BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2151–2152
doi:10.1093/bioinformatics/btr338

Sequence analysis Advance Access publication June 8, 2011

RNASAlign: RNA Structural Alignment System

Thomas K. F. Wong[1], Kwok-Lung Wan[1], Bay-Yuan Hsu[2], Brenda W. Y. Cheung[1], Wing-Kai Hon[2], Tak-Wah Lam[1] and Siu-Ming Yiu[1,*]

[1]Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong and [2]Department of Computer Science, National Tsing Hua University, Taiwan

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Structural alignment of RNA is found to be a useful computational technique for idenitfying non-coding RNAs (ncRNAs). However, existing tools do not handle structures with pseudoknots. Although algorithms exist that can handle structural alignment for different types of pseudoknots, no software tools are available and users have to determine the type of pseudoknots to select the appropriate algoirthm to use which limits the usage of structural alignment in identifying novel ncRNAs.

Results: We implemented the first web server, RNASAlign, which can automatically identify the pseudoknot type of a secondary structure and perform structural alignment of a folded RNA with every region of a target DNA/RNA sequence. Regions with high similarity scores and low e-values, together with the detailed alignments will be reported to the user. Experiments on more than 350 ncRNA families show that RNASAlign is effective.

Availability: http://www.bio8.cs.hku.hk/RNASAlign.

Contact: smyiu@cs.hku.hk

Received on March 14, 2011; revised on May 21, 2011; accepted on May 30, 2011

1 INTRODUCTION

Structural alignment is an important topic in studying RNAs especially for non-coding RNAs (ncRNAs) which are found to be evolutionary conserved in terms of primary sequences and secondary structures. To check if a given sequence (target) contains regions that are possibly candidates of a particular ncRNA family, a useful computational approach is to compute the structural alignment between a folded ncRNA (query) of the family and every unfolded region in the target sequence where the alignment score represents their sequence and structural similarity. The approach has been shown to be effective (Han *et al.*, 2008; Wong *et al.*, 2011). Some existing software such as RSEARCH (Klein and Eddy, 2003) and FASTR (Zhang *et al.*, 2005) belongs to this category.

However, the approach is not being widely used partially due to the followings. First, the available tools (RSEARCH and FASTR) do not support query with pseduoknots. The secondary structure of an ncRNA is said to contain a pseudoknot if there are two base pairs at positions (i,j) and (i',j'), where $i<j$ and $i'<j'$, such that $i<i'<j<j'$ or $i'<i<j'<j$. Secondary structures including pseudoknots are found to be critical in some biological functions (Adams *et al.*, 2004; Dam *et al.*, 1992). Secondly, although there are algorithms for computing the structural alignment for different types of pseudoknots, no software tools are available. Thirdly, each of these algorithms is designed for a specific pseudoknot type, the user has to know the pseudoknot type of the query to select the appropriate algorithm for performing the structural alignment. When the query is long, it may not be a trivial problem for the user.

In this note, we introduce the first web server RNASAlign designed for this purpose. The user only requires to input a query RNA sequence together with its secondary structure and a target DNA/RNA sequence. The system will automatically identify the pseudoknot type of the query structure and perform the structural alignment between the query and every region of the target sequence. The system supports regular structure (without pseudoknots) and a wide range of pseudoknots such as standard pseudoknot, embedded standard pseudoknot, recursive standard pseudoknot, simple non-standard pseudoknot and recursive simple non-standard pseudoknot [for the definitions of different pseudoknot types, please refer to Wong *et al.* (2011); Zhang *et al.* (2005) or the help page of the web server]. According to our experiment of more than 350 ncRNA families, RNASAlign is effective in identifying ncRNAs and achieves an average sensitivity of 95% when considering the top 100 reported candidates.

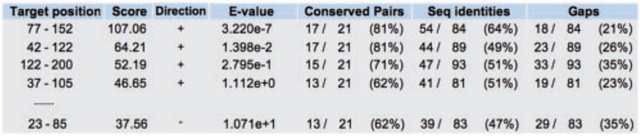
2 METHODS

There are three major components of the system. The system first analyzes the secondary structure of the input query to classify the structure as a regular structure or one of the known pseudoknot types. This is not a trivial problem. We designed efficient algorithms to solve this classification problem (see the help page for more details). We implemented all algorithms for structural alignments of different pseudoknot types (Wong *et al.*, 2011, ?). All these algorithms use dynamic programming approach and give an optimal solution for the alignment. We use ‘RIBOSUM85-60’ as the scoring matrix (Klein and Eddy, 2003). Once the pseudoknot type is determined, the system will perform the structural alignment between the folded query and every unfolded region (with similar length as the query) of the target sequence using the appropriate program. Besides the structural alignment score, we also compute the corresponding e-value according to the method suggested by (Klein and Eddy, 2003). The alignment scores are assumed to follow the Gumbel distribution and the e-value represents the expected number of hits with score greater than the resulting score. Finally, the system will report all regions with e-value below a user-defined threshold (the default is 0.1) together with the detailed alignment of the region and the query.

3 USAGE EXAMPLE

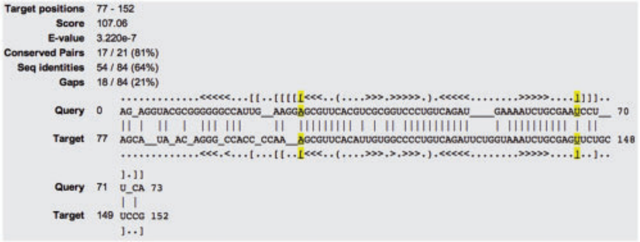
The input to the system includes a query sequence with its secondary structure and a target sequence. We use a member of the family

*To whom correspondence should be addressed.



Target position	Score	Direction	E-value	Conserved Pairs	Seq identities	Gaps
77 - 152	107.06	+	3.220e-7	17 / 21 (81%)	54 / 84 (64%)	18 / 84 (21%)
42 - 122	64.21	+	1.398e-2	17 / 21 (81%)	44 / 89 (49%)	23 / 89 (26%)
122 - 200	52.19	+	2.795e-1	15 / 21 (71%)	47 / 93 (51%)	33 / 93 (35%)
37 - 105	46.65	+	1.112e+0	13 / 21 (62%)	41 / 81 (51%)	19 / 81 (23%)
....						
23 - 85	37.56	-	1.071e+1	13 / 21 (62%)	39 / 83 (47%)	29 / 83 (35%)

Fig. 1. Summary of the scores and e-values of the reported target regions.



```
Target positions  77 - 152
           Score  107.06
         E-value  3.220e-7
 Conserved Pairs  17 / 21 (81%)
  Seq identities  54 / 84 (64%)
            Gaps  18 / 84 (21%)
                     .............<<<<<...[[..[[[[[<<<..(....>>>.>>>>>.).<<<<<.......>>>>>....]]]]..
           Query  0  AG_AGGUACGCGGGGGGCCAUUG__AAGGAGCGUUCACGUCGCGGUCCCUGUCAGAU____GAAAAUCUGCGAAUCCU__ 70
                     || |  ||  |  ||| |||     ||  ||||||||| | | || |||||||||||    | |||||||||| | ||
          Target 77  AGCA__UA_AC_AGGG_CCACC_CCAA__AGCGUUCACAUUGUGGCCCCUGUCAGAUUCUGGUAAAUCUGCGAGUUCUGC 148
                     .............<<<.<...[...[[..[<<<..(....>>>.>.>>>.).<<<<<.......>>>>>....]..]..

                     ].]]
           Query 71  U_CA 73
                     | |
          Target 149 UCCG 152
                     ]..]
```

Fig. 2. The alignment between the query and each target region.

Mammalian CPEB3 ribozyme (RF00622) as a query example and the structure is of a recursive simple non-standard pseudoknot. We use a 200 nt long sequence as an example target sequence. Figure 1 shows the summary of the result with the alignment scores and the e-values for a set of regions in the target sequence. The results are sorted in decreasing alignment scores. The summary also shows the percentages of conserved base pairs, sequence identities and gaps between the query and the region.

RNASAlign also lists the detailed alignment between the query and each reported regions (Fig. 2). The alignment indicates the conserved sequence between the query and the region by using symbol '|'. Also, when the user moves the curser over the base pair, the base pair will be highlighted in yellow if the base pair appears in both query and the target region, while it will be highlighted in red if it only appears in the query but not in the target region.

RNASAlign is installed in a server with one quad-core CPU and 16 G memory. It can handle four simultaneous queries at the same time. As shown in Table 1, a query with length 70–90 usually takes around half an hour for structurally aligning with a target sequence of length 500 or around 1 h for a target of length 1000. It will take longer if the query includes a non-standard pseudoknotted structure.

4 EXPERIMENTAL RESULTS

We selected a set of families in Rfam 10.1 database for which the lengths of the consensus structures are $\leq$150 long and they have >100 members. There are over 350 families. For each family, we randomly picked one of the members as the query and we constructed a long random sequence of length about 200 times the length of the query sequence. The random sequence was generated with same probability for all characters A,C,G,U. Then we embedded 100 of the members into this long random sequence in arbitrary positions. This is our target sequence. For each family, we inputted the query and the target into RNASAlign, and we checked the first 100 non-overlapping[1] regions and counted how many of them are the real members of the family. The average percentage of real members shown in top 100 non-overlapping regions for all around 350 families is 95. Table 2 shows the results of some families (see the help page for a full list). We also compare our programs with those without considering pseudoknots. We repeated the experiment on BLAST (with default parameters), a sequence-only software, and FASTR, a pseudoknot-free structural alignment method. Results are shown in Table 2. Note that that both BLAST and FASTR run very fast (<1 s).

Table 1. Summary of alignment time (in seconds)

Structure	Query length	
	70	90
Regular	0.4	0.5
Standard pseudoknot	51.0	95.6
Recursive standard pseudoknot	16.0	32.4
Embedded standard pseudoknot	39.5	74.7
Recursive simple non-standard	1151.9	2959.6
Simple non-standard pseudoknot	4696 (1.3 h)	10303 (2.9 h)

Table 2. Summary of the experimental results

Family	Structure	% of members in top 100 regions		
		RNASAlign	BLAST	FASTR
RF00021	Regular	100	99	100
RF00094	Rec. simple non-standard	100	38	96
RF00523	Standard pseudoknot	96	26	90
RF00622	Rec. simple non-standard	97	91	97
RF01084	Rec. standard pseudoknot	97	67	97

Funding: This research is supported in parts by the Seed Funding Programme for Basic Research (201011159115) of HKU.

Conflict of Interest: none declared.

REFERENCES

Adams,P.L. *et al*. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45–50.

Dam,E. *et al*. (1992) Structural and functional aspects of RNA pseudoknots. *Biochemistry*, **31**, 11665–11676.

Han,B. *et al*. (2008) Structural alignment of pseudoknotted RNA. *J. Comput. Biol.*, **15**, 489–504.

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

Wong,T.K.F. *et al*. (2011) A Memory Efficient Algorithm for Structural Alignment of RNAs with Pseudoknots. In *IEEE/ACM TCBB* [Epub ahead of print, doi:10.1109/TCBB.2011.66.].

Wong,T.K.F. *et al*. (2011) Structural alignment of RNA with complex pseudoknot structure. *J. Comput. Biol.*, **18**, 97–108.

Zhang,S. *et al*. (2005) Searching genomes for noncoding RNA using FastR. *IEEE/ACM TCBB*, **2**, 4.

[1] Two regions are regarded as overlapped with each other if they have >50% in common.