Homology-modeling of complex structural RNAs

Wei Wang[a], Matthieu Barba[b][1], Philippe Rinaudo[a], Alain Denise[a,b] and Yann Ponty[c]

a. LRI, Université Paris-Sud, CNRS,Université Paris-Saclay. b. I2BC, Université Paris-Sud, CNRS,Université Paris-Saclay. c. LIX, CNRS, Ecole Polytechnique, INRIA, Université Paris-Saclay.

Abstract

Aligning macromolecules such as proteins, DNAs and RNAs in order to reveal, or conversely exploit, their functional homology is a classic challenge in bioinformatics, with far-reaching applications in structure modelling and genome annotations. In the specific context of complex RNAs, featuring pseudoknots, multiple interactions and non-canonical base pairs, multiple algorithmic solutions and tools have been proposed for the structure/sequence alignment problem. However, such tools are seldom used in practice, due in part to their extreme computational demands, and because of their inability to support general types of structures. Recently, a general parameterized algorithm based on tree decomposition of the query structure has been designed by Rinaudo et al. We present an implementation of the algorithm within a tool named LiCoRNA. We compare it against state-of-the-art algorithms. We show that it both gracefully specializes into a practical algorithm for simple classes pseudoknot, and offers a general solution for complex pseudoknots, which are explicitly out-of-reach of competing softwares.

Introduction

Since Thomas R. Cech discovered that RNA is able to catalyze chemical reaction (Cech 1985), increasing exciting experimental results demonstrated the versatility of RNA and its importance in many cellular processes. Non-coding (nc) RNA has increasingly been shown to be a major player in all cell processes, notably in gene regulation (Zimmerman and Dahlberg 1996) (Sleutels, Zwart, and Barlow 2002)(Willard and Salz 1997)(Eddy 2001). Like proteins, ncRNAs molecules fold into complex three-dimensional structures which are essential to their function. Therefore, one cannot fully understand the biological process without a structurally-aware annotation of ncRNAs.

Briefly, modeling RNA structure relies on two complementary approaches, homology modeling and *ab initio* modeling. Here we focus on homology modeling, also known as sequence-structure alignment. In recent years, there has been an increasing amount of literatures

[1] Now at European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

on RNA sequence-structure alignment for secondary structure prediction including pseudoknots. Matsui et al. (Matsui, Sato, and Sakakibara 2004) proposed pair stochastic tree adjoining grammars (PSTAG) for aligning and predicting RNA structure including Lyngso & Pederson (L&P) pseudoknots (Lyngsø and Pedersen 2000). Han et al. (Han, Dost, and Bafna 2008) contributed an algorithm for Jabbari & Condon (J&C) pseudoknots (Jabbari et al. 2007). However, the resulting algorithm was complex, and practical tools such as PAL (Han, Dost, and Bafna 2008) only support L&P pseudoknots. Another proposed method is based on profile context-sensitive hidden Markov models (profile-csHMMs) by Yoon et al. (Yoon and Vaidyanathan 2008). Profile-csHMMs have been proven more expressive, and were shown to handle J&C pseudoknot. Here, we developed a fully general method for the sequence-structure comparison, which is able to take as input any type of pseudoknotted structures. In the following we briefly present the algorithm, its implementation and some preliminary results in comparison with other programs.

Material & Methods

Model and algorithmic foundations

Since details of the LiCoRNA sequence-structure alignment algorithm has been published in (Rinaudo et al. 2012), we only briefly describe the algorithm. A query RNA sequence/structure (A) is represented as a general arc-annotated sequence, in which vertices represent nucleotides, and edges represent canonical interactions and backbone adjacencies. Our goal is then to align a query RNA A to a target RNA sequence (B), in a way that minimizes an overall cost function, depending on sequence similarity, base-pair similarity, and structure conservation. More specifically, our objective cost function includes terms for base and interaction substitutions, calculated with RIBOSUM85-60 as described by Klein and Eddy (Klein and Eddy 2003). Gap penalties are computed using two affine cost functions for loops and helices. Since helices are generally more conserved than loop regions, an optimal alignment is less likely to feature gaps in stacked regions. Accordingly, the opening gap penalty within stacked regions is set to twice that of loop regions (200 and 100, respectively). The elongation gap penalty are set to 50 and 20 respectively.

Our alignment algorithm critically relies on the concept of tree decomposition, which we now remind.

Definition 1 (Tree decomposition of an arc-annotated sequence). Given an arc-annotated sequence A=(S, P), a tree decomposition of A is a pair (X,T) where $X=\{X_1, ..., X_N\}$ is a family of subsets of positions {i, $i \in [1,n]$}, n = length(S), and T is a tree whose nodes are the subsets X_r (called bags), satisfying the following properties:

1. Each position belongs to a bag: $\cup_{r\in[1,N]} X_r=[1,n]$.
2. Both ends of an interaction are present in a bag: $\forall (i,j) \in P, \exists r \in [1, N], \{i, j\} \in X_r$.
3. Any two consecutive positions are both present in a bag: $\forall i \in [1, n-1], \exists r \in [1, N], \{i, i+1\} \in X_r$.
4. For every X_r and X_s, $r, s \in [1, N]$, $X_r \cap X_s \subset X_t$ for all X_t on the shortest path between X_r and X_s.

Figure 1 shows a tree decomposition for a pseudoknot-free and L&P pseudoknot arc-annotated sequence. Once a tree decomposition has been built for the query RNA, a dynamic

programming algorithm is used to find the optimal alignment between the query and a given target sequence.

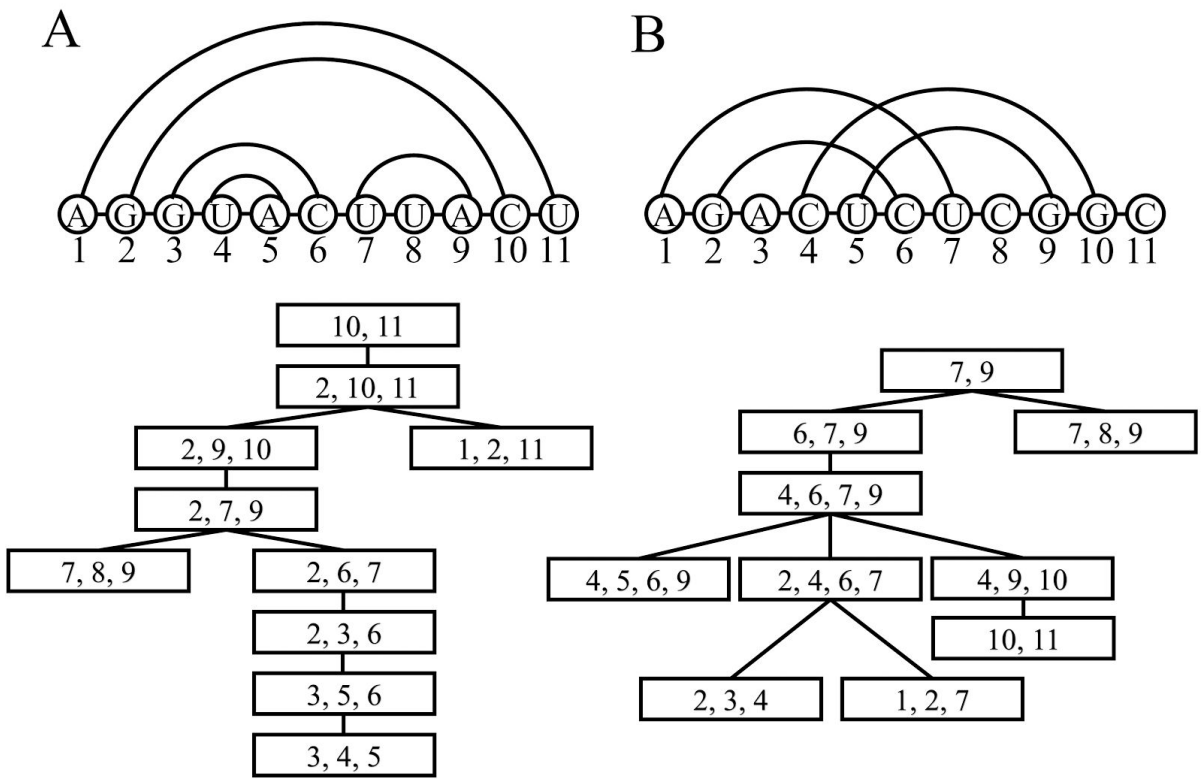

Figure 1. Two arc-annotated sequences and corresponding tree decompositions. A) pseudoknot-free. B) L&P pseudoknot.

Implementation aspects

Finding the optimal tree width and tree decomposition for a general graph is a NP-hard problem (Bodlaender and al. 1986). Fortunately, efficient heuristic algorithms algorithms have been proposed for computing upper/lower bounds on the treewidth in a constructive fashion (Bodlaender and Koster 2010)(Bodlaender and Koster 2011). We used LibTW, a Java library implementing various tree decomposition algorithms (http://www.treewidth.com). In particular, we used the GREEDYDEGREE and GREEDYFILLIN heuristics (van Dijk, van den Heuvel, and Slob 2006) because of their dominant behavior observed upon previous empirical studies.

To accelerate the dynamic programming algorithm without losing accuracy, we implemented the M constraints as illustrated by Uzilov et al (Uzilov, Keegan, and Mathews 2006). The M parameter reduces the computational cost through restricting the scanned region instead of the whole length in the target sequence for a particular base. To be more precise, suppose L_1 and L_2 are respectively the total length of the query structure A and target sequence B, and m and n are the nucleotide indices in A and B respectively. The following inequality must be satisfied for allowing m and n to be aligned together:

$$\left|\frac{L_2}{L_1}m\text{-}n\right| \le M$$

The default value for M is the difference between the two sequence lengths; and if the difference is less than 6, we set M to be 6.

Benchmark and applications

First, we compared our generic program LiCoRNA with the three available state-of-the-art programs which handle pseudoknots: PSTAG (Matsui, Sato, and Sakakibara 2004), PCSHMM (Yoon and Vaidyanathan 2008) and PAL (Han, Dost, and Bafna 2008). The dataset used in our experiments combines data from the RFAM database (Nawrocki, Burge, and Bateman 2014) and the PseudoBase database (Taufer et al. 2009). RFAM is a collection of RNA families in which all the sequences are aligned and all families are annotated with secondary structures using covariance model (CM) method. However, CM cannot effectively model pseudoknots and therefore, there is no reliable pseudoknot annotation for each family. On the other hand, PseudoBase provides reliable pseudoknot annotations for single sequences. Infernal (Nawrocki and Eddy 2013) was used here to find the most similar RFAM families with default E-value cutoff of 0.0001 for each pseudoknot sequence in PseudoBase database. Pseudoknot annotations were added into the corresponding RFAM families. Overall, 14 families with different kinds of pseudoknot were obtained. For each family in our dataset, we chose in turn each of its members, along with its pseudoknotted consensus, as the query sequence to predict the secondary structure of the other members.

Rfam	TW	Len	Pseudo	LiCoRNA(%)			PAL(%)			PSTAG(%)			PCSHMM(%)		
				SN	SP	AFI	SN	SP	AFI	SN	SP	AFI	SN	SP	AFI
RF00165	3	60~65	PKB255	88.6	93.2	81.4	91.3	**96.3**	**82.4**	**92.8**	93.8	79.4	92.6	93.7	80.5
RF00381	3	56~60	PKB177	**92.0**	**96.1**	**94.3**	90.1	95.4	92.8	89.7	93.6	89.5	90.4	94.7	90.6
RF00521	3	71~81	PKB345	80.6	89.2	**89.3**	**88.2**	**90.3**	81.9	86.1	87.7	67.6	85.6	87.0	70.7
RF00523	3	41~47	PKB206	88.2	93.7	**91.3**	**89.3**	**100**	89.1	83.0	91.2	88.0	84.2	93.2	89.7
RF01072	3	31~32	PKB55	**100**	**100**	**99.1**	**100**	**100**	**99.1**	**100**	**100**	98.7	99.8	99.8	98.7
RF01074	3	39~41	PKB44	**96.8**	**100**	**95.0**	**96.8**	**100**	89.6	93.0	94.6	92.3	95.8	98.3	**95.0**
RF01076	3	73	PKB218	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
RF01093	3	59~61	PKB258	92.1	96.5	**96.5**	92.1	96.5	95.5	**93.5**	**96.9**	87.0	91.5	94.7	87.5
RF01097	3	50~53	PKB107	**96.8**	**100**	**96.5**	**96.8**	**100**	94.9	92.6	94.7	85.9	95.0	97.6	88.0
RF01099	3	49	PKB273	**94.7**	**100**	**100**	88.6	97.8	97.0	94.4	99.5	99.5	94.6	99.8	99.8
RF01077	3	67~69	PKB57	**99.4**	**99.9**	**99.5**				98.9	99.4	98.0	98.9	99.4	99.0
RF00041	4	120~123	PKB169	**91.8**	**96.8**	**97.3**									
RF00094	4	90~96	PKB75	**88.7**	**92.1**	**74.1**									
RF00140	4	114~120	PKB71	**94.0**	**96.7**	**92.3**									

Table 1. 14 Pseudoknotted RFAM families, and their support within LiCoRNA, PAL, PSTAG and PCSHMM. The highest values of parameter SN, SP, AFI for each family are labeled in bold. TW: treewidth (calculated by GREEDYFILLIN), SN: Sensitivity, SP: Specificity, AFI: Average fractional identity, Len: the length range.

Table 1 reports the comparison of the three state-of-the-art implementations and our software for each RFAM pseudoknotted family in our benchmark set. Average fractional identity (AFI) of pairwise alignment and Sensitivity/specificity analysis have been performed, assuming correctness of the RFAM alignment. The fractional identity represents the alignment identity between the test and reference alignment, that is the number of identities divided by the length of the alignment. This parameter is calculated by the tool CompalignP that is distributed with BRAliBase 2.1 (Wilm, Mainz, and Steger 2006). Good alignment performance is demonstrated by being close to 1. The predicted structure is evaluated by Specificity = TP / (TP + FP) and Sensitivity = TP / (TP + FN), where TP (true positive) represents the number of correctly predicted base pairs, FP (false positive) represents the number of predicted base pairs which are not in the annotated structure, and FN (false negative) represents the number of base pairs in the annotated structure that are not predicted. The fact that the parameter Specificity and Sensitivity are close to 1 indicates good performances.

Table 1 shows that LiCoRNA results are generally equivalent or better than results of its competitors. Notably, the AFI is almost always better for LiCoRNA than for any of the other programs. It must be noted that some of the structures predicted by PAL were corrected, since they contained non Watson-Crick and non Wobble base pairs, even though the reference structures contained only Watson-Crick and Wobble basepairs. Moreover, LiCoRNA is the only program which gives an alignment for the last three structures whose pseudoknots belong to the Rivas & Eddy (R&E) class (Rivas and Eddy 1999), the most complex class of pseudoknots.

Use case : Realignment of PK RFAM families

LiCoRNA can also be used to curate RFAM pseudoknotted alignments. Indeed, RFAM alignments are retrieved and aligned using covariance models, a grammar formalism that discards crossing interactions. Some pseudoknots are recovered *post-facto* by an iterative modeling strategy, but some are very likely missed due to the lack of structural awareness of the initial alignment. However, once an experimental structure is known, it can be mapped onto the alignment, and each sequence be realigned in the light of the new structural constraint.

For instance, we took the query sequence X63538.1/1434-1495 (A) and target sequence DQ445912.1/27322-27383 (B) (Figure 3A) in the Corona_pk3 RFAM family (RF00165). Sequence A has a validated L&P pseudoknot in PseudoBase with PKB255. Our realignment reveals some discrepancy with the initial RFAM alignment, but the overall predicted structure for sequence B is conserved. However, in Figure 3B, there is quite large discrepancy between RFAM and our realignment. The query and target sequence are AAWR02040610.1/2027-2086 (PseudoBase number PKB258) and ACTA01044722.1/650-709 in RF_site5 family, respectively. We hope that a systematic realignment will allow to reveal or refute an evolutionary pressure towards the preservation of a functional pseudoknot.

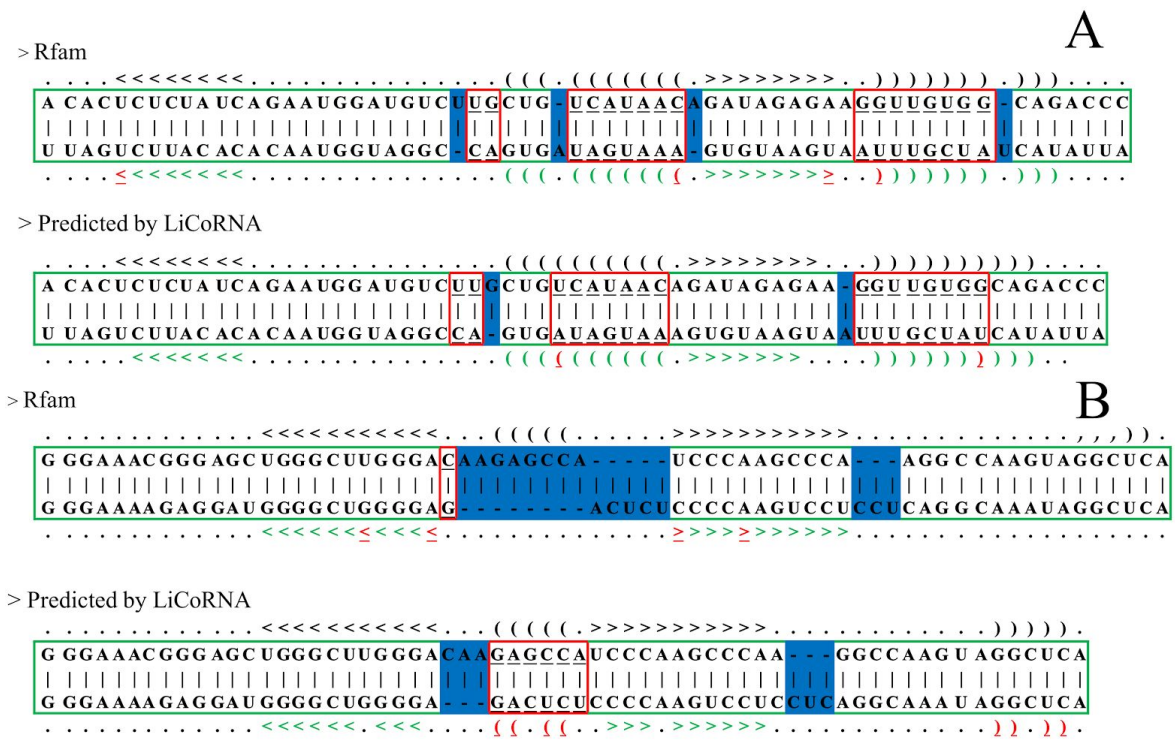

Figure 3. The detailed sequence structural alignment comparison. A) X63538.1/1434-1495 and DQ445912.1/27322-27383 in Corona_pk3 family. B) AAWR02040610.1/2027-2086 and ACTA01044722.1/650-709 in RF_site5 family. The green blocks indicate the same alignment, red blocks with underline represent different alignment and blue blocks denote gaps. Green brackets indicate the same structures while red brackets with underline denote different structures for target sequence.

Conclusion

In this work, we introduced the LiCoRNA software (aLignment of Complex RNAs), a program that gives a sequence-structure alignment for two RNAs in the presence of arbitrary pseudoknots. The program is based on a tree decomposition of query sequences and a general parameterized algorithm to compute the optimal alignment. Notably, and contrarily to other state-of-the-art programs, LiCoRNA supports any type of pseudoknotted structure as the query. Besides use-cases mentioned in this abstract, interesting applications would include scanning for homologs of a single structured RNA sequence within whole genomes, possibly from unassembled NGS data. Besides, by incorporating the scoring function (Stombaugh et al. 2009) using all canonical and non-canonical base pairs, we could extend our algorithm to 3D structure alignment.

References

Bodlaender, Hans Leo, and Others. 1986. *Classes of Graphs with Bounded Tree-Width*. Department of Computer Science, University of Utrecht.

Bodlaender, Hans L., and Arie M. C. A. Koster. 2010. “Treewidth Computations I. Upper Bounds.” *Information and Computation* 208 (3). Elsevier: 259–75.

———. 2011. “Treewidth Computations II. Lower Bounds.” *Information and Computation* 209 (7). Elsevier: 1103–19.

Cech, T. R. 1985. “Self-Splicing RNA: Implications for Evolution.” *International Review of Cytology* 93. Elsevier: 3–22.

Eddy, Sean R. 2001. “Non–coding RNA Genes and the Modern RNA World.” *Nature Reviews. Genetics* 2 (12). Nature Publishing Group: 919–29.

Han, B., B. Dost, and V. Bafna. 2008. “Structural Alignment of Pseudoknotted RNA.” *Journal of Computational*. online.liebertpub.com. http://online.liebertpub.com/doi/abs/10.1089/cmb.2007.0214.

Jabbari, Hosna, Anne Condon, Ana Pop, Cristina Pop, and Yinglei Zhao. 2007. “HFold: RNA Pseudoknotted Secondary Structure Prediction Using Hierarchical Folding.” In *Algorithms in Bioinformatics*, 323–34. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Klein, R. J., and S. R. Eddy. 2003. “RSEARCH: Finding Homologs of Single Structured RNA Sequences.” *BMC Bioinformatics*. bmcbioinformatics.biomedcentral. .... http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-4-44.

Lyngsø, Rune B., and Christian N. S. Pedersen. 2000. “Pseudoknots in RNA Secondary Structures.” In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 201–9. RECOMB ’00. New York, NY, USA: ACM.

Matsui, H., K. Sato, and Y. Sakakibara. 2004. “Pair Stochastic Tree Adjoining Grammars for Aligning and Predicting Pseudoknot RNA Structures.” *Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference*. ieeexplore.ieee.org. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1332442.

Nawrocki, E. P., S. W. Burge, and A. Bateman. 2014. “Rfam 12.0: Updates to the RNA Families Database.” *Nucleic Acids*. Oxford Univ Press. http://nar.oxfordjournals.org/content/early/2014/11/11/nar.gku1063.short.

Nawrocki, E. P., and S. R. Eddy. 2013. “Infernal 1.1: 100-Fold Faster RNA Homology Searches.” *Bioinformatics* . Oxford Univ Press. https://bioinformatics.oxfordjournals.org/content/29/22/2933.full.

Rinaudo, Philippe, Yann Ponty, Dominique Barth, and Alain Denise. 2012. “Tree Decomposition and Parameterized Algorithms for RNA Structure-Sequence Alignment Including Tertiary Interactions and Pseudoknots.” In *Algorithms in Bioinformatics*, 149–64. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Rivas, E., and S. R. Eddy. 1999. “A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots.” *Journal of Molecular Biology* 285 (5). Elsevier: 2053–68.

Sleutels, Frank, Ronald Zwart, and Denise P. Barlow. 2002. “The Non-Coding Air RNA Is Required for Silencing Autosomal Imprinted Genes.” *Nature* 415 (6873). nature.com: 810–13.

Stombaugh, Jesse, Craig L. Zirbel, Eric Westhof, and Neocles B. Leontis. 2009. “Frequency and Isostericity of RNA Base Pairs.” *Nucleic Acids Research* 37 (7): 2294–2312.

Taufer, M., A. Licon, R. Araiza, and D. Mireles. 2009. “PseudoBase++: An Extension of PseudoBase for Easy Searching, Formatting and Visualization of Pseudoknots.” *Nucleic Acids*. Oxford Univ Press. https://nar.oxfordjournals.org/content/37/suppl_1/D127.full.

Uzilov, Andrew V., Joshua M. Keegan, and David H. Mathews. 2006. “Detection of Non-Coding RNAs on the Basis of Predicted Secondary Structure Formation Free Energy Change.” *BMC Bioinformatics* 7 (March): 173.

van Dijk, T., J. P. van den Heuvel, and W. Slob. 2006. “Computing Treewidth with LibTW.” Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.1128&rep=rep1&type=pdf.

Willard, H. F., and H. K. Salz. 1997. “Remodelling Chromatin with RNA.” *Nature* 386 (6622). cat.inist.fr: 228–29.

Wilm, Andreas, Indra Mainz, and Gerhard Steger. 2006. “An Enhanced RNA Alignment Benchmark for Sequence Alignment Programs.” *Algorithms for Molecular Biology: AMB* 1 (October). almob.biomedcentral.com: 19.

Yoon, B. J., and P. P. Vaidyanathan. 2008. “Structural Alignment of RNAs Using Profile-csHMMs and Its Application to RNA Homology Search: Overview and New Results.” *Automatic Control,*

IEEE. ieeexplore.ieee.org. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4439827.
Zimmerman, R. A., and A. E. Dahlberg. 1996. “Ribosomal RNA.” CRC Press, Boca Raton.