Computational Biology and Chemistry 32 (2008) 264–272

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Detecting conserved secondary structures in RNA molecules using constrained structural alignment

Mugdha Khaladkar[a], Vandanaben Patel[a], Vivian Bellofatto[b], Jeffrey Wilusz[c], Jason T.L. Wang[a,b,*]

[a] Bioinformatics Program and Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA
[b] Department of Microbiology and Molecular Genetics, University of Medicine and Dentistry of New Jersey-New Jersey Medical School, International Center for Public Health, 225 Warren Street, Newark, NJ 07103, USA
[c] Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO 80523, USA

ARTICLE INFO

Article history:
Received 20 November 2007
Received in revised form 21 March 2008
Accepted 24 March 2008

Keywords:
Constrained structural alignment
RNA motif
Viruses and protozoa

ABSTRACT

Constrained sequence alignment has been studied extensively in the past. Different forms of constraints have been investigated, where a constraint can be a subsequence, a regular expression, or a probability matrix of symbols and positions. However, constrained structural alignment has been investigated to a much lesser extent. In this paper, we present an efficient method for constrained structural alignment and apply the method to detecting conserved secondary structures, or structural motifs, in a set of RNA molecules. The proposed method combines both sequence and structural information of RNAs to find an optimal local alignment between two RNA secondary structures, one of which is a query and the other is a subject structure in the given set. The method allows a biologist to annotate conserved regions, or constraints, in the query RNA structure and incorporates these regions into the alignment process to obtain biologically more meaningful alignment scores. A statistical measure is developed to assess the significance of the scores. Experimental results based on detecting internal ribosome entry sites in the RNA molecules of hepatitis C virus and *Trypanosoma brucei* demonstrate the effectiveness of the proposed method and its superiority over existing techniques.

1. Introduction

In recent years, it is becoming clear that post-transcriptional processes at the RNA level play a major role in determining the complexity of the proteome along with a significant amount of regulation of gene expression (McKee and Silver, 2007; Sanchez-Diaz and Penalva, 2006). Numerous examples of co-regulation of sets of transcripts in RNA regulons have also been described (Keene, 2007). The identification characterization of RNA sequence and structural regulatory elements, therefore, is of fundamental importance to molecular biology (Ambros et al., 2003; Griffiths-Jones et al., 2003).

Inspired by the success of proteomics using sequence-based techniques, researchers anticipated achieving the same level of success in RNA study. Unfortunately, till now the accomplishment is far from what had been expected. A typical example is with RNA motif exploration: unlike protein motif searching which can be accomplished through the development of sophisticated amino acid substitution matrices and sequence alignment tools, detecting RNA motifs is still at a primitive stage without broadly accepted methods in the literature. One important reason for the failure of substitution matrices-based alignment methods in analyzing RNA sequences is that nucleotide bases do not carry as much functional information as amino acid residues do (Gautheret and Lambert, 2001). To properly characterize an RNA motif, information concerning both distant base interactions and sequential nucleotide composition is required to define its structure, and hence its function.

At the sequence level, one important topic is to measure the similarity of two biosequences (Bork et al., 1992; Green et al., 1993). The next step is to find an alignment between two sequences or among several sequences. Tools capable of performing sequence alignments include BLAST (Altschul et al., 1990), FASTA (Pearson and Lipman, 1988), ClustalW (Thompson et al., 1994), with their primary goal of detecting homologs from sequence databases.

However, biological activities of many molecules, such as non-coding functional RNAs, are largely dependent on their secondary or tertiary structures. Furthermore, it has been observed that myriad functions involved in post-transcriptional gene regulation are accomplished by RNA–protein-binding mechanisms, which

* Corresponding author at: Bioinformatics Program and Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA.
Tel.: +973 596 3396; fax: +973 596 5777.
E-mail address: jason.t.wang@njit.edu (J.T.L. Wang).

1476-9271/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2008.03.013

require conserved structural RNA motifs be present at the binding sites. Thus, it is biologically justifiable that conserved RNA motifs in the form of secondary or tertiary structure could be more important and informative than those in the primary sequence format (Eddy, 2002).

In this paper we propose a new approach to RNA secondary structure alignment and present an application of our approach to searching for conserved secondary structures, or structural motifs, in RNAs. The problem we tackle here is defined as follows: given a query structure Q and a set of RNA subject structures, find the subject structures that are most similar to the query structure where the similarity between the query structure Q and a subject structure S is measured by the score of local matches between Q and S. When the query structure is a structural motif or a conserved secondary structure, the problem becomes finding those subject structures containing the conserved secondary structure and displaying the locations of the conserved secondary structure in those subject structures.

Central to our approach is an efficient constrained structural alignment (CSA) method for comparing two RNA secondary structures with quadratic time and space complexities. The CSA method allows the user to annotate a portion of the query structure, or the entire query structure, as conserved, and then uses this information, or *constraint*, to align the query structure Q with each subject structure S in the given set. The constraint guides the alignment process, which dynamically varies the alignment scores between portions of Q and S to obtain a more accurate alignment between the two structures.

There are two groups of work that are closely related to ours. The first group is concerned with constrained sequence alignment (He and Arslan, 2005; Lu and Huang, 2005), which requires the incorporation of new scoring formulas into the recurrence equations of the dynamic programming algorithms employed by the existing sequence alignment tools. Different forms of constraints have been investigated, where a constraint can be a subsequence, a regular expression, or a probability matrix of symbols and positions.

The second group is concerned with RNA secondary structure alignment. Shapiro and Zhang (1990) and Wang et al. (1996), Wang et al. (1998) used a tree-based model to coarsely represent RNA secondary structures as trees, and compared these trees based on edit distance. The RNA structures are obtained by folding RNA sequences using either mfold (Zuker, 2003) or RNAfold (Hofacker, 2003). Jiang et al. (2002) considered a general edit distance for comparing RNA secondary structures. RNAforester (Hochsmann et al., 2003) extended the tree model to a forest model.

Corpet and Michot (1994) designed RNAlign to provide more rigorous RNA structural comparisons at the cost of computing efficiency: $O(n^4)$ in space and $O(n^5)$ in time where n is the length of the RNA structures to be compared. Several other tools are available that carry out RNA folding and alignment at the same time, such as Dynalign (Mathews and Turner, 2002) and FOLDALIGN (Gorodkin et al., 2001). These tools can achieve better structure prediction and alignment at the expense of computing time. In addition, algorithms using derivative-free optimization techniques, such as genetic algorithms and simulated annealing (Kim et al., 1996; Notredame et al., 1997), have been proposed to increase the accuracy in structure-based RNA alignment. Most of these methods suffer from high time complexities, making the structure-based RNA tools much less efficient than sequence-based tools.

There are pattern-matching methods for RNA analysis (Gautheret and Lambert, 2001; Laferriere et al., 1994; Pesole et al., 2000). Pesole et al. (2000) proposed a sequence-scanning technique, called PatSearch. The pattern present in an RNA secondary structure is depicted by a series of pattern description units. The sequences in a dataset are scanned one by one to decide whether the given pattern can match these sequences. In another related study (Gautheret and Lambert, 2001), a profile-based sequence-scanning algorithm was proposed and implemented under the name ERPIN by Gautheret and Lambert. Like most statistical model based methods, ERPIN requires a multiple alignment of sequences with secondary structure annotation and infers a statistical secondary structure profile (SSP). This SSP is then matched with the sequences in the dataset by using a dynamic programming algorithm to calculate scores of the best matches.

Some probabilistic models, such as stochastic context-free grammars (SCFGs) (Sakakibara et al., 1994) and covariance models (CMs) (Eddy and Durbin, 1994), have been applied to RNA structural alignment. A model is first trained by a set of manually curated sequences with known structural similarities. The trained model is then used to compare with other related RNA structures. Since a prior multiple sequence alignment (with structural annotation) is needed to train the model, its applicability is limited to RNA types for which structures of a large number of sequences are available, such as snoRNA and tRNA (Lowe and Eddy, 1999; Sakakibara et al., 1994). Klein and Eddy (2003) extended the SCFGs to find homologs of structured RNA sequences using RIBOSUM substitution matrices derived from ribosomal RNAs to score the matches in single-stranded (ss) and double-stranded (ds) regions. Motivated by SCFGs, Holmes and Rubin (2002) proposed a pairwise SCFG approach for comparing RNA structures directly. They introduced an idea of "fold envelope" to improve efficiency by confining the search space involved in calculations. However, the pairwise SCFG method requires computing time as high as $O(n^3)$ (Klein and Eddy, 2003). More recently, better algorithms based on the probabilistic models have been developed (Holmes, 2005; Yao et al., 2006). However these methods do not deal with constrained alignments as described in the next section.

2. Methods

Constrained structural alignment (CSA) constructs the alignment between a query RNA structure and a subject RNA structure based upon the knowledge of the conserved region in the query structure. We have implemented our CSA method in a web server, called RADAR (Khaladkar et al., 2007), accessible at http://datalab.njit.edu/software.

Fig. 1 shows the input interface of RADAR for aligning a query structure with a set of subject structures. The structures are represented in the Vienna style Dot Bracket format (Hofacker, 2003). Each position of the conserved region in the query RNA structure is marked using a special character "*". Fig. 2 shows the output obtained from the input data in Fig. 1, where RADAR compares the query structure with each subject structure using the proposed CSA method and ranks the subject structures in the dataset based upon their similarities to the query structure. The top ranked subject structure is most similar to the query structure, with the maximum alignment score. The score diminishes as the quality of the alignment decreases. A statistical measure, namely a p-value, is associated with each alignment score, which indicates the significance of the score. The smaller the p-value, the more significant and more reliable the score is. We describe below how to calculate the alignment scores and their p-values.

2.1. Extended Loop and Structural Component

The proposed CSA method is built on our previously developed RSmatch algorithm for RNA structural alignment (Liu et al., 2005). We model RNAs using a structural decomposition scheme similar to the loop-decomposition method commonly used in RNA structure prediction algorithms (Zuker, 2003). Thus pseudoknots are not allo-

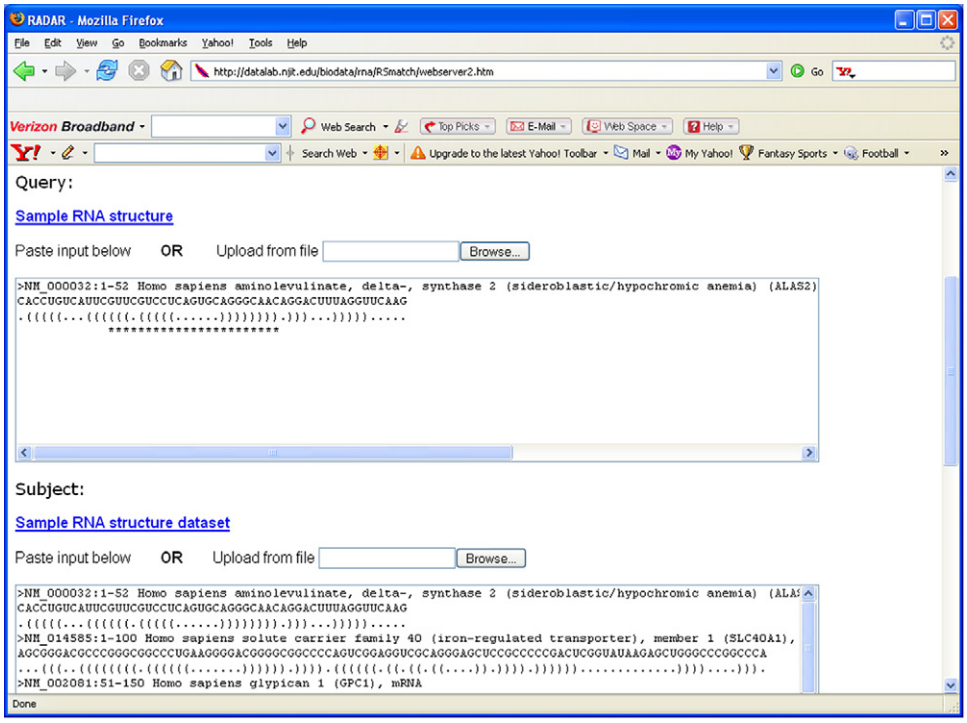

Fig. 1. The input interface of RADAR for constrained structural alignment. The first text box contains the query structure. The constrained region in the query is marked with "*". The second text box lists the subject RNA structures that form the dataset.

wed. An RNA secondary structure is completely decomposed into units called *extended loops*; cf. Fig. 3(A). An extended loop, or simply a loop when the context is clear, is a set of structural components (single bases or base-pairs), which are reachable from one another by traversing within the loop without crossing any bond. The extended loops considered in this paper differ from the commonly used loops described by Zuker (2003) in that the extended loops can be part of a stem in an RNA secondary structure.

The above-obtained extended loops can be organized into a hierarchical tree according to their relative positions in the secondary

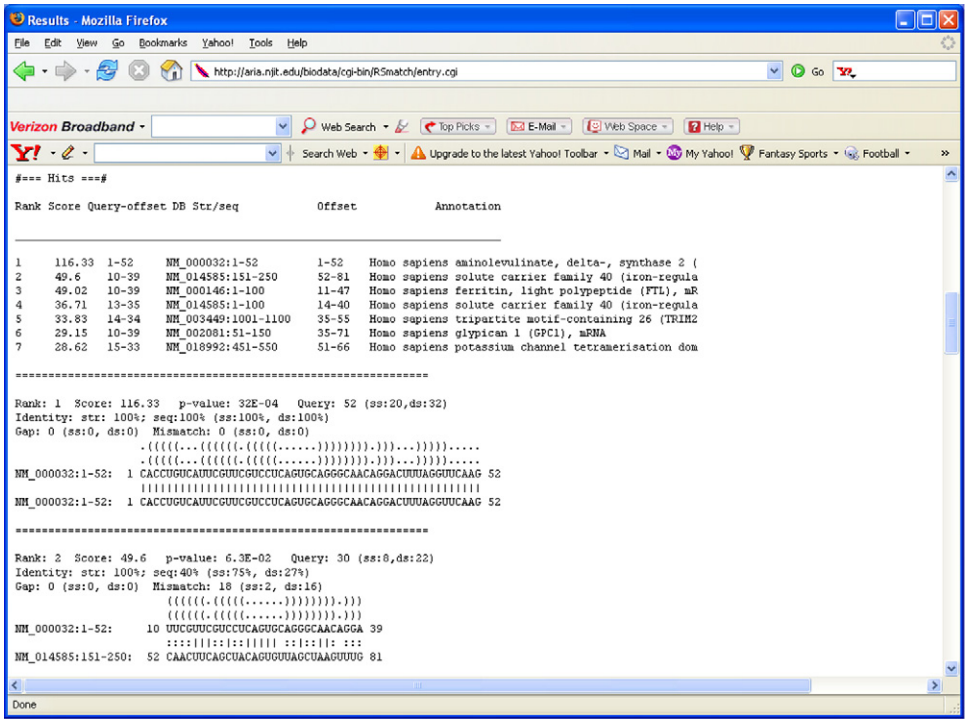

Fig. 2. The output obtained after performing the constrained structural alignment between the query structure and subject structures in Fig. 1. It first lists down a summary of the top ranked alignments including the score, subject structure name, region aligned, and so on, for each of the alignments. Then each of the alignments is shown one after the other starting with the best alignment.

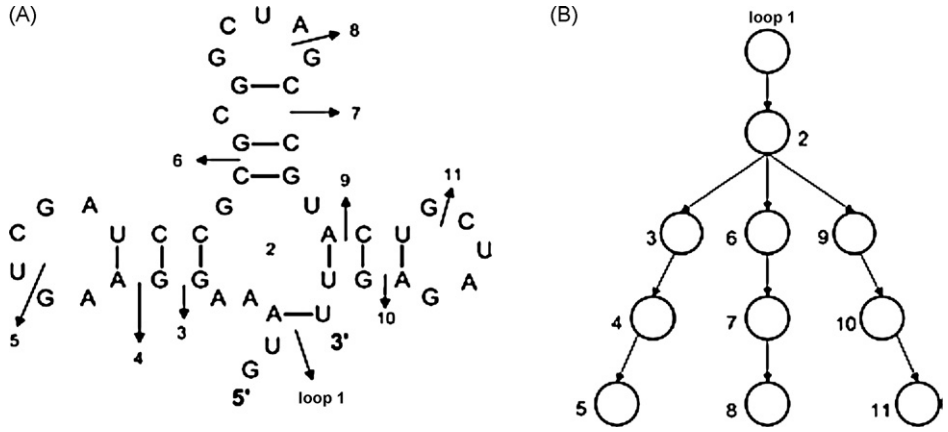

Fig. 3. (A) A hypothetical RNA secondary structure is decomposed into extended loops. (B) The hierarchical tree comprising the extended loops for the RNA secondary structure in (A).

structure, where each node in the tree corresponds to an extended loop, cf. Fig. 3(B). The tree construction is as follows. The root node is established as the extended loop containing the 5′-most and 3′-most bases. Within the root loop, each base-pair r is used to form a subtree (or child tree) whose root corresponds to another extended loop containing r. This process is iteratively performed until no further extended loop can be found and the tree is completely constructed. Furthermore, we require that the nucleotide pairs be processed from 5′ to 3′ within the extended loops. Consequently, the final tree is an ordered tree in which the order among sibling nodes is important.

In describing the relative positions between two structural components (single base or base pairs), we consider the precedence and hierarchical relationships between them. Let c_1 and c_2 be two structural components in an RNA sequence and its secondary structure. We say c_1 precedes c_2 if at the sequence level the 3′-base of c_1 is closer to the sequence's 5′-end than the 3′-base of c_2. To specify the hierarchical relationship of c_1 and c_2, we need to establish a mapping from the structural components to extended loops in the tree that represents the RNA secondary structure. It is obvious that each single base component can be mapped to a unique loop. However, a base pair component can be mapped to up to two alternate loops where one is an ancestor of the other. To resolve this ambiguity, we choose the ancestor loop as the base pair's mapping target. Suppose c_1 is mapped to loop e_1 and c_2 is mapped to loop e_2. The hierarchical relationship between c_1 and c_2 is one of the following: (1) c_1 is hierarchically identical to c_2 if e_1 and e_2 are the same; (2) c_1 is an ancestor (descendant, respectively) of c_2 if e_1 is an ancestor (descendant, respectively) of e_2; or (3) c_1 and c_2 are cousins if e_1 and e_2 are cousins or siblings in the tree.

2.2. Partial Structure

A structural component is either a single base or a base pair. The *partial structure* induced by a structural component α, is a set of structural components S_α such that for any structural component c $c \in S_\alpha$ the following three conditions are satisfied: (1) c precedes α; (2) c is not an ancestor of α; and (3) α itself belongs to S_α. Furthermore, since a base pair could appear in two extended loops, the partial structure induced by a base pair could be divided into two smaller substructures: *parent structure* and *child structure* (Fig. 4). Formally, if the structural component α is a base-pair, its parent structure is a set of components $P_\alpha \subset S_\alpha$ (excluding α) such that for any component c $c \in P_\alpha$, c′s 3′-base is always 5′ upstream of α′s 5′-base; its child structure contains a set of components $C_\alpha \subseteq S_\alpha$ (including α itself) such that for any component $c \in C_\alpha$, c′s 5′-base is always 3′ downstream of α′s 5′-base. It can be verified that $P_\alpha \cup C_\alpha = S_\alpha$ and $P_\alpha \cap C_\alpha = \varphi$.

Using the concept of partial structures, we progressively align two given RNA secondary structures using a dynamic programming (DP) algorithm by initially aligning smaller partial structures and expanding each partial structure one structural component at a time. Ultimately, the two partial structures will become the two overall structures, and the DP scoring table will be fully filled with alignment scores from which we can find the optimal local alignment between the two given RNA secondary structures.

2.3. Scoring Scheme

To measure the quality of an alignment, a scoring scheme must be provided. The proposed CSA method leaves great latitude for the choice of various scoring schemes. One important aspect of a scoring scheme is to define an alignment function of two structural components to measure the quality of matching one component to the other.

The other important aspect is a penalty parameter, which punishes the action of aligning structural component(s) to gap(s).

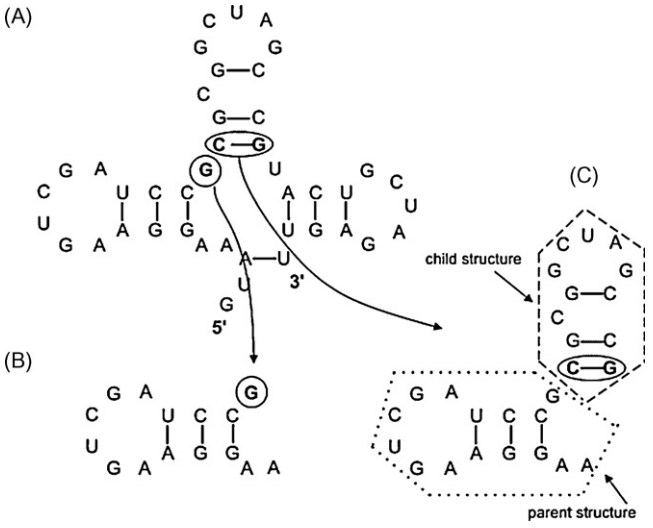

Fig. 4. (A) A hypothetical RNA secondary structure is used to illustrate how partial structures are determined. (B) The partial structure induced by the single base G is shown. (C) The partial structure induced by the base pair C–G consists of two parts, a parent structure and a child structure. The base pair is included in the child structure.

During the course of computation, one structural component (single base or base pair) could be matched to a gap; or one parent substructure or child substructure could also be matched to a big gap. Intuitively, the bigger the gap, the heavier the penalty is. In our implementation, we set a basic penalty for the smallest gap involving only one base. Then the larger gap will be punished proportionally to the number of bases involved in the gap. We use μ to denote the basic penalty in the following discussions. Let x be a structural component in the query structure and let y be a structural component in the subject structure. Let $h(x, y)$ denote the alignment score between x and y. We can extend this function to represent the alignment score between two substructures D_Q, D_S from the query structure Q and the subject structure S, respectively, as follows:

$$\varphi(D_Q, D_S) = \sum_{\substack{i \in D_Q \\ j \in D_S}} h(i, j) + \mu G \tag{1}$$

where G represents the total number of gaps in aligning D_Q and D_S.

In calculating the alignment function h, we need to consider the constraint, or conserved region, annotated in the query structure. Refer to Fig. 1. Each position of the conserved region in the query RNA structure is marked using a special character '*' underneath the position. This is termed *binary 0/1 conservation* since any position in the query RNA structure is treated to be either 100% conserved (if it is marked with '*') or not conserved at all. If it is found, from wet lab experiments or other sources, that a particular RNA structure contains a motif that we want to search for in other RNA structures in a data set, then that particular RNA structure can be used as a query structure and that motif region can be marked by '*' as conserved in the query structure.

Let $g(\alpha,\beta)$ be the alignment score between two structural components α,β where no constraint is involved. In our implementation presented here, $g(\alpha,\beta)$ is similar to that defined in (Liu et al., 2005), as shown below:

$$g(\alpha, \beta) = \begin{cases} 1 & \text{if } \alpha, \beta \text{ are single bases and } \alpha = \beta \\ -1 & \text{if } \alpha, \beta \text{ are single bases and } \alpha \neq \beta \\ -2 & \text{if } \alpha \text{ is a single base and } \beta \text{ is a gap, or vice versa} \\ 3 & \text{if } \alpha, \beta \text{ are base pairs and } \alpha = \beta \\ 1 & \text{if } \alpha, \beta \text{ are base pairs and } \alpha \neq \beta \\ -4 & \text{if } \alpha \text{ is a base pair and } \beta \text{ is a gap, or vice versa} \end{cases} \tag{2}$$

The alignment function h in Eq. (1) is calculated by

$$h(x, y) = \begin{cases} \lambda g(x', y) & \text{if } x \text{ is constrained} \\ g(x, y) & \text{otherwise} \end{cases} \tag{3}$$

where x (y, respectively) is a structural component in the query RNA structure (subject RNA structure, respectively), and λ is used to increase or diminish the score to take into account the conserved region in the query structure. When x is constrained, we use x' to represent the corresponding structural component without the constraint.

With binary 0/1 conservation, λ is defined as

$$\lambda = 1 + \frac{L}{N} \tag{4}$$

where L is the length of the conserved region and N is the total length of the query RNA structure.

2.4. Recurrence Formulas

In this section we present scoring formulas for aligning partial structures induced by structural components from the query structure Q and the subject structure S, respectively. The recurrence formulas in the proposed dynamic programming algorithm take into account the constraint occurring in the query structure. Notice that when a structural component involved in an alignment is a base pair, we only need to consider the child and partial structures induced by the base pair (Liu et al., 2005). The reason is that the parent structure induced by a base pair can always be derived as a partial structure induced by another structural component and hence is considered when the alignment score of that structural component is calculated (Liu et al., 2005).

Given the query RNA structure Q and the subject structure S, the proposed CSA method is a dynamic programming (DP) algorithm that matches partial structures from Q and S, respectively. Let x be a single base in Q and let y be a single base in S. Let x^p denote the structural component that precedes x. In matching the partial structure S_x with the partial structure S_y there are three cases: (i) x is aligned with y; (ii) x is aligned with a gap; and (iii) y is aligned with a gap. Thus the score of matching S_x with S_y can be calculated by the following equation:

$$\varphi(S_x, S_y) = \max \begin{cases} \varphi(S_{x^p}, S_{y^p}) + h(x, y) \\ \varphi(S_{x^p}, S_y) + \mu \\ \varphi(S_x, S_{y^p}) + \mu \end{cases} \tag{5}$$

where $h(x,y)$ is defined in Eq. (3) and $\mu = -2$ is the basic penalty for aligning a base with a gap, cf. Eq. (2).

Next, we consider the situation where x is a base pair and y is a single base. (The situation where x is a single base and y is a base pair is similar and hence omitted.) As discussed before, besides the partial structure S_x we have to consider the child structure C_x for the base pair x. We first calculate the structural alignment score between the child structure C_x and the partial structure S_y. There are two cases: (i) the single base component y is aligned with a gap and (ii) the base pair x is aligned with a gap. Therefore we have

$$\varphi(C_x, S_y) = \max \begin{cases} \varphi(C_x, S_{y^p}) + \mu \\ \varphi(S_{x^p}, S_y) + 2\mu \end{cases} \tag{6}$$

In aligning the partial structure S_x with the partial structure S_y, there are three cases: (i) the single base y matches with a gap; (ii) the partial structure S_y matches with the child structure C_x; (iii) the partial structure S_y matches with the parent structure P_x. Thus

$$\varphi(S_x, S_y) = \max \begin{cases} \varphi(S_x, S_{y^p}) + \mu \\ \varphi(C_x, S_y) + |P_x|\mu \\ \varphi(P_x, S_y) + |C_x|\mu \end{cases} \tag{7}$$

Then we consider the situation where x is a base pair and y is also a base pair. We need to compute four alignment scores because each base pair corresponds to two structures: one child structure and one partial structure. While aligning the child structure C_x with the child structure C_y, it is clear that

$$\varphi(C_x, C_y) = \max \begin{cases} \varphi(S_{x^p}, S_{y^p}) + h(x, y) \\ \varphi(S_{x^p}, C_y) + 2\mu \\ \varphi(C_x, S_{y^p}) + 2\mu \end{cases} \tag{8}$$

since both x and y are the last components in the respective child structures.

Eq. (9) gives the alignment score between the partial structure S_x and the child structure C_y:

$$\varphi(S_x, C_y) = \max \begin{cases} \varphi(S_x, S_{y^p}) + 2\mu \\ \varphi(P_x, C_y) + |C_x|\mu \\ \varphi(C_x, C_y) + |P_x|\mu \end{cases} \tag{9}$$

The first case corresponds to that y is aligned with a gap. If y does not match with a gap, it can be shown that, the second and third cases in Eq. (9) cover all possible situations. Similarly, we can

calculate the score of aligning the child structure C_x and the partial structure S_y as shown in Eq. (10):

$$\varphi(C_x, S_y) = \max \begin{cases} \varphi(S_{x^{\mathrm{p}}}, S_y) + 2\mu \\ \varphi(C_x, P_y) + |C_y|\mu \\ \varphi(C_x, C_y) + |P_y|\mu \end{cases} \quad (10)$$

In aligning the partial structure S_x with the partial structure S_y, there are five cases: (i) the parent structure P_x is matched with the parent structure P_y and the child structure C_x is matched with the child structure C_y; (ii) the child structure C_x is matched with gaps; (iii) the child structure C_y is matched with gaps; (iv) the parent structure P_x is matched with gaps; and (v) the parent structure P_y is matched with gaps. Therefore

$$\varphi(S_x, S_y) = \max \begin{cases} \varphi(P_x, P_y) + \varphi(C_x, C_y) \\ \varphi(P_x, S_y) + |C_x|\mu \\ \varphi(S_x, P_y) + |C_y|\mu \\ \varphi(C_x, S_y) + |P_x|\mu \\ \varphi(S_x, C_y) + |P_y|\mu \end{cases} \quad (11)$$

It can be shown that this CSA method for aligning the query structure Q and the subject structure S allowing constraints to exist in Q has a polynomial time complexity of $O(mn)$ where m is the length of the query structure and n is the length of the subject structure.

2.5. p-Value

To determine what match is likely or unlikely to occur by chance, we have incorporated the computation of a statistical measure, namely a p-value, into our CSA method (cf. Fig. 2). Karlin and Altschul (1990) showed that in the case of a gapless alignment, the distribution of alignment scores of random sequences is the Gumbel or extreme value distribution (Gumbel, 1958). However for a gapped alignment, there is no theory that predicts the distribution of alignment scores for random sequences. It has been conjectured based on numerical evidence that the score distribution is still of the Gumbel form (Altschul and Gish, 1996; Collins et al., 1988; Smith et al., 1985). We adopt this assumption while computing the statistical measure. For the comparison of random sequences of sufficient lengths m and n, the number of distinct local alignments with score at least x is approximately Poisson distributed, with mean

$$E(x) = Kmn\,\mathrm{e}^{-\lambda x} \quad (12)$$

where λ and K can easily be calculated (Karlin and Altschul, 1990). The optimal alignment score S' follows an extreme-value distribution with

$$\mathrm{Prob}\,(S' \geq x) = 1 - \mathrm{e}^{-E(x)} \quad (13)$$

Accurate estimation of λ and K is essential to using these equations. We have used the Island method (Altschul et al., 2001; Olsen et al., 1999) to do the estimation. As suggested by this method, we first compute constrained structural alignments of biologically occurring RNA secondary structures chosen randomly from Rfam (Griffiths-Jones et al., 2003). While performing the alignment between two RNA secondary structures, we annotate one of the structures as constrained. Thus the scores obtained from the alignments are consistent with the proposed constrained structural alignment scoring scheme. The local alignment results are several locally optimal matches, each being comparable to an island in the large sequence. All the scores that are greater than a threshold c are selected. In the study presented here, the c value is set to 10. The threshold value is chosen such that it is a reasonable score obtained when aligning short RNA motifs of the commonly occurring length. Let us assume that the set I_c of such local alignment islands has cardinality R_c and the mean score in excess of c for these islands is S_c:

$$S_c = \frac{\sum_{i \in I_c} [S(i) - c]}{R_c} \quad (14)$$

where $S(i)$ is the score of island i. Then the maximum-likelihood estimator for λ is

$$\lambda_c = \ln\left(1 + \frac{1}{S_c}\right) \quad (15)$$

The maximum-likelihood estimator for K is

$$K_c = \frac{R_c\,\mathrm{e}^{\lambda_c \cdot c}}{A} \quad (16)$$

where A is the aggregate "area" of the search space from where the local alignments are taken. If a single pair of structures of length m and n is used, then $A = mn$. If B such comparisons are performed, then $A = Bmn$. Once λ_c and K_c are determined, we use these values to calculate the p-value for an alignment score x by plugging λ_c and K_c in Eqs. (12) and (13). The p-value is the probability, by chance, that there is another alignment with a similarity score greater than or equal to the score x. The p-value is a measure of the reliability of the score x. The smaller the p-value, the more reliable x is.

3. Experiments and Results

We tested the proposed constrained structural alignment method by detecting internal ribosome entry sites in the RNA sequences of *T. brucei* and hepatitis C virus, respectively. An internal ribosome entry site (IRES) is a nucleotide sequence which functions to allow for translation initiation in the non-coding region of an mRNA sequence (Hellen and Sarnow, 2001). An IRES element is able to attract the eukaryotic ribosome to close vicinity of a start codon and thus to initiate its translation. The secondary structure of an internal ribosome entry site in *T. brucei* mRNA sequences is portrayed in Fig. 5.

There are two different datasets used in our experiments. For the first dataset D_1, we extracted 20 non-redundant untranslated regions (UTRs) of *T. brucei* mRNA sequences containing internal ribosome entry sites from UTRdb (Pesole and Liuni, 1999). These IRES-containing UTR sequences, listed in Table 1, formed the positive data for the dataset D_1. Their lengths are in the range 85–993 nt. Also shown in the table are the start and end positions of the IRES element in each *T. brucei* UTR sequence. The presence of IRESs in these UTR sequences was suggested by UTRscan (Pesole and Liuni, 1999) which is a sequence analysis tool provided by

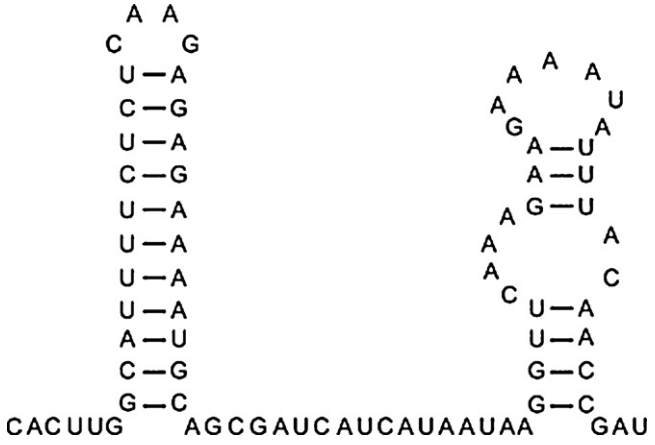

Fig. 5. The secondary structure of an internal ribosome entry site in *T. brucei* mRNA sequences.

Table 1
The 20 IRES-containing *T. brucei* UTR sequences used as positive data in D_1

EMBL accession number	Description	IRES start	IRES end
AB033824	5′UTR in *T. brucei* GPI10 mRNA for GPI anchor biosynthesis protein, complete cds	5	92
AF007547	5′UTR in *T. brucei* Trab5B mRNA, complete cds	73	158
AF049901	5′UTR in *T. brucei* rhodesiense prohibitin mRNA, complete cds	72	166
AF068705	5′-UTR in *T. brucei* rhodesiense transferring-binding protein (ESAG 6-d) mRNA, complete cds	475	558
AF101480	5′UTR in *T. brucei* pf20 homolog (TWD1) mRNA, complete cds	1	101
AF189284	5′UTR in *T. brucei* nucleolar G-protein NOG1 (NOG1) mRNA, complete cds	168	254
AF226674	5′UTR in *T. brucei* 20S proteasome beta 5 subunit (PSB5) mRNA, complete cds	267	346
AF301417	5′UTR in *T. brucei* procyclin-associated gene 2 polypeptide (PAG2), procyclin-associated gene 4 polypeptide (PAG4), GU2 (GU2), and GU1 (GU1) genes, complete cds	9178	9265
AF404116	5′UTR in *T. brucei* proteasome regulatory non-ATP-ase subunit 8 (Rpn8) mRNA, complete cds	135	235
AJ242519	5′UTR in *T. brucei* mRNA for cyclin 2 (CYC2 gene)	6	103
AM159084	5′UTR in *T. brucei* mRNA for RNA polymerase I subunit RPA12 (RPA12 gene)	3	97
AM159570	5′UTR in *T. brucei* mRNA for RNA polymerase I subunit RPC40 (RPC40 gene)	213	308
AY157028	5′UTR in *T. brucei* putative G1 cyclin CycE2 mRNA, complete cds	124	217
AY157032	5′UTR in *T. brucei* putative mitotic B-type cyclin CycB3 mRNA, complete cds	142	239
AY370775	5′UTR in *T. brucei* strain Lister 427 Rab23 mRNA, complete cds	22	116
K02198	5′UTR in *T. brucei* spliced leader mRNA (pSLc4) from procyclic stage	11	109
K02945	5′UTR in *T. brucei* gambiense calmodulin mRNA 2 with a spliced leader sequence	15	104
L03777	5′UTR in *T. brucei* protein kinase (nrkB) allele nrkB-2 mRNA, complete non-functional cds and alleles nrkB-1 and nrkB-3	901	993
U18329	5′UTR in *T. brucei* small GTP-binding protein mRNA, clone rtb9, complete cds	75	157
U80910	5′UTR in *T. brucei* ribonucleotide reductase large subunit (RNR1) mRNA, complete cds	8	85

UTResource. UTRscan analyzes user-submitted sequences for the functional elements defined in the UTRsite database of UTResource. Notice that even though the 20 *T. brucei* UTR sequences contain internal ribosome entry sites, there are no known conserved secondary structures, or structural motifs, in the IRES-containing UTRs (Palenchar and Bellofatto, 2006; Palenchar et al., 2006). We also added 30 other sequences from UTRdb which were not known to contain internal ribosome entry sites. These 30 sequences formed the negative data for the dataset D_1. All these 50 sequences were folded using RNAfold (Hofacker, 2003). Finally, 5 of the 20 IRES-containing *T. brucei* UTR sequences were randomly selected and the IRES-containing region in each of the 5 sequences was extracted. These IRES-containing regions were separately folded using RNAfold and they formed the query structures in our experiment involving D_1.

The second dataset D_2 was made up of 20 non-redundant hepatitis C virus (HCV) sequences, which contained internal ribosome entry sites, from Rfam (Griffiths-Jones et al., 2003). These sequences belong to the IRES_HCV family in Rfam. Table 2 lists these sequences, which formed the positive data for the dataset D_2. Their lengths are in the range 190–267 nt. In Rfam, these 20 HCV sequences share a consensus or conserved secondary structure. We added 30 other sequences taken from UTRdb to the dataset D_2. These 30 sequences did not belong to the hepatitis C virus, and were not known to contain internal ribosome entry sites. These 30 sequences formed the negative data for the dataset D_2. All these 50 sequences were folded using RNAfold. Separately, we also randomly selected 5 of the 20 IRES-containing HCV sequences from the dataset D_2 and extracted just the IRES-containing region from each of the 5 sequences. These IRES-containing regions were folded using RNAfold. The resulting five structures formed the query structures in our experiment involving D_2.

On these two datasets D_1 and D_2, we applied our constrained structural alignment (CSA) method with binary 0/1 conservation (0/1 constraints), by aligning each of the five selected query structures one by one with the RNA secondary structures in D_1 and D_2, respectively. (Here, every base in a query structure was marked with "*".) For comparison purposes, we also applied two other methods to the same datasets. They were the regular pairwise structural

Table 2
The 20 IRES-containing HCV sequences used as positive data in D_2

EMBL accession number	Description	IRES start	IRES end
AF021888	HCV strain GE 174 5′ non-coding region type 1a	1	190
AF021898	HCV strain GE 56 5′ non-coding region type 4	1	190
AF021904	HCV strain SL 34 5′ non-coding region type 1a	1	190
AF034628	HCV type 3 5′ noncoding region, partial sequence	2	253
AF041264	HCV isolate 498 5′ untranslated region	1	191
AF041266	HCV isolate 611 5′ untranslated region	1	191
AF041267	HCV isolate 614 5′ untranslated region	1	191
AF041300	HCV isolate 966 5′ untranslated region	1	191
AF055303	HCV type 1a strain CHCH3 5′ untranslated region, partial sequence	1	240
AF055305	HCV type 1a strain CHCH5 5′ untranslated region, partial sequence	1	239
AF041309	HCV isolate 982 5′ untranslated region	1	191
AF041329	HCV type 2c isolate 760 5′ noncoding sequence and core protein gene, partial cds	1	267
AF056005	HCV type 1b strain CHCH6 5′ untranslated region, partial sequence	1	237
AF055301	HCV type 1a 5′ untranslated region, partial sequence	1	238
AF057147	HCV type 2b strain CHCH13 5′ untranslated region, partial sequence	1	240
AF057150	HCV type 3a strain CHCH16 5′ untranslated region, partial sequence	1	237
AF077228	HCV isolate patient 20 5′ non-coding region, partial sequence	1	250
AF141989	HCV isolate 8-63 polyprotein mRNA, 5′ untranslated region, partial sequence	1	195
AF216795	HCV isolate SOM1 5′UTR, partial sequence	3	205
AF217298	HCV clone Sot10 5′UTR sequence	1	256

Table 3
The average error rate calculated by using five *T. brucei* queries against the dataset D_1

Query	CSA with 0/1 constraints		RSmatch		RNAforester	
	TP	FP	TP	FP	TP	FP
Q1	14	6	15	5	11	9
Q2	11	9	10	10	11	9
Q3	11	9	10	10	12	8
Q4	12	8	11	9	10	10
Q5	12	8	10	10	12	8
Average error rate	0.40		0.44		0.44	

The table shows the number of true positives and the number of false positives respectively that occur in the top 20 hits of a search.

alignment method without constraints offered in RSmatch (Liu et al., 2005) and the RNAforester structural alignment method (Hochsmann et al., 2004). Thus, a database search was carried out with each of these alignment methods by aligning the corresponding query structures one by one with the subject structures in D_1 and D_2, respectively. Then, from the top 20 hits, i.e. the top 20 RNA subject structures with the largest alignment scores, in a search result, we found out how many hits were true positives and how many were false positives. True positives are those hits in which an internal ribosome entry site is actually present. False positives are those hits that appear in the search result as containing internal ribosome entry sites, though in reality they are not known to contain internal ribosome entry sites. We use the error rate (e), defined below, to evaluate the effectiveness of an alignment method:

$$e = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}} \tag{17}$$

where TP is the number of true positives, FP is the number of false positives, and TP + FP = 20 in our experiments.

Table 3 shows the results and presents the average error rate obtained from using the five different *T. brucei* query structures for each alignment method. Table 4 presents this data for the five different queries belonging to the HCV dataset. As can be seen from the tables, the proposed CSA method with 0/1 constraints gives the lowest average error rate, outperforming the other two alignment techniques. These results were obtained by using the optimal structure for each sequence. We also compared the alignment algorithms by using 20% of suboptimal structures for each sequence, and the qualitative conclusion remains the same.

It was observed that there is little similarity shared by the IRES-containing *T. brucei* sequences. The average pairwise sequence identity for the 20 IRES-containing *T. brucei* sequences is 29%. This explains why the three alignment algorithms have high error rates for the *T. brucei* dataset (Table 3). On the other hand, the 20 IRES-containing HCV sequences are conserved at both the sequence and the secondary structure level. The average pairwise sequence identity for the 20 IRES-containing HCV sequences is 88%. Under this circumstance, all the three alignment algorithms have good performance; the algorithms have much lower error rates for the HCV dataset (Table 4) than for the *T. brucei* dataset (Table 3).

Table 4
The average error rate calculated by using five HCV queries against the dataset D_2

Query	CSA with 0/1 constraints		RSmatch		RNAforester	
	TP	FP	TP	FP	TP	FP
Q1	18	2	18	2	16	4
Q2	20	0	20	0	17	3
Q3	19	1	17	3	18	2
Q4	20	0	20	0	18	2
Q5	19	1	19	1	17	3
Average error rate	0.04		0.06		0.14	

The table shows the number of true positives and the number of false positives respectively that occur in the top 20 hits of a search.

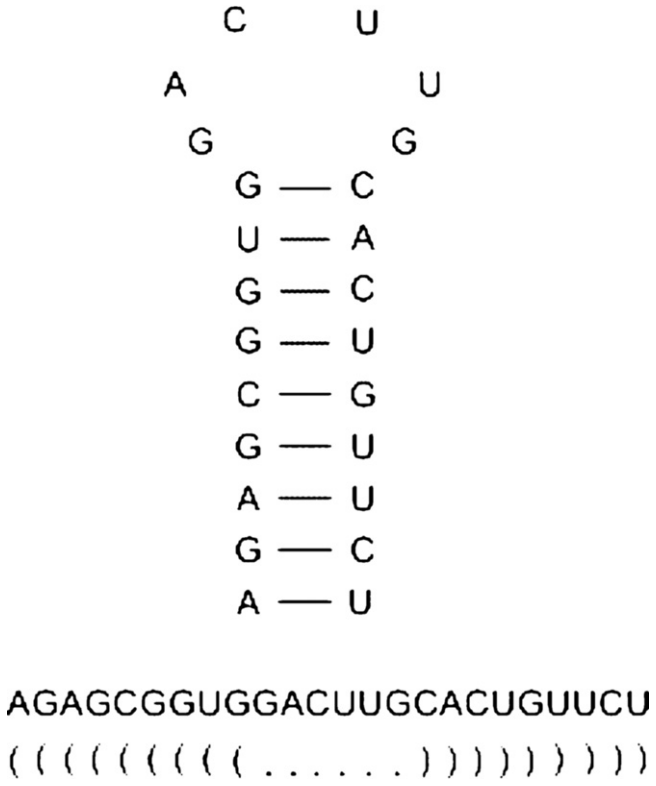

Fig. 6. A putative structural motif in *T. brucei* UTRs obtained from the multiple structural alignment of the top 10 positive structures that occurred in the search result of query Q1 in Table 3 using the proposed CSA method with 0/1 constraints.

From Table 3, it can be seen that among the top 20 hits, the CSA method with 0/1 constraints found 11–14 positive structures and 6–9 negative structures. The consensus of the found positive structures may suggest a conserved secondary structure or structural motif in the *T. brucei* UTRs. Fig. 6 shows the consensus secondary structure together with its Vienna style Dot Bracket representation of the top 10 positive structures most similar to query Q1 in Table 3 according to the CSA method with 0/1 constraints. The consensus secondary structure is computed by the multiple structural alignment (MSA) function of our RADAR tool (Khaladkar et al., 2007). For the HCV data in Table 4, the consensus secondary structure found by the proposed constrained structural alignment method in combination with RADAR's MSA function is consistent with that documented in Rfam.

4. Conclusions

In this paper, we presented constrained structural alignment algorithms for matching two RNA secondary structures. The algorithms have been implemented in the RADAR server, which can be downloaded from http://datalab.njit.edu/software. We developed a statistical measure for assessing the significance of alignment scores. We then applied the proposed techniques to searching for internal ribosome entry sites in RNA sequences of *T. brucei* and hepatitis C virus, respectively. For the HCV sequences, there is a known consensus secondary structure in them, as documented in Rfam, and our method accurately detected the consensus secondary structure in the HCV data. For the *T. brucei* molecules, there is little

similarity shared by their IRES-containing UTR sequences, and our experimental results suggested the possible existence of a conserved secondary structure in these IRES-containing UTR sequences. The results also showed the superiority of the proposed techniques over existing methods. In future work, we plan to apply the constrained structural alignment algorithms to testing other functional elements in the RNA molecules of *T. brucei* and HCV, and to searching for structural motifs in other organisms as suggested in (Babak et al., 2007; Xue and Liu, 2007).

Acknowledgements

We thank Marvin Nakayama, Bruce Shapiro and Bin Tian for helpful conversations during the preparation of this paper. This work was partially supported by National Science Foundation grant IIS-0707571 and National Institutes of Health grant R01 AI053835.

References

Altschul, S.F., Bundschuh, R., Olsen, R., Hwa, T., 2001. The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res. 29, 351–361.

Altschul, S.F., Gish, W., 1996. Local alignment statistics. Methods Enzymol. 266, 460–480.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G., Tuschl, T., 2003. A uniform system for microRNA annotation. RNA 9, 277–279.

Babak, T., Blencowe, B.J., Hughes, T.R., 2007. Considerations in the identification of functional RNA structural elements in genomic alignments. BMC Bioinformatics 8, 33.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E., 1992. Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. Protein Sci. 1, 1677–1690.

Collins, J.F., Coulson, A.F.W., Lyall, A., 1988. The significance of protein sequence similarities. Comput. Appl. Biosci. 4, 67–71.

Corpet, F., Michot, B., 1994. RNAlign program: alignment of RNA sequences using both primary and secondary structures. Comput. Appl. Biosci. 10, 389–399.

Eddy, S.R., 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics 3, 18.

Eddy, S.R., Durbin, R., 1994. RNA sequence analysis using covariance models. Nucleic Acids Res. 22, 2079–2088.

Gautheret, D., Lambert, A., 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J. Mol. Biol. 313, 1003–1011.

Gorodkin, J., Stricklin, S.L., Stormo, G.D., 2001. Discovering common stem-loop motifs in unaligned RNA sequences. Nucleic Acids Res. 29, 2135–2144.

Green, P., Lipman, D., Hillier, L., Waterson, R., States, D., Claverie, J.M., 1993. Ancient conserved regions in new gene sequences and the protein databases. Science 259, 1711–1716.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R., 2003. Rfam: an RNA family database. Nucleic Acids Res. 31, 439–441.

Gumbel, E.J., 1958. Statistics of Extremes. Columbia University Press, New York.

He, D., Arslan, A.N., 2005. A space-efficient algorithm for the constrained pairwise sequence alignment problem. Genome Informat. 16, 237–246.

Hellen, C.U.T., Sarnow, P., 2001. Internal ribosome entry sites in eukaryotic mRNA molecules. Genes Dev. 15, 1593–1612.

Hochsmann, M., Toller, T., Giegerich, R., Kurtz, S., 2003. Local similarity in RNA secondary structures. In: Proceedings of the IEEE Computational Systems Bioinformatics Conference, Stanford, August 11–14. IEEE Computer Society, Los Alamitos, pp. 159–168.

Hochsmann, M., Voss, B., Giegerich, R., 2004. Pure multiple RNA secondary structure alignments: a progressive profile approach. IEEE/ACM Trans. Comput. Biol. Bioinformat. 1, 53–62.

Hofacker, I.L., 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31, 3429–3431.

Holmes, I., 2005. Accelerated probabilistic inference of RNA structure evolution. BMC Bioinformatics 6, 73.

Holmes, I., Rubin, G.M., 2002. Pairwise RNA structure comparison with stochastic context-free grammars. In: Proceedings of the Seventh Pacific Symposium on Biocomputing, Hawaii, January 3–7. World Scientific, Singapore, pp. 163–174.

Jiang, T., Lin, G., Ma, B., Zhang, K., 2002. A general edit distance between RNA structures. J. Comput. Biol. 9, 371–388.

Karlin, S., Altschul, S.F., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U.S.A. 87, 2264–2268.

Keene, J.D., 2007. RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet. 8, 533–543.

Khaladkar, M., Bellofatto, V., Wang, J.T.L., Tian, B., Shapiro, B.A., 2007. RADAR: a web server for RNA data analysis and research. Nucleic Acids Res. 35, W300–W304.

Kim, J., Cole, J.R., Pramanik, S., 1996. Alignment of possible secondary structures in multiple RNA sequences using simulated annealing. Comput. Appl. Biosci. 12, 259–267.

Klein, R.J., Eddy, S.R., 2003. RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics 4, 44.

Laferriere, A., Gautheret, D., Cedergren, R., 1994. An RNA pattern matching program with enhanced performance and portability. Comput. Appl. Biosci. 10, 211–212.

Liu, J., Wang, J.T.L., Hu, J., Tian, B., 2005. A method for aligning RNA secondary structures and its application to RNA motif detection. BMC Bioinformatics 6, 89.

Lowe, T., Eddy, S.R., 1999. A computational screen for methylation guide snoRNAs in yeast. Science 283, 1168–1171.

Lu, C.L., Huang, Y.P., 2005. A memory-efficient algorithm for multiple sequence alignment with constraints. Bioinformatics 21, 20–30.

Mathews, D.H., Turner, D.H., 2002. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J. Mol. Biol. 317, 191–203.

McKee, A.E., Silver, P.A., 2007. Systems perspectives on mRNA processing. Cell Res. 17, 581–590.

Notredame, C., O'Brien, E.A., Higgins, D.G., 1997. RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Res. 25, 4570–4580.

Olsen, R., Bundschuh, R., Hwa, T., 1999. Rapid assessment of extremal statistics for gapped local alignment. In: Lengauer, T., Schneider, R., Bork, P., Brutlag, D.L., Glasgow, J.I., Mewes, H.-W., Zimmer, R. (Eds.), Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology. Heidelberg, Germany, August 6–10. AAAI Press, Menlo Park, pp. 211–222.

Palenchar, J.B., Bellofatto, V., 2006. Gene transcription in trypanosomes. Mol. Biochem. Parasitol. 146, 135–141.

Palenchar, J.B., Liu, W., Palenchar, P.M., Bellofatto, V., 2006. A divergent transcription factor TFIIB in trypanosomes is required for RNA polymerase II-dependent spliced leader RNA transcription and cell viability. Eukaryot. Cell 5, 293–300.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85, 2444–2448.

Pesole, G., Liuni, S., 1999. Internet resources for the functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNA. Trends Genet. 15, 378.

Pesole, G., Liuni, S., D'Souza, M., 2000. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. Bioinformatics 16, 439–450.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R.C., Haussler, D., 1994. Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res. 22, 5112–5120.

Sanchez-Diaz, P., Penalva, L.O., 2006. Post-transcription meets post-genomic: the saga of RNA binding proteins in a new era. RNA Biol. 3, 101–109.

Shapiro, B.A., Zhang, K., 1990. Comparing multiple RNA secondary structures using tree comparisons. Comput. Appl. Biosci. 6, 309–318.

Smith, T.F., Waterman, M.S., Burks, C., 1985. The statistical distribution of nucleic acid similarities. Nucleic Acids Res. 13, 645–656.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Wang, J.T.L., Shapiro, B.A., Shasha, D., Zhang, K., Chang, C.-Y., 1996. Automated discovery of active motifs in multiple RNA secondary structures. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, August. AAAI Press, Menlo Park, pp. 70–75.

Wang, J.T.L., Shapiro, B.A., Shasha, D., Zhang, K., Currey, K.M., 1998. An algorithm for finding the largest approximately common substructures of two trees. IEEE Trans. Pattern Anal. Mach. Intell. 20, 889–895.

Xue, C., Liu, G.P., 2007. RScan: fast searching structural similarities for structured RNAs in large databases. BMC Genomics 8, 257.

Yao, Z., Weinberg, Z., Ruzzo, W.L., 2006. CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics 22, 445–452.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415.