Profiling and Searching for RNA Pseudoknot Structures in Genomes

Chunmei Liu[1], Yinglei Song[1], Russell L. Malmberg[2], and Liming Cai[1]

[1] Department of Computer Science, University of Georgia, Athens GA 30602, USA
{chunmei, song, cai}@cs.uga.edu
[2] Department of Plant Biology, University of Georgia, Athens GA 30602, USA
russell@plantbio.uga.edu

Abstract. We developed a new method that can profile and efficiently search for pseudoknot structures in noncoding RNA genes. It profiles interleaving stems in pseudoknot structures with independent Covariance Model (CM) components. The statistical alignment score for searching is obtained by combining the alignment scores from all CM components. Our experiments show that the model can achieve excellent accuracy on both random and biological data. The efficiency achieved by the method makes it possible to search for structures that contain pseudoknot in genomes of a variety of organisms.

1 Introduction

Searching genomes with computational models has become an effective approach for the identification of genes. During recent years, extensive research has been focused on developing computationally efficient and accurate models that can find novel noncoding RNAs and reveal their associated biological functions. Unlike the messenger RNAs that encode the amino acid residues of protein molecules, noncoding RNA molecules play direct roles in a variety of biological processes including gene regulation, RNA processing, and modification. For example, the human 7SK RNA binds and inhibits the transcription elongation factor P-TEFb [17][25] and the RNase P RNA processes the 5’ end of precursor tRNAs and some rRNAs [7]. Noncoding RNAs include more than 100 different families [23]. Genome annotation based on models constructed from homologous sequence families could be a reliable and effective approach to enlarging the known families of noncoding RNAs.

The functions of noncoding RNAs are, to a large extent, determined by the secondary structures they fold into. Secondary structures are formed by bonded base pairs between nucleotides and may remain unchanged while the nucleotide sequence may have been significantly modified through mutations over the course of evolution. Profiling models based solely on sequence content such as Hidden Markov Model (HMM) [12] may miss structural homologies when directly used to search genomes for noncoding RNAs containing complex secondary structures. Models that can profile noncoding RNAs must include both the content and the structural information from the homologous sequences. The Covariance Model (CM) developed by Eddy and Durbin [6] extends the profiling HMM by allowing the coemission of paired nucleotides on certain states to model base pairs, and introduces bifurcation states to emit parallel stems. The CM is capable of modeling secondary structures comprised of nested and parallel stems. However, pseudoknot structures, where at least two structurally interleaving stems are involved, cannot be directly modeled with the CM and have remained computationally intractable for searching [1][13][14][18][19][20][21][24].

So far, only a few systems have been developed for profiling and searching for RNA pseudoknots. One example is ERPIN developed by Gautheret and Lambert [8][15]. ERPIN searches genomes by sequentially looking for single stem loop motifs contained in the noncoding RNA gene, and reports a hit when significant alignment scores are observed for all the motifs at their corresponding locations. Since ERPIN does not allow the presence of gaps when it performs alignments, it is computationally very efficient. However, alignments with no gaps may miss distant homologies and thus result in a lower sensitivity.

Brown and Wilson [2] proposed a more realistic model comprised of a number of Stochastic Context Free Grammar (SCFG) [3][22] components to profile pseudoknot structures. In their model, the interleaving stems in a pseudoknot structure are derived from different components; the pseudoknot structure is modeled as the intersection of components. The optimal alignment score of a sequence segment is computed by aligning it to all the components iteratively. The model can be used to search sequences for simple pseudoknot structures efficiently. However, a generic framework for modeling interleaving stems and carrying out the search was not proposed in their work. For pseudoknots with more complex structure, more than two SCFG components may be needed and the extension of the iterative alignment algorithm to k components may require $k!$ different alignments in total since all components are treated equally in their model.

In this paper, we propose a new method to search for RNA pseudoknot structures using a model of multiple CMs. Unlike the model of Brown and Wilson, we use independent CM components to profile the interleaving stems in a pseudoknot. Based on the model, we have developed a generic framework for modeling interleaving stems of pseudoknot structures; we propose an algorithm that can efficiently assign stems to components such that interleaving stems are profiled in different components. The components with more stems are associated with higher weights in determining the overall conformation of a sequence segment. In order to efficiently perform alignments of the sequence segment to the model, instead of iteratively aligning the sequence segment to the CM components, our searching algorithm aligns it to each component independently following the descending order of component weights. The statistical log-odds scores are computed based on the structural alignment scores of each CM component. *Stem contention* may occur such that two or more base pairs obtained from different components require the participation of the same nucleotide. Due to the conformational constraints inherently imposed by the CM components, stem contentions occur infrequently (less than 30%) and can be effectively resolved based on the conformational constraints from the alignment results on components with higher weight values. The algorithm is able to accomplish the search with a worst case time complexity of $O((k-1)W^3L)$ and a space complexity of $O(kW^2)$, where k is the number of CM components in the model, W and L are the size of the searching window and the length of the genome respectively.

We used the model to search for a variety of RNA pseudoknots inserted in randomly generated sequences. Experiments show that the model can achieve excellent sensitivity (SE) and specificity (SP) on almost all of them, while using only slightly more computation time than searching for pseudoknot-free RNA structures. We then applied the model and the searching algorithm to identify the pseudoknots on the 3' untranslated region in several RNA genomes from the corona virus family. An exact match between the locations found by our program and the real locations is observed. Finally, in order to test the ability of our program to cope with noncoding RNA genes with complex pseudoknot structures, we carried out an experiment where the complete DNA genomes of two bacteria were searched to find the locations of the tmRNA genes. The results show that our program identified the location with a reasonable amount of error (with a right shift of around 20 nucleotide bases) for one bacterial genome and for the other bacteria search was perfect. To the best of our knowledge, this is the first experiment where a whole genome of more than a million nucleotides is searched for a complex structure that contains pseudoknots.

2 Experiments and Results

To test the performance of the model, we developed a search program in C language and carried out searching experiments on a Sun/Solaris workstation. The workstation has 8 dual processors and 32GB main memory. We evaluated the accuracy of the program on both real genomes and randomly generated sequences with a number of RNA pseudoknot structures inserted. The RNAs we choose to test the model are shown in Table 1. Model training and testing are based on the multiple alignments downloaded from the Rfam database [10]. For each RNA pseudoknot, we divided the available data into a training set and a testing set, and the parameters used to model it are estimated based on multiple structural alignments among $5-90$ homologous training sequences with a pairwise identity less than 80%. The emission probabilities of all nucleotides for a given state in a CM component are estimated by computing their frequencies to appear in the corresponding column in the multiple alignment of training sequences; transition probabilities are computed similarly by considering the relative frequencies for different types of transitions that occur between the corresponding consecutive columns in the alignment. Pseudocounts, dependent on the number of training sequences, are included to prevent overfitting of the model to the training data.

To measure the sensitivity and specificity of the searching program within a reasonable amount of time, for each selected pseudoknot structure, we selected $10-40$ sequence segments from the set of testing data and inserted them into each of the randomly generated sequences of 10^5 nucleotides. In order to test whether the model is sensitive to the base composition of the background sequence, we varied the C+G concentration in the random background. The program computes the log-odds, the logarithmic ratio of the probability of generating sequence segment s by the null (random) model R to that by our model M. It reports a hit when the Z-score of s is greater than 4.0. The computation of Z-scores requires knowing the mean and standard deviation for the distribution of log-odd scores of random sequence segments; both of them can be

Table 1. Information on training sequences used for the estimation of model parameters

RNA	Number of training sequences	Number of nucleotides	Pseudocount
tmRNA−pk12	36	$130-250$	1.5
tmRNA−pk34	89	$90-120$	2.4
srpRNA	24	$30-50$	1.2
telomerase−vert	13	$90-200$	0.9
corona−pk3	14	$60-70$	0.9
HDV−ribozyme	15	$90-100$	1.0
tombus−3−IV	17	$90-100$	1.0
alpha−RBS	9	$100-120$	0.8
antizyme−FSE	13	$50-60$	0.9
IFN−gamma	5	$160-180$	0.6

determined with methods similar to the ones introduced by Klein and Eddy [11] before the search starts.

As can be seen in Table 2, the program correctly identifies more than 80% of inserted sequence segments with excellent specificity in most of the experiments. The only exception is the srpRNA, where the program misses more than 50% inserted sequence segments in one of the experiments. The relatively lower sensitivity in that particular experiment can be partly ascribed to the fact that the pseudoknot structure of srpRNA contains fewer nucleotides; thus its structural and sequence patterns have a larger probability to occur randomly. The running time for srpRNA, however, is also significantly shorter than that needed by most of other RNA pseudoknots due to the smaller size of the model. Additionally, while the alpha−RBS pseudoknot has a more complex structure and three CM components are needed to model it, our searching algorithm efficiently identifies more than 95% of the inserted pseudoknots with high specificities. A higher C+G concentration in the background does not adversely affect the specificity of the model; it is evident from Table 2 that the program achieves better overall performance in both sensitivity and specificity in a background of higher C+G concentrations. We therefore conjecture that the specificity of the model is partly determined by the base composition of the genome and is improved if the base composition of the target gene is considerably different from its background.

To test the accuracy of the program on real genomes, we performed experiments to search for particular pseudoknot structures in the genomes for a variety of organisms. Table 3 shows the genomes on which we have searched with our program and the locations annotated for the corresponding pseudoknot structures. The program successfully identified the exact locations of known 3'UTR pseudoknot in four genomes from the family of corona virus. This pseudoknot was recently shown to be essential for the replication of the viruses in the family [9].

In addition, the genomes of the bacteria, *Haemophilus influenzae* and *Neisseria meningitidis MC58*, were searched for their tmRNA genes. The *Haemophilus influenzae* DNA genome contains about 1.8×10^6 nucleotides and *Neisseria meningitidis MC58* DNA genome contains about 2.2×10^6 nucleotides. The tmRNA functions in the trans-translation process to add a C-terminal peptide tag to the incomplete protein product of

Table 2. The performance of the model on different RNA pseudoknots inserted into a background (of 10^5 nucleotides) randomly generated with different C+G concentrations. TN is the total number of pseudoknotted sequence segments inserted; CI is the number of sequence segments correctly identified by the program (with a positional error less than ±3 bases); NH is the number of sequence segments returned by the program; SE and SP are sensitivity and specificity respectively. The thresholds of log-odds score are predetermined using the Z-score value of 4.0

RNA	TN	CI	NH	SE(%)	SP(%)	Running time(hr)	Background C+G (%)
tmRNA−pk12	25	20	24	80.0	83.3	56.33	57.0
tmRNA−pk34	27	26	31	96.0	84.0	59.36	57.0
srpRNA	29	13	16	44.8	81.3	4.79	57.0
telomerase−vert	14	14	15	100.0	93.3	68.83	57.0
corona−pk3	37	37	39	100.0	94.8	2.89	57.0
HDV−ribozyme	37	37	37	100.0	100.0	6.54	57.0
tombus−3−IV	13	13	13	100.0	100.0	15.45	57.0
alpha−RBS	24	24	25	100.0	96.0	27.85	57.0
antizyme−FSE	28	28	28	100.0	100.0	0.94	57.0
IFN−gamma	10	10	10	100.0	100.0	31.24	57.0
tmRNA−pk12	24	24	25	100.0	96.0	55.57	67.0
tmRNA−pk34	27	27	30	100.0	90.0	56.42	67.0
srpRNA	25	17	19	68.0	89.4	4.76	67.0
telomerase−vert	13	13	14	100.0	92.9	67.80	67.0
corona−pk3	33	33	34	100.0	97.1	2.90	67.0
HDV−ribozyme	37	37	37	100.0	100.0	6.52	67.0
tombus−3−IV	20	20	20	100.0	100.0	16.63	67.0
alpha−RBS	18	18	18	100.0	100.0	27.79	67.0
antizyme−FSE	28	28	29	100.0	96.6	0.94	67.0
IFN−gamma	10	10	10	100.0	100.0	33.15	67.0
tmRNA−pk12	26	26	29	100.0	90.0	55.45	77.0
tmRNA−pk34	25	25	33	100.0	75.7	53.55	77.0
srpRNA	29	22	23	75.9	95.7	4.78	77.0
telomerase−vert	16	16	16	100.0	100.0	66.07	77.0
corona−pk3	37	37	37	100.0	100.0	3.13	77.0
HDV−ribozyme	37	37	37	100.0	100.0	6.57	77.0
tombus−3−IV	20	20	20	100.0	100.0	16.94	77.0
alpha−RBS	22	22	22	100.0	100.0	28.86	77.0
antizyme−FSE	28	28	28	100.0	100.0	0.96	77.0
IFN−gamma	10	10	10	100.0	100.0	32.55	77.0
tmRNA−pk12	24	24	25	100.0	96.2	55.09	87.0
tmRNA−pk34	27	27	28	100.0	96.4	52.39	87.0
srpRNA	26	25	25	96.2	100.0	4.81	87.0
telomerase−vert	17	17	17	100.0	100.0	70.60	87.0
corona−pk3	37	37	37	100.0	100.0	3.17	87.0
HDV−ribozyme	37	37	37	100.0	100.0	6.64	87.0
tombus−3−IV	20	20	20	100.0	100.0	16.94	87.0
alpha−RBS	24	23	23	95.8	100.0	29.08	87.0
antizyme−FSE	26	26	26	100.0	100.0	0.94	87.0
IFN−gamma	10	10	10	100.0	100.0	32.84	87.0

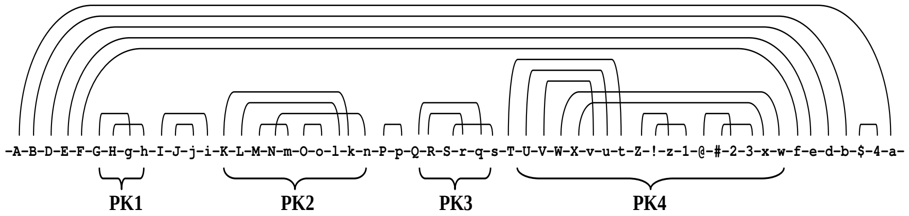
-A-B-D-E-F-G-H-g-h-I-J-j-i-K-L-M-N-m-O-o-l-k-n-P-p-Q-R-S-r-q-s-T-U-V-W-X-v-u-t-Z-!-z-1-@-#-2-3-x-w-f-e-d-b-$-4-a-
PK1
PK2
PK3
PK4

Fig. 1. Diagram of the pairing regions on the tmRNA gene. Upper case letters indicate base sequences that pair with the corresponding lower case letters. The four pseudoknots constitute the central part of the tmRNA gene and are called Pk1, Pk2, Pk3, Pk4 respectively

a defective mRNA [16]. The central part of the secondary structure of tmRNA molecule consists of four pseudoknot structures. Figure 1 shows the pseudoknot structures on the tmRNA molecule.

In order to search the bacterial DNA genomes efficiently, the combined pseudoknots 1 and 2 were used to search the genome first; the program searches for the whole tmRNA gene only in the region around the locations where a hit for Pk1 and Pk2 is detected. We cut the genome into segments with shorter lengths (around 10^5 nucleotide bases for each), and ran the program in parallel on ten of them in two rounds. The result for *Neisseria meningitidis MC58* shows that we successfully identified the exact locations of tmRNA. However, the locations of tmRNA obtained for *Haemophilus influenzae* have a shift of around 20 nucleotides with respect to its real location (7% of the length of the tmRNA). This slight error can probably be ascribed to our "hit-and-extend" searching strategy to resolve the difficulty arising from the complex structure and the relatively much larger size of tmRNA genes; positional errors may occur during different searching stages and accumulate to a significant value. Our experiment on the DNA genomes also demonstrates that, for each genome, it is very likely there is only one tmRNA gene in it, since our program found only one significant hit. To our knowledge, this is the first computational experiment where a whole genome of more than a million nucleotides was successfully searched for a complex structure that contains pseudoknot structures.

3 Models and Algorithms

The Covariance Model (CM) proposed by Eddy and Durbin [6][5] can effectively model the base pairs formed between nucleotides in an RNA molecule. Similarly to the emission probabilities in HMMs, the emission probabilities in the CM for both unpaired nucleotides and base pairs are positional dependent. The profiling of a stem hence consists of a chain of consecutive emissions of base pairs. Parallel stems on the RNA sequence are modeled with bifurcation transitions where a bifurcation state is split into two states. The parallel stems are then generated from the transitions starting with the two states that result respectively.

Table 3. The results obtained with our searching program on the genomes of a variety of organisms. GA is the accession number of the genome; RL specifies the real location of the pseudoknot structure in the genome; SL is the one returned by the program; RT is the running time needed to perform the searching in hours; GL is the length of the genome in its number of bases. The genome of Haemophilus searched in our experiment is the reversed complementary DNA strand

GA	Organism	ncRNA	RL	SL	RT(hr)	GL(bs)
NC000907	Haemophilus	tmRNA	$472210 - 472575$	$472177 - 472542$	170.00	1.83×10^6
NC003112	Neisseria meningitidis	tmRNA	$1241197-1241559$	$1241197-1241559$	170.00	2.2×10^6
NC003045	Bovine CoronaVirus	3'UTR pk	$30798 - 30859$	$30798 - 30859$	1.24	31028
NC002645	Human CoronaVirus	3'UTR pk	$27063 - 27125$	$27063 - 27125$	1.12	27317
NC001846	Murine HepatitusVirus	3'UTR pk	$31092 - 31153$	$31092 - 31153$	1.27	31357
NC003436	Porcine DiarrheaVirus	3'UTR pk	$27820 - 27882$	$27820 - 27882$	1.17	28033

The genome is scanned by a window with an appropriate length. Each location of the window is scored by aligning all subsequence segments contained in the window to the model with the CYK algorithm. The maximum log-odds score of them is determined as the log-odds score associated with the location. A hit is reported for a location if the computed log-odds score is higher than a predetermined threshold value.

Pseudoknot structures are beyond the profiling capability of a single CM due to the inherent context sensitivity of pseudoknots. Models for pseudoknot structures require a mechanism for the description of their interleaving stems. Previous work by Brown and Wilson [2] and Cai *et al.* [4] has modeled the pseudoknot structures with grammar components that intersect or cooperatively communicate. A similar idea is adopted in this work; a number of independent CM components are combined to resolve the difficulty in profiling that arises from the interleaving stems. Interleaving stems are profiled in different CM components and the alignment score of a sequence segment is determined based on a combination of the alignment scores on all components.

However, the optimal conformations from the alignments on different components may violate some of the conformational constraints that a single RNA sequence must follow. For example, a nucleotide rarely forms two different base pairs simultaneously with other nucleotides in an RNA molecule. This type of restriction is not considered by the independent alignments carried out in our model and thus may lead to erroneous searching results if not treated properly. In our model, *stem contention* may occur. We break the contention by introducing different priorities to components; base pairs determined from components with the highest priority win the contention. We hypothesize that, biochemically, components profiling more stems are likely to play more dominant roles in the formation of the conformation and are hence assigned higher priority weights.

3.1 Model Generation

In order to profile the interleaving stems in a pseudoknot structure with independent CM components, we need an algorithm that can partition the set of stems on the RNA sequence into a number of sets comprised of stems that mutually do not interleave. Based on the consensus structure of the RNA sequence, an undirected graph $G = (V, E)$ can be constructed where V, the set of vertices in G, consists of all stems on the sequence. Two vertices are connected with an edge in G if the corresponding stems are in parallel or nested. The set of vertices V needs to be partitioned into subsets such that the subgraph induced by each subset forms a clique.

We use a greedy algorithm to perform the partition. Starting with a vertex set S initialized to contain a arbitrarily selected vertex, the algorithm iteratively searches the neighbors of the vertices in S and computes the set of vertices that are connected to all vertices in S. It then randomly selects one vertex v that is not in S from the set and modifies S by assigning v to S. The algorithm outputs S as one of the subsets in the partition when S can not be enlarged and randomly selects an unassigned vertex and repeats the same procedure. It stops when every vertex in G has been included in a subset. Although the algorithm does not minimize the number of subsets in the partition, our experiments show that it can efficiently provide optimal partitions of the stems on pseudoknot structures of moderate structural complexity.

The CM components in the profiling model are generated and trained based on the partition of the stems. The stems in the same subset are profiled in the same CM component. For each component, the parameters are estimated by considering the consensus structure formed by the stems in the subset only.

3.2 Searching Algorithm

The optimal alignments of a sequence segment to the CM components are computed with the dynamic programming based CYK algorithm. As we have mentioned before, higher priority weights are assigned to components with more stems profiled. The component with the maximum number of stems thus has the maximum weight and is the *dominant component* in the model. The algorithm performs alignments in the descending order of component weights. It selects the sequence segment that maximizes the log-odds score from the dominant component. The alignment scores and optimal conformations of this segment on other components are then computed and combined to obtain the overall log-odds score for the segment's position on the genome.

More specifically, we assume that the model contains k CM components M_0, M_1, ..., M_{k-1} in descending order of component weights. The algorithm considers all possible sequence segments s_d that are enclosed in the window and uses Equation (1) to determine the sequence segment s to be the candidate for further consideration, where W is the length of the window used in searching, and Equation (2) to compute the overall log-odds score for s. We use sm_i to denote the parts of s that are aligned to the stems profiled in CM component M_i. Basically, $Log\_odds(sm_i|M_i)$ accounts for the contributions from the alignment of sm_i to M_i. The log-odds score of sm_i is counted in both M_0 and M_i and must be subtracted from the sum.

$$s = \arg \max_{0<|s_d|<W} \{Log\_odds(s_d|M_0)\}. \qquad (1)$$

$$Log\_odds(s|M) = Log\_odds(s|M_0) + \sum_{i=1}^{k-1} \sum_{sm_i \in M_i} (Log\_odds(sm_i|M_i) - Log\_odds(sm_i|M_0)). \qquad (2)$$

3.3 Stem Contention

The conformations corresponding to the optimal alignments of a sequence segment to all CM components are obtained by tracing back the dynamic programming matrices and checking to ensure that no stem contention occurs. Since each nucleotide in the sequence is represented with a state in a CM component, the CM inherently imposes constraints on the optimal conformations of sequence segments aligned to it. We hence expect that stem contention occurs with a low frequency. In order to verify this intuition, we tested the model on sequences randomly generated with different base compositions and evaluated the frequencies of stem contentions for pseudoknot structures on which we have performed an accuracy test; the results are shown in Figure 2.

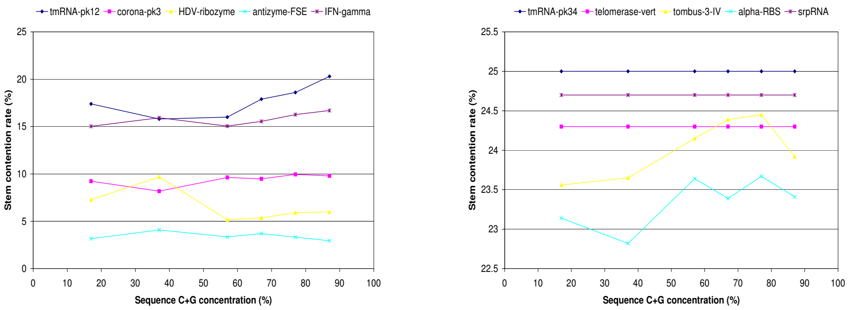

Fig. 2. 4000 random sequences were generated at each given base composition and aligned to the corresponding profiling model. The sequences are of about the same length as the length of the pseudoknot structure. The stem contention rates for each pseudoknot structure were measured and plotted. They were the ratio of the number of random sequences in which stem contentions occurred to the number of total random sequences. Left: plots of profiling models observed to have a stem contention rate lower than 20%, right: plots of these with slightly higher stem contention frequencies. The experimental results demonstrate that, in all pseudoknots where we have performed accuracy tests, stem contention occurs with a rate lower than 30% and is insensitive to the base composition of sequences

The presence of stem contention increases the running time of the algorithm, because the alignment of one of the involved components must be recomputed to resolve the contention. Based on the assumption that components with more stems contribute more to the stability of the optimal conformation, we resolve the contention in favor of such components. We perform recomputation on the component with a lower number of stems by incorporating conformational constraints inherited from components with more stems into the alignment algorithm, preventing them from forming the contentious stems.

Specifically, we assume that stem $S_j \in M_i$ and stem contention occurs between S_j and other stems profiled in M_{i-1}; the conformational constraints from the component M_{i-1} are in the format of (l_1, l_2) and (r_1, r_2). In other words, to avoid the stem contention, the left and right parts of the stem must be the subsequences of indices (l_1, l_2) and (r_1, r_2) respectively. The dynamic programming matrices for S_j are limited to the rectangular region that satisfies $l_1 \leq s \leq l_2$ and $r_1 \leq t \leq r_2$.

The stem contention frequency depends on the conformational flexibilities of the components in the covariance model. More flexibilities in conformation may improve the sensitivity of the model but cause higher contention frequency and thus increase the running time for the algorithm. In the worst case, recomputation is needed for all non-dominant components in the model and the time complexity of the algorithm becomes $O((k-1)W^3L)$, where k is the number of components in the model, W and L are the window length and the genome length respectively.

4 Conclusions and Future Work

In this paper, we have introduced a new model that serves as the basis for a generic framework that can efficiently search genomes for the noncoding RNAs with pseudoknot structures. Within the framework, interleaving stems in pseudoknot structures are modeled with independent CM components and alignment is performed by aligning sequence segments to all components following the descending order of their weight values. Stem contention occurs with a low frequency and can be resolved with a dynamic programming based recomputation. The statistical log-odds scores are computed based on the alignment results from all components. Our experiments on both random and biological data demonstrate that the searching framework achieves excellent performance in both accuracy and efficiency and can be used to annotate genomes for noncoding RNA genes with complex secondary structures in practice.

We were able to search a bacterial genome for a complete structure with a pseudoknot in about one week on our Sun workstation. It would be desirable to improve our algorithm so that we could search larger genomes and databases. The running time, however, could be significantly shortened if a filter can be designed to preprocess DNA genomes and only the parts that pass the filtering process are aligned to the model. Alternatively, it may be possible to devise alternative profiling methods to the covariance model that would allow faster searches.

References

1. T. Akutsu, “Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots.”, *Discrete Applied Mathematics*, 104: 45-62, 2000.
2. M. Brown and C. Wilson, “RNA Pseudoknot Modeling Using Intersections of Stochastic Context Free Grammars with Applications to Database Search.”, *Pacific Symposium on Biocomputing*, 109-125, 1995.
3. M. Brown, “Small subunit ribosomal RNA modeling using stochastic context-free grammars.”, *Proc. of Int. Conf. Intel. Syst. Mol. Biol.*, 56: 57-66, 2000.
4. L. Cai, R. L. Malmberg, and Y. Wu, “Stochastic Modeling of Pseudoknot Structures: A Grammatical Approach.”, *Bioinformatics*, 19, $i66 - i73$, 2003.
5. R. Durbin, S. R. Eddy, A. Krogh, and G. J. Mitchison, “Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.”, *Cambridge University Press*, 1998.
6. S. Eddy and R. Durbin, “RNA sequence analysis using covariance models.”, *Nucleic Acids Research*, 22: 2079-2088, 1994.
7. D. N. Frank and N. R. Pace, “Ribonuclease P: unity and diversity in a tRNA processing ribozyme.”, *Annu Rev Biochem.*, 67: 153-180, 1998.
8. D. Gautheret and A. Lambert, “Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles.”, *Journal of Molecular Biology*, 313: 1003-1011, 2001.
9. S. J. Geobel, B. Hsue, T. F. Dombrowski, and P. S. Masters, “Characterization of the RNA components of a Putative Molecular Switch in the 3’ Untranslated Region of the Murine Coronavirus Genome.”, *Journal of Virology*, 78: 669-682, 2004.
10. S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, “Rfam: an RNA family database.”, *Nucleic Acids Research*, 31: 439-441, 2003.
11. R. J. Klein, S. R. Eddy, “RSEARCH: Finding Homologs of Single Structured RNA Sequences.”, *BMC Bioinformatics*, 4:44, 2003.
12. A. Krogh, M. Brown, IS. Mian, K. Sjolander, and D. Haussler, “Hidden Markov models in computational biology. Applications to protein modeling.”, *Journal of Molecular Biology*, 235: 1501-1531, 1994.
13. D. Lee, K. Han, “Prediction of RNA Pseudoknots-Comparative Study of Genetic Algorithms.”, *Genome Informatics*, 13: 414-415, 2002.
14. R. B. Lyngso and C. N. S. Pederson, “RNA pseudoknot prediction in energy based models.”, *Journal of Computational Biology*, 7: 409-428, 2000.
15. T. Macke, D. Ecker, R. Gutell, and D. Gautheret, D. Case, R. Sampath, “RNAMotif, an RNA secondary structure definition and search algorithm.”, *Nucleic Acids Research*, 29: 4724-4735, 2001.
16. N. Nameki, B. Felden, J. F. Atkins, R. F. Gesteland, H. Himeno, A. Muto, “Functional and structural analysis of a pseudoknot upstream of the tag-encoded sequence in E. coli tmRNA.”, *Journal of Molecular Biology*, 286(3): 733-744, 1999.
17. V. T. Nguyen, T. Kiss, A. A. Michels, O. Bensaude, “7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes.”, *Nature* 414: 322-325, 2001.
18. J. Reeder and R. Giegeritch, “Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics.”, *BMC Bioinformatics*, 5: 104, 2004.
19. E. Rivas and S. Eddy, “The language of RNA: a formal grammar that includes pseudoknots.”, *Bioinformatics*, 16: 334-340, 2000.
20. E. Rivas and S. Eddy, “A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots.”, *Journal of Molecular Biology*, 285: 2053-2068, 1999.
21. J. Ruan, G. D. Stormo, and W. Zhang, “An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots.”, *Bioinformatics*, 20: 58-66, 2004.

22. Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjolander, R. C. Underwood, and D. Maussler, “Stochastic Context-Free Grammars for tRNA Modeling.”, *Nucleic Acids Research*, 22: 5112-5120, 1994.
23. G. Storz, “An expanding universes of noncoding RNAs.”, *Science*, 296(5571): 1260-1263, 2002.
24. Y. Uemura, A. Hasegawa, Y. Kobayashi, T. Yokomori, “Tree adjoining grammars for RNA structure prediction.”, *Theoretical Computer Science*, 210: 277-303, 1999.
25. Z. Yang, Q. Zhu, K. Luo, and Q. Zhou, “The 7SK small nuclear RNA inhibits the Cdk9/cyclin T1 kinase to control transcription.”, *Nature* 414: 317-322, 2001.