

StructMiner: A Tool for Alignment and Detection of Conserved Secondary Structure

Qian Yang*

qian@mcb.mcgill.ca

Mathieu Blanchette

blanchette@mcb.mcgill.ca

McGill Center for Bioinformatics, McGill University, Montreal, Quebec, H3A 2B4, Canada

Abstract

Functional RNA molecules typically have structural patterns that are highly conserved in evolution. Here we present an algorithmic method for multiple alignment of RNAs, taking into consideration both structural similarity and sequence identity. Furthermore, our window-sized comparative analysis corrects the misaligned structure within a distance threshold and identifies the conserved substructures. Based on this new algorithm, StructMiner outperforms existing approaches, which ignore structure information for the alignment and lack the effective means to adjust the misalignments in the analysis phase. In addition, StructMiner is efficient in terms of CPU time and memory usage, making it suitable for structural analysis of very long sequences.

Keywords: RNA, structural alignment, conserved substructure, window-sized comparative analysis

1 Introduction

Recent discovery of a variety of functions for non-coding RNA has brought increased demand for suitable tools for RNA structure analysis. RNA molecules exhibit a close interplay between structure and function. As many functional classes of RNA molecules including tRNA, rRNA, SRP rna etc. are highly conserved in secondary structure but share little sequence similarity, our traditional methods of multiple alignments need to be extended to take structure information into account. Almost all RNA molecules form secondary structure. However, the presence of secondary structure doesn't imply functional significance. Therefore, it is important for us to identify the potential functional parts of a secondary structure before investigating what function it has.

Functional structures are often conserved among related species. Several algorithms exist for finding the conserved structures among RNA molecules. Classic Sankoff's algorithm that simultaneously fold and align RNA molecules is computationally very expensive ($O(n^6)$ in CPU and $O(n^4)$ in memory), making it impractical for any but the smallest problems. Stochastic context-free grammars (SCFG) present an alternative approach to the structure-alignment problem. SCFGs do not utilize the energy information in forming structure, but rely on production rules derived from a training set. The major limitation of this method is the estimation of complex parameters and local maximum associated with EM algorithm. Available tools such as alidot and RNAalifold use the result of sequence only alignment to get a list of candidate base pairs and sort the list by credibility checking. Clearly, methods of this kind suffer from the absence of significant sequence similarity and reliable alignment.

StructMiner combines the thermodynamic structure information with comparative analysis of probability matrices, a connection between structure genomics and comparative genomics. The basic idea is: Secondary structure elements that are consistently present in a group of sequences in spite of weak sequence identity are most likely the result of functional stabilization, not the consequence of a high degree of sequence homology. Such structure elements are conserved substructures with

*Corresponding author

potential functional meaning. Based on the above observation, we designed an algorithm that can align the sequences using structural homology as a major consideration without neglecting sequence similarity. Compared with existing structural analysis approaches, we extend the single pair-specific comparative analysis to window-sized comparison of segment pairs in order to reduce the propagation of the alignment error into the actual detection of conserved structures. A flow diagram of our method is shown in Figure 1.

As an example for the quality of the prediction, we apply StructMiner to two sets of 5S rRNA sequences. The conserved secondary structures it predicts are both identical to the published ones. We avoid missing correct structures by using the probability matrices instead of single predicted structure from Vienna RNA Package since base pairing probabilities contain information about a large number of plausible structures. As a matter of fact, we do find disagreement between our prediction of conserved structure and the optimal structure generated by Vienna (e.g. *Agrobacterium tumefaciens* with accession code X02627). The conserved substructure was already confirmed by crystallography study, and the error lies in Vienna. Vienna is popular software for structure prediction. Because of hardness of the structure prediction problem itself, the accuracy of the Vienna's prediction drops quickly when the ambiguity of pairing probability and the length of sequence increase. As the implementation of the new structure-sequence alignment algorithm and window-sized comparative analysis, StructMiner provides an effective approach not only in detecting conserved substructure, but also in improving the whole structure prediction through the systematic analysis of thermodynamic pairing probability among related species.

2 Method and Results

2.1 Preliminaries

RNA secondary structures obey a “nested” constraint: for any two based pairs (i, j) and (k, l) , where $i < k$, it must satisfy either $i < j < k < l$ or $i < k < l < j$. Under this property, the structural alignment requires the simultaneous alignment of two nucleotides involved in the base pairs. Given two RNA sequences A_1 and A_2 , we say the segment $S_1[1, \dots, N]$ of sequence A_1 is structurally aligned with segment $S_2[1, \dots, M]$ of sequence A_2 if for any paired nucleotides (i, j) in S_1 , there must exist a corresponding pair (k, l) in S_2 , such that i is aligned with k and j with l . Unpaired nucleotide in one segment is aligned with either an unpaired element in the other segment or a gap.

2.2 Structure-Sequence Alignment

The input of the alignment is base pairing probability matrices P^{A_1} and P^{A_2} for sequence A_1 and A_2 , predicted by means of McCaskill's algorithm. (McCaskill, 1990 [13]) (The algorithm is implemented in the RNAfold program of Vienna RNA package.) The independent computation of thermodynamic pairing probability allows alternative choices of probability matrices, e.g. probability for kinetic energy. The problem now becomes the alignment of two probability matrices, a kind of threading problem known to be NP-hard in the general case. Dimension reduction is a technique frequently used in multidimensional data mining and we successfully applied it here to reduce the complexity of the problem. The notation of upstream and downstream probabilities was introduced in Bonhoeffer *et al.* (1993) [2]. Given P_{ij} denoting the probability of nucleotide at position i paired with nucleotide at position j , upstream probability for i is the probability of being paired upstream $p^<(i) = \sum_{j>i} P_{ij}$, and downstream of i is $p^>(i) = \sum_{j<i} P_{ji}$. With this definition, we construct an array $p_{A_1}^<$ containing the upstream probability for all the nucleotides in sequence A_1 , and an array $P_{A_1}^>$ for all the downstream probabilities. $p_{A_2}^<$ and $P_{A_2}^>$ for sequence A_2 can also be obtained. We do lose specific pairing information when we convert the probability matrix into two linear vectors. However the upstream and downstream vectors still catch the trend of pairing capability pretty well. A similar heuristic people often use in analyzing secondary structure is: when they align the mountain plot of the secondary structure, they care more about the shape (or slope in the other word) of the mountain instead of its

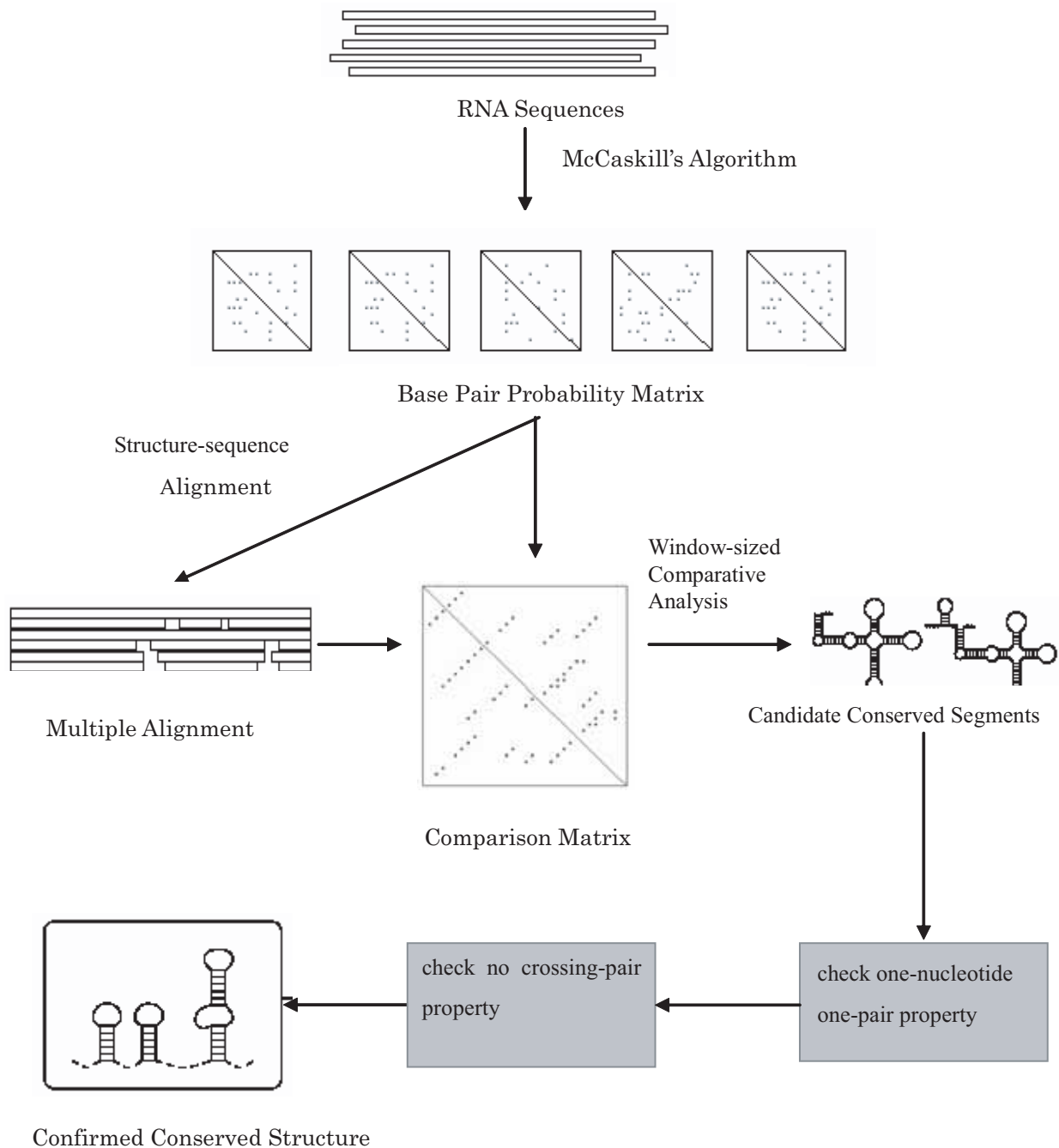


Figure 1: Flow diagram of the algorithm StructMiner. Base pairing probability calculated through McCaskill's algorithm is used in generating multiple alignment. The alignment result is then combined with probability matrix to form the alignment of matrices, and comparison matrices are created pairwise. Window-sized comparative analysis corrects misalignment and detects conserved structure at the same time. The technique of filtration and the nested property check are applied before the output of conserved structure.

exact height. Mountain plot is an alternative way of describing secondary structure. Figure 2 gives an example of mountain plot, and Figure 3 illustrates the alignment of mountain plots.

We look for an alignment of the sequences A_1 and A_2 such that

$$\alpha \sum (p_{A_1}^<(i) \cdot p_{A_2}^<(k) + P_{A_1}^>(j) \cdot P_{A_2}^>(l) + \mu) + N_{open} \cdot \omega_{open} + N_{ext} \cdot \omega_{ext} + \sum \delta(A_1(i), A_2(k)) \rightarrow \max$$

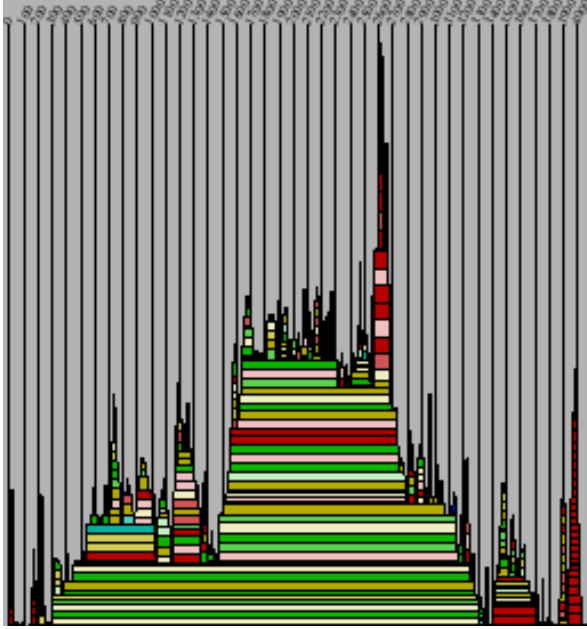


Figure 2: Mountain plot of comovirus RNA2. Horizontal striations upon a particular peak are bonds between paired bases, and vertical links between the horizontal striations represent stems

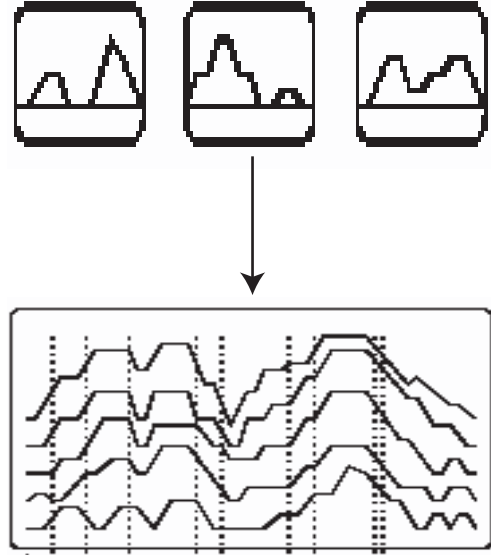


Figure 3: Alignment of 3 mountain plots. The slopes of the mountains on the dotted vertical lines are the same and therefore aligned.

The upper line describes the score for structural alignment. (i, j) is a base pair in A_1 aligned with (k, l) in A_2 . Here α is a ratio of structural counts over sequence similarity. As our alignment is structure-oriented, we set $\alpha = 5$. $\mu < 0$ is the penalty for aligning two nucleotides at least one of which doesn't have pairing potential, and we consider this as mismatches in structural only alignment. The rest part is the score for sequence similarity and gap penalty. For better result of alignment, we adapt the affine gap penalty. Thus, $\omega_{open} < 0$ is the opening gap penalty and N_{open} is the number of gap opens, correspondingly $\omega_{ext} < 0$ and N_{ext} for gap extension. Like what we did in traditional sequence alignment, $\delta(A_1(i), A_2(k)) > 0$ when it is a match and less than 0 when mismatch occurs. We use a similar kind of scoring scheme for matches and mismatches as in [17].

Let $V(i, k)$ be the score of the best alignment for sequence $A_1[1, \dots, i]$ and $A_2[1, \dots, k]$, the dynamic programming recursion can be obtained as follows with the initial condition $V(i, 0) = V(0, k) = 0$.

$$V(i, k) = \max[E(i, k), F(i, k), G(i, k)]$$

if $A_1(i)A_2(k)$ structurally match

$$G(i, k) = V(i - 1, k - 1) + p_{A_1}^<(i) \cdot p_{A_2}^<(k) + P_{A_1}^>(i) \cdot P_{A_2}^>(k) + \delta(A_1(i), A_2(k))$$

if $A_1(i)A_2(k)$ structurally mismatch

$$G(i, k) = V(i - 1, k - 1) + \mu + \delta(A_1(i), A_2(k))$$

$$E(i, k) = \max[E(i, k - 1), V(i, k - 1) - \omega_{open}] - \omega_{ext}$$

$$F(i, k) = \max[F(i - 1, k), V(i - 1, k) - \omega_{open}] - \omega_{ext}$$

Here $G(i, k)$ is the alignment type when $A_1(i)$ and $A_2(k)$ are aligned opposite to each other. $E(i, k)$ aligns $A_1(i)$ to the left of $A_2(k)$ and $F(i, k)$ aligns $A_1(i)$ to the right of $A_2(k)$. Finally $V(i, k)$ is defined as the maximum values of the three terms $E(i, k)$, $F(i, k)$ and $G(i, k)$.

When we extend the pair-wise alignment to multiple alignment, the gap penalties at existing gaps are lowered. The basic idea is: If there are already gaps at a position, then the gap opening penalty ω_{open} and gap extension penalty ω_{ext} are reduced in proportion to the number of sequences with a gap at this position. So the new gap opening penalty and extension penalty are recalculated as:

$$\begin{aligned}\omega_{open} &= \omega_{open} \cdot (\text{no. of sequences without a gap} / \text{no. of sequences}) \\ \omega_{ext} &= \omega_{ext} \cdot (\text{no. of sequences without a gap} / \text{no. of sequences})\end{aligned}$$

Experiments on both 5S and 16S rRNA show that on average our alignment result is 10~15% more accurate than CLUSTAL W. The improvement becomes more obvious when dissimilarity of sequences increases.

2.3 Identification of Conserved Secondary Structure

The results of multiple alignments and the original probability matrices are integrated to detect the conserved secondary structure. First, we filter out base pairs with a probability less than 10^{-3} as they are very unlikely to be part of an important structure. Due to the “nested” property, a valid structure forms a line (we called it a “valid line”) perpendicular to the diagonal of the matrix (Figure 4). In addition, any conserved pair segment has at least two contiguous base pairs meaning two contiguous spot along a valid line. So we filter out orphan pairs along each valid line in the second step. In the next step, gaps in the alignment are inserted into corresponding probability matrices. If the length of the alignment is l , we now get $n \times l \times l$ matrices numbered $M_{A1}, M_{A2}, M_{A3}, \dots, M_{An}$ for RNA sequences $A1, A2, A3, \dots, An$. M_{Ai} ($i = 1, \dots, n$) are symmetric and the lower triangle of each matrix is redundant. From M_{A1}, M_{A2} , we construct our comparison matrix M_{A1A2} by combining upper triangle of M_{A1} with lower triangle of M_{A2} . Similarly, we get M_{A3A4}, M_{A5A6} , etc.

If all the base pair segments are aligned correctly, we should be able to detect all the conserved substructures between A_1 and A_2 by comparing the symmetric elements along the valid lines of comparison matrix A_{12} . The reality is: Structural alignment requires simultaneous alignment of i and k , j and l for matching base pairs (i, j) and (k, l) . This is a stronger restriction than aligning each nucleotide separately. Sequence only alignment can’t avoid the misalignments. Neither can the structural alignment. However, our structural alignment algorithm not only gives a more accurate alignment compared with traditional alignment tools like CLUSTAL W, but also captures the shape of the structure pretty well. As a result, most of the misalignments are kept within certain distance. In other words, at some point, misalignment may occur and pairs matching to each other are shifted, but the shift won’t last long until a strong upstream/downstream picks up the correct alignment again. The strong upstream/downstream usually happen at the start/end of a well conserved segment, which means the misalignment of one conserved segment is usually not propagated to the next, and the error is limited inside a block. Therefore, we introduce the window-sized comparative analysis to correct the misalignment within a block and identify the conserved segments at the same time. A misalignment inside the comparison matrix is displayed as a shift of contiguous segment pairs from one valid line to another. An up/down shift from the valid line corresponds to the left/right shift of the opening pair of the segment in the alignment, and a left/right shift corresponds to the left/right shift for the ending pair of the segment. So we define a valid strip centered at a valid line with a certain window size s to cover all the possible shifts in four directions (Figure 4). The comparative analysis is performed inside the strip, so that even if the conserved substructures are not aligned so well as to be symmetric along the valid lines, we can still catch it inside a valid strip. We choose $s = 3$ based on the fact that at least three nucleotides exist between two substructures to form a loop, so there should be at most one valid substructure inside one strip. We can set the window size to a larger number in order to correct the

misalignment with further distance. Then there could be more than one conserved substructure inside one strip and we need to identify the matching parts between the upper triangle and lower triangle. Minimum distance can be a criteria and the modification is trivial.

Conserved structures can be identified from the segment pairs symmetric inside the valid strip. For example, $p_1, p_2, p_3, \dots, p_t$ are contiguous pairs in M_{A1} and $q_1, q_2, q_3, \dots, q_t$ also form contiguous pairs in M_{A2} . Furthermore, these two segments of pairs are symmetric in comparison matrix M_{A1A2} . Then, they are potentially important pairs and are output to the upper triangle of a new matrix $M_{A1A2A3A4}$ by setting the pairing probability as $(p_i + q_i)/2$, the lower triangle of the new matrix is filled by potentially conserved pairs from M_{A3A4} using the similar kind of filtration illustrated for M_{A1A2} . We repeat this procedure and finally get $M_{A1A2, \dots, An}$ which contains the potentially conserved pairs among all the sequences. The candidate pairs may still violate one or both of the following condition: (i) no nucleotide takes part in more than one base pair (ii) base pairs never cross. As we guarantee that the above two conditions are maintained inside one segment by using the concept of valid line and valid strip, we only need to see if there are conflicts between different segments. For condition (i), we check if there are two segments involved in the same row or column. If so, we choose the segment of pairs having the largest probability $\prod p_i (i = 1, \dots, t$ and t is the number of contiguous elements). The same criterion is used to filter out those pairs violating condition (ii), before we output the final list of conserved base pairs.

2.4 Complexity

Multiple alignment requires $O(nL^2)$ time, with n being the number of sequence and L being the length of the alignment. The comparative analysis takes $O(\log n \cdot L^2)$. So the overall running time is $O(nL^2)$ and memory usage is $O(L^2)$. To get a rough idea, running StructMiner on four 16sRNA with about 2000 nucleotides long for each sequence need only a few seconds on a Linux PC with P4 1.1Ghz.

2.5 Results

Our test data is taken from 5S rRNA database at <http://www.rna.icmb.utexas.edu/>. All the sequences from the database share a conserved structure pattern shown in Figure 5(a). We randomly chose four sequences (*G. Stearothermophilus*, *M. luteus*, *A. tumefaciens* and *E. coli*) with accession code M25591, K02682, X02627 and V00336. StructMiner's prediction is identical to the published one except for two orphan pairs at segment A and E are missing. Orphan pairs are usually not stable and it is acceptable to exclude them in conserved structures. Because of high similarity (around 90%) among the four sequences, alidot and MARNA (Siebert & Backofen 2003) also predict a similar kind of structure. To make a further comparison, we replace K02682 and V00336 with two manually selected sequences X67579 (*S.cerevisiae*) and AF034620 (*H. Marismortui*). Both share less than 50% sequence identity with all the other sequences. The result is displayed in Figure 5(c)-(d). And we can see the advantage of StructMiner is more apparent when structures converge but sequences diverge, which are often true for many functional RNA molecules

In order to see the performance of StructMiner on longer sequences, we applied it to four 16S rRNA sequences randomly chosen from Archaea category (*M.formicicum*-M36508, *H.marismortui* rrnB-X61689, *A.pernix*-AP000062 and *M.vannielii*-M36507). Each sequence has about 2000 nucleotides long and has a similar kind of complex secondary structure (Figure 6). The conserved substructures predicted by StructMiner are marked as red lines. As a result, it successfully predicted more than 85% base pairs with very little false positive highlighted as blue lines.

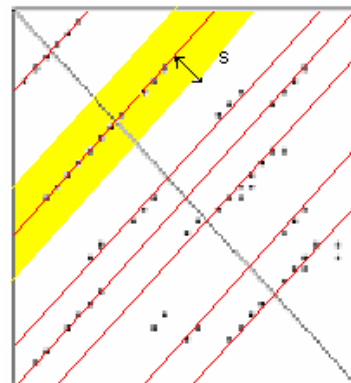
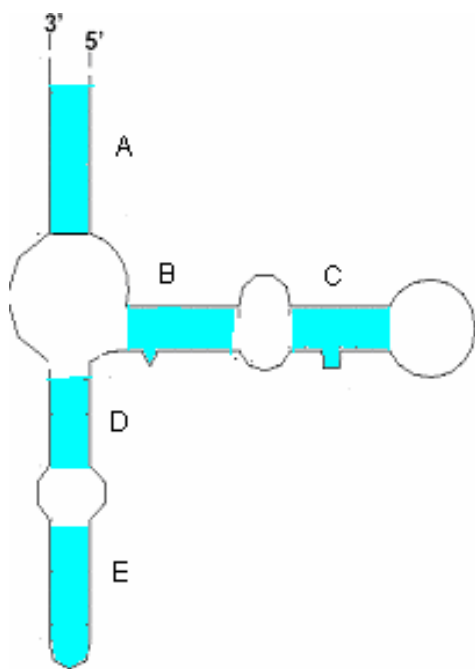
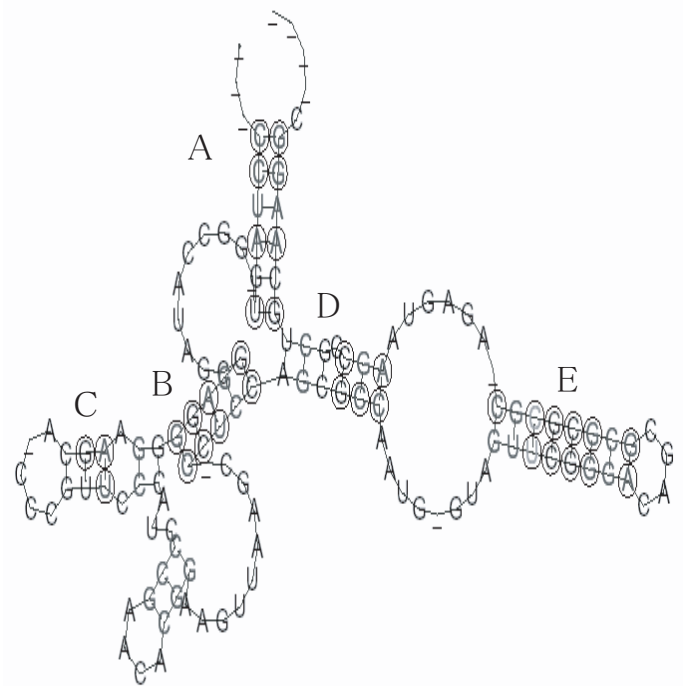


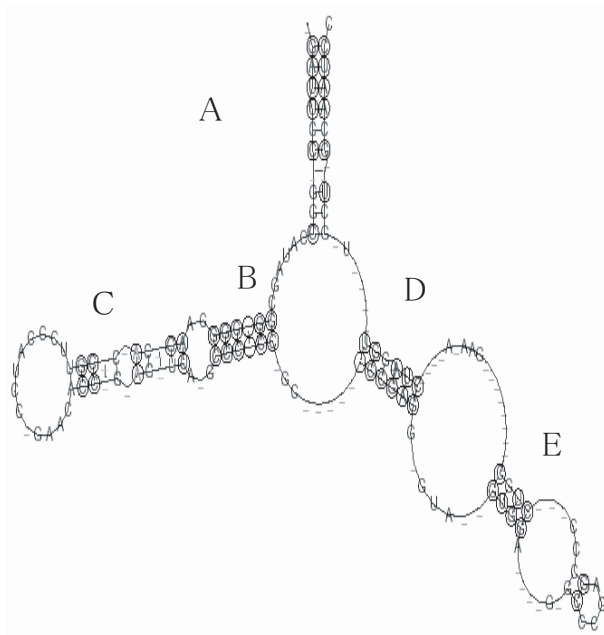
Figure 4: Comparison matrices with valid lines (red) and valid strip (yellow block), s is the window size.



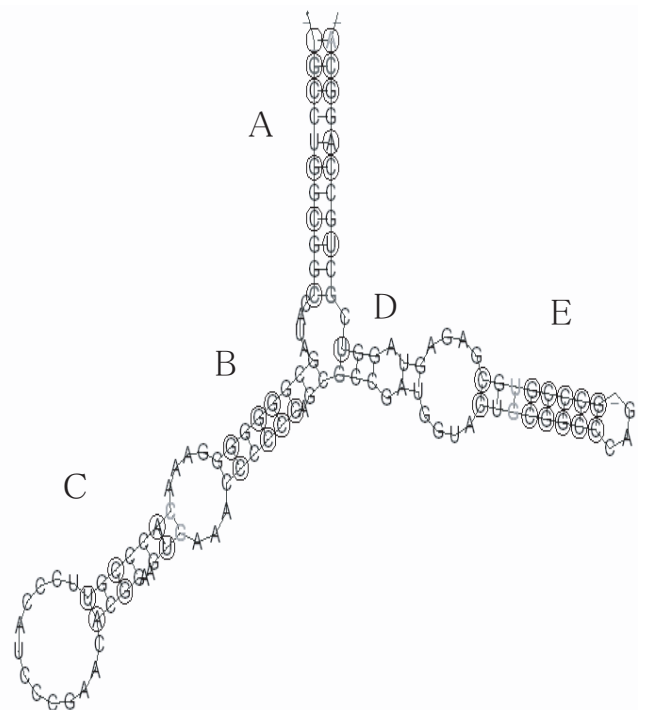
(a) Conserved structure pattern of 5S rRNA.



(b) Conserved structure of 5S rRNA predicted by Alidot.



(c) Conserved structure of 5S rRNA predicted by MARNA .



(d) Conserved 5S rRNA predicted by StructMiner.

Figure 5: Confirmed structure pattern of 5S rRNA (a) and the comparison of prediction by Alidot (b) using clustal w alignment <http://www.es.embnet.org/Doc/phylogendron/clustal-form.html>, Marna (c) and StructMiner (d). Because of low quality alignment of clustalW, alidot's prediction is far away from (a). MARNA takes structure information into consideration and yields an acceptable structure pattern. Finally, StructMiner corrects the redundant loop in segment E of Marna result and produces the conserved structure closest to (a).

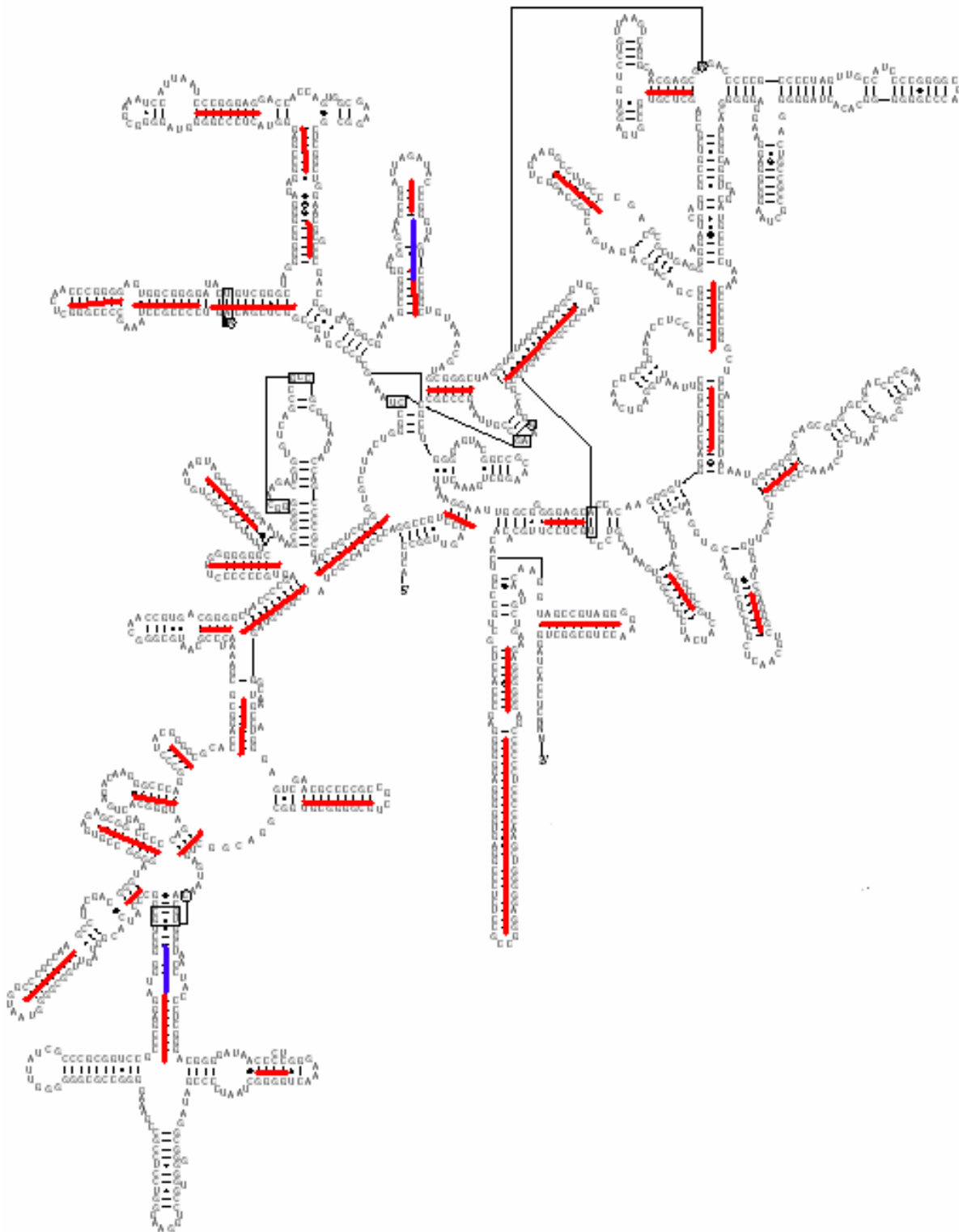


Figure 6: Result of StructMiner's prediction on 16S rRNA (*M.formicicum*-M36508, *H.marismortui* rrnB-X61689, *A.pernix*-AP000062 and *M.vannielii*-M36507). Red lines indicate the correct prediction of segment pairs. Blue lines are false positive prediction.

3 Discussion

The main idea of StructMiner is that we abstract the two-dimensional search space of pairing probability matrix into one dimension, and integrate both structure and sequence information into dynamic programming algorithm. The multiple alignment procedure can be separated as an individual tool, and the result can be analyzed further to obtain edit distance, or derive phylogenetic relationship when combined with other parsimony programs. In addition, we introduce a window-sized comparative analysis approach to correct the misalignments when detecting the conserved substructures. Future work is ongoing to effectively use more matrix property into the step of automatic comparative analysis. The multidimensional character of tertiary structure for RNA and protein gives lots of challenge to researchers. Dimension reduction of 3D probability data to a DAG (directed acyclic graph) could be a feasible way to tackle the problem, and a variety of graph searching algorithm may be utilized for the alignment of tertiary structures.

References

- [1] Backofen, R. and Will, S., Local sequence-structure motifs in RNA, *J. Bioinfo. Comput. Biol.*, 15–33, 2004.
- [2] Bonhoeffer, S., McCaskill, J.S., Stadler, P.F., and Schuster, P., RNA multi-structure landscape: A study based on temperature dependent partition functions, *Eur. Biophys. J.*, 22:14–24, 1993.
- [3] Dirks, R.M. and Pierce, N.A., An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots, *J. Comput. Chem.*, 25:1295–1304, 2004.
- [4] Eddy, S.R., A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure, *BMC Bioinformatics*, 3:18, 2002.
- [5] Gusfield, D., *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press, 1997.
- [6] Hochsmann, M., Toller, T., Giegerich, R., and Kurtz, S., Local similarity in RNA secondary structures, *Proc. Comput. Systems Bioinfo.*, 159–169, 2003.
- [7] Hofacker, I.L., The vienna RNA secondary structure server, *Nucleic Acids Res.*, 31:3429–3431, 2003.
- [8] Hofacker, I.L., Bernhart, S.H., and Stadler, P.F., Alignment of RNA base pairing probability matrices, *Bioinformatics*, 20(14):2222–2227, 2004.
- [9] Hofacker, I.L., Fekete, M., and Stadler, P.F., Secondary structure prediction for aligned RNA sequences, *J. Mol. Biol.*, 319:1059–1066, 2002.
- [10] Hofacker, I.L. and Stadler, P.F., Automatic detection of conserved base pairing patterns in RNA virus genomes, *Comput. Chem.*, 23:401–414, 1999.
- [11] Holmes, I. and Rubin, G.M., Pairwise RNA structure comparison using stochastic context-free grammars, *Pacific Symposium on Biocomputing (PSB 2002)*, World Scientific, Singapore, 163–174, 2002.
- [12] Knudsen, B. and Hein, J.J., RNA secondary structure prediction using stochastic context-free grammars and evolutionary history, *Bioinformatics*, 15:446–454, 1999.
- [13] McCaskill, J.S., The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers*, 29:193–216, 1999.

- [14] Pavesi, G., Mauri, G., Stefani, M., and Pesole, G., RNAProfile: An algorithm for finding conserved secondary structure motifs in unaligned RNA sequences, *Nucleic Acids Res.*, 32(10):3285–3269, 2004.
- [15] Rauscher, S., Flamm, C., Mandl, C.W., Heinz, F.X., and Stadler, P.F., Secondary structure of the 3'-non-coding region of flavivirus genomes, *RNA Journal*, 3(7):779–791, 1997.
- [16] Sankoff, D., Simultaneous solution of the RNA folding, alignment and proto-sequence problems, *SIAMJ. Appl. Math.*, 45:810–825, 1985.
- [17] Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W., Human-mouse alignments with BLASTZ, *Genome Res.*, 13:103–107, 2003.
- [18] Siebert, S. and Backofen, R., MARNA: A server for multiple alignment of RNAs, *Proc. German Conf. Bioinfo. GCB2003*, 1:135–153, 2003.
- [19] Thompson, J.D., Higgins, D.G., and Gibson, T.J., CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weigh matrix choice, *Nucleic Acids Res.*, 22:4673–4680, 1994.
- [20] Witwer, C., Rauscher, S., Hofacker, I.L., and Stadler, P.F., Conserved RNA secondary structures in picornaviridae genomes, *Nucleic Acids Res.*, 29:5079–5089, 2001.
- [21] <http://www.bio.inf.uni-jena.de/Software/MARNA/>
- [22] <http://www.es.embnet.org/Doc/phylodendron/clustal-form.html>
- [23] <http://rna.tbi.univie.ac.at/>