July 2005, Vol.20, No.4, pp.454–464 J. Comput. Sci. & Technol.

RNA Structural Homology Search with a Succinct Stochastic Grammar Model

Ying-Lei Song[1*] (宋英磊), Ji-Zhen Zhao[1] (赵继贞), Chun-Mei Liu[1] (刘春梅), Kan Liu[1] (刘　戡), Russell Malmberg[2], and Li-Ming Cai[1,*] (蔡黎明)

[1] *Department of Computer Science, University of Georgia, Athens GA30602, U.S.A.*

[2] *Department of Plant Biology, University of Georgia, Athens GA30602, U.S.A.*

E-mail: song@cs.uga.edu; jizhen@cs.uga.edu; chunmei@cs.uga.edu; kan@cs.uga.edu; russell@plantbio.uga.edu; cai@cs.uga.edu

Revised November 4, 2004.

Abstract An increasing number of structural homology search tools, mostly based on profile stochastic context-free grammars (SCFGs) have been recently developed for the non-coding RNA gene identification. SCFGs can include statistical biases that often occur in RNA sequences, necessary to profile specific RNA structures for structural homology search. In this paper, a succinct stochastic grammar model is introduced for RNA that has competitive search effectiveness. More importantly, the profiling model can be easily extended to include pseudoknots, structures that are beyond the capability of profile SCFGs. In addition, the model allows heuristics to be exploited, resulting in a significant speed-up for the CYK algorithm-based search.

Keywords RNA structural homology search, stochastic context-free grammar (SCFG), structure-sequence alignment, pseudoknot

1 Introduction

Stochastic context-free grammars (SCFGs), a natural extension from the hidden Markov model (HMM), are suitable for modeling stem-loops in RNA secondary structure[1,2]. In particular, context-free production rules are capable of modeling the nested base pairs that may involve nucleotides separated by a number of nucleotides in an RNA sequence and subsequently, conformations formed by parallel stems. Compared to Tinico's thermodynamic model[3], SCFGs can more easily include statistical biases that often occur in RNA sequences, making it suitable for them to profile specific RNA structures in structural homology search. Indeed, there have been a number of structure search tools, for example, tRNAscan-SE and Rsearch[4,5] that were developed based on the Covariance Model (CM)[2], a special type of SCFG. A CM provides a position-specific structural template and can be very effective in searching. The application of these search tools has demonstrated the successful role of grammar-based structure search in non-coding RNA gene finding[6−8], especially when the secondary structure is the most conserved characteristic of functional RNAs[9].

In order to search for more complex RNA structures at a larger (e.g., genome) scale, a number of important issues need to be fully addressed. In particular, position-specific profiling with SCFGs usually yields models large in size, making the profiling a tedious, challenging task. Smaller grammar models would also make it feasible to combine stochastic information from other sources (e.g., phylogeny)[10,11]. Moreover, a search based on the CYK alignment algorithm runs in time $O(mnw^2)$ for size m of the model and size w of the window scanning through the target sequence of length n[12]. Hence, the search could become very slow even for moderately large profiles or long target sequences (e.g., genomes)[13]. In addition, SCFGs are incapable of describing pseudoknotted structures[4,12,14] which have been frequently observed in functional RNAs[9,11,15]. Therefore it is desirable to develop stochastic grammar models that can facilitate profiling, speed up the search, and include pseudoknots.

The design of a stochastic grammar model needs to address two different but related issues: model toplogy (i.e., production rules) and parameter estimation. In spite of the success that has been achieved by SCFGs[1,11,16,17], there has been limited development of automated probability parameter estimation after the grammar (or an initial grammar) topology is devised[1,2]. A further difficulty is that automatically learning the topology of an SCFG model from training data may lead to the local maxima, as in the case of learning HMMs, by allowing all possible transitions[12]. Thus successful SCFGs are constructed mainly based on the immediate problem of interest. For example, the CM[2,12] describes alignment probability with designated position (and position pair)-specific production rules of match, insertion, and deletion. More recently, small and simple SCFGs have also been investigated[10] which can combine other stochastically expressible information from additional sources such as evolutionary models of comparative RNA sequence analysis.

In this paper, we introduce a new method to profile RNA secondary structure with succinct stochastic grammars. The method is based on a region-specific view of RNA secondary structure and consequently leads to a simple design for grammar topology. In particular, RNA secondary structure is considered to be a list of

Regular Paper

*Corresponding authors

regions that are either non-structural loops or pairing nucleotides that form stems. Each region (or a pair of regions forming a stem) is described with a designated set of a few rules, including a direct-recursion rule. The number of production rules in a succinct grammar thus only depends on the number of regions in the structure, a significantly reduced number compared to that of the CM. In order to avoid the possible loss of crucial position-specific information, the succinct model also allows the presence of position-specific rules to account for highly conserved sequence regions.

We have conducted theoretical and experimental analyses on the effectiveness of the new model. In spite of the geometric length penalty imposed by the direct-recursion rules, with very few exceptions the optimal structure obtained with the succinct model remains the same as the optimal structure obtained with the CM. When applied to searching for tRNAs and 5S rRNAs, both models achieve high sensitivity and specificity. The new model can improve its effectiveness even closer to that of the CM when the alignment is scored with a new technique the succinct model enables.

The simplicity of the succinct grammar method also makes it easy to include pseudoknots. Based on the grammatical techniques developed by Cai *et al.*[18], we are able to profile pseudoknotted structures with succinct stochastic grammars that include special nonterminals as the description of crossing stems. The profiling has been tested on two pseudoknots in tmRNAs. Our experiments have consistently shown that Z-scores of the pseudoknots exceed those of the random sequences with the same base composition, demonstrating the great potential of the succinct model in RNA pseudoknot homology search. In order to evaluate the ability of the model to recognize the profiled pseudoknots on real biological sequences, we performed experiments to search for the pseudoknots on the 3'UTR region in the genomes of tobacco mosaic virus and related viruses. The results demonstrate that the succinct profiling model can correctly identify most of the structural signals on genomes where the presence of pseudoknots has been experimentally verified[19] and, in addition, provide reasonable predictions of their locations on other genomes. The succinct model allows heuristics to be exploited to speed up the alignment and search. In particular, stochastic productions modeling a region penalize alignments where the length of the region is larger than the expected value. The possible lengths of a region can therefore be restricted to a certain range according to a quality control parameter determined by the penalty function. Such restrictions make it possible to avoid the computation of useless probabilities in the full dynamic programming-based alignment. Experimental results on tRNAs have shown that, without adversely affecting the quality of the alignment, the heuristics can increase the speed of the CYK algorithm several fold even under the requirement of high quality alignment.

2 Succinct Profile SCFGs

Based on the SCFG[1,2] developed for profiling RNA sequences, we propose a succinct stochastic context free grammar (SSCFG) to model RNA secondary structure from the perspective of regions (Durbin *et al.*[12] provides a detailed survey of SCFG). RNA secondary structure consists of stems and loops. A *loop* is a non-structural subsequence of nucleotides. A *stem* is a set of stacked base pairs, formed by two (possibly distant) subsequences which we call *base pairing regions*. The region between the two halves of a stem is called the *span* of the stem. For some stems the span is simply a (hairpin) loop while for others it can fold into more complex structure, for example a secondary structure that contains other stems and loops.

As shown in Fig.1(a), we define the RNA secondary structure as a list of base region units: $(r_1, r_2, \ldots, r_m)$, where r_i is either a base pairing region (i.e., a half of a stem) or a non-structural loop. A base pairing region forms a stem with another base pairing region; neither of them can contribute to additional stems. In particular, given a list of regions $(r_i, r_{i+1}, \ldots, r_j)$, $i \leqslant j$, the secondary structure $S(i,j)$ (without pseudoknots) defined by the list can be interpreted as follows, recursively from $5'$ (left) to $3'$ (right).

(a) if $i = j$, then r_i is a loop;

(b) if $i < j$ and r_i is a loop, then $S(i,j)$ consists of two

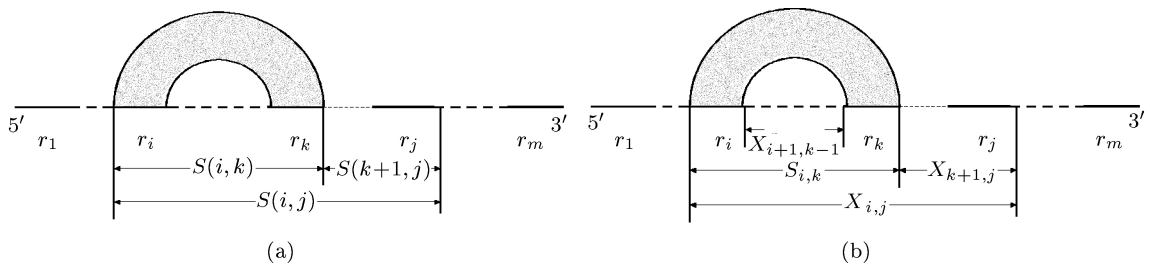

Fig.1. Modeling RNA secondary structure without pseudoknots using region-based SCFG. (a) An RNA sequence may contain a list of base region units: $(r_1, r_2, \ldots, r_m)$; $S(i,j)$ denotes the secondary structure formed on the region list $(r_i, r_{i+1}, \ldots, r_j)$. (b) Regions can be modeled with nonterminals and productions in SCFG; a deriving procedure starting with $X_{i,j}$ can generate the subsequence between regions i and j; $S_{i,k}$ represents the subsequence formed by the pairing regions i and k of a stem and its span; unstructured loop region r_j is derived from the nonterminal L_j with a direct-recursion production.

substructures: loop r_i, followed by the secondary structure $S(i+1,j)$ defined by list $(r_{i+1},\ldots,r_j)$;

(c) if r_i pairs with r_k to form a stem, for some k, $i<k\leqslant j$, then $S(i,j)$ consists of three substructures: the stem formed by regions r_i and r_k, it spans $S(i+1,k-1)$ defined by list $(r_{i+1},\ldots,r_{k-1})$, and the secondary structure $S(k+1,j)$ defined by list $(r_{k+1},\ldots,r_j)$.

Therefore, a consensus secondary structure defined by list $(r_1,r_2,\ldots,r_m)$ can be profiled with stochastic context-free grammar productions. Fig.1(b) shows the nonterminals used in a region-based profiling SCFG. Specifically, we let $X_{i,j}$ start the derivation where the structure $S(i,j)$ in the list $(r_i,\ldots,r_j)$ can be generated and L_i represents loop region r_i; nonterminal $S_{i,j}$ models the stem formed between base pairing regions r_i and r_j with span $S(i+1,k-1)$. According to definitions (a), (b) and (c), the structure $(r_i,\ldots,r_j)$ can be profiled as:

1) $X_{i,j}\rightarrow L_i$;

2) $X_{i,j}\rightarrow L_iX_{i+1,j}\,|\,X_{i+1,j}$, $L_i\rightarrow xL_i\,|\,x$;

3) $X_{i,j}\rightarrow S_{i,k}X_{k+1,j}\,|\,X_{i+1,k-1}X_{k+1,j}|X_{k+1,j}$, $S_{i,k}\rightarrow xI_{i,k}y$, $I_{i,k}\rightarrow xI_{i,k}y\,|\,X_{i+1,k-1}$.

Note that the profile can avoid overfitting by allowing the absence of substructures (which is achieved with the "optional" notation "|"). In particular, in 2), loop L_i may or may not appear. Also in 3), the stem $S_{i,k}$ may not appear. Furthermore, the span $X_{i+1,k-1}$ is independent of the presence of $S_{i,k}$ and may be absent.

For productions in 2), loop L_i is profiled with a direct-recursion production to allow the alignment of any subsequence with a length greater than one. Symbol x represents the four normal nucleotides. Note that alternatively, productions in 2) can be replaced with productions $X_{i,j}\rightarrow L_iX_{i+1,j}$ and $L_i\rightarrow L_i\,|\,\emptyset$, where $\emptyset$ represents the empty object, such that the loop may be absent from the secondary structure obtained in some alignments. In 3) stem $S_{i,k}$ is profiled to contain the first base pair and the rest of the stem $I_{i,k}$ that is profiled with a recursion rule. The derivation of the stem $S_{i,k}$ ends with the presence of nonterminal $X_{i+1,k-1}$. The pair $x-y$ represents all the 16 possible base pairs (or 24 pairs including the pairings between bases and gap Δ to allow for left and right bulges).

The probability parameters for these productions can be estimated from a set of training set sequences that have been aligned to the consensus structure. In general, approaches to estimating production probabilities need to consider two possible types of productions. First, the probabilities for productions that have $X_{i,j}$ on the left hand side in 2) can be estimated by counting the frequency of loop L_i in the training sequences. Additionally, to account for the different percentages of four normal nucleotides in the loop base composition, we compute the frequency of each normal nucleotide to appear in the loop and assign the value to the production where the nucleotide is generated. The probabilities associated with the productions in 3) can be estimated similarly from the frequency of stem $S_{i,k}$, and that of the span $X_{i+1,k-1}$ if $S_{i,k}$ does not appear. Moreover, the probability for each production that generates a given base pair is the relative frequency of the base pair in the stem. Pseudocounts are included in order to prevent overfitting of the model to the training sequences.

Second, the statistics for lengths of the recursively defined substructures (i.e., loops and stems) are considered to follow the geometric distribution. The probability for recursion rule $L_i\rightarrow xL_i$ thus must be multiplied by an additional factor of $\beta/(\beta+1)$, where β is the length mean of the subsequences in all the training sequences aligned to the loop L_i. Similarly, an additional factor of $(\alpha-1)/\alpha$ must be incorporated into the probabilities for recursion rules in the format of $I_{i,k}\rightarrow xI_{i,k}y$, where α is the length mean of the subsequences in all the training sequences aligned to the stem $S_{i,k}$.

We have shown that imposing the geometric distribution on stems and loops would not cause anomalies in the formation of a stable stem when the stem length α and the lengths of its two neighboring loops β_1,β_2 satisfy one of the properties as follows. The details of the proof are shown in Appendix A.

1) $\max\{\beta_1,\beta_2\}\leqslant(\alpha-1)\big(1+\sqrt{1+\frac{1}{\alpha-1}}\big)$;

2) $\beta_1\geqslant t(\alpha-1)$ and $\beta_2\leqslant\alpha-1+\frac{\alpha}{t-1}$, for any $t\geqslant 1+\sqrt{1+\frac{1}{\alpha-1}}$.

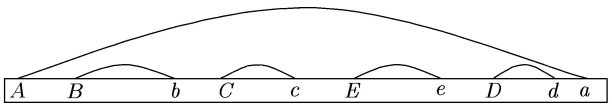

Fig.2. Diagram of the base pairing regions of tRNA molecules. Upper case letters indicate base regions that pair with the corresponding lower case letters. Stem E-e appears only in some sequences.

However, the actual length distribution of stems and loops in the training sequences may not be geometric. The geometric distribution tends to favor shorter regions, the above properties only show that it is suitable for modeling loops and stems in sequences with stable secondary structure. For sequences that contain many noncanonical base pairs in their stems, the geometric distribution may fail to identify the structural signals carried in stems when they are surrounded by a random background. Alternative distributions should be considered to alleviate the preference of shorter regions resulting from the geometric distribution. Hence, we have developed a more appropriate distribution to describe the statistics of the lengths of loops and stems. In particular, let S be a substructure (e.g., a stem) of length mean α defined with recursion rules. We would like motifs of length close to α to be penalized less than those with a larger deviation in length from α. The penalty is thus a function in $|l-\alpha|$ for the motif of length l. The function can simply be geometric or a more sophisticated one.

The computation of the new alignment penalty would be difficult to do with the full dynamic programming used by the CYK algorithm. The algorithm must

be slightly modified to incorporate the more centralized distribution. Assume the substructure stem S (without bulges) is defined with rules:

$$X \to SY, \quad S \to xIy, \quad I \to xIy|T \tag{1}$$

where the recursion rule $I \to xIy$ is used to generate base pairs other than the first pair in the stem S, and T is the span of the stem. Let $f(l, \alpha)$ be the penalty function. The probability of aligning motif $s[i..j]$ to stem S can be computed with the formula:

$$\begin{aligned} P(S,i,j) = & p(s[i],s[j])P(I,i+1,j-1)\times \\ & f(D(I,i+1,j-1)+1,\alpha) \end{aligned} \tag{2}$$

where $P(I, i+1, j-1)$, the probability of admitting the stem substructure excluding the first pair is computed with:

$$\begin{aligned} P(I,i+1,j-1) = \max\{ & p(s[i+1],s[j-1])\times \\ & P(I,i+2,j-2), P(T,i+1,j-1)\} \end{aligned} \tag{3}$$

and integer function $D(I, i+1, j-1)$, the length of segment in $s[i+1..j-1]$ aligned to the stem S (except the first base pair) is computed along with (3) as:

$$\begin{aligned} & D(I,i+1,j-1) = 1 + D(I,i+2,j-2) \quad \text{or} \\ & D(I,i+1,j-1) = 0 \end{aligned} \tag{4}$$

3 Structural Homology Search and Sensitivity Improvement

We now evaluate the performance of the succinct model in structural homology search based on the experiments conducted on tRNAs and 5S rRNAs. For tRNA search, 43 *Homo sapiens* tRNA sequences (from 70 to 90 nucleotides) were selected as the training data from http://rna.wustl.edu/GtRDB/Hs/Hsalign.html. All pairs of the selected sequences had less than 80% identity value. A pseudocount value 1.0 was used for the model construction. Two models were constructed. The CM for tRNAs had 210 nonterminals, 9 bifurcation rules, and 420 rules in total. The target sequence was a random background sequence with 10^5 nucleotides where 27 *Drosphila melanogaster* tRNA sequences (selected from http://rna.wustl.edu/GtRDB/Dm/Dmalign.html) were inserted. The random background was generated with four different types of base compositions. The scoring scheme was log-odds, in other words, the logarithm of the ratio between the probability of the alignment on the structural model and that of the null model. In addition, in order to study the possible effect of the length distribution on the searching accuracy of the model, we performed experiments on both the geometric and centralized length penalty functions. Fig.2 provides a schematic description of the secondary structure in tRNA molecules.

Table 1 compares the search effectiveness and efficiency achieved by the CM and those by the succinct model that uses the geometric length penalty function. It shows that while the CM is almost perfect in accuracy, with values of sensitivity and specificity larger than 95% in all experiments, the succinct model also achieves values of more than 80% in sensitivity and specificity. It is interesting to note that, for both models, improvement in accuracy is observed when the base composition of the real structures differs to a larger extent from that of their background. This property could be utilized in identifying noncoding RNA genes as previous work[8] has shown that the difference in base composition between RNA genes and other parts of the genomes can be significant in some cases.

In addition, Table 1 shows the results of similar experiments using 5S rRNA sequences with the succinct model and the CM respectively. The succinct model was trained with a set of training sequences downloaded from the Rfam[20] database. With a pair-wise identity value less than 80%, the training set is comprised of 30 sequences with each of them containing 115 to 130 nucleotides. The resulting CM for 5S rRNAs contained 162 nonterminals, 12 bifurcation rules and 122 rules in total, while the succinct model had 36 nonterminals, 12 bifurcation rules, and 54 rules in total. The target sequence was randomly generated with a length of 10^5 nucleotides into which was inserted 35 5S rRNAs different from the training sequences.

We examined the effectiveness discrepancy (especially in the sensitivity) between the succinct model and the CM. Part of the problem was the use of the simple geometric distribution, in general, the search is performed with the CYK algorithm applied to the window

Table 1. Performance Comparison Between the Succinct SCFG (SSCFG) and the CM in tRNA and 5S rRNA Structure Homology Search (The base compositions for the target sequences were obtained from the base frequencies of the inserted structures, for type k, $1 \leqslant k \leqslant 4$, with pseudocounts $0.05(1-k)$ for A and T, and $0.05(k-1)$ for G and C. The target sequence was of length 10^5 and the scanning window of size 160. The sensitivity and specificity are in percentage and time in second.)

Base composition	Model	tRNA			5S rRNA		
		Sensitivity	Specificity	Time	Sensitivity	Specificity	Time
C+G= 57.0%	CM	100	100	9,538	80	100	8,626
	SSCFG	81.5	91.7	4,791	74.3	100	6,007
C+G= 67.0%	CM	96.3	96.1	9,692	91.2	100	8,626
	SSCFG	81.5	88.0	4,749	77.1	100	6,019
C+G= 77.0%	CM	100	100	9,586	100	100	8,610
	SSCFG	96.3	96.3	4,728	100	100	6,107
C+G= 87.0%	CM	100	100	9,559	100	100	8,745
	SSCFG	100	100	4,643	100	100	6,027

used to scan across the target sequence. A structure motif aligned to the profile is considered significant if its score exceeds the pre-defined threshold. On a sample sequence (e.g., randomly sampled from the target sequence and with 5% to 10% of its contents), alignment scores are computed from all windows and the distribution of scores can be determined (and is usually assumed to be Gaussian). A search success must have sufficient statistical significance, which leads to the definition that a nonstructured random sequence segment receives a score higher than the threshold with a probability less than a small number P (We used $P = 0.001$ in all experiments). Due to the use of geometric distribution for lengths of stems and loops in the succinct model, a motif of length k aligned to a stem (of length mean α) receives a penalty score proportional to the amount of $(\frac{\alpha-1}{\alpha})^k$. Therefore, on random sequences that in general do not carry significant structural signals for stems at their corresponding locations, the SSCFG tends to select shorter sequence segments. Additionally, without the positional dependent states and probabilities that essentially comprise the CM, the SSCFG does not have a mechanism to penalize sequence segments signficantly shorter than the training sequences. This results in a larger mean and a relatively narrower distribution of scores on random sequences (data not shown). Moreover, due to the raised random background in an SSCFG based search, a structural motif with the profiled secondary structure would receive a score statistically less signficant than it is able to achieve on the corresponding CM model, which may give rise to the lower sensitivity we have observed in the experiments. We then replaced the geometric distribution with the centralized distribution we have developed and performed structure search experiments again on the tRNAs and 5S rRNAs using SSCFG. Table 2 presents the comparison of the searching accuracy in terms of sensitivity and specificity between the two different length distribution functions. The results show that a centralized distribution provides a more appropriate description on the statistics of region lengths and thus improves the searching accuracy. In addition, we expect this centralized distribution to achieve more significant improvement on secondary structure when a considerable number of noncanonical base pairs are involved.

4 Profile and Search for Pseudoknot Structures

In this section, we extend the succinct profiling model to include pseudoknots. Pseudoknotted structures consist of crossing stems that cannot be modeled with SCFGs and require a context-sensitive grammar. Our previous work[18] extended SCFG to include special nonterminal symbols for the modeling of crossing stems. Since the general problem of pseudoknot determination is computationally intractable[21], our method deals with restricted cases of such structures. A secondary structure including pseudoknots can also be defined with a list of base regions $(r_i, \ldots, r_j)$, $i \leqslant j$. However, due to the presence of crossing stems, such a list of base regions may contain a *sticky* base region that would pair with another sticky region outside of the structure. Therefore, there are two types of secondary structures: with and without a sticky region.

Table 2. Comparison of the Performance of the SSCFG on tRNA and 5S rRNA Structural Homology Search Using both the Geometric and Centralized Alignment Penalty Calculation. The test data are the same as those used in the experiments to obtain the results in Table 1. SE and SP are sensitivity and specificity in percentage respectively)

Base composition	tRNA				5S rRNA			
	Geometric		Centralized		Geometric		Centralized	
	SE	SP	SE	SP	SE	SP	SE	SP
C+G= 57.0%	81.5	91.7	88.0	92.3	74.3	100	77.1	100
C+G= 67.0%	81.5	88.0	88.0	96.0	77.1	100	85.7	100
C+G= 77.0%	96.3	96.3	100	100	100	100	100	100
C+G= 87.0%	100	100	100	100	100	100	100	100

As shown in Figs.3(a)–3(d), given a list of base regions $(r_i, \ldots, r_h, \ldots, r_j)$, $i \leqslant j$, the secondary structure $D(i, j, h)$ containing sticky region r_h defined by the list can be interpreted recursively from $5'$ (left) to $3'$ (right).

(a) If $i = h$, i.e., r_i is a sticky region, then $D(i, j, h)$ consists of two substructures: sticky region r_i, followed by the secondary structure $S(i + 1, j)$ defined by list $(r_{i+1}, \ldots, r_j)$ (see Section 2) that does not contain a sticky region.

(b) If $i < h$ and r_i is a loop, then $D(i, j, h)$ consists of two substructures: loop r_i followed by the secondary structure $D(i + 1, j, h)$ containing the sticky region r_h defined by list $(r_{i+1}, \ldots, r_h, \ldots, r_j)$;

(c) If r_i pairs with r_k to form a stem, for some k, $i < k \leqslant h$, then $D(i, j, h)$ consists of three substructures: the stem formed by regions r_i and r_k, its span $S(i+1, k-1)$ defined by list $(r_{i+1}, \ldots, r_{k-1})$ that does not contain a sticky region, and the secondary structure $D(k+1, j, h)$ defined by list $(r_{k+1}, \ldots, r_h, \ldots, r_j)$ that contains the sticky region r_h; and

(d) If r_i pairs with r_k to form a stem, for some k, $h < k \leqslant j$, then $D(i, j, h)$ consists of three substructures: the stem formed by regions r_i and r_k, its span $D(i + 1, k - 1, h)$ containing the sticky region r_h defined by list $(r_{i+1}, \ldots, r_h, \ldots, r_{k-1})$, and the secondary structure $S(k + 1, j)$ not containing a sticky region defined by list $(r_{k+1}, \ldots, r_j)$.

Given a list of regions $(r_i, \ldots, r_j)$, the secondary structure $Pk(i, j)$ including pseudoknots defined by the list can be interpreted as consisting of three substructures: secondary structure $D(i, k, h)$ containing the sticky region r_h, secondary structure $D(k + 1, j, l)$ containing the sticky region r_l, for $h \leqslant k < l$, and the crossing stem formed by r_h and r_l.

A consensus secondary structure $(r_1, r_2, \ldots, r_m)$ including pseudoknots can be profiled with the following stochastic grammar. In addition to the nonterminal symbols given in Section 2, let $Y_{i,j,h}$ be the nonterminal for secondary structure $D(i, j, h)$ containing the sticky region r_h. Let $T_{i,j,h}$ be the nonterminal for the stem defined by base pairing regions r_i and r_j with span $D(i + 1, k - 1, h)$. Then in addition to the rules for sec-

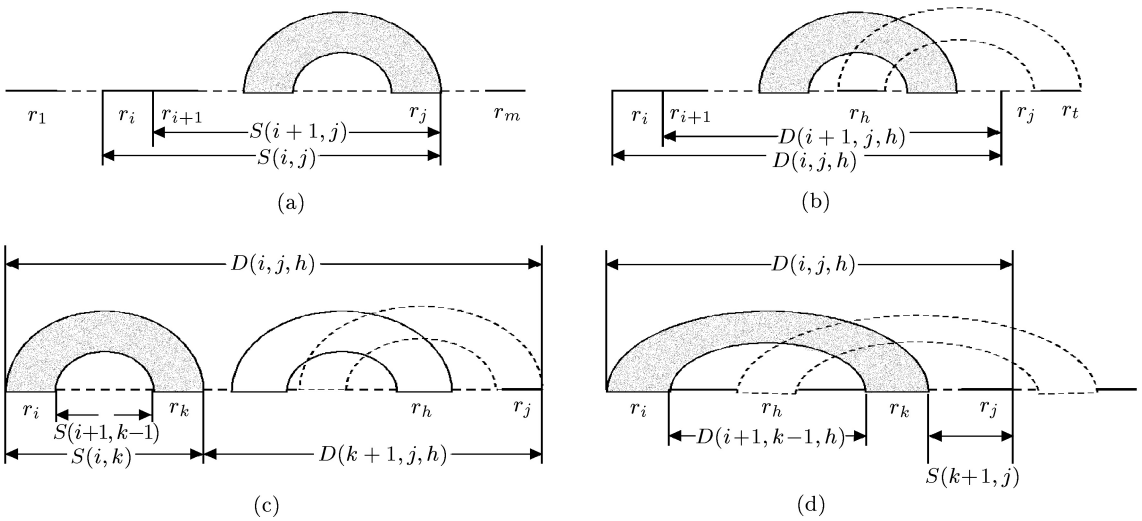

Fig.3. Modeling secondary structure with pseudoknots using region-based SCFG. (a) $i = h$, $D(i,j,h)$ contains two substructures: r_i and $S(i+1,j)$. (b) $i < h$, $D(i,j,h)$ contains two substructures: r_i and $D(i+1,j,h)$. (c) r_i pairs with r_k in a stem, for $i < k \leqslant j$, $D(i,j,h)$ contains three substructures: the stem formed by r_i and r_k, $S(i+1,k-1)$ and $D(k+1,j,h)$. (d) r_i pairs with r_k in a stem, for $h < k \leqslant j$, $D(i,j,h)$ contains three substructures: the stem formed by r_i and r_k, $D(i+1,k-1,h)$, and $S(k+1,j)$.

ondary structures without pseudoknots given in Section 2, rules for pseudoknots are:

4) $X_{i,j} \to Q_i Y_{i+1,j,l}$; and

5) $X_{i,j} \to T_{i,j,h} Y_{k+1,j,l} | Y_{i+1,k-1,h} Y_{k+1,j,l}$

$T_{i,k,h} \to x J_{i,k,h} y$

$J_{i,k,h} \to x J_{i,k,h} y | Y_{i+1,k-1,h}$

where the rules for secondary structures with a sticky region are:

6) $Y_{i,j,h} \to Q_i X_{i+1,j}$;

7) $Y_{i,j,h} \to L_i Y_{i+1,j,h}$;

8) $Y_{i,j,h} \to S_{i,k} Y_{k+1,j,h}$; and

9) $Y_{i,j,h} \to T_{i,k,h} X_{k+1,j}$.

The grammar specifies sticky region r_i with special nonterminal Q_i. As implicitly defined by the rules in 4) and 5), two such sticky regions, contained in two neighboring substructures, form a (crossing) stem. To define probability parameters for crossing stems, the following explicit rules are needed for the pair of nonterminals Q_h, Q_l to define a crossing stem:

10) $S_{h,l} \to x I_{h,l} y$, $I_{h,l} \to x I_{h,l} y | \emptyset$

where $\emptyset$ is the empty object. Notice the difference between the above rules and the rules of 3) in Section 2 for stems.

To evaluate the performance of the succinct grammar model for pseudoknot profiling, we aligned tmRNA sequences and random sequences to the pseudoknot profile and examined its ability to recognize real sequence structures from randomly shuffled ones. The alignment algorithm was constructed based on the techniques we previously developed for stochastic grammar-based pseudoknot structural alignment. From the tmRNA database (http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html) 85 tmRNAs were downloaded. The tmRNA molecules have up to 4 pseudoknots in their structure. We modeled pseudoknots 1 and 2, together with the region in between them (as shown in Fig.3). In the dataset we used, pseudoknot 1 has an average length of about 30 nucleotide bases, pseudoknot 2 has an average length of about 150. Not all the pairing regions indicated in Fig.4 are found in all sequences; in particular, in pseudoknot 2, stem $L-l$ is presented in 60% of the sequences and stem $O-o$ is presented in only 14% of the sequences.

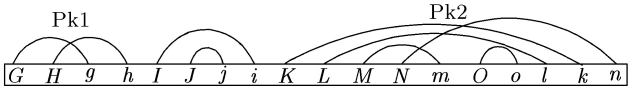

Fig.4. Diagram of the pairing regions of tmRNA pseudoknots 1 and 2 and the sequence between them. Upper case letters indicate base sequences that pair with the corresponding lower case letters. Not all structures are found in all sequences. This substructure of tmRNA, which contains 150–250 nucleotides, is called Pk1-2.

We constructed a phylogenetic tree of these sequences, then used this tree as the basis for dividing the data set into two, so that the two halves each sampled the evolutionary diversity of the data. One half set was used as a training set to estimate the required probabilities for the succinct grammar profile. The second half set was used as an evalution set. On average, a sequence in the training set has 68.3% identity with its closest match in the evaluation set. We compared the results obtained from a given sequence in the evaluation set with randomly permuted sequences of the same length and the same base composition. We took the same tmRNA pseudoknot 1 and 2 sequences used to test the structural alignment and then randomized the order of nucleotides in each sequence 50 times. We used the alignment algorithm to align random sequences to the profile. A distribution of 50 probability scores was obtained and the Z-scores were computed for both original sequences and random sequences being aligned to

the pseudoknot profile.

Fig.5 shows the comparison of Z-scores between the original sequences and the random sequences. It can be inferred from the Figure that, on average, the structural signals obtained on real sequences significantly stand out from the randomized background, which has an invariant Z-score value of around 0.7–0.9. In particular, the average Z-score for pseudoknot 1 structural alignments was 2.22, and for pseudoknot 2 the value is 3.75. In a few cases, the Z-score value of aligning the real sequence to the model was less than that of the random background obtained by averaging the Z-scores on the 50 randomly shuffled sequences. This happened 4 out of 38 times for pseudoknot 1, and 1 out of 42 times for pseudoknot 2. (We ascribe this type of error to the geometric penalty function imposed by recursion rules. The scale of the errors should be substantially reduced by employing the centralized distribution introduced in Section 2.) Thus, although random RNA sequences frequently reveal suggestive stem-loop patterns, in general, a distinct signal exists in real sequences that contain the profiled pseudoknot and methods based on our model are able to detect this signal.

To test the searching capability of the succinct stochastic grammar model on real biological data, we designed experiments that use the model to search the genomes from tobacco mosaic virus and related viruses for a domain that folds into a pseudoknot structure in the 3′UTR region and consists of five simple pseudoknots with each pseudoknot structures containing around 30 nucleotide bases. We use Pk1-5 to represent them respectively. Due to the considerable amount of running time the program needs on long sequences, we trained the grammar model with the only 5 available sequences we have and we divided the pseudoknot structure into four pieces where each piece contains one or two simple pseudoknots; the genomes are then searched for each piece. We look for hits from several of the sub-structures in the same sequence neighborhood and consider a real hit as comprised of hits that are from the results for different pieces and contiguous in locations on the genome. Table 3 shows the results of our experiments.

To summarize, the algorithm based on the succinct model successfully identified a complex multiple pseudoknot structure in viral genomes at essentially the correct location; however, some portions of these complex structures were not found correctly. It can be seen from Table 3 that the searching algorithm is able to recognize most of the structural signals of the pseudoknot structures in this particular domain. The Pk1 is not found on genomes of TMVC, TVV and RV at the corresponding locations where it is presented in those of TMVF and TMV. Sequence alignments performed manually also demonstrate that it is difficult to identify Pk1 in the part that contiguously precedes Pk2 and Pk3. For the genomes of BVQ, CMV and OPV, our results predict they should have a similar pseudoknot structure to TMV in their 3′UTR regions. However, in these genomes, the searching algorithm fails to find Pk4 in the region between Pk2-3 and Pk5, this may suggest that the structure on this region has been significantly changed by mutations or it is the only true negative of the searching program. In addition, we observed from the results that the Pk1 is not identified on locations that contiguously precedes Pk2 and Pk3 on the genomes of BVQ, CMV and OPV. However, for two of them, the BVQ and OPV, the program finds a hit on locations close to Pk2 and Pk3.

5 Speed up the Searching

Table 1 shows the running time discrepancy between the models CM and SSCFG. Because the number of bifurcation rules is the same for both models and dominates the running time of the CYK algorithm, the speed-up can be theoretically derived. In particular, let k be the number of bifurcation rules in both models, w be the window size, and N the length of the target sequence. The running time of the CYK-based search is $O(kw^2N + (m-k)wN)$ for a total number of m rules in

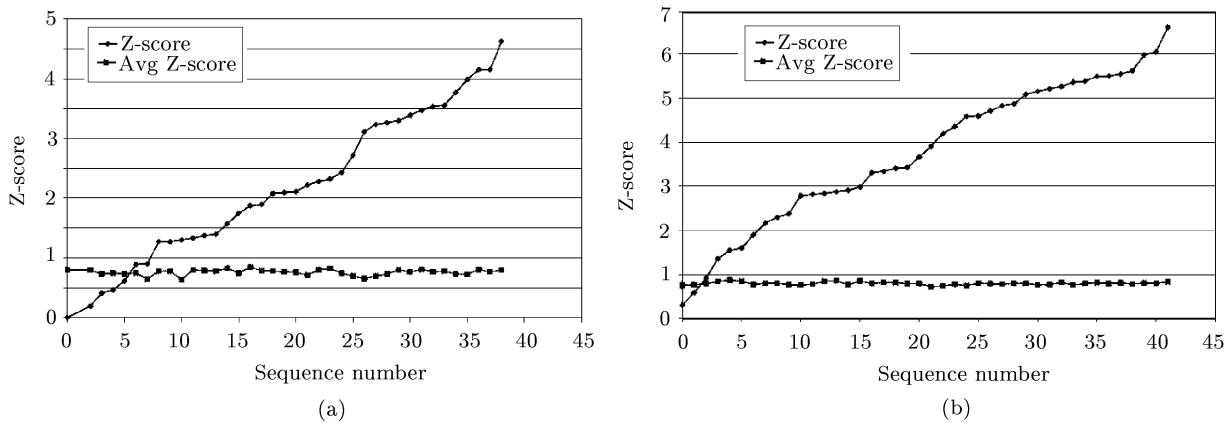

Fig.5. Structural alignment Z-scores of tmRNA sequences vs. the structural alignment Z-scores of reshuffled sequences. (a) Pk1 structure. (b) Pk2 structure. Z-scores in both plots are sorted in ascending order. It is evident from both plots that most of the real sequences have structural alignment Z-scores significantly larger than the background. Only a few exceptions are observed.

Table 3. Searching Results on the Genomes from Tobaco Mosaic Virus (TMV) Family (3′UTR pseudoknot structure is divided into four pieces with shorter lengths (less than 70 nucleotides each). For each genome, only one set of hits that are close or contiguous in locations is found. TMV is the Tobaco Mosaic Virus; BVQ is the Beet Virus Q; CMV is the Cucumber Mottle Virus; OPV is the Obuda Pepper Virus; RV is the Ribgrass Virus; TMVC and TMVF represent the TMV Crucifier and Fujian respectively. TVV is the Turnip Vein Virus; RL specifies the corresponding real location of each piece; we use N/A to mark the unavailable real location data; SL denotes the location found by the algorithm; RT is the running time; GL is the length of a genome in the number of base residues respectively)

Organism	SL(Pk1)	SL(Pk2-3)	SL(Pk4)	SL(Pk5)	RT (h)	GL(bs)
TMV	6,183–6,237	6,233–6,290	6,291–6,356	6,358–6,395	6.53	6,395
RL	6,182–6,237	6,238–6,289	6,290–6,357	6,358–6,395		
BVQ	5,922–5,978	5,798–5,857	Missing	5,963–6,003	6.12	6,003
RL	N/A	N/A	N/A	N/A		
CMV	Missing	6,262–6,319	Missing	6,387–6,424	6.72	6,424
RL	N/A	N/A	N/A	N/A		
OPV	6,161–6,215	6,342–6,401	Missing	6,469–6,506	6.81	6,506
RL	N/A	N/A	N/A	N/A		
RV	5,638–5,692	6,139–6,198	6,200–6,263	6,264–6,300	6.48	6,301
RL	N/A	6,145–6,197	6,198–6,263	6,264–6,301		
TMVC	Missing	6,142–6,201	6,203–6,266	6,267–6,303	6.48	6,304
RL	N/A	6,153–6,205	6,206–6,271	6,272–6,304		
TMVF	6,183–6,237	6,233–6,290	6,291–6,357	6,358–6,395	6.37	6,395
RL	6,182–6,237	6,238–6,289	6,290–6,357	6,358–6,395		
TVV	Missing	6,150–6,209	6,211–6,274	6,275–6,311	6.51	6,311
RL	N/A	6,156–6,208	6,209–6,274	6,275–6,311		

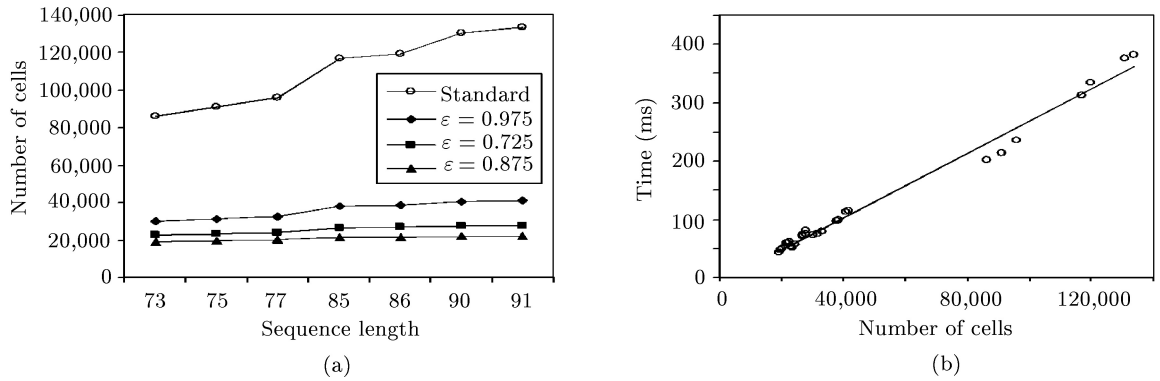

Fig.6. Performance evaluation for the speed up technique. (a) Number of cells computed by the standard and the three improved algorithms (with quality control parameter $\epsilon = 0.975, 0.925$, and 0.875). (b) (Linear) relationship between the number of computed cells and the running time.

the succinct model. For the CM, the total number of rules is about $4w$; the running time is thus $O(kw^2N + (4w-k)wN)$. The ratio of the two running times is approximately $(k-1)/(k+3)$ considering $w \gg k, m$. Note that with a careful design, the number k of bifurcation rules can be the same as the number of parallel stems in the modeled secondary structure. In the case of tRNAs, this is 5 and the ratio $(k-1)/(k+3) = \frac{1}{2}$. In our experiments, the SCFG grammar was not "optimally" crafted since it has 9 bifurcation rules as does the CM. But as illustrated by the formula, the ratio of their running time remains approximately $\frac{1}{2}$ as shown in Table 1.

Much more speed-up is possible with the succinct model. In particular, the recursion rules make it easier to exploit heuristics that may avoid the full dynamic programming in the CYK algorithm. We have observed that, for a given nonterminal X in the SSCFG, only a subset of all possible subsequences in the target sequence can be aligned to it with a score of sufficient magnitude and useful in the later computation. We thus employ an idea to identify, for each nonterminal in the SSCFG, the set of subsequences on which the alignment needs to be peformed. In other words, we identify the set of (i, j)'s for X such that the probability $P(X, i, j)$ is *valid*. This technique can effectively reduce the number of matrix cells for probabilities that need to be determined and is thus expected to signficantly improve the efficiency of the algorithm. We have developed an algorithm that can preprocess the profile SSCFG and produce a list of probability matrix cells that need to be computed. Computational details are shown in Appendix B.

We have implemented the improved CYK algorithm and conducted experiments on tRNAs to evaluate its performance on efficiency by comparing with the standard CYK algorithm for structural alignment. Twenty tRNA sequences were chosen from the website http://rna.wustl.edu/tRNAdb/ as the training data to build the succinct SCFG model consisting of 32 nonterminals and 68 rules. Among the training sequences, 16

sequences contained 4 stems and the rest have 5 stems. Seven different tRNA sequences from the same website were selected, mainly based on their diverse lengths (ranging from 73 to 91). Fig.6(a) shows the test results for different quality control parameter ϵ. Compared to the standard CYK, the number of valid cells drastically decreases in the improved algorithm. For example, the number of valid cells is reduced to the fracton of 1/8 for $\epsilon = 0.875$. Fig.6(b) shows that the running time of the improved algorithm is (linearly) proportional to the number of valid cells.

To evaluate the performance of the speed-up method in discriminating tRNAs from random structures, we tested both algorithms on the random sequences as well. We downloaded 108 tRNAs from the same website and randomly shuffled them 10 times to obtain in total $1,080$ random sequences. They were aligned to the tRNA profile (with quality control parameter $\epsilon = 0.9$); their logodds scores ranged from -29.59 to -4.71 with a mean value of -17.70 and a standard deviation of 3.85. Using the logodds score with a Z-score value of 2.33 as a threshold, 107 out of 108 real tRNA sequences were recognized from the randomly shuffled sequences. The experimental results indicate that the efficient alignment algorithm may achieve excellent accuracy in structural homology search.

6 Conclusion

We have introduced a succinct stochastic grammar model to profile RNA secondary structure to address some important issues concerning structural homology search with profile SCFGs. In particular, while being simple and small, our new model retains the qualities of being effective in alignment and search. The nature of the model also allows heuristics to be exploited so that alignment and search with the model can be significantly speeded up. In addition, the new model makes it possible to include pseudoknots that cannot be profiled with SCFGs. To evaluate its performance, especially to demonstrate the above advantages of the new model, we have also presented detailed theoretical and experimental analyses.

As the goal of future research, a number of additional issues related to the succinct model may be investigated. Firstly, in addition to the region-specific rules, the new model allows position-specific ones to be included to profile highly conserved sequence regions. However, the profiles constructed in our experiments were all region-specific rules. Secondly, in general, the construction of profiling models needs training data sets that contain a large number of sequences. It may thus be necessary to incorporate additional stochastic information from other sources into the model when the number of available training sequences is not sufficiently large. Given the simplicity of the model, it would be interesting to study how to combine it with evolutionary or biophysical models. Finally, the efficiency techniques enabled by the succinct model (see Section 5) can be immediately applied to the structural alignment and search for structures with pseudoknots, tasks that are still forbidden for profiles of even a moderate size.

Acknowledgments We would like to thank the anonymous referees for their constructive remarks on an earlier version of this paper.

References

[1] Sakakibara Y, Brown M, Hughey R *et al.* Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 1994, 22: 5112–5120.

[2] Eddy S R, Durbin R. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 1994, 22: 2079–2088.

[3] Tinico I, Borer P N, Dengler B *et al.* Improved estimation of secondary structure in ribonucleic acids. *Nature New Biology*, 1973, 246: 40–41.

[4] Lowe T M, Eddy S R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genes in genomic sequences. *Nucleic Acids Research*, 1997, 25: 955–964.

[5] Klein R J, Eddy S R. Rsearch: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 2003, 4(1): 44.

[6] Rivas E, Eddy S R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2001, 2(8).

[7] Rivas E, Klein R J, Jones T A, Eddy S R. Computational identification of non-coding RNAs in *E. coli* by comparative genomics. *Curr. Biol.*, 2001, 1(1): 1369–1373.

[8] Rivas E, Eddy S R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 2000, 16: 583–605.

[9] Eddy S R. Non-coding RNA genes and the modern RNA world. *Nature Genetics*, 2001, 2: 919–929.

[10] Dowell R D, Eddy S R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 2004, 5(1): 71.

[11] Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 1999, 15: 446–454.

[12] Durbin R, Eddy S R, Krogh A, Mitchison G J. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 1998.

[13] Weinberg Z, Ruzzo W L. Faster genome annotation of non-coding RNA families without loss of accuracy. In *Proc. the Eighth Annual Int. Conf. Research in Computational Molecular Biology*, 2004, 243–251.

[14] Brown M, Wilson C. RNA pseudoknot modeling using intersections of stochastic context-free grammars with applications to database search. In *Pacific Symposium on Biocomputing*, 1996.

[15] Felden B, Massire C, Westhof E *et al.* Phylogenetic analysis of tmRNA genes within a bacterial subgroup reveals a specific structural signature. *Nucleic Acids Research*, 2001, 29: 1602–1607.

[16] Brown M P. Small subunit ribosomal RNA modeling using stochastic context-free grammars. In *Proc. Int.*

Conf. Intelligent Systems in Molecular Biology, 2000, 8: 57–66.

[17] Holmes I, Rubin D H. Pairwise RNA structure comparison with stochastic context-free grammars. In *Pacific Symposium on Biocomputing*, 2002, pp.191–203.

[18] Cai L, Malmberg R L, Wu Y. Stochastic modeling of RNA pseudoknotted structures: A grammatical approach. In *Proceedings of the 11th Intelligent Systems for Molecular Biology*, also *Bioinformatics*, 2003, 19: 66–73.

[19] Zeenko V V, Ryabova L A, Spirin A S *et al.* Eukaryotic elongation factor 1A interacts with the upstream pseudoknot domain in the 3′ untranslated region of tobacco mosaic virus RNA. *Journal of Virology*, 2002, 76(11): 5678–5691.

[20] Griffiths-Jones S, Bateman A, Marshall M *et al.* Rfam: An RNA family database. *Nucleic Acids Research*, 2003, 31(1): 439–441.

[21] Lyngso R B, Pedersen C N S. RNA pseudoknot prediction in energy based models. *Journal of Computational Biology*, 2000, 7: 409–428.

[22] Sprinzl M, Horn C, Brown M *et al.* Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, 1998, 26(1): 148–153.

[23] Tanaka Y, Hori T, Tagaya M *et al.* Imino proton NMR analysis of HDV ribozymes: Nested double pseudoknot structure and Mg2+ion-binding site close to the catalytic core in solution. *Nucleic Acids Research*, 2002, 30: 766–774.

Ying-Lei Song received his B.S. degree in physics from Tsinghua University in 1998, and his M.S. degree in computer science from Ohio University in 2003. He is at present a Ph.D. candidate in the Department of Computer Science at University of Georgia. His research interests concentrate on designing efficient algorithms for predicting and studying secondary and tertiary structures of RNAs and proteins.

Ji-Zhen Zhao received his M.S. degree in biology from Peking University in 1997. He is currently a Ph.D. candidate in the Department of Computer Science at University of Georgia. His research interests focus on modeling of RNA secondary structures and biological networks.

Chun-Mei Liu received her B.E. and M.E. degrees in computer science and engineering from Anhui University in 1999 and 2002 respectively. She is currently a Ph.D. candidate in the Department of Computer Science at the University of Georgia. Her research interests include secondary and tertiary structures of RNAs and proteins, graph theory, and theory of computation.

Kan Liu received his B.S. degree in engineering mechanics from Beijing Institute of Technology, China, in 1995 and his M.S. degree in computer science from the University of Georgia in 2004. He is a Ph.D. candidate in the University of California, Riverside. His research focuses on developing efficient algorithms and software for computational problems in molecular biology and genomics.

Russell L. Malmberg is a professor in the Plant Biology Department, University of Georgia, USA. He received his Ph.D. degree from the University of Wisconsin in Genetics, then did postdoctoral work at Michigan State University and Cold Spring Harbor Laboratory, before moving to the University of Georgia. His current research interests are in bioinformatics and in evolutionary genetics.

Li-Ming Cai is an associate professor in the Department of Computer Science at University of Georgia. He received his Ph.D. degree in computer science from Texas A&M University in 1994. He also holds B.S. and M.S. degrees in computer science awarded by Tsinghua University. His current research interests include algorithms, computational biology, and theory of computation.

Appendix A

In this appendix, we show that imposing the geometric distribution on stems and loops would not cause anomalies, such as changing the regions that pair to form a stem, in the formation of the profiled secondary structure. We study a simple SSCFG that models a stem and the two neighboring loops. We show that, under certain conditions, base pairs are preferred to unpaired nucleotides when the length distributions of all regions follow the geometric distribution. We first examine it from the probability point of view.

Let L_1SL_2 in the profile be a substructure consisting of stem S (of length mean α) and its two neighboring loops L_1 and L_2 (of length mean β_1 and β_2 respectively). Let

$$s = y_1 \cdots y_{l_1} c_1 \cdots c_k a_1 \cdots a_h x_1 \cdots x_l$$
$$b_h \cdots b_1 d_k \cdots d_1 z_1 \cdots z_{l_2}$$

be some sequence aligned to the structure L_1SL_2, in which regions $y_1 \cdots y_{l_1}$ and $z_1 \cdots z_{l_2}$, $l_1, l_2 \geqslant 0$, are aligned to the loops L_1 and L_2 respectively, base pairs $\langle a_1, b_1 \rangle \cdots \langle a_h, b_h \rangle, h \geqslant 2$ are aligned to the stem S, and subsequence $x_1 \cdots x_l$ is aligned to the span of S. Considering the probability contributions from stems and loops due to the geometric distribution, the ratio γ between the probability of SSCFG to generate a stem comprised of base pairs

$\langle c_1, d_1\rangle \cdots, \langle c_k, d_k\rangle$ and that of the null model is:

$$\gamma = \frac{\left(\frac{\alpha-1}{\alpha}\right)^k}{\left(\frac{\beta_1}{\beta_1+1}\right)^k\left(\frac{\beta_2}{\beta_2+1}\right)^k}\prod_{i=1}^{k}\frac{p(c_i,d_i)}{p(c_i)p(d_i)} \tag{5}$$

where $p(c_i, d_i)$ is the probability to generate base pair $\langle c_i, d_i\rangle$ in the SSCFG and $q(c_i), q(d_i)$ are the probabilities to generate independent bases c_i and d_i in the null model. Note that a relatively larger value of the odds $\frac{p(c_i,d_i)}{p(c_i)p(d_i)}$ indicates that forming a base pair between c_i and d_i can increase the overall alignment probability and thus is preferred in the SSCFG model. Therefore, without counting the influence of the geometric length distribution, the total odds $\omega = \prod_{i=1}^{k}\frac{p(c_i,d_i)}{p(c_i)p(d_i)} \geqslant 1$ would imply the inclusion of the stack to the stem. To maintain such alignment under the succinct grammar model, the ratio γ needs to be greater than or equal to 1, which implies that:

$$\frac{\left(\frac{\alpha-1}{\alpha}\right)^k}{\left(\frac{\beta_1}{\beta_1+1}\right)^k\left(\frac{\beta_2}{\beta_2+1}\right)^k} \geqslant 1. \tag{6}$$

Straightforwardly, we obtain:

$$\frac{\alpha-1}{\alpha} \geqslant \frac{\beta_1}{\beta_1+1}\frac{\beta_2}{\beta_2+1}. \tag{7}$$

Therefore, imposing the geometric length distribution on stems and loops will not cause anomalies if their length means satisfy one of the two following properites:

1) $\max\{\beta_1, \beta_2\} \leqslant (\alpha-1)\left(1+\sqrt{1+\frac{1}{\alpha-1}}\right)$;

2) $\beta_1 \geqslant t(\alpha-1)$ and $\beta_2 \leqslant \alpha - 1 + \frac{\alpha}{t-1}$, for any $t \geqslant 1+\sqrt{1+\frac{1}{\alpha-1}}$.

Therefore, before we use the model based on the geometric distribution to profile RNA sequences from a particular family, we need to verify that the lengths of stems and their corresponding loops in the RNA sequences satisfying one of the above properties. For example, before we used our model to search for tRNA genes, we examined the multiple alignment of 3,491 sequences of cytoplasmic tRNA genes out of more than 5,300 sequences, which do not contain modified or non-standard bases used for the purpose of multiple alignment. Table 4 shows the length means for the four stems and their neighboring loops in the tRNAs. It is evident from the data that more than 90.0% of the sequences in the multiple alignment satisfy the property (1). This provides further justification for using the geometric distribution in SSCFG for profiling the tRNA genes.

Appendix B

This appendix presents the computational details on improving the efficiency of the CYK algorithm that aligns RNA sequences to the SSCFG model, we start with computing the set of all possible lengths of subsequences that can be derived from every nonterminal X in the SSCFG. In particular, we need to determine the set $Size(X)$ defined as:

$$Size(X) = \{k : \exists s, |s| = k, \quad \text{such that} \quad X \Rightarrow^* s\}. \tag{8}$$

Based on the dependency relationship graph for distinct nonterminals, *Size* can be computed in a bottom-up fashion for all nonterminals not involved in recursion rules. For nonterminals X involved in recursion rules, the cardinality of the $Size(X)$ would be infinite. However, since direct recursion rules describe only stems and loops, we can remove the size values in $Size(X)$ that have a high penalty score due to the length distribution, especially the size values much larger or smaller than the mean of the substructure (i.e. stem or loop) derived from X. This can be achieved by assuming a certain distribution for the sizes such that only $1-\epsilon$ fraction of sizes are removed, ϵ is called the *quality control parameter*. The standard CYK may be regarded as the case where $\epsilon = 1$.

We then compute the *outside offset pairs* for every nonterminal X. It is defined as:

$$\begin{aligned} OF(X) = \{(l, r) : \exists \alpha_1, \alpha_2, |\alpha_1| = l, |\alpha_2| = r,\\ S_0 \Rightarrow^* \alpha_1 X \alpha_2\} \end{aligned} \tag{9}$$

where S_0 is the start nonterminal in the grammar. Computing $OF(X)$ for every nonterminal X in the grammar can be done in a top-down fashion based on the dependency relationship graph, starting from the start nonterminal S_0, $OF(S_0) = \{(0, 0)\}$. We then determine the set of dynamic programming matrix cells needed for nonterminal X by:

$$\begin{aligned} VC(X) = \{(i, j) : j - i + 1 \in Size(X) \quad \text{and}\\ (i-1, n-j) \in OF(X)\}. \end{aligned} \tag{10}$$

The CYK algorithm can be modified such that it only computes probabilities $P(X, i, j)$ for $(i, j) \in VC(X)$. It is necessary to point out the computation of sizes, outside offset pairs, and valid cells only depends on the grammar model and the length of the target sequence and is independent of the content of the target sequence.

Table 4. Compliance of Cytoplasmic tRNAs with Our Theoretical Conclusion (The tRNAs were obtained from http://www.uni-bayreuth.de/departments/biochemie/sprinzl/index. html[22]. For all the 4 stems, length means for all neighboring loops are calculated. Inside loops are those loops in the span of a stem. The total odds ω (see Appendix A for details) was computed based on the stack of all base pairs in the stem excluding the first or last three pairs.)

Stem	Length mean					Compliance (%)	
	α	(outside) β_1	(outside) β_2	(inside) β_1	(inside) β_2	$\omega \geqslant 1$	Property (1)
1	7	0	2.045	2	0	99.9	yes
2	4	2	1	8.322	8.322	100	yes
3	5	1	5.11	7	7	92.2	yes
4	5	5.11	0	7	7	100	yes
Overall						98	yes