Memory efficient alignment between RNA sequences and stochastic grammar models of pseudoknots

Yinglei Song* and Chunmei Liu

Department of Computer Science,
413 Boyd Graduate Research Center,
University of Georgia, Athens, GA 30602, USA
E-mail: song@cs.uga.edu E-mail: chunmei@cs.uga.edu

Russell L. Malmberg

Department of Plant Biology,
Miller Plant Sciences Building,
University of Georgia, Athens, GA 30602, USA
E-mail: russell@plantbio.uga.edu

Congzhou He and Liming Cai*

Department of Computer Science,
413 Boyd Graduate Research Center,
University of Georgia, Athens, GA 30602, USA
E-mail: he@cs.uga.edu E-mail: cai@cs.uga.edu
*Corresponding authors

Abstract: Stochastic Context-Free Grammars (SCFG) has been shown to be effective in modelling RNA secondary structure for searches. Our previous work (Cai et al., 2003) in Stochastic Parallel Communicating Grammar Systems (SPCGS) has extended SCFG to model RNA pseudoknots. However, the alignment algorithm requires $O(n^4)$ memory for a sequence of length n. In this paper, we develop a memory efficient algorithm for sequence-structure alignments including pseudoknots. This new algorithm reduces the memory space requirement from $O(n^4)$ to $O(n^2)$ without increasing the computation time. Our experiments have shown that this novel approach can achieve excellent performance on searching for RNA pseudoknots.

Keywords: RNA pseudoknot; Stochastic Parallel Communicating Grammar Systems (SPCGS); sequence-structure alignment.

Reference to this paper should be made as follows: Song, Y., Liu, C., Malmberg, R.L., He, C. and Cai, L. (2006) 'Memory efficient alignment between RNA sequences and stochastic grammar models of pseudoknots', *Int. J. Bioinformatics Research and Applications,* Vol. 2, No. 3, pp.289–304.

Biographical notes: Yinglei Song received his BS Degree in Physics from Tsinghua University in 1998, and his MS Degree in Computer Science from Ohio University in 2003. He is at present a PhD candidate in the Department of Computer Science at the University of Georgia. His research interests concentrate on designing efficient algorithms for predicting and studying secondary and tertiary structures of RNAs and proteins.

Chunmei Liu received her BE and ME Degrees in Computer Science and Engineering from Anhui University in 1999 and 2002 respectively. She is currently a PhD 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.

Russell L. Malmberg is a Professor in the Plant Biology Department, University of Georgia, USA. He received his PhD Degree from the University of Wisconsin in Genetics, then did Postdoctoral work at Michigan State University and Cold Spring Harbour Laboratory, before moving to the University of Georgia. His current research interests are in bioinformatics and in evolutionary genetics.

Congzhou He is currently a PhD candidate in the Department of Computer Science at the University of Georgia. She also holds a PhD in Linguistics from Shanghai International Studies University. Her current research interests include natural language processing, algorithms and database management.

Liming Cai is an Associate Professor in the Department of Computer Science at the University of Georgia. He received his PhD Degree in Computer Science from Texas A&M University in 1994. He also holds BS and MS Degrees in Computer Science awarded by Tsinghua University. His current research interests include algorithms, computational biology and theory of computation.

1 Introduction

Secondary structure based search of RNAs is one of the viable approaches to finding and annotating new non-coding RNAs. In such an approach, a sophisticated alignment algorithm can be used to identify nucleotide sequences that possess a structure similarity to the structural profile modelling the RNA of interest. For modelling of the stem loop structure, which is comprised of only parallel and nested stems, SCFG have been demonstrated effective (Durbin et al., 1998; Eddy and Durbin, 1994; Brown, 2000; Holmes and Rubin, 2002; Sakakibara et al., 1994). Along with the associated CYK alignment algorithm, SCFG models have been extensively used in the search for non-coding RNA genes. Because the stem-loop structure does not contain pseudoknots, the computational costs for the search are moderate, with $O(n^4)$ time and $O(n^2)$ memory space consumptions.

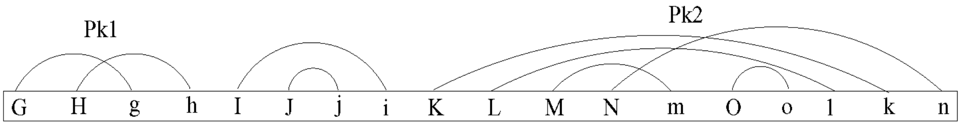
Pseudoknot structures contain structurally crossing stems in addition to parallel and nested stems. Although not as common as stem loops, more and more pseudoknots have been found as they have been looked for. Such structures are biologically important in non-coding RNAs and involved in translation (Felden et al., 1994), viral genome structure (Paillart et al., 2002), ribozyme active sites (Tanaka et al., 2002), and other functionalities (Kolk et al., 1998). The modelling of crossing patterns of base pairings in pseudoknots requires non-trivial rules other than the context-free rewriting productions (Akutsu, 2000). For example, a formal grammar system (Rivas and Eddy, 2000) has been devised to describe the legal structures identified by a thermodynamic based dynamic programming algorithm for predicting pseudoknots (Rivas and Eddy, 1999). The grammar is based on a number of auxiliary symbols used to reorder the strings generated by another context-free grammar. On the other hand, tree adjoining grammars have also been used for pseudoknot modelling (Uemura et al., 1999). However, searching with either model appears much more difficult to implement than with an SCFG.

Alternatively, our previous work modelled RNA pseudoknots based on the notion of Parallel Communicating Grammar Systems (PCGS) (Cai, 1995). We developed a new stochastic grammar model (SPCGS) that consists of a context free grammar and a few regular grammar components. In this model, crossing patterns of base pairings are modelled by the auxiliary regular grammars coordinated with the master context-free grammar. The technical advantage of this modelling is that the coordinated auxiliary regular grammars can be removed and replaced by probabilistic matrices that model the base pairings in the crossing stems. To further facilitate the SPCGS model based structure search, we also developed a dynamic programming alignment algorithm that can compute the optimal structural alignment of a sequence to the model (Cai et al., 2003). The algorithm, however, requires a large amount of memory $O(n^4)$ to accommodate the full dynamic programming table. The memory requirement prevents the application of the algorithm to searching for RNA pseudoknots of even a small length (150 nucleotides).

Due to the inherent computational intractability of the RNA pseudoknot prediction (Lyngso and Pederson, 2000; Reeder et al., 2004; Ruan et al., 2004), the computational resource consumptions may only be reduced through the use of heuristics. In this paper, we address the issue of memory reduction by introducing a heuristic approach to performing the alignment between sequences and the stochastic grammar model. In particular, based on the SPCGS model, the new approach can efficiently identify one of the crossing stems involved in a pseudoknot structure, avoiding the exhaustive search that otherwise must be carried out by the original full dynamic programming algorithm. It reduces the space requirement from $O(n^4)$ to $O(n^2)$, theoretically making it possible to align sequences of 1000 nucleotides to the structural model. The new approach does not requires additional computation time to achieve the memory efficiency.

We used several different approaches to evaluate the performance of the memory efficient approach. We downloaded the sequences for pseudoknot 1 and 2 in the bacterial tmRNA secondary structure (see Figure 1) from the tmRNA database (Knudsen et al., 2001) and divided the sequences into a training data set and a testing data set. We then constructed structural models for Pk1, Pk2, and their combination Pk12 from the sequences in the training data set. The accuracy of the memory efficient approach is evaluated by aligning the sequences in the testing data set to their structural models and comparing the results with the ones from the original algorithm. To study the structural specificity of the model, we randomly changed the order of nucleotides for each sequence in the testing data set and aligned the resulting sequences to their corresponding structural models. Reshuffled sequences provide the background distribution of the alignment scores for each sequence and hence the basis for the evaluation of the specificity. We then used the program as a searching tool to search for pseudoknot structures on both random and biological data. Our results demonstrate that the new approach can effectively detect the structural signals of the pseudoknot structures and accurately identify their locations on biological genomes.

Figure 1 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 Pk12



2 Models and algorithms

2.1 The Stochastic Parallel Communicating Grammar Systems (SPCGS) model

We first review some basics of PCGS and the SPCGS modelling of RNA pseudoknots. We refer the reader to Paun and Santean (1990), Cai (1995) and Cai et al. (2003) for more detailed descriptions of PCGS and SPCGS.

A *PCGS* (Paun and Santean, 1990) *T* consists of more than one Chomsky grammars $\Gamma = (\mathbf{G_0}, \mathbf{G_1}, \ldots, \mathbf{G_k})$ (called *components*, one of which, $\mathbf{G_0}$, is called the *master*). These grammars can share an alphabet and a set of non-terminals. Additionally, there are some special non-terminals, called *query symbols*, for communication between the grammars. The derivation of the system consists of the rewriting of every grammar component in parallel and synchronously through queries. For instance, if query symbol Q_j is present in the string ω_i produced by component $\mathbf{G_i}$, Q_j in ω_j will be replaced in the next step by the string ω_j currently produced by component $\mathbf{G_j}$. Querying a string has priority over component-wise parallel derivations (provided the string being queried does not further contain query symbols nor non-terminal symbols that are *not derivable* in the component making the query). The communication protocol allows each grammar to return to its start symbol after being queried. The language produced by the system is the set of strings produced by the master grammar component (Paun and Santean, 1990; Cai, 1995).

To discuss the PCGS modelling of RNA secondary structure, we need to define some terms.

Given an RNA sequence, a *base region* is a segment of contiguous nucleotides in the sequence. Two *base-paired* regions are two base regions that are base-paired to form a stem. In such a case, each base region is said to *contribute* to the stem.

Let *t* be a segment in sequence *s,* a *potential pairing region* in *t* is a base region such that

- it contributes to some stem in *s*
- but it does not contribute to any stem in *t.*

Then *t* is a *P-structure* if it contains a potential pairing region in it.

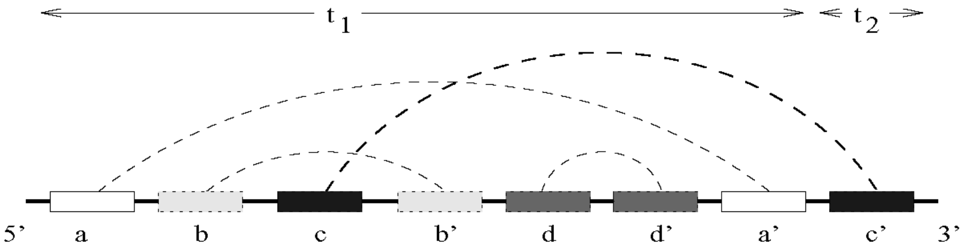
A *P*-structure is called *non-trivial* if it contains a potential pairing region that lies *between* two base-paired regions in *t.*

An RNA sequence *s* is a *pseudoknot* if *s* contains two non-overlapped *P*-structures t_1 and t_2 such that

- either t_1 or t_2 is a non-trivial *P*-structure
- the two potential pairing regions contained (respectively) within t_1 and t_2 form a stem (called *a crossing stem*).

Figure 2 describes Pk2, one of the pseudoknots in bacterial tmRNA, Pk2, in terms of *P*-structures, a part of which is shown in Figure 1.

Figure 2 The description of Pk2, one of the pseudoknots in the structure given in Figure 1, in terms of two *P*-structures t_1 and t_2, where t_1 is non-trivial, and *c* in t_1 and *c'* in t_2 are two potential pairing regions



The PCGS modelling of RNA pseudoknots are based on the above definition of pseudoknots. Table 1 shows a PCGS model of RNA pseudoknots. The model includes a master context-free component $\mathbf{G_0}$ and three auxiliary regular grammar components $\mathbf{G_1}$, $\mathbf{G_2}$, and $\mathbf{G_3}$. In $\mathbf{G_0}$, non-terminal Pk; defines a pseudoknot as consisting of two (side-by-side) *P*-structures P_1 and *Ph*. Non-terminal *prim* defines an unpaired region or loop. Respectively, non-terminals P_1 and *Ph* can derive query symbols Q_1 and Q_2. In this example, Q_1 and Q_2 are used, through querying the strings produced by the auxiliary regular grammar components, to yield two potential pairing regions that can base-pair to form a crossing stem (e.g., regions *c* and *c′* in Figure 2). Parallel derivations of the grammar components generate such two base pairing potential pairing regions modelled by Q_1 and Q_2.

Table 1 A PCGS model of RNA pseudoknots, where a, c, g, and u are terminal symbols and ϵ is the empty word

$\mathbf{G_1}$: S_1	$\to Q_2$	$\mathbf{G_2}$: S_2	$\to T$	$\mathbf{G_3}$: S_3	$\to A$	$\mathbf{G_0}$: S_0	$\to Pk$
T	$\to T_1$	T	$\to Q_3$	S_3	$\to C$	Pk	$\to P_1\ P_2$
T_1	$\to Q_3$	A	$\to Q_3\mathrm{u}$	S_3	$\to G$	P_1	$\to prim\ Q_1\ prim$
A	$\to \mathrm{a}Q_3$	C	$\to Q_3\mathrm{g}$	S_3	$\to U$	P_2	$\to prim\ Ph\ prim$
C	$\to \mathrm{c}Q_3$	G	$\to Q_3\mathrm{c}$	S_3	$\to H$	Ph	$\to \mathrm{a}\ Ph\ \mathrm{u}$
G	$\to \mathrm{g}Q_3$	U	$\to Q_3\mathrm{a}$			Ph	$\to \mathrm{c}\ Ph\ \mathrm{g}$
U	$\to \mathrm{u}Q_3$					Ph	$\to \mathrm{g}\ Ph\ \mathrm{c}$
						Ph	$\to \mathrm{a}\ Ph\ \mathrm{u}$
						Ph	$\to Q_2$
						H	$\to \epsilon$

For example, parallel derivations of the auxiliary grammars can yield base-paired regions acg and cgu. As shown in Table 2, the synchronisation between $\mathbf{G_1}$ and $\mathbf{G_2}$ is accomplished by production $S_1 \rightarrow Q_2$ because $\mathbf{G_1}$ has to wait until $\mathbf{G_2}$ starts (deriving non-terminal T) and T is copied to the current string of $\mathbf{G_1}$. After $\mathbf{G_2}$ is queried, it returns to the start symbol S_2. Then $\mathbf{G_1}$ and $\mathbf{G_2}$ make queries to $\mathbf{G_3}$ at the same time. However, the same symbol copied to $\mathbf{G_1}$ and to $\mathbf{G_2}$ invokes different derivations to generate a base and its complement, eventually producing two base-paired regions in $\mathbf{G_1}$ and $\mathbf{G_2}$, respectively. Through queries to regular components $\mathbf{G_1}$ and $\mathbf{G_2}$, the master context-free component can generate sequence … acg … caa … cgu … uug … that contains a pseudoknot with acg and cgu forming a crossing stem (Figure 3).

Table 2 Parallel derivations of base-paired regions acg and cgu. Components $\mathbf{G_1}$ and $\mathbf{G_2}$ yield acg$H$ and Hcgu, respectively

$S_1 \Rightarrow Q_2$	$S_2 \Rightarrow T$	$S_3 \Rightarrow A$
$\Rightarrow T$	$\Rightarrow S_2$	$\Rightarrow A$
$\Rightarrow T_1$	$\Rightarrow T$	$\Rightarrow A$
$\Rightarrow Q_3$	$\Rightarrow Q_3$	$\Rightarrow A$
$\Rightarrow A$	$\Rightarrow A$	$\Rightarrow S_3$
$\Rightarrow \mathtt{a}Q_3$	$\Rightarrow Q_3\mathtt{u}$	$\Rightarrow C$
$\Rightarrow \mathtt{a}C$	$\Rightarrow C\mathtt{u}$	$\Rightarrow S_3$
$\Rightarrow \mathtt{ac}Q_3$	$\Rightarrow Q_3\mathtt{gu}$	$\Rightarrow G$
$\Rightarrow \mathtt{ac}G$	$\Rightarrow G\mathtt{gu}$	$\Rightarrow S_3$
$\Rightarrow \mathtt{acg}Q_3$	$\Rightarrow Q_3\mathtt{cgu}$	$\Rightarrow H$
$\Rightarrow \mathtt{acg}H$	$\Rightarrow H\mathtt{cgu}$	$\Rightarrow S_3$

Figure 3 A derivation of sequence … acg … caa … cgu … uug … by the master context-free component of the PCGS model. Strings acg$H$ and Hcga yielded by Q_1 and Q_2 are queried from components $\mathbf{G_1}$ and $\mathbf{G_2}$, forming a crossing stem after rule $H \rightarrow \epsilon$ is applied

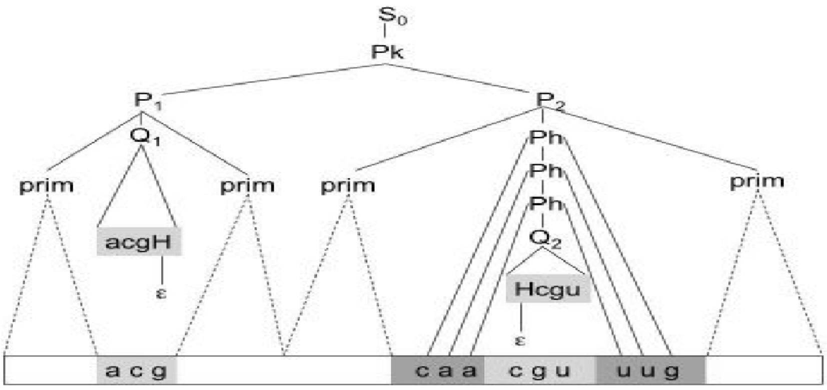


Figure 3 resembles a derivation of a context-free grammar except for the parts involving queries. In fact, when the stochastic version of the PCGS model is considered, all the auxiliary regular components can be removed and replaced by a probabilistic distribution of the base pairs in the crossing stems. This technically results in a stochastic PCGS (SPCGS) as simple as an SCFG (Cai et al., 2003).

2.2 Sequence-model alignment

The sequence-model alignment algorithm we previous developed presumes that the given SPCGS is simply an SCFG containing query symbols, such as $\mathbf{G_0}$ in Table 1. These query symbols are designated such that they exist in pairs, with each pair describing a crossing stem. In addition, for each such pair of query symbols, there is a 5×5 probability matrix defining the probability distribution for pairings among nucleotides a, c, g and u and the gap Δ.

Non-terminals in an SPCGS are grouped into 4 types: Pk-structure, *P*-structure, *N*-structure, and query. As denned in the previous section, each *P*-structure non-terminal defines a sequence contains a 'sticky' base region that can potentially pair with another outside of sequence. In contrast, *N*-structure non-terminals define sequences that do not contain a potential pairing region. The Pk-structure is a special case of the *N*-structure. Table 3 shows the non-terminal types and the corresponding abstract productions in the Chomsky normal form. In the stochastic form, a model specific probability distribution is associated with the grammar production rules.

Table 3 The types of non-terminals and corresponding productions in the master component of a SPCGS model for RNA pseudoknots. Terminal *a* represents one of the four nucleotides

Non-terminal type	*Productions*
Pk-Structure: K_r	$K_r \rightarrow P_sP_t$
P-sturcture: P_r	$P_r \rightarrow N_sP_t$, $P_r \rightarrow P_sN_t$, $P_r \rightarrow Q_s$
N-structure: N_r	$N_r \rightarrow N_sN_t$, $N_r \rightarrow a$
Query symbol: Q_r	–

The alignment between an RNA sequence and an SPCGS pseudoknot model consists of testing the derivation of the sequence with the model. As with the SCFG modelling of RNA stem loops, the probability corresponding to the derivation is the product of the probabilities of all the rules involved in the derivation. The *optimal alignment* is the one with the maximum probability.

A dynamic programming algorithm has been designed for finding the optimal alignment between a given RNA and a SPCGS model, similar to the CYK algorithm designed for optimising the statistical alignment score with an SCFG. Table 4 provides a detailed description of the recurrences used by the dynamic programming algorithm. Compared with the CYK algorithm, this algorithm needs two additional integer indices for *P*-structure non-terminals to store the probabilities associated with all possible potential pairing regions. Therefore, the algorithm has space and time complexities of $O(n^4)$ and $O(n^6)$ respectively (Cai et al., 2003).

2.3 A memory efficient alignment approach

As shown in Table 4, the original alignment algorithm, employs an exhaustive search over all possible combinations of potential pairing region pairs. However, most of the potential pairing region pairs may not need to be considered since, biochemically, only a few of them can form stable stem structures. Our heuristic approach is to identify the most likely locations of the potential pairing regions to avoid the exhaustive search required by the full dynamic programming.

Table 4 The dynamic programming alignment recurrences for all types of non-terminals in the SPCGS model. The alignment of *N*-structure non-terminals simply follows the CYK algorithm

NT	*DP recurrences*
K_r	$\alpha[K_r,i,j] = \max\limits_{i\le p\le q\le u\le v\le j}\{\beta[P_s,i,k,p,q]\times\beta[P_t,k+1,j,u,v]\times\gamma(p,q,u,v)\times Prob(K_r\rightarrow P_sP_t)\}$
P_r	$\beta[P_r,i,j,u,v] = \max\limits_{i\le k\le u}\{\alpha[N_s,i,k]\times\beta[P_t,k+1,j,u,v]\times Prob(P_r\rightarrow N_sP_t)\}$
	$\beta[P_r,i,j,u,v] = \max\limits_{v\le k\le j}\{\beta[P_s,i,k,u,v]\times\alpha[N_t,k+1,j]\times Prob(P_r\rightarrow P_sN_t)\}$
	$\beta[P_r,i,j,u,v] = \beta[Q_s,i,j,u,v]\times Prob(P_r\rightarrow Q_s)$
N_r	$\alpha[N_r,i,j] = \max\limits_{i\le k\le j}\{\alpha[N_s,i,k]\times\alpha[N_t,k+1,j]\times Prob(N_r\rightarrow N_sN_t)\}$
	$\alpha[N_r,i,j] = Prob(N_r\rightarrow a)$ for $S[i]=a$
Q_r	$\beta[Q_r,u,v,u,v]=1$
	$\beta[Q_r,i,j,u,v]=0$ for all $i\ne u$ or $j\ne v$

Since a pseudoknot structure is modelled with a bifurcation production, $K_r \rightarrow P_sP_t$ which implicitly assumes a stem across the two *P*-structures, the location of this stem can be determined with a *local complementary alignment* between the two partitioned segments respectively conformed to P_j and Pk. (Technically, one of the segments needs to be reversed and complemented to convert the base pairs in a stem into matches for the alignment.) The possible locations of the partitioning point are exhaustively searched.

Our heuristic approach involves an early detection of the pseudoknot potential pairing regions in a way that simplifies the subsequent resolution of the grammar model. For a given partitioning point, we can consider the crossing stem to be comprised of the two potential pairing regions with the maximum local complementary alignment score, which can be predetermined with a variant of the Needleman and Wunsch's algorithm. This strategy reduces the complexity of the parsing of the grammar model. A possible difficulty is that it leaves open the possibility of the local complementary alignment overlapping or conflicting with a pairing region required by the grammar. We resolve this difficulty by restricting the possible locations used in the local complementary alignment (Needleman and Wunsch, 1970). To implement this, we introduce the concept of an *offset pair* for a *P*-structure non-terminal.

The set of offset pairs for a *P*-structure non-terminal constitutes pairs of integers (*l*, *r*). An offset pair (*l*, *r*) indicates that the left and right ends of the potential pairing region yielded from a query symbol must have distance no less than *l* and *r* from the left and right ends of the subsequence on which the local complementary alignment is performed. Every offset pair provides one possible positional restriction for the potential pairing region modelled with the query symbol, and can be recursively determined from the productions in the master context-free component of the SPCGS model. Table 5 shows the recursive computation of offset pairs for *P*-structure non-terminals.

Table 5 Recursion relationships used to compute the offset pairs of *P*-structure non-terminals. The set of offset pairs for every query symbol is set to be {(0,0)}

NT	*Productions*	*Recursive computing offset pairs*
N_r	$N_r \rightarrow N_s N_t$	$len(N_r) = len(N_r) \cup \{i + j : i \in len(N_s), j \in len(N_t)\}$
	$N_r \rightarrow a$	$len(N_r) = 1$
P_r	$P_r \rightarrow N_s P_t$	$op(P_r) = op(P_r) \cup \{(i + k, j) : k \in len(N_s), (i, j) \in op(P_t)\}$
	$P_r \rightarrow P_s N_t$	$op(P_r) = op(P_r) \cup \{(i, k + j) : k \in len(N_t), (i, j) \in op(P_s)\}$
	$P_r \rightarrow Q_s$	$op(P_r) = op(P_r) \cup op(Q_s)\}$
Q_r	–	$op(Q_r) = \{(0,0)\}$

The offset pairs are determined for all *P*-structure non-terminals to integrate the conformational constraints to the memory efficient approach such that the optimal local complementary alignment is performed only on regions allowed by the model. Offset pairs are constraints imposed by the model and thus are independent of the length of the sequence.

This approach does not compute the probabilistic scores of a *P*-structure for all possible locations of the potential pairing region it contains; it only needs to compute the probability scores associated with the involved *P*-structures after the location of the potential pairing region has been determined with the local complementary alignment. This can be easily carried out with a dynamic programming approach by setting up the constraint that a query symbol must yield the predetermined region. The recurrences for performing this computation are shown in Table 6.

Table 6 The recurrences for the memory efficient alignment algorithm. For every partitioning point *k*, the location [*p*..*q*] and [*u*..*v*] of the most likely crossing stem is predetermined by the local complementary alignment

NT	*DP recurrences*
K_r	$\alpha[K_r, i, j] = \max\limits_{i \le k \le j}\{\beta_{p,q}[P_s, i, k] \times \beta_{u,v}[P_t, k+1, j] \times \gamma(p, q, u, v) \times Prob(K_r \rightarrow P_s P_t)\}$
P_r	$\beta_{u,v}[P_r, i, j] = \max\limits_{i \le k \le u}\{\alpha[N_s, i, k] \times \beta_{u,v}[P_t, k+1, j] \times Prob(P_r \rightarrow N_s P_t)\}$
	$\beta_{u,v}[P_r, i, j] = \max\limits_{u \le k \le j}\{\beta_{u,v}[P_s, i, k] \times \alpha[N_t, k+1, j] \times Prob(P_r \rightarrow P_s N_t)\}$
	$\beta_{u,v}[P_r, i, j] = \beta_{u,v}[Q_s, i, j] \times Prob(P_r \rightarrow Q_s)$

3 Experiments and results

To evaluate our approach, we have conducted RNA pseudoknot search on genomes with SPCGS and with the new alignment algorithm. We present the experimental results in this section.

To build the structural model, we need a set of sequences that have been aligned and annotated with respect to stem-loop and pseudoknot structures. The tmRNA database (Knudsen et al., 2001) provided such a dataset. We downloaded 85 aligned and structurally annotated sequences from the database. We constructed a phylogenetic tree of these sequences, then used this tree as the basis for dividing the dataset into two parts of equal size, so that the two halves each sampled the evolutionary diversity of the data. We used one half set as a training set to estimate the required probabilities for base pairs and for the grammar production rules. We used the second half set as an evaluation set to compare the results of our alignment program with the structural annotations provided by the database.

The tmRNA molecules have up to 4 pseudoknots in their structure. Figure 1 shows pseudoknot 1 (Pk1), pseudoknot 2 (Pk2) and the sequence region between them. Pk2 is complex in structure and may contain a stem loop that is present only in some training sequences. Pk1 and Pk2 are connected by a sequence region that contains two parallel sets of nested stems. In the dataset we used, Pk1 and Pk2 have average lengths of about 35 and 65 bases respectively. The two structures together with their separating regions are called Pk12 and contain about 150–250 bases.

3.1 Alignment accuracy

We assume that the region lengths of both stems and loops follow the geometric distribution[1] and the probabilities associated with productions that extend the regions of loops and stems can be computed from their average lengths, frequencies of bases and base pairs for loops and stems respectively. We then align sequences in the testing set to the model and compare the conformations obtained from the computed alignment results with their conformations annotated in the database.

Table 7 shows the results of simply counting the number of mismatches between the aligned conformation and the correct conformation as downloaded from the tmRDB database. On this basis, a comparison with the original optimal structural alignment algorithm is made and shows that the memory efficient approach can be 80% as accurate as the original one. The mismatch scores for both approaches have a correlation of around 84%, indicating that difficulties arise on the same sequences for both algorithms.[2] An examination of the individual structure predictions shows that both approaches predict most structures similarly; however, on the individual sequence structure predictions that give both methods difficulty, the new approach tends to have many more mismatch errors. The results demonstrate that the SPCGS model captures most of the conformational information in the sequences and the new approach can be as effective as the original one.

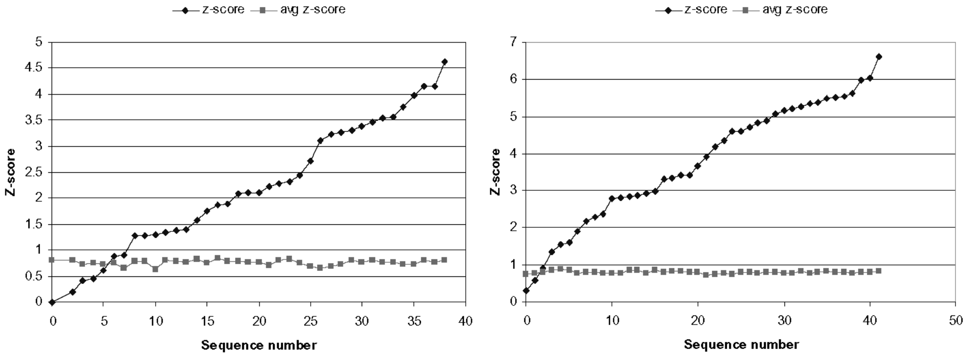
Table 7 A comparison of the mismatch scores between the optimal alignment algorithm and the memory efficient one on a few tested structures. The Pk12 is not determined (n.d.) for the optimal algorithm due to the unrealistic storage space requirement

			Mismatch score	
Sequence	*Length* (nts)	*Algorithm*	*Average*	*Standard Deviation*
Pk1	30–40	Optimal	5.0	5.6
		Memory-efficient	6.8	6.1
Pk2	60–70	Optimal	8.2	10.0
		Memory-efficient	9.6	10.0
Pk12	150–250	Optimal	n.d.	n.d.
		Memory-efficient	31	15

3.2 The sensitivity and specificity of Stochastic Parallel Communicating Grammar Systems (SPCGS)

As a means of evaluating the structural specificity of the SPCGS model, we compare the optimal alignment scores between sequences of a given structure and those of the same length and same base composition randomly reshuffled. The base order of each sequence from the testing data for Pk1 and Pk2 was randomised 50 times; the resulting sequences were then aligned to the SPCGS model with the memory efficient approach. This procedure provides the background distribution of optimal alignment scores for each sequence in the testing data set. The Z-score associated with each sequence can thus be computed from its background distribution. Figure 4 shows the Z-scores for each sequence in both testing data sets for Pk1 and Pk2 and, for comparison, their background Z-scores. It is evident from Figure 4 that most of the sequences are statistically significant with respect to the SPCGS model. The SPCGS model is capable of capturing the structural signal specific to the pseudoknot structures it models and providing a good level of specificity for structural alignments.

Figure 4 The structural alignment Z-scores of tmRNA sequences vs. the structural alignment Z-scores of reshuffled sequences. Left: Pk1 structure, and right: 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



In addition to the structural alignment, the SPCGS model may provide a possible approach to searching biological genomes for the structural signals of non-coding RNAs including pseudoknot structures. In order to test its capability on searching, we inserted sequences from testing data set for Pk1, Pk2, and Pk12 into background sequences of 10^4 nucleotides; the background sequences are randomly generated with different base compositions. In particular, we use a window to scan through a genome sequence and use the memory efficient approach to align all sequence segments in the window to a SPCGS model that specifies the searched structural pattern. We select the maximum log-odds score of all sequence segments as the score for a given position in the genome sequence. It is then compared with a threshold predetermined with a similar method as the one used in Klein and Eddy (2003) (computed based on a Z – score of 2.0). A hit is reported if the score is greater than the threshold. Table 8 shows an evaluation of the results in terms of sensitivity and specificity. It can be seen from the table that, this alignment algorithm can achieve excellent accuracy (generally around 85–95% for both sensitivity and specificity) in recognising the pseudoknot structures inserted into a random background. It is also clear that the performance of the approach does not vary significantly with the base composition of the background. Moreover, a slight improvement in both sensitivity and specificity is observed when the concentration of nucleotides C and G increases in the background.

Table 8 Results of experiments using the SPCGS model to find structural signals of pseudoknot structures inserted into the random sequences. TP is the true positives; SE and SP are sensitivity and specificity respectively.

Pseudoknot	*Total*	*TP*	*Reported*	*SE*(%)	*SP*(%)	*Time* (hr)	*C + G*(%)
Pk1	32	29	34	90.6	85.3	2.78	57.0
Pk2	28	25	31	89.3	80.6	69.52	57.0
Pk12	25	22	22	88.0	100.0	70.12	57.0
Pk1	32	31	35	96.9	88.6	2.86	67.0
Pk2	28	26	31	92.9	83.9	70.23	67.0
Pk12	24	22	22	91.7	100.0	71.45	67.0
Pk1	32	31	34	96.9	91.2	2.63	77.0
Pk2	28	26	29	92.9	89.7	71.19	77.0
Pk12	26	24	24	92.3	100.0	70.53	77.0
Pk1	32	31	32	96.9	96.9	2.56	87.0
Pk2	28	27	29	96.4	93.1	70.69	87.0
Pk12	25	24	24	96.0	100.0	70.42	87.0

3.3 Searching genomes for pseudoknots

We used the SPCGS model and the alignment algorithm to search the genomes from the family of Tobaco Mosaic Virus (TMV) RNAs for a domain that folds into a pseudoknot structure in the 3'UTR region. Figure 5 shows the structure consisting of five simple pseudoknots with each pseudoknot structure containing around 30 nucleotide bases (Zeenko et al., 2002). We use Pk1-5 to represent them respectively. Due to the considerable running time the program needs on long sequences, we trained the SPCGS model with the only five 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 consider a real hit as comprised of hits that are from the results for different pieces and contiguous in locations on the genome. Table 9 shows the results of our experiments.

Our approach 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 9 that the searching algorithm is able to recognise most of the structural signals of the pseudoknot structures in this particular domain. Pk1 is not found on genomes of TMVC, TVV and RV at the corresponding locations where it is present in those of TMVF and TMV. Sequence alignments performed manually also demonstrate that Pk1 is unlikely to appear 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 program fails to find the 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 a true negative for 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. It is likely that the hits are false positives or the repeated structural patterns of Pk1 in locations nearby.

Figure 5 The five pseudoknot structures in the 3'UTR region of the Tobaco Mosaic Virus (TMV) RNA family; upper case letters indicate base sequences that pair with the corresponding lower case letters. Diagram was made based on Zeenko et al. (2002)

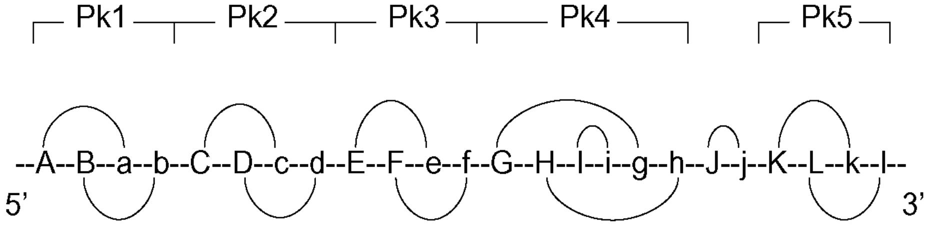

Pk1
Pk2
Pk3
Pk4
Pk5
--A--B--a--b--C--D--c--d--E--F--e--f--G--H--I--i--g--h--J--j--K--L--k--l--
5'
3'

Table 9 The searching results on a few biological genomes. RL is the real location, the real location, SL denotes the location found by the program, RT is the amount of computation time in hours, GL is the length of the genomes in the number of base residues they contain

OR	*SL*(*Pk1*)	*SL*(*Pk2-3*)	*SL*(*Pk4*)	*SL*(*Pk5*)	*RT*	*GL*
TMV	6183 – 6237	6233 – 6290	6291 – 6356	6358 – 6395	6.53	6395
RL	6182 – 6237	6238 – 6289	6290 – 6357	6358 – 6395	–	–
BVQ	5922 – 5978	5798 – 5857	Missing	5963 – 6003	6.12	6003
RL	N/A	N/A	N/A	N/A	–	–
CMV	Missing	6262 – 6319	Missing	6387 – 6424	6.72	6424
RL	N/A	N/A	N/A	N/A	–	–
OPV	6161 – 6215	6342 – 6401	Missing	6469 – 6506	6.81	6506
RL	N/A	N/A	N/A	N/A	–	–
RV	5638 – 5692	6139 – 6198	6200 – 6263	6264 – 6300	6.48	6301
RL	N/A	6145 – 6197	6198 – 6263	6264 – 6301	–	–
TMVC	Missing	6142 – 6201	6203 – 6266	6267 – 6303	6.48	6304
RL	N/A	6153 – 6205	6206 – 6271	6272 – 6304	–	–
TMVF	6183 – 6237	6233 – 6290	6291 – 6357	6358 – 6395	6.37	6395
RL	6182 – 6237	6238 – 6289	6290 – 6357	6358 – 6395	–	–
TVV	Missing	6150 – 6209	6211 – 6274	6275 – 6311	6.51	6311
RL	N/A	6156 – 6208	6209 – 6274	6275 – 6311	–	–

4 Techniques for speeding up

Although the space complexity needed for structural alignment has been significantly reduced, the worst case time complexity remains prohibitively high (of $O(n^6)$) and may become an serious issue when longer sequences are present. However, the program can be significantly speeded up during the implementation by avoiding recomputation. First, the aggregate time for local complementary alignments when the partitioning point changes contiguously can be reduced, because the computation of the optimal local alignments for a given partitioning point can be computed based on the results obtained for the preceding points. Second, the local complementary alignment results may have some locality. In other words, the location determined for the stem modelled with the query symbol may remain unchanged for a considerable number of partitioning points. This property of locality enables us to cache the probabilistic scores for *P*-structure non-terminals since it is very likely that the same local complementary alignment result would be obtained for the same partitioning point in the upcoming steps and the result can be directly used without recomputation.

5 Conclusions

In this paper, we have presented a memory efficient heuristic approach that can perform the sequence-structure alignment between RNA sequences and the SPCGS structural model of pseudoknots. Based on the biochemical stability of the stem modelled with the query symbol in the SPCGS model, the approach is able to determine the location of the crossing stem with a local complementary alignment approach and thus avoids the exhaustive search used by the original alignment algorithm. We have used the approach to study the alignment accuracy of the SPCGS model and search randomly generated sequences with pseudoknot structures inserted and a few biological genomes to annotate the structural signals of non-coding RNAs that contain pseudoknots. Our investigations have also provided a further understanding of both the SPCGS model and the pseudoknot structures and demonstrated that the SPCGS model may serve as a good basis for the profiling and searching for RNA pseudoknots.

References

Akutsu, T. (2000) 'Dynamic programming algorithms for RNA secondary prediction with pseudoknots', *Discrete Applied Mathematics*, Vol. 104, pp.45–62.

Brown, M.P. (2000) 'Small subunit ribosomal RNA modelling using Stochastic Context-Free Grammars', *Proceedings of International Conference on Intelligent Systems in Molecular Biology*, La Jolla, California, Vol. 8, pp.57–66.

Cai, L. (1995) 'Computational complexity of PCGS with regular components', *Proceedings of the 2nd International, Conference On Development of Language Theory*, pp.209–219.

Cai, L., Malmberg, R.L. and Wu, Y. (2003) 'Stochastic modelling of RNA pseudoknotted structures: a grammatical approach', *Proceedings of ISMB 2003, Bioinformatics*, Vol. 19, pp.i66–i73.

Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G.J. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, England.

Eddy, S.R. and Durbin, R. (1994) 'RNA sequence analysis using covariance models', *Nucleic Acids Research*, Vol. 22, pp.2079–2088.

Felden, B., Massire, C., Westhof, E., Atkins, J.F. and Gesteland, R.F. (1994) 'Phylogenetic analysis of tmRNA genes within a bacterial subgroup reveals a specific structural signature', *Nucleic Acids Research*, Vol. 22, pp.2079–2088.

Holmes, I. and Rubin, D.H. (2002) 'Pairwise RNA structure comparison with stochastic context-free grammars', *Pacific Symposium on Biocomputing*, Wailea, Maui, pp.191–203.

Klein, R.J. and Eddy, S.R. (2003) 'RSEARCH: finding homologs of single structured RNA sequences', *BMC Bioinformatics*, Vol. 4, p.44.

Knudsen, B., Wower, J., Zwieb, C. and Gorodkin, J. (2001) 'tmRDB (tmRNA database)', *Nucleic Acids Research*, Vol. 29, No. 1, pp.171–172.

Kolk, M.H., van der Graff, M., Wijmenga, S.S., Pleij, C.W.A., Heus, H.A. and Hilbers, C.W. (1998) 'NMR structure of a classical pseudoknot: interplay of single- and double-stranded RNA', *Science*, Vol. 280, pp.434–438.

Lyngso, R.B. and Pederson, C.N.S. (2000) 'RNA pseudoknot prediction in energy based models', *Journal of Computational Biology*, Vol. 7, pp.409–428.

Needleman, S.B. and Wunsch, C.D. (1970) 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *Journal of Molecular Biology*, Vol. 48, pp.443–453.

Paillart, J.C., Skripkin, E., Ehresmann, B., Ehresmann, C. and Marquet, R. (2002) 'In vitro evidence for a long range pseudoknot in the 5'-untranslated and matrix coding regions of HIV-1 genomic RNA', *Journal of Biological Chemistry,* Vol. 277, pp.5995–6004.

Paun, G. and Santean, L. (1990) 'Further remarks on Parallel Communicating Grammar Systems', *International Journal on Computational Mathematics*, Vol. 34, pp.187–203.

Reeder, J. and Giegerich, R. (2004) 'Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics', *BMC Bioinformatics*, Vol. 5, p.104.

Rivas, E. and Eddy, S.R. (1999) 'A dynamic programming algorithm for RNA structure prediction including pseudoknots', *Journal of Molecular Biology*, Vol. 285, pp.2053–2068.

Rivas, E. and Eddy, S.R. (2000) 'The language of RNA: a formal grammar that includes pseudoknots', *Bioinformatics*, Vol. 16, pp.334–340.

Ruan, J., Stormo, G.D. and Zhang, W. (2004) 'An iterative loop matching approaches to the prediction of RNA secondary structures with pseudoknots', *Bioinformatics*, Vol. 20, pp.58–66.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R.C. and Haussler, D. (1994) 'Stochastic Context-Free Grammars for tRNA modelling', *Nucleic Acids Research*, Vol. 22, pp.5112–5120.

Song, Y., Zhao, J., Liu, C., Liu, K., Malmberg, R. and Cai, L. (2005) 'RNA structural homology search with a succinct stochastic grammar model', *Journal of Computer Science and Technology, Special Issue in Bioinformatics*, Vol. 20, No. 4, Springer, pp.454–464.

Tanaka, Y., Hori, T., Tagaya, M., Sakamoto, T., Kurihara, Y., Katahira, M. and Uesugi, S. (2002) '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*, Vol. 30, pp.766–774.

Uemura, Y., Hasegawa, A., Kobayashi, Y. and Yokomori, T. (1999) 'Tree adjoining grammars for RNA structure prediction', *Theoretical Computer Science*, Vol. 210, pp.277–303.

Zeenko, V.V., Ryabova, L.A., Spirin, A.S., Rothnie, H.M., Hess, D., Browning, K.S. and Hohn, T. (2002) 'Eukaryotic elongation factor 1A interacts with the upstream pseudoknot domain in the 3' untranslated region of tobacco mosaic virus RNA', *Journal of Virology*, Vol. 76, No. 11, pp.5678–5691.

Notes

[1] Alternative probability distributions can be used, which may result in more accurate alignments (Song *et al*., 2005).

[2] In our another investigation on the issue (Song *et al*., 2005), we attributed the difficulty to the geometric distribution assumption on the lengths of stems and loops.