Benchmarking RNA secondary structure comparison algorithms*

Julien Allali[1†], Yves d'Aubenton-Carafa[2], Cédric Chauve[5], Alain Denise[4,6], Christine Drevet[6], Pascal Ferraro[1], Daniel Gautheret[6], Claire Herrbach[4,6], Fabrice Leclerc[8], Antoine de Monte[7], Aida Ouangraoua[1], Marie-France Sagot[3], Cédric Saule[4], Michel Termier[6], Claude Thermes[2], Hélène Touzet[7]

[1] LaBRI, UMR 5800, Université de Bordeaux I, 351, cours de la Libération, F-33405 Talence cedex, France
[2] Centre de Génétique Moléculaire, CNRS, UPR 2167, 91198 Gif-sur-Yvette Cedex, France
[3] Inria Rhône-Alpes and LBBE UMR 5558, Université Claude Bernard, Lyon I, 69622 Villeurbanne, France
[4] LRI, UMR 8623, Université Paris-Sud 11 et CNRS, Orsay, France
[5] Department of Mathematics, Simon Fraser University, Burnaby (BC) Canada
[6] IGM, UMR 8621, Université Paris-Sud 11 et CNRS, Orsay, France
[7] LIFL, UMR 8022, Université Lille 1 and INRIA, France
[8] MAEM, Université Henti Poincaré et CNRS, Nancy, France

Abstract: *Since the last ten years several tools have been proposed for RNA secondary structure pairwise comparison. These tools use different models (ordered tree or forest, arc annotated sequence, multi-level tree) and methods (edit distance, alignment). We present a first benchmark on these tools. For various RNA families, we built two sets of secondary structures. The first, called the reference set, is composed of a small number of RNAs with their known structures. The second is composed of sequences folded using Mfold and RNAshapes. Some of these sequences correspond to structural RNAs of the same families (true events), other correspond to noise. We study the ability of each tool to find the true events using the reference set. In particular we focus on the results in term of sensibility/specificity, on the spreading of the scores and on the time of computation.*

Keywords: RNA, secondary structure, comparison, benchmark

We propose a first evaluation of all available algorithms programs for RNA secondary structure pairwise comparison. Our benchmark is based on the 'real-life' problem of being able to distinguish sequences of non coding RNAs of a given family from other sequences. Typically, this problem occurs when one aims to find, in a genome or a set of sequences, non-coding RNAs that are similar to some given RNAs from a same family. Another application is the automatic classification of a whole set of putative non-coding RNAs into different functional families. We study the ability for each program to distinguish alignments within a same family from alignments between two different families. Of course, the quality of a comparison depends on the ability for the folding tool to find the correct secondary structure. Thus the comparison programs are been tested in presence of 'real-life' noisy data.

We compare six tools: RNAforester [2], MiGaL [1], TreeMatching [3], gardenia [4], NestedAlign [5], and RNAStrAT [8]. These tools use different models of secondary structure. We also include BLAST.

* This work was supported by ANR project BRASERO ANR-06-BLAN-0045.

† corresponding author : allali@labri.fr

RNAforester is a ordered trees local/global alignment algorithm. It uses a special tree encoding that allows the breaking of nucleotide links under certain conditions.

MiGaL use a multi-level representation of the secondary structure composed by four layers coded by rooted ordered trees. These layers model structure from multiloop network to nucleotides. The algorithm used is an adapted edit distance applied successively to each layer.

TreeMatching is based on a quotiented tree representation of the secondary structure which is an auto-similar structure composed of two rooted ordered trees on two different scales (nucleotides and structural elements). The core of the method relies on the comparison of both scales simultaneously: it computes an edit distance between quotiented trees at the macroscopic scale using edit costs defined as edit distances between subtrees at the microscopic scale.

gardenia and NestdAlign use an arc-annotated based representation, that allows for complex edit operations, such as arc-breaking or arc-altering. They allow local and global alignment features. gardenia notably allows affine gap scores. NestedAlign notably implements an original local alignment algorithm.

RNAStrAT makes the comparison in two steps. First, it compares stems of the two structures using an alignment algorithm with complex edit operations. Then it finds an optimal mapping between the different stems.

We applied the following benchmarking protocol on several families of structural RNAs (tRNA, SRP, ...). For each family we use about five references. These references are secondary structure obtained on reputed databases. Then we take about 75 sequences of RNAs that are known to be of that family (true events). We add 200 sequences that are randomly peek into viral genomes (false events). All sequences (true and false) are folded using MFold (keeping optimal and suboptimal) and RNAshapes, this defines the data set. For each tool, we compare each structures of a same sequence of the data set with all structure of the reference set. We store the best score obtained. At the end, we sort the sequences of the data set according to the score obtained. This give us a array of true and false events. Using this array, we draw the ROC curve, that is the sensibility ($\frac{nb\ of\ true\ positive}{nb\ true\ event}$) and specificity ($\frac{nb\ of\ false\ negative}{nb\ false\ event}$) curve. We also represent the spreading of the scores and the computation times.

With this work, we propose to help a user to decide which tool is the most adapted to his situation (time performance, sequence length, ...). This benchmark is the first one of a set of benchmarks on RNA secondary structure pairwise comparison tools called BRASERO.

References

[1] J. Allali and M-F. Sagot, A multiple layer model to compare RNA secondary structures *Software: Practice and Experience*, to appear.

[2] M. Höchsmann, T. Töller, R. Giegerich, S. Kurtz, Local Similarity in RNA Secondary Structures, *Proceedings of the IEEE Bioinformatics Conference 2003* pp 159-168, 2003.

[3] A. Ouangraoua, P. Ferraro, L. Tichit, S. Dulucq, Local similarity between quotiented ordered trees, *Journal of Discrete Algorithms* 1(5):23-35, 2007.

[4] G. Blin and H. Touzet. How to compare arc-annotated sequences: The alignment hierarchy. In SPIRE, *Lecture Notes in Computer Science* 4209, pp 291-303, 2006.

[5] C. Herrbach, Etude algorithmique et statistique de la comparaison de structures secondaires d'ARN, *Université de Bordeaux 1*, 26 Sept. 2007.

[6] D.H. Mathews, J. Sabina, M. Zuker, D.H. Turner Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure, *J. Mol. Biol.*, vol. 288,pp 911-940, 1999.

[7] P. Steffen, B. Voß, M. Rehmsmeier, J. Reeder, R. Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes, *Bioinformatics*, 22(4), 2006.

[8] V. Guignon, C. Chauve , S. Hamel, An edit distance between RNA stem-loops. *SPIRE*, LNCS 3772:334-345, 2005.