社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－MPS－52　(22)
2004／12／21

シュードノットを含むRNA二次構造の効率的アラインメント手法

関　新之助　　木島　篤志
小林　聡　　三瓶　厳一
電気通信大学大学院 電気通信学研究科 情報工学専攻

Abstract

本論文は、シュードノットを含むRNA二次構造の効率的アラインメント手法を提案する事を目的とする。我々が提案する手法は与えられた個々のRNA二次構造を重複のない部分構造に分割する事で階層木と呼ばれる木構造に変換し、それらに対して既存の木構造アラインメント手法を適用する事で結果を得る。生物学的ないし情報工学的観点より、我々が提案する階層アラインメント手法は次の２つの利点を有する。▪構造的相同性検出能の向上および▪探索空間削減によるアラインメント手法の効率の向上。実験データにより我々の手法によるアラインメント結果は既存の手法によるものに比べて生物学的により有望である事が示唆されている。

Efficient Alignment Method for RNA Secondary Structures Including Pseudoknots

Shinnosuke Seki,　　Satoshi Kobayashi,
Atsushi Kijima,　　Gen-ichi Sanpei
Department of Computer Science,
Department of Applied Physics and Chemistry,
University of Electro-Communications,

Abstract

In this paper, we will propose an efficient alignment method for RNA secondary structures including pseudoknots; which converts given RNA secondary structures into hierarchical trees by decomposing each structure into pairwise disjoint substructures, and then applies an existing tree alignment method to these hierarchical trees. From biological and computational standpoints, the hierarchical alignment has two advantages: (i) The decomposition makes possible to detect structural homologies even between phylogenetically less-related RNAs; and (ii) reduces search space of alignments, which enables us to devise computationally efficient algorithms. Experimental results suggest that alignments of RNAs by our method are more biologically plausible than by conventional ones.

1　Introduction

Many interesting RNA regions on the genome conserve molecular conformations generally called "structural homology" more than they conserve their nucleotide sequences (S.R.Eddy[3]). The assumption is now widely accepted, and identifying these regions has much scope to be investigated.

In this paper, we will propose a notion of *hierarchical alignment* of RNA structures including pseudoknots. The hierarchical alignment first decomposes given RNA structures each into *h-components*; and then defines a relation between them. Consequently, the RNA structures are converted into trees whose nodes are h-components, called *h-trees*. Then, the h-trees are aligned with dynamic programming algorithm.

2 Hierarchical Alignment of RNA Structures

Let (S, P) be an arc-annotated sequence [6]. For a base r, two bases i_1 and i_2 *surround* r if $i_1 < r < i_2$. For an arc (j_1, j_2), two bases i_1 and i_2 *surround* (j_1, j_2) if $i_1 < j_1 < j_2 < i_2$. In case of $(i_1, i_2) \in P$, we say that (i_1, i_2) surrounds (j_1, j_2). A base r is *accessible from* i_1 and i_2 if i_1 and i_2 surround r and there exists no arc (k_1, k_2) satisfying $i_1 < k_1 < r < k_2 < i_2$. An arc (j_1, j_2) is *accessible from* i_1 and i_2 if i_1 and i_2 surround (j_1, j_2) and there exists no arc (k_1, k_2) satisfying $i_1 < k_1 < j_1 < j_2 < k_2 < i_2$. In case of $(i_1, i_2) \in P$, we say that (j_1, j_2) is accessible from (i_1, i_2). For two arcs $p_1, p_2 \in P$, if p_2 is the only one accessible arc from p_1, then we write $p_2 <_s p_1$.

We define a *decomposition* $\mathbf{C}$ of (S, P) as a set of pairwise disjoint substructures of (S, P) whose union is (S, P). For components $c_1, c_2 \in \mathbf{C}$, c_1 *surrounds* c_2 iff there exist two bases i, j in c_1 and a base k in c_2 satisfying $i < k < j$. We say that c_2 is *accessible from* c_1 iff c_1 surrounds c_2 and there exists no component $c' \in \mathbf{C}$ such that c_1 surrounds c' and c' surrounds c_2. For a decomposition $\mathbf{C}$ of (S, P), we define a directed graph $G(\mathbf{C}, (S, P)) = (V, E)$ by $V = \mathbf{C}$ and $E = \{(c_1, c_2) \in \mathbf{C} \times \mathbf{C} \mid c_2 \text{ is accessible from } c_1\}$. When $G(\mathbf{C}, (S, P))$ is a forest, $\mathbf{C}$ is called a *hierarchical* decomposition of (S, P), and an element of $\mathbf{C}$ *hierarchical component* or "*h-component*".

We define a function ϕ which for a given input arc-annotated sequence (S, P), outputs a hierarchical decomposition of (S, P). Let $as_1 = (S_1, P_1)$ and $as_2 = (S_2, P_2)$ be arc-annotated sequences. Let Aln be an alignment between as_1 and as_2. For a base b in as_i, by $Aln_\phi(b)$, we denote the substructure in $\phi(as_j)$ $(j \neq i)$ containing the base b' which b is aligned to. (In case that such b' does not exist, $Aln_\phi(b)$ is undefined.) A substructure c in $\phi(as_i)$ is said to be *overlapping* in Aln if there exist two bases b_1 and b_2 in c such that $Aln_\phi(b_1)$ and $Aln_\phi(b_2)$ are defined and $Aln_\phi(b_1) \neq Aln_\phi(b_2)$ holds. An alignment Aln is *valid* for ϕ if every substructure in $\phi(as_i)$ $(i = 1, 2)$ is not overlapping in Aln.

Hierarchical alignment problem (**HAP**) is defined as follows:

Hierarchical Alignment Problem
[**Input**] A pair (as_1, as_2) of arc-annotated sequences, and a hierarchical decomposition function ϕ
[**Output**] An alignment Aln between $G(\phi(as_1), as_1)$ and $G(\phi(as_2), as_2)$ of minimum score such that Aln is valid for ϕ.

In this paper, we adopt a function ϕ which decomposes a given arc-annotated sequence (S, P) into three kinds of h-components: *L-arc*, *P-knot* and *S-loop* each defined below.

For two arcs $p_1 = (i_1, i_2) \in P$ and $p_2 = (j_1, j_2) \in P$, we write $p_1 \succ p_2$ iff $i_1 < j_1 < i_2$ or $i_1 < j_2 < i_2$ holds. $\succ^*$ denotes the reflexive and transitive closure of $\succ$. We define $p_1 \equiv_p p_2$ iff both $p_1 \succ^* p_2$ and $p_1 \prec^* p_2$ hold. Note that $\equiv_p$ is an equivalence relation over P.

L-arc is a sequence $p_1, \ldots, p_k$ of arcs of *maximal length* in P such that $p_1 <_s \cdots <_s p_k$. *P-knot* is an equivalence class in $P/\equiv_p$ whose cardinality is more than one. Note that P-knots correspond conventional pseudoknots. *S-loop* is a sequence of adjacent unpaired bases of maximal length.

Theorem 1 *An arc-annotated sequence can be decomposed into L-arcs, P-knots and S-loops in a unique manner. Furthermore, this decomposition is always a hierarchical decomposition.*

For calculating alignment between two h-components, our dynamic programming algorithm for HAP employs an original heuristic method for the case when they are both P-knots, and T.Jiang's method[6] otherwise. For two inputs of length n and m, respectively, the time complexity of our hierarchical alignment method is $O(n^2m^2)$.

3 Experimental Results

We have performed two kinds of experiments for comparing our method with T.Jiang's one[6], which can align pseudoknotted v.s. pseudoknot-free secondary structures.

Table 1: Comparisons of alignment accuracies[%] for three RNA families.

	Our method		T. Jiang's method		
RNA family name	average (± SD)	worst	average (± SD)	worst	pseudoknot
Tombus_3_IV	96.9 ± 3.3	88.5	97.1 ± 3.4	88.5	12.00
HDV_ribozyme	92.9 ± 5.4	81.6	97.4 ± 2.3	89.4	21.43
Corona_pk3	91.8 ± 8.0	68.8	83.3 ± 15.9	56.3	44.44

First, we have tested the prediction accuracy of the two methods for three RNA families including pseudoknots taken from Rfam[5]. Note that the column "pseudoknot" in Table 1 shows the ratio of arcs which we should remove for applying T.Jiang's method. The results in Table 1 show that our method achieves high and *stable* accuracy around 93%, whereas in T. Jiang's one, the accuracy decreases when we should remove more crossing arcs from the data.

Next, we compared the performances of structurally aligning two pairs of RNA group I introns taken from Gutell et al.[2]: (x) *Cryptendozyla hypophloia*, (y) *Metarhizium anisopliae* and (z) *Tetrahymena thermophila*. The pair (x, y) is phylogenetically more closely related than the pair (x, z). Figure 1 represents a piece of the alignment of the pair (x, z) by our method. The notations [P2.1], $\cdots$ in the figure correspond to those in [2]. We conclude that the whole results ensure the significance of our method in the following three points: (1) Our method exactly aligns pseudoknot regions P3 and P7 because of its pseudoknot admissibility. (2) Our method involves tolerances for long strand indels because hierarchical decompositions help localize the effect of indels. In T.Jiang's alignment results, an insertion of a long stacked pairs probably causes significant decrease of accuracy of aligning surrounding regions. (3) It is the most significant feature of hierarchical alignment that our method can detect structurally homologous regions even between RNA secondary structures of the phylogenetically less-related organisms. We should focus attention on P5 regions in Figure 1. Our method obtains plausible alignments even in the phylogenetically less-related case, while T.Jiang's method misaligns structurally homologous P5, P5a, P5b and P5c regions in this case. In evolutive process of RNAs, it is highly probable that structural homologies are preferred to be conserved more than sequence ones, and our method depends on structural homologies. On the other hand, T.Jiang's method more heavily depends on sequence homologies than our method. This is presumably the main reason why T.Jiang's method does not show good alingment results in the phylogenetically less-related case.

```
[  P3  ]    [ P4 ]   [P5]        [   P5a   ]                                         [     P5a     ]      [P5]    [ P4  ][  P6   ]   [  P6a ]      [ P6a
(((( (((    ((((((    (((        (   ( (( ((                                         ))     ) ) )  )       )))    )) ))))(    ( ((   ( ( ((((      ))))
CACCAGCUGG-UACAGGGGAC-GCCU-AA----C---C-GU-CU-----------------------------UAUAG-------GG-----ACC-G--G--C----GGCCCGAUC-CUGUG----G-CG-A-G-C-CCGAGUAGCGUCGG-
GACC-GUCAAAUUGCGGGAAAGGGGUCAACAGCCGUUCAGUACCAAGUCUCAGGGGAAACUUUGAGAUGGCCUUGCAAAGGGUAUGGUAAUAAGCUGACGGACAUGGUCCUA-ACCACGCAGCCAAGUCCUAAGUCAACAGAUCUU-CUGUU
(((( (((    ((((((    (((      (((( (((((((((((((((((((    )))))))))))(((     )))    )))    ) ))))))))     )))    )) ))))(((  ((((   ((((((((      )))))
[  P3  ]    [ P4 ]   [P5]      [     P5a   ][   P5b   ]    [   P5b   ][P5c] [P5c]    [     P5a       ]    [P5]    [ P4  ][P6] [P6a]  [  P6b ]      [ P6b

  ]  [   P6  ]   [  P7 ]    [ P3  ] [ P8  ]     [  P8 ]        [ P7 ][P9.0][ P9 ]      [ P9 ][ P9.1 ]    [P9.1a]      [P9.1a]      [ P9.1 ]
) )  )) )    )   ( (((((    ))))))) (((((((     )))))))        )))))) ((   ((((((      ))))))((((((((      (((           )))       ))))))))(((((((  (((
G-C--CG-U----CGCAACGCACGC-UAAGCGGUG-GGCUCUCAACA-GAGGGCU-UAAGGUACGUGCUAAU---CCACGGGAUAA-CCGUGGCCGCCUUGACU-A-GUUCCGAGUUUACGAAGC-A-AGGCGGGGUGGACAGGCGCUGCA
GAUAUGGAUGCAGUUCACAGACUAAAU-GUCGGUCGGGGAAGAUGUAUUCUUCUCAUAAGAUAUAGUCG-GACCU-CUCCU--UAAUGGGAG-CUA-GCGGA-UGAAGUG---A-UGCA--ACACUGGAG--CCGC-UGG-----------
)))  ))))  )))   ( (((((    )))))))((((((((     ))))))))       )))))) ((    (((((      ))))) ((( ((((     ((((           ))))      )))) )))
  ] [P6a] [P6]   [  P7 ]    [ P3  ][  P8  ]     [  P8  ]       [ P7 ][P9.0] [P9 ]      [ P9] [ P9.1 ]    [P9.1a]      [P9.1a]      [ P9.1 ]
```

Figure 1: Alignments by our method between phylogenetically less-related RNA group 1 intron data.

4 Related Works and Discussions

There exist several related works to our approach. Most related work is an algorithm for computing substring-preserving edit distance(Evans and Wareham[4]). Hierarchical alignment is similar to their approach, but differs from theirs in the respect of recurrence relations for calculating an alignment cost of two subtrees. Other related works are a method for computing similarity between RNA structures(Zhang et al.[9]) with extension to handle H-type pseudoknots and a method on a general edit distance(Jiang et al.[6]). Our method can handle a general class of pseudoknots properly including H-type. Significant comparisons between our method and the latter are discussed throughout this paper.

References

[1] H. Asakawa. Parsing Universal Tree Adjoining Grammars using Structural Information. Master's thesis, Dept. of Comp. Sci., Univ. of Electr.-Communi., 2004. (in Japanese).

[2] J. J. Cannone, S. Subramanian, M.N. Schnare, J.R. Collett, L.M. D'Souza, Y. Du, B. Feng, N. Lin, L.V. Madabusi, K.M. Muller, N. Pande, Z. Shang, N. Yu, and R.R. Gutell. The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and Other RNAs. *BioMed Central Bioinformatics.* 3:2, 2002. http://www.rna.icmb.utexas.edu.

[3] S.R.Eddy. Computational genomics of noncoding RNA genes. *Cell*, 109: pages 137-140, 2002.

[4] P.A. Evans and H.T. Wareham. Exact algorithms for computing pairwise alignments and 3-medians from structure-annotated sequences. *Pacific Symposium on Biocomputing* 6: pages 559-570, 2001.

[5] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucleic Acids Research*, 31, 1, pages 439-441, 2003. http://www.sanger.ac.uk/Software/Rfam/.

[6] T. Jiang, G.H. Lin, B. Ma, and K. Zhang. A General Edit Distance between RNA Structures. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB'01)*, pages 211-220, 2001.

[7] T. Jiang, L. Wang and K. Zhang. Alignment of trees-an alternative to tree edit. *Theoretical Computer Science 148*, pages 137-148, 1995.

[8] Y. Uemura, A. Hasegawa, S. Kobayashi and T. Yokomori. Tree adjoining grammars for RNA structure prediction. *Theoretical Computer Science 210*, pages 277-303, 1999.

[9] K. Zhang, L. Wang and B. Ma. Computing similarity between RNA structures. In *Proceedings of 10th Annual Symposium on Combinatorial Pattern Matching(CPM'99)*, LNCS 1645, Springer, pages 281-293, 1999.