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. A new method is developed 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 the pseudoknot structures in genomes of a variety of organisms.

1 Introduction

Searching genomes with computational models has become an effective approach to 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 functionalities. Functionalities 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 over evolution while the nucleotides of a sequence have been significantly modified through mutations. Profiling models based solely on sequence content such as Hidden Markov Model (HMM) [9] may miss important homologies when directly used to search genomes for noncoding RNAs containing complex secondary structures. Therefore, models that can profile noncoding RNAs must include both the content and structural information from the homologous sequences. The Covariance Model (CM) developed by Eddy and Durbin [5] extends the profiling HMM by allowing the coemission of pairing nucleotides on certain states to model base pairs, and introduces bifurcation states to emit parallel stems in the conformation. 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 remained computationally intractable for searching [1][10][13][14][15].

So far, only a few systems have been developed for profiling and searching for RNA pseudoknots. One example of related work includes ERPIN, developed by Gautheret and Lambert [6][11]. ERPIN searches genomes by sequentially looking for single stem

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3515, pp. 968–975, 2005.

loop motifs contained in the noncoding RNA gene. 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. Another example is from Brown and Wilson [2]. They proposed a more realistic model comprised of a number of Stochastic Context Free Grammar (SCFG) [5][16] components to profile pseudoknot structures. In their model, different components are used to derive the interleaving stems in a pseudoknot structure. 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 with simple pseudoknot structures efficiently. However, for pseudoknots with more complex structure, more than two SCFG components may be needed and the extension of the iterative alignment algorithm on k components may need to perform $k!$ 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, independent CM components are used 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, 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 on each CM component. 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, two DNA genomes of bacterias were searched to find the location of the tmRNA genes. The results show that our program identified the locations with a small amount of error (with a right shift of around 20 nucleotide bases). To the best of our knowledge, these are the first experiments where a whole genome of more than a million nucleotides is successfully searched for a complex structure that contains pseudoknots.

2 Experiments and Results

To test the performance of the model, we developed a searching 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. To test the model, we chose 10 RNAs from 9 RNA families, tmRNA, srpRNA, Telomerase−vert, Corona−pk3, HDV−ribozyme, Tombus−3−IV, Alpha−RBS, Antizyme−FSE and IFN−gamma, which have pseudoknot annotations in Rfam [8].

Model training and testing are based on the multiple alignments downloaded from the Rfam database. For each RNA pseudoknot, we divide 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 pairwise identity less than 80%. 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 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 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 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. Our results demonstrate that higher C+G concentration in the background does not adversely affect the specificity of the model. The program achieves better overall performance in both sensitivity and specificity on the 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 can be 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 1 shows the genomes on which we have searched with our program and the locations annotated for the corresponding pseudoknot structures. It is evident from the results that the program successfully identified the exact locations of known 3'UTR

Table 1. 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

pseudoknot in the 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 [7]. In addition, we performed an experiment where the genomes of bacterias, *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 importantly in the trans-translation process to add a C-terminal peptide tag to the incomplete protein product of a broken mRNA [12]. The central part of the secondary structure of tmRNA molecule consists of four pseudoknot structures. Figure 1 shows the pseudoknot structures on the tmRNA gene.

In order to search the DNA genomes efficiently, the combined pseudoknots 1 and 2 were used to search the genome first and 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 run 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. This 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 experiment where a whole genome of more than a million nucleotides is successfully searched for a complex structure that contains pseudoknot structures.

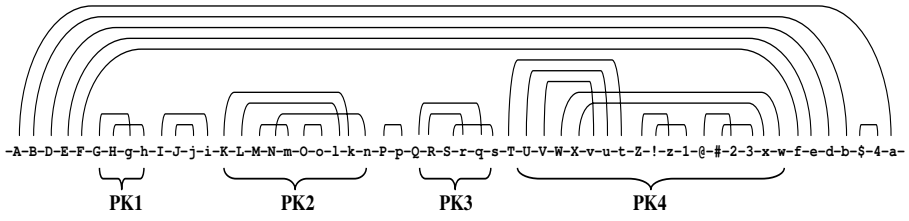

Fig. 1. Diagram of the pairing regions on the tmRNA gene. Upper case letters indicate base sequences that pair with the corresponding lower case letters. The four pseudoknots constitute the central part of the tmRNA gene and are called Pk1, Pk2, Pk3, Pk4 respectively

3 Models and Algorithms

The Covariance Model (CM) proposed by Eddy and Durbin [5][4] can effectively model the base pairs formed between nucleotides in an RNA molecule. Similar to the emission probabilities in HMMs, the emission probabilities in the CM for both unpaired nucleotides and base pairs are position 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 resulting states respectively.

The genome is scanned through by a window with an appropriate length. The ending 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 the segments 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.* [3] 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 such that two or more base pairs obtained from different components require the participation of the same nucleotide. We break the contention by introducing different priorities to components; base pairs determined from components with the highest priority win

the contention. We consider 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 its corresponding position on the genome.

More specifically, we assume that the model contains,in descending order of component weights, k CM components $M_0, M_1, \cdots, M_{k-1}$. 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.

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

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

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 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 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.

Supplementary Material

We have put RNA information we used for the estimation of model parameters, the experimental results of the model on the RNA pseudoknots, and the detailed stem contention explanation and experimental results for stem contention rates on the web. They are available at http://www.uga.edu/RNA-Informatics/pksearch/.

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. L. Cai, R. Malmberg, and Y. Wu: Stochastic Modeling of Pseudoknot Structures: A Grammatical Approach. *Bioinformatics*, 19, $i66 - i73$, 2003
4. 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
5. S. Eddy and R. Durbin: RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22: 2079-2088, 1994
6. 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
7. 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
8. 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
9. 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
10. R. B. Lyngso and C. N. S. Pederson: RNA pseudoknot prediction in energy based models. *Journal of Computational Biology*, 7: 409-428, 2000
11. 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
12. 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
13. E. Rivas and S. Eddy: The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics*, 16: 334-340, 2000
14. E. Rivas and S. Eddy: A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots. *Journal of Molecular Biology*, 285: 2053-2068, 1999
15. 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
16. 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