ParaRNA: A parametric tool for aligning two RNA structures

Lusheng Wang[1] Wangsen Feng[1,2] Hao Zhao[1] Kaizhong Zhang[3] Jianping Li[1] Xiaowen Liu[1] *

[1]Dept. of Computer Science, City University of Hong Kong
[2] Department of Computer Science, Peking University, P.R. China
[3]Dept. of Computer Science, University of Western Ontario, London, Ont. N6A 5B7, Canada

Abstract

Alignment of RNA structures is very important in biological research. Similar to pair-wise sequence alignment, there is often disagreement about how to weight matches, mismatches, indels, and gaps when we compare two RNA structures. Here, we develop a parametric tool for aligning two RNA structures. With this tool, the users can see explicitly and completely the effect of parameter choices on the optimal alignments of RNA structures.
Availability: The software is available at http://www.cs.cityu.edu.hk/~lwang/software/ParaRNA
Contact: cswangl@cityu.edu.hk

1 Introduction

Ribonucleic Acid (RNA) plays an very important role in biological systems. It is well known that RNA regulates some viruses' functions (e.g.HIV) since the genetic information is contained in RNA instead of DNA. RNA has recently received more and more attention in biological research because of its catalytic properties. In general, it is presupposed by biologists that RNAs with similar molecular structures also have similar biological functions. Consequently, the comparison of RNA structures is useful for the classification and taxonomy of bacteria, viruses, etc(Chen *et al.*2000, Sakakibara *et al.*1999 , Sancoff 1985).

Similar to pair-wise sequence comparison, there is often disagreement about how to weight matches, mismatches, indels and gaps when comparing two trees. The study of setting parameters for sequence alignment started long time ago. For example, Kruskal and Sankoff investigated the setting of weights for gaps, substitutions and other operations for RNA sequences (Kruskal and Sankoff, 1983). Parametric alignment attempts to avoid the problem of choosing fixed parameter settings by computing the optimal alignment as a function of variable parameters for weights and penalties. The goal is to partition the parameter space into regions such that in each region one alignment is optimal. For sequence comparison, the parametric sequence alignment tools have been developed (Gusfield *et al.*, 1994; Gusfield and Stelling, 1996; Vingron and Waterman, 1994; Waterman *et al.*, 1992; Zimmer and Lengauer, 1997). It allows the users to see explicitly and completely the effect of parameter choices on the optimal sequence alignments. A software for parametric alignment of ordered trees was developed in (Wang and Zhao, 2003). Ordered trees can be used to describe RNA secondary structures (Jiang *et al.*, 1995).

Recently, new measures have been proposed for comparison of RNA structures. Those new measures allow users directly compare bases and base-pairs. The *edit distance* between RNA structures was proposed in (Zhang, 1998; Ma *et al.*, 2002). When both structures are tertiary, then the problem is NP-hard (Zhang, 1998). If one of the structures is secondary, we can use the algorithm in (Ma *et al.*, 2002) to solve the problem. An algorithm that considers affine gaps was given in (Collins *et al.*, 2000). Another measure is the *alignment distance* between RNA structures (Wang and Zhang, 2001). Again, if one of the structures is secondary, we can solve the problem efficiently (Wang and Zhang, 2001). The algorithm for alignment distance is faster than that of edit distance for RNA structures. Here, in this paper, we adopt the alignment approach for RNA structures (Wang and Zhang, 2001). In order to combine with the algorithm for paramatric space decomposition in (Gusfield *et al.*, 1994; Gusfield and Stelling, 1996), we propose the maximiza-

*Correspondence author: liuxw@cs.cityu.edu.hk

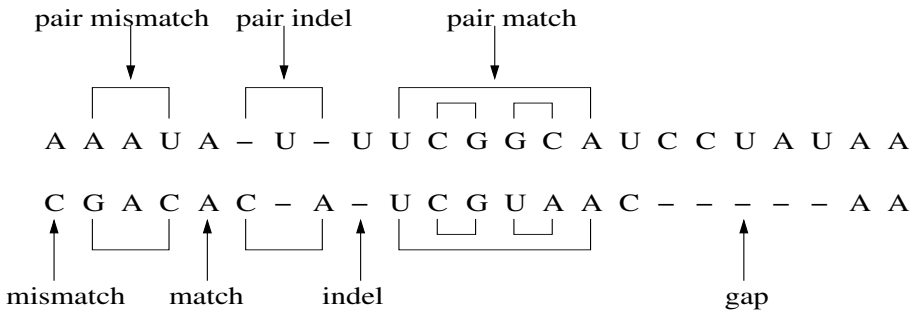

Figure 1: Edit operations

tion version of alignment between RNA structures. We extend the algorithm in (Wang and Zhang, 2001) to work for the maximization version of alignment between RNA structures. We develop a software, ParaRNA, that allows users to see the effect of the parameter choices on optimal alignments of RNA structures.

2 Algorithms

Our software contains two algorithms, the algorithm for computing the optimal alignment between two RNA structures (Wang and Zhang, 2001) and the algorithm for decomposing the parametric space (Gusfield *et al.*, 1994; Gusfield and Stelling, 1996). We will discuss the two algorithms in the following sections.

2.1 Computing the optimal alignment of RNA Structures

Since aligning two tertiary RNA structures is NP-hard, we assume that at least one of the two RNA structures is secondary structure.

The alignment of two RNA structures was first studied in (Wang and Zhang, 2001). Let R be an RNA sequence on $\{A, U, G, C\}$. $r[i]$ represents the i-th nucleotide. An RNA structure $R(P)$ contains an RNA sequence R and a set of base pairs $P \subset \{1, 2, \cdots, |R|\}^2$, where (i, j), $i < j$, represents a base pair $(r[i], r[j])$ in R. When there is no confusion, we use R instead of $R(P)$ to represent an RNA structure.

Given two RNA structures R_1 and R_2, an alignment of R_1 and R_2 can be obtained by inserting spaces into the two sequences R_1 and R_2 such that the obtained sequences R'_1 and R'_2 has the same length and the following conditions are satisfied.

1. If $r'_1[i]$ is an unpaired base in R'_1, then either $r'_2[i]$ is an unpaired base in R'_2 or $r'_2[i]$ is a space. If $r'_2[i]$ is an unpaired base in R'_2, then either $r'_1[i]$ is an unpaired base in R'_1 or $r'_1[i]$ is a space..
2. If $(r'_1[i], r'_1[j])$ is a base pair in R'_1, then either $(r'_2[i], r'_2[j])$ is a base pair in R'_2 or both $r'_1[i]$ and $r'_1[j]$ are spaces.

The alignment thus obtained is denoted as $A(R'_1, R'_2)$.

The edit operations

Consider an alignment $A(R'_1, R'_2)$. (1)If $r'_1[i]$ and $r'_2[i]$ are unpaired bases in R_1 and R_2 and $r'_1[i] = r'_2[i]$, then there is a *match* at the i-th position. (2) If $r'_1[i]$ and $r'_2[i]$ are unpaired bases in R_1 and R_2 and $r'_1[i] \neq r'_2[i]$, then there is a *mismatch* at the i-th position. (3) If $r'_1[i]$ is an unpaired base and $r'_2[i] =' -'$, then this is a *deletion* at the i-th position. (4) If $r'_2[i]$ is an unpaired base and $r'_1[i] =' -'$, then this is a *insertion* at the i-th position. We do not distinguish insertion and deletion and thus use *indel* to indicate both insertion and deletion. Suppose that $(r'_1[i], r_1\prime[j])$ and $(r'_2[i], r_2\prime[j])$ are two base pairs in R_1 and R_2, respectively. (5)If $r'_1[i] = r'_2[i]$ and $r'_1[j] = r'_2[j]$, then there is a *base pair match*. (6) If $r'_1[i] \neq r'_2[i]$ or $r'_1[j] \neq r'_2[j]$, then there is a *base pair mismatch*. (7)If $(r'_1[i], r'_1[j])$ is a pair in R_1 while $r'_2[i] = r'_2[j] =' -'$, then there is a *base pair insertion/deletion* (base pair indel). Figure 1 illustrates the above edit operations.

The alignment value

For simplicity, in our software, we only have three kinds of operations, match, mismatch and indel. Each base pair operation is counted as two base operations. A *gap* in an alignment $A(R'_1, R'_2)$ is

a consecutive subsequence of spaces in either R'_1 or R'_2 with maximal length.

Let mt_A, ms_A, mi_A and gp_A be the numbers of matches, mismatches, indels and gaps, in $A(R'_1, R'_2)$, respectively. We define the value of the alignment to be

$$V_A \equiv \alpha \times mt_A - \beta \times ms_A - \gamma \times mi_A - \delta \times gp_A \quad (1)$$

where α, β, γ and δ are the values for a single match, mismatch indel and gap, respectively. In order to use the parametric space decomposition algorithm in (Gusfield and Stelling, 1996), we have to adopt the maximization version. That is, all α, β, γ and δ should be non-negative. Thus, only matches give positive contribution to the value and the other three operations contribute negative values. Modifying the algorithm in (Wang and Zhang, 2001), we have an algorithm to compute an optimal alignment between two RNA structures for the maximization version. (In (Wang and Zhang, 2001), the algorithms are for minimization version.) The time complexity of the algorithm is $O(|R_1||R_2|S_1S_2)$, where $|R_1|$ and $|R_2|$ are the lengths of the two RNA sequences and S_1 and S_2 are the numbers of stems in R_1 and R_2, respectively.

2.2 The algorithm for computing a polygonal decomposition

We consider the parametric problem where two parameters in equation (1) are fixed and the other two are variable. In our software, α and γ are fixed while β and δ are variable. The default value of α is 1 and the default value of γ is 0.5. The users are allowed to set their own α and γ values. Our goal is to partition the parametric space for β and δ into some regions such that the whole region has the same optimal alignment. The users are allowed to see explicitly the effect of parametric choices on computing the optimal alignment. In (Gusfield *et al.*,1994), it was proved that

Lemma 1 *For sequence alignment, the parameter space is decomposed into convex polygons such that any alignment that is optimal for some β, δ point in the interior of a polygon P is optimal for all points in P and nowhere else.*

Since the proof of Lemma 1 is only related to equation (1), Lemma 1 also holds for RNA structure alignment. (Gusfield and Stelling, 1996) gave an efficient algorithm for computing a polygonal decomposition. Since the algorithm is only related to equation (1), we can directly use the algorithms for alignment of RNA structures. The time complexity of the algorithm is $O(S \cdot T)$, where S is the number of polygons, T is the time required for optimally aligning two RNA structures.

3 Implementation

ParaRNA is developed in C++. The input of ParaRNA is two RNA structures. At most one of them can be of tertiary structure. ParaRNA can decompose the parametric space into convex polygons. In our program, α is fixed as 1. A value of γ should be provided by the user.. The default value of γ is 0.5. The user has to choose different value ranges for β and δ. The β-δ plane is the parametric space. ParaRNA generates a graphical output of the final polygonal decomposition of the parametric space. Figure 2 illustrates such a decomposition. When the user click any point in a region, ParaRNA computes an optimal alignment between the two RNA structures using the parameters at this point. Some real RNA structures are provided in ParaRNA as examples for the use to test our program.

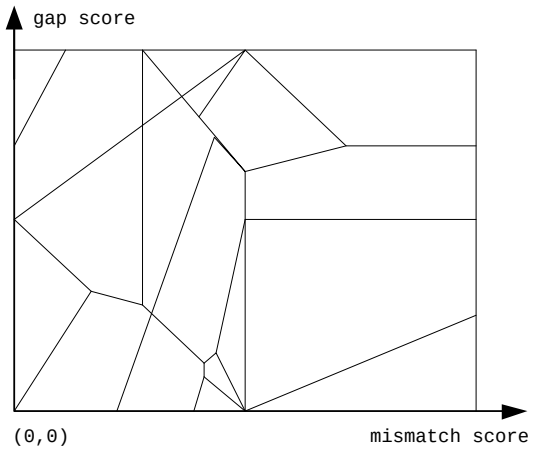

Figure 2: An decomposition example. When the user click a point in the polygon, ParaRNA computes and outputs an optimal alignment at this point. The user can click the "view" button to get all the optimal alignment corresponding to the polygons.

4 Acknowledgements

The work is fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China[Project No. CityU 1070/02E].

References

[1] Chen, J.H., Le, S.Y. and Maizel, J. V., (2000) Prediction of common secondary structures of RNAs: a genetic algorithm approach, *Nucleic Acids Research*, **28**, 991-999.

[2] Collins, G., Le, S. Y., and Zhang, K., (2000) A new method for computing similarity between RNA structures, *Proceedings of the Second International workshop on Biomolecular Informatics*, 761-765, Atlantic City.

[3] Gusfield, D., Balasubramanian, K., and Naor, D., (1994) Parametric optimization of sequence alignment, *Algorithmica*, **12**, 312-326.

[4] Gusfield, D. and Stelling, P. (1996) Parametric and inverse-parametric sequence alignment with XPARAL. In R. F. Doolittle, editor, *Methods in Enzymology*, **266**, Computer Methods for Macromolecular Sequence Analysis, 481-491.

[5] Kruskal J., and Sankoff, D., (1983) An anthology of algorithms and concepts for sequence comparison, in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, D. Sankoff and J. Kruskal (eds), Addison-Wesley, 265-310.

[6] Ma, B., Wang, L, and Zhang, K., (2002) Computing similarity between RNA structures, *Theoretical Computer Science*, **276**, 111-132.

[7] Sakakibara, Y.M., Trown, Hughey, R., Mian, I.S., KSjolander, Underwood,R., and Haussler, D., (1994) Stochastic context-free grammar for tRNA modeling, *Nucleic aids research*, **22**, 5112-5120.

[8] Vingron, M. and Waterman, M. (1994) Sequence alignment and penalty choice. *J. Mol. Biol.*, **235**, 1-12.

[9] Wang, L., and Zhao, J., (2003) Parametric Alignment of Ordered Trees, *Bioinfomatics* **19**, 2237-2245.

[10] Wang, Z., and Zhang, K., (2001) Alignment between two RNA Structures, *MFCS* 690-702.

[11] Waterman, M., Eggert, M. and Lander, E. (1992) Parametric sequence comparisons. *Proc. Natl. Academy Science*, **89**, 6090-6093.

[12] Zhang, K., (1998) "Computing similarity between RNA secondary structures", *Proceedings of IEEE International Joint Symposia on Intelligence and Systems*, 126-132, Rockville, Maryland.

[13] Zimmer, R. and Lengauer, T. (1997) Fast and numerically stable parametric alignment of biosequences. *RECOMB*, 44-353.