Detecting Conserved RNA Secondary Structures in Viral Genomes: The RADAR Approach

Mugdha Khaladkar and Jason T.L. Wang

Bioinformatics Program and Department of Computer Science
New Jersey Institute of Technology, Newark, NJ 07102, USA
jason.t.wang@njit.edu

Abstract. Conserved regions, or motifs, present among RNA secondary structures serve as a useful indicator for predicting the functionality of the RNA molecules. Automated detection or discovery of these conserved regions is emerging as an important research topic in health and disease informatics. In this short paper we present a new approach for detecting conserved regions in RNA secondary structures by the use of constrained alignment and apply the approach to finding structural motifs in some viral genomes. Our experimental results show that the proposed approach is capable of efficiently detecting conserved regions in the viral genomes and is comparable to existing methods. We implement our constrained structure alignment algorithm into a web server, called RADAR. This web server is fully operational and accessible on the Internet at http://datalab.njit.edu/biodata/rna/RSmatch/server.htm.

1 Introduction

RNA molecules play various roles in the cell [1-3]. Their functionality depends not only on the sequence information but to a large extent on their secondary structures. It would be more cost effective if one were able to determine RNA structure by computational means rather than by using biochemical methods. So, the development of computational predictive approaches of RNA structure is essential [4-7]. RNA structure prediction is usually based on the thermodynamics of RNA folding [8-10] or phylogenetic conservation of base-paired regions [11-14].

In this short paper, we present a methodology for the detection of conserved regions, or motifs, in RNA secondary structures. We adopt a comparative approach by carrying out RNA structure alignment. The alignment has been designed so as to be constrained based upon the properties of the RNA structure. Our constrained alignment algorithm is an upward extension of our previously developed RSmatch method for comparing RNA secondary structures, which did not consider constraints in the structures being aligned [7]. We have implemented the constrained structure alignment algorithm into a web server, called RADAR (acronym for *R*NA *D*ata *A*nalysis and *R*esearch) [15]. This web server provides multiple capabilities for RNA structure alignment data analysis, which includes pairwise structure alignment, multiple structure alignment, constrained structure alignment and consensus structure prediction. Our aim behind developing this web server is to develop a versatile tool that provides a computationally efficient platform for performing several tasks related to RNA

D. Zeng et al. (Eds.): BioSurveillance 2007, LNCS 4506, pp. 222–227, 2007.

secondary structure. RADAR has been developed using Perl-CGI and Java. The web server has a user friendly interface that is easy to understand even for novice users.

The rest of this short paper is organized as follows. Section 2 presents the constrained structure alignment algorithm. Section 3 reports implementation efforts, showing some features of the RADAR server, and describes experimental results. Section 4 concludes the paper and points out some directions for future research.

2 Constrained Structure Alignment

In practice, biologists favor integrating their knowledge about conserved regions into the alignment process to obtain biologically more meaningful similarity scores between RNAs. This motivates us to develop algorithms for constrained structure alignment (CSA). Our CSA method is developed by modifying the recurrence formulas of the dynamic programming algorithm used in RSmatch [7] to take into account the constraints, or conserved regions, in the structures being aligned. The time complexity of RSmatch is $O(mn)$, where m and n are the sizes of the two RNA structures respectively that are being aligned. The CSA method has the same time complexity. In practice, the CSA method can align two RNA structures in less than two seconds in wall clock time.

2.1 Method

The input of the CSA method is a query RNA structure and a database of RNA secondary structures. The method constructs the alignment between the query structure and each database structure based upon the knowledge of conserved regions in the query structure. The alignment score is dynamically varied so as to utilize the information of conserved regions. The alignment computed this way is able to detect structural similarity more accurately. The method comprises two main parts:

(I) Annotating a region in the query RNA structure as conserved
Each position of the conserved region in the query RNA structure is marked using a special character "*" underneath the position. This is termed as *binary conservation* since any position in the query RNA structure is treated to be either 100% conserved (if it is marked with "*") or not conserved at all. If it is found, from wet lab experiments or other sources, that a particular RNA structure contains a motif that we want to search for in other RNA structures in a database, then that particular RNA structure can be used as a query structure and that motif region can be marked to be conserved in the query structure. Under this circumstance, the user can adopt binary 0/1 conservation.

Another technique of applying constraints to structure alignment is using the concept of sequence logos [16]. First, the multiple sequence alignment of the RNA sequences under analysis is obtained using, for example, ClustalW (http://www.ebi.ac.uk/clustalw/). Then the frequency of each base at each position of an RNA sequence is calculated, denoted by $f(b,l)$. Using this information the uncertainty at each position is calculated by $H(l) = - \sum f(b,l) \log_2 f(b,l)$, where $H(l)$ is the uncertainty at position l, b is one of the bases (A,C,G,U), and $f(b,l)$ is the frequency of base b at position *l*. The information content at each position is represented by the decrease in uncertainty as the binding site is located (or aligned), i.e. $R_{sequence}(l) = 2 - H(l)$, where $R_{sequence}(l)$ is the amount of information present at position *l* and 2 is the maximum uncertainty at any given position. Thus, the information content is a value between 0 and 2. We scale down this value to get a value between 0 and 1 which is then converted to the percentage conservation at each position of the RNA sequence and its secondary structure.

(II) Utilization of information about the conserved region

Two cases occur as we compute the alignment score between the query structure and a database structure where the query structure contains marked conserved regions.

1. *Comparison between non-conserved regions:* In this case the score assigned is the regular score that is derived from the scoring matrix used by RSmatch.
2. *Comparison involving conserved regions:* Here, we multiply the score obtained from the scoring matrix used by RSmatch by a factor λ that will cause the score to either increase or decrease by the λ value. This factor λ is determined by the type of conservation as discussed in more detail in the subsection on "Scoring Scheme".

2.2 Scoring Scheme

The factor by which the score should get magnified or diminished to take into account the conserved region is decided based upon the following: (i) the length of the conserved region; (ii) the length of the whole RNA sequence; (iii) the type of conservation that has been indicated; and (iv) any special conditions/preferences decided by the user.

In the default scenario, where knowledge about conservation is not used, the score is directly taken from the scoring matrix employed in RSmatch. For the binary conservation case, the default value for the factor λ is $\lambda = 2 - L/N$ where *L* is the length of the conserved region and *N* is the length of the whole RNA sequence. This ratio is then subtracted from a constant value (2, arbitrarily chosen) so that the bonus/penalty is inversely proportional to the length of the conserved region. If the conservation information is spread over 0-100%, as described earlier, these percentage values are passed along with the query RNA structure to the scoring engine and the alignment score varies based on these values.

3 Implementation and Experimental Results

We have implemented the proposed constrained structure alignment algorithm into the RADAR server. This web server together with a standalone downloadable version is freely available at http://datalab.njit.edu/biodata/rna/RSmatch/server.htm. RADAR accepts, as input data, either RNA sequences in the standard FASTA format or RNA secondary structures represented by the Vienna style Dot Bracket format [8]. For performing the constrained structure alignment between a query RNA structure and a database structure, we require the user to annotate the query RNA structure to indicate

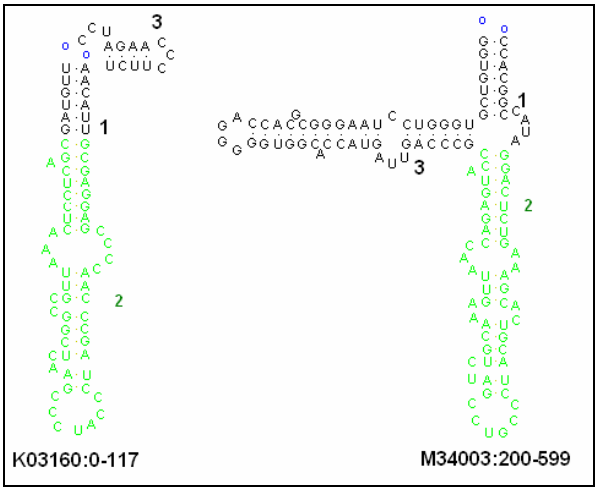

Fig. 1. An example showing the common region of two RNA secondary structures where the local matches in the two structures are highlighted with the green color

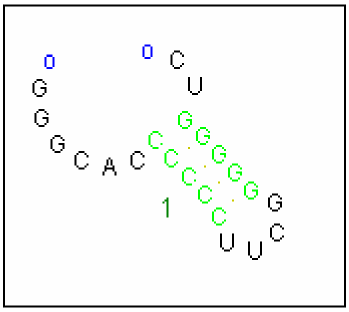

Fig. 2. The structure of a GC–rich hairpin that has been found to be conserved in the *Leviviridae* family [18]

which region is conserved by marking the region with "*". Thus, the input consists of an annotated query RNA structure and a database of RNA structures against which the constrained alignment is carried out. Upon completion of a constrained alignment job, RADAR presents the alignment results on a web page. In Figure 1, the common region of two RNA secondary structures given in an alignment result is portrayed using RnaViz [17], where the local matches in the two structures are highlighted with the green color.

We have conducted several experiments to evaluate the performance of the proposed constrained structure alignment algorithm by applying this method to finding structural motifs in viral genomes. Study of viral genomes has shown that they often contain functionally active RNA structural motifs that play an important role in the different stages of the life cycle of the virus [18]. Detection of such motifs or conserved regions would greatly assist the study of these viruses.

One of our experiments was to search for a short GC–rich hairpin (tetraloop) which follows an unpaired GGG element, shown in Figure 2, present at the 5' end of the *Levivirus* genome [18]. Here we applied the proposed constrained structure alignment algorithm, with the binary conservation option, to a dataset comprising 6838 RNA structures each with length 200 nt formed from 10 *Levivirus* genomes and 4 other viral genomes. The query structure used was the GC–rich hairpin. There were 10 structures in this dataset containing the region of interest. Our algorithm was able to correctly identify 8 out of the 10 structures. The same experiment was repeated using the non-constrained alignment method given in RSmatch, which identified 6 out of the 10 structures. These 6 structures were part of the 8 structures found by the constrained structure alignment (CSA) algorithm. This shows that the CSA method improves upon the performance of the existing RSmatch method and has a better sensitivity. We also applied the Infernal tool (http://infernal.janelia.org/) [19] to this same viral genome dataset. Infernal identified 6 out of the 10 structures. Again, these 6 structures were part of the 8 structures found by the CSA method.

4 Conclusion

The proposed constrained structure alignment algorithm is an upward extension of our previously developed RSmatch method. This new algorithm allows the user to annotate some region in a query structure as conserved so as to produce biologically more meaningful similarity scores. We have implemented this algorithm into the RADAR server accessible on the Internet. The application of this algorithm to viral genomes demonstrates the use of the algorithm in RNA informatics research and its ability to detect conserved regions in RNA secondary structures. Other functions of RADAR, not described here due to space limitations but can be accessed on the Internet, include multiple structure alignment and the prediction of the consensus structure for a set of RNA sequences.

The work presented here is part of a large project aiming to build a cyber infrastructure (http://datalab.njit.edu/bioinfo) for RNA structural motif discovery in human, virus and trypanosome mRNAs. Human immunodeficiency virus type 1 is the causative agent of AIDS and is related to many cancers in humans. Hepatitis C virus is related to hepatocellular cancer in humans. *Trypanosoma brucei* causes African trypanosomaiasis, or sleeping sickness, in humans and animals in Africa. RNA motifs or conserved structures have been shown to play various roles in post-transcriptional control including mRNA translation, mRNA stability, and gene regulation, among others. This cyber infrastructure will contribute to integrated genomic and epidemiological data analysis, by enabling access, retrieval, comparison, analysis, and discovery of biologically significant RNA motifs through the Internet as well as the integration of these motifs with online biomedical ontologies.

Acknowledgements

We thank the anonymous reviewers whose comments help to improve the presentation and quality of this paper.

References

1. Wang, J.T.L., Zaki, M.J., Toivonen, H.T.T. and Shasha, D., eds. (2005) *Data Mining in Bioinformatics*. Springer, New York.
2. Wang, J.T.L., Wu, C.H. and Wang, P.P., eds. (2003) *Computational Biology and Genome Informatics*. World Scientific, Singapore.
3. Wang, J.T.L., Shapiro, B.A. and Shasha, D., eds. (1999) *Pattern Discovery in Biomolecular Data: Tools, Techniques and Applications*. Oxford University Press, New York.
4. Wang, J.T.L., Shapiro, B.A., Shasha, D., Zhang, K. and Chang, C.-Y. (1996) Automated discovery of active motifs in multiple RNA secondary structures. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 70-75.
5. Wang, J.T.L. and Wu, X. (2006) Kernel design for RNA classification using support vector machines. *International Journal of Data Mining and Bioinformatics*, **1**, 57-76.
6. Bindewald, E. and Shapiro, B.A. (2006) RNA secondary structure prediction from sequence alignments using a network of *k*-nearest neighbor classifiers. *RNA,* **12**, 342-352.
7. Liu, J., Wang, J.T.L., Hu, J. and Tian, B. (2005) A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics*, **6**, 89.
8. Hofacker, I.L. (2003) RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429-3431.
9. Schuster, P., Fontana, W., Stadler, P.F. and Hofacker, I.L. (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Roy. Soc. (London) B*, **255**, 279-284.
10. Zuker, M. (1989) Computer prediction of RNA structure. *Methods Enzymol.*, **180**, 262-288.
11. Akmaev, V.R., Kelley, S.T. and Stormo, G.D. (2000) Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics*, **16,** 501-512.
12. Gulko, B. and Haussler, D. (1996) Using multiple alignments and phylogenetic trees to detect RNA secondary structure. In *Proceedings of the 1st Pacific Symposium on Biocomputing*, pp. 350-367.
13. Knudsen, B. and Hein J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423-3428.
14. Rivas, E. and Eddy, S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053-2068.
15. Khaladkar, M., Bellofatto, V., Wang, J.T.L., Tian, B. and Zhang, K. (2006) RADAR: an interactive web-based toolkit for RNA data analysis and research. In *Proceedings of the 6th IEEE Symposium on Bioinformatics and Bioengineering*, pp. 209-212.
16. Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*., **18,** 6097-6100.
17. Rijk, P.D., Wuyts, J. and Wachter, R.D. (2003) RnaViz2: an improved representation of RNA secondary structure. *Bioinformatics,* **19**, 299-300.
18. Hofacker, I.L., Stadler, P.F. and Stocsits, R.R. (2004) Conserved RNA secondary structures in viral genomes: a survey. *Bioinformatics*, **20**, 149.
19. Eddy, S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.