed to make self-transitions in order to model multiple insertions. The delete state D_k is used to represent the case when the k-th symbol in the original alignment is not present in the observed symbol sequence. Sometimes, the observation sequence may be shorter than the average length of the alignment that was used to construct the profile-csHMM. In such cases, there will be one or more gaps when we align the observed sequence to the given alignment. Each of these gaps are modeled using the delete states. Note that D_k is a non-emitting state that is simply used as a place-holder to interconnect other states. Fig. 1 (c) shows the final structure of the profile-csHMM after adding the insert states and delete states to the ungapped profile-csHMM in Fig. 1 (b).

2.3. Searching for similar sequences

Once we have constructed the profile-csHMM that reflects the common characteristics of the sequences in the given alignment, we can use this model to look for 'similar' sequences in a database. An essential problem in performing a similarity search is how we can quantitatively measure the similarity between a new observed sequence and the statistical model at hand. A widely used approach is to compute the optimal probability of the observation based on the given model, and use it as a similarity measure. Let $\mathbf{x} = x_1 \ldots x_L$ be an observed symbol sequence and let us denote its underlying state sequence as $\mathbf{y} = y_1 \ldots y_{L_s}$. Note that the length of the state sequence L_s can be larger than the length L of the observed sequence, when there exist deleted symbols. We also define Θ, which is the set of model parameters of the profile-csHMM at hand. The similarity score $S(\mathbf{x}, \Theta)$ between the observation $\mathbf{x}$ and the profile-csHMM can be computed as follows

$$S(\mathbf{x}, \Theta) = \max_{\mathbf{y}} S(\mathbf{x}, \mathbf{y}|\Theta) = S(\mathbf{x}, \mathbf{y}^*|\Theta), \qquad (1)$$

where $S(\mathbf{x}, \mathbf{y}|\Theta)$ is the score for $\mathbf{x}$ whose underlying state sequence is $\mathbf{y}$. Note that $\mathbf{y}^*$ is the optimal state sequence that maximizes the

[2] An RNA can be simply viewed as sequence of four symbols (or bases) A, C, G, and U, which are read from the so-called 5'-end to the 3'-end.

[3] A and U (and also C and G) can form a hydrogen-bonded base-pair, hence these bases are said to be complementary to each other.

similarity score $S(\mathbf{x}, \mathbf{y}|\Theta)$. If this similarity score is larger than a predefined threshold λ such that $S(\mathbf{x}, \Theta) \geq \lambda$, we can view the observed sequence as a good candidate that is likely to be a new member of the same sequence family. On the contrary, if $S(\mathbf{x}, \Theta) < \lambda$, we can conclude that $\mathbf{x}$ is unlikely to be a member of the given family. Therefore, when we search a database to find new members, only those sequences that satisfy $S(\mathbf{x}, \Theta) \geq \lambda$ will be reported as a "match".

When using profile-csHMMs to represent sequence families, we can utilize the *sequential component adjoining (SCA) algorithm* [4] for finding $\mathbf{y}^*$ and computing $S(\mathbf{x}, \Theta)$. The SCA algorithm is the counterpart of the Viterbi algorithm, which can be used for decoding profile-csHMMs. The computational complexity of the SCA algorithm is variable, and it depends on the correlation structure of the profile-csHMM [4]. For example, the complexity for computing $S(\mathbf{x}, \Theta)$ for typical RNA pseudoknots ranges between $O(L^4)$ and $O(L^6)$, which can be very large for long RNA sequences.

3. FAST SEARCH USING PRE-SCREENING FILTERS

One advantage of using profile-csHMMs in a similarity search is the increased specificity. When computing the similarity score, profile-csHMMs combine contributions from sequence similarity as well as structural similarity (in terms of symbol correlations). This makes it possible to reject false candidates that look similar to the reference sequences in the sequence-level, but do not preserve the original correlation structure.

However, when performing a similarity search, there will be typically many sequences that look very different from the reference sequences in the sequence-level, such that their similarity scores cannot exceed the threshold λ even after combining the contributions from their structural similarity. As the measure of sequence-level similarity can be quickly computed using a simpler model, such as the profile-HMM, we do not have to use a profile-csHMM in such cases.

Based on this observation, we propose a practical strategy that can make the database search much faster, compared to the search based on profile-csHMM alone.[4] The proposed approach is as follows. In the first place, we construct a pre-screening filter using a profile-HMM. The profile-HMM will have the same size as the original profile-csHMM, and its model parameters Θ_p will be derived from the parameters Θ of the profile-csHMM. When given a new observation sequence, we first compute the sequence-level similarity score $S_p(\mathbf{x}, \Theta_p)$ using the pre-screening filter. This score is compared with a new threshold λ_p to decide whether the overall similarity score $S(\mathbf{x}, \Theta)$ can exceed the original threshold λ after combining the contributions from the structural similarity. If this is possible, the sequence is handed over to the full-blown profile-csHMM to compute $S(\mathbf{x}, \Theta)$. Otherwise, the observation will be rejected. The overall algorithm is illustrated in Fig. 2.

3.1. Constructing the pre-screening filter

Now the question is how to choose the model parameters of the profile-HMM and how we should choose the threshold λ_p so that there will be no degradation in the prediction accuracy. In order to answer this question, let us first define several notations. Firstly, we

[4]This is conceptually similar to the approach proposed in [7] that was used to expedite a CM (covariance model) search. CMs can be viewed as profile-SCFGs (stochastic context-free grammars) that can represent nested correlations but not crossing correlations [1].

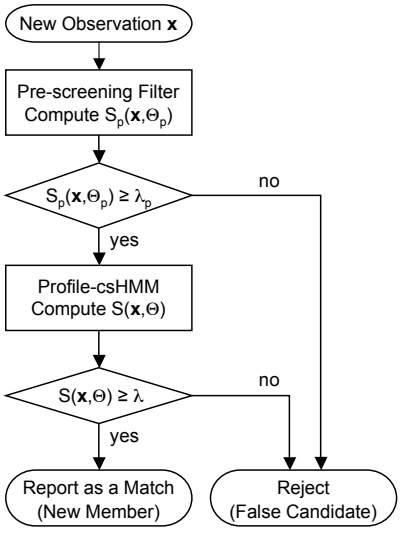

Fig. 2. Illustration of the proposed algorithm.

define the set $\mathcal{K} = \{k|\, M_k \text{ is a context-sensitive state}\}$. At single-emission states and pairwise-emission states, we denote the emission score of a symbol x at state v as $s_e(x|v)$. At context-sensitive states, the emission score of a symbol x_c at state v is denoted as $s_e(x_c|v, x_p)$, where x_p is the symbol that was previously emitted at the corresponding pairwise-emission state. A typical choice of the emission scores would be the logarithm of the emission probabilities, but we can also use other scoring schemes. The transition score from state v to state w is defined as $s_t(v, w|m)$ for $w = D_k, M_k, I_{k-1}$, where $k \in \mathcal{K}$. The variable $m \in \{0, 1\}$ indicates whether the memory associated with the context-sensitive state M_k is empty ($m = 0$) or not ($m = 1$). For all other w, the transition score is simply defined as $s_t(v, w)$. A typical choice for the transition scores would be the logarithm of the transition probabilities, but we can also use other scores.

Based on the parameters of the original profile-csHMM defined above, we choose the parameters of the pre-screening filter as follows. In the first place, the emission scores are chosen as

$$s_e^p(x|v) = \begin{cases} \min_{x_p}[s_e(x|v, x_p)] & \text{for } v = M_k \ (k \in \mathcal{K}) \\ s_e^p(x|v) = s_e(x|v) & \text{for other emitting states } v. \end{cases}$$

We also define $\Delta_e(k)$ for $k \in \mathcal{K}$ as follows

$$\Delta_e(k) = \max_x \left(\max_{x_p}[s_e(x|v, x_p)] - \min_{x_p}[s_e(x|v, x_p)] \right). \quad (2)$$

In the second place, the transition score of the profile-HMM for a transition from state v to state w is chosen as follows

$$s_t^p(v, w) = \begin{cases} s_t(v, D_k|m = 0) & w = D_k \ (k \in \mathcal{K}) \\ s_t(v, M_k|m = 1) & w = M_k \ (k \in \mathcal{K}) \\ \min_m s_t(v, I_{k-1}|m) & w = I_{k-1} \ (k \in \mathcal{K}) \\ s_t(v, w) & \text{otherwise.} \end{cases}$$

In addition to this, we define $\Delta_t(k)$ for $k \in \mathcal{K}$

$$\Delta_t(k) = \left(\max_m[s_t(v, I_{k-1}|m)] - \min_m[s_t(v, I_{k-1}|m)] \right). \quad (3)$$

Finally, we choose the threshold of the pre-screening filter to be $\lambda_p = \lambda - \Delta$, where $\Delta = \sum_{k \in \mathcal{K}} [\Delta_e(k) + \Delta_t(k)]$.

3.2. No degradation in the prediction accuracy

Based on the pre-screening filter constructed as described in Sec. 3.1, we can compute the sequence-level similarity score as follows

$$S_p(\mathbf{x}, \Theta_p) = \max_{\mathbf{y}} S_p(\mathbf{x}, \mathbf{y}|\Theta_p) = S_p(\mathbf{x}, \mathbf{y}^*|\Theta_p), \qquad (4)$$

where $\mathbf{y}^*$ is the optimal state sequence. Using the score in (4) with the threshold λ_p guarantees that there will be no loss in the prediction accuracy. This can be shown as follows.

Theorem For an observed sequence $\mathbf{x}$, if the score $S_p(\mathbf{x}, \Theta_p)$ computed from the pre-screening filter is smaller than λ_p, its score $S(\mathbf{x}, \Theta)$ from the original profile-csHMM cannot exceed λ.

Proof If $S_p(\mathbf{x}, \Theta_p) < \lambda_p$, we have

$$\max_{\mathbf{y} \in \mathcal{Y}} S_p(\mathbf{x}, \mathbf{y}|\Theta_p) \leq \max_{\mathbf{y}} S_p(\mathbf{x}, \mathbf{y}|\Theta_p) < \lambda_p,$$

where $\mathcal{Y}$ is the set of all feasible state sequences in the original profile-csHMM. Then we have

$$\begin{aligned} S(\mathbf{x}, \Theta) &= \max_{\mathbf{y}} S(\mathbf{x}, \mathbf{y}|\Theta) = \max_{\mathbf{y} \in \mathcal{Y}} S(\mathbf{x}, \mathbf{y}|\Theta) \\ &\leq \underbrace{\max_{\mathbf{y} \in \mathcal{Y}} S_p(\mathbf{x}, \mathbf{y}|\Theta_p)}_{<\lambda_p} + \underbrace{\max_{\mathbf{y} \in \mathcal{Y}} \Big[S(\mathbf{x}, \mathbf{y}|\Theta) - S_p(\mathbf{x}, \mathbf{y}|\Theta_p) \Big]}_{\leq \Delta} \\ &< (\lambda - \Delta) + \Delta = \lambda. \end{aligned}$$

■

This shows that the pre-screening filter will reject only those sequences that are guaranteed to be rejected by the original profile-csHMM, hence there will be no degradation in the prediction accuracy.

4. EXPERIMENTAL RESULTS

To demonstrate the proposed idea, we carried out an experiment using real RNA sequences. We first constructed a profile-csHMM for the CORONA-PK3 RNA family in the Rfam database [8].[5] Note that the secondary structure of CORONA-PK3 contains pseudoknots, hence they cannot be modeled using CMs (or SCFGs). Based on the constructed profile-csHMM, we have built a profile-HMM pre-screening filter by following the procedure elaborated in Sec. 3.

After constructing the models, we evaluated the performance of the profile-csHMM search and that of the proposed pre-screening approach. For evaluation, we used a database that consists of real CORONA-PK3 RNA sequences and 10,000 random RNA sequences. As expected, the pre-screening filter did not miss any RNA that was reported as a "match" by the original profile-csHMM. Consequently, the prediction accuracies of both methods were identical. The average CPU time used by the pre-screening filter to compute the similarity score $S_p(\mathbf{x}, \Theta_p)$ was 0.0093 sec, which is much smaller than 28.1 sec of the profile-csHMM.[6] The rejection rate of the pre-screening filter was around 98.8%, hence only 1.2% of the inspected RNAs was passed to the profile-csHMM in the second stage. As a result, the average CPU time used by the proposed method was around 0.34 sec, which is around eighty times faster than the search method based on a profile-csHMM alone.

[5] For constructing the model, we used the 'seed alignment' which provides a reasonably reliable structural annotation of the given RNA family.

[6] We used a fixed search region size of $D = 7$.

It is important to note that the rejection rate of the pre-screening filter has a crucial impact on the overall reduction in the search time. Ideally, the pre-screening filter should reject most sequences that will be rejected by the profile-csHMM, and pass only a small fraction to the second stage for further inspection. However, there can be also occasions when the rejection rate is quite small, in which case the reduction in the search time will not be significant. For many applications, the criterion used in Sec. 3 for deriving the parameters of the pre-screening filter and the threshold λ_p will be too stringent, and it may be beneficial to relax it a little bit to make the search faster, at a slight loss of the prediction accuracy.

5. CONCLUDING REMARKS

In this paper, we proposed a method that can make a database search based on a profile-csHMM significantly faster. The proposed method uses a pre-screening filter based on profile-HMMs, whose computational complexity is much lower than that of the profile-csHMMs. As shown in the paper, this pre-screening filter rejects only those sequences that are guaranteed to be rejected by the original profile-csHMM. This leads to a considerable reduction in the overall search time without any degradation in the prediction accuracy of the search. Important topics for future research include optimizing the parameters of the pre-screening filter to reduce the search time further without affecting the prediction accuracy, and finding heuristic methods for choosing the filter parameters that will make the search even faster with a small trade-off in the prediction accuracy.

6. REFERENCES

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

[2] O. Gotoh, "Significant improvement in accuracy of multiple protein alignments by iterative refinement as assessed by reference to structural alignments", *Journal of Molecular Biology*, vol. 264, pp.823-838, 1996.

[3] S. R. Eddy, "Hidden Markov models", *Current Opinion in Structural Biology*, vol. 6, pp. 361-365, 1996.

[4] B.-J. Yoon and P. P. Vaidyanathan, "Profile context-sensitive HMMs for probabilistic modeling of sequences with complex correlations", *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, May 2006.

[5] B.-J. Yoon and P. P. Vaidyanathan, "Context-sensitive hidden Markov models for modeling long-range dependencies in symbol sequences", *IEEE Trans. Signal Processing*, vol. 54, pp. 4169-4184, Nov. 2006.

[6] 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.

[7] 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.

[8] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy and A. Bateman, "Rfam: annotating non-coding RNAs in complete genomes", *Nucleic Acids Research*, vol. 33, pp. D121-D124, 2005.