Benchmarking RNA secondary structures 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, Bordeaux [2]CGM, Gif s/Yvette [3]INRIA, LBBE, Lyon [4]LRI, Orsay [5]Univ. S. Fraser, Canada [6]IGM, Orsay [7]LIFL, Lille [8]MAEM, Nancy

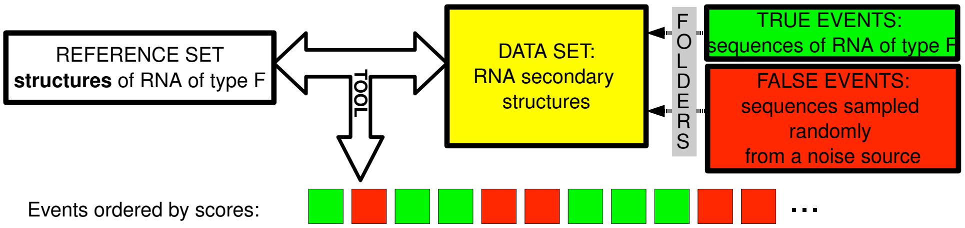
This work was supported by the ANR project BRASERO ANR-06-BLAN-0045.

Abstract:

In 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 online benchmark for comparing 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), others correspond to noise. We studied the ability of each tool to find the true events using the reference set.

Protocol:

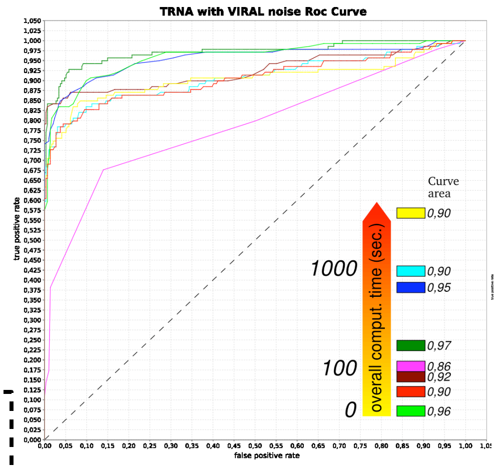
For each run, two sets of RNA secondary structures are built. The reference set is composed of 4 to 6 RNAs of a same family using the structures provided in the literature. The data set is composed of structures obtained by folding sequences of RNAs of the same family (true events) and sequences of the same length as the references but supposed not to belong to that family (called noise or false events).

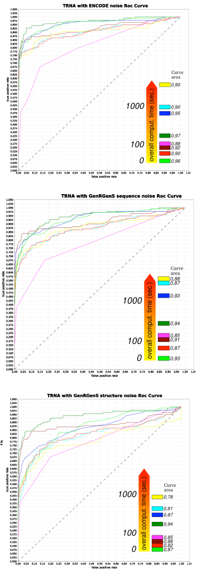


For each event, the best score obtained between the references and all its possible structures (optimal and suboptimal structures found by mfold or rnashape) is retained. Given all events sorted by their best scores, a ROC curve (False Positive Rate; True Positive Rate) is plotted.

tRNA

Nb references	6
Average references length	70,6
min ref. length	67
max ref. length	73
Nb True Events (TE)	139
Average TE Length	73
Min TE length	30
Max TE length	111
Nb False Events	300
Data Set Size	>5000





·Four different sources of noise are used:
·Encode II sequences (10 800 overall comparisons)
Viral genomes (10 776)
RNA structures generated using *GenRGenS* (7 152)
GenRGenS RNA sequences refolded as for true event sequences (16 944)

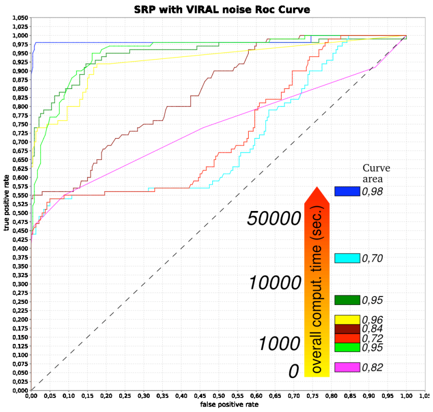
We observe that the results are mostly the same independently of the noise source.

In the case of GenRGenS[1], since the noise is built in such a way that the structures look like tRNAs, the various tools have more difficulty in distinguishing true tRNAs from noise.

[1] Y. Ponty, M. Termier and A. Denise
GenRGenS: Software for generating random genomic sequences and structures, *Bioinformatics*, 2006

SRP:

Nb references	4
Average references length	368
min ref. length	302
max ref. length	523
Nb True Events (TE)	100
Average TE Length	218
Min TE length	74
Max TE length	534
Nb False Events	300
Data Set Size	>1400



Conclusion:

We present a general protocol to evaluate the scoring capabilities of methods for comparing RNA secondary structures.

In particular, the data and the software used in this benchmark are freely available at: http://brasero.labri.fr.

These results are preliminary since most of these tools were run with their default parameters. Hence, this work represents only a starting point for a general benchmark. The impact of the various parameters common or specific to each tool (scoring function, matrices...) will be studied in future.

Currently, we considered data sets for three families: tRNA, SRP and 16S. We will add Intron Group I and II, RNAseP and 23S to the final benchmark.

Finally, another benchmark will be added to analyse the quality of the RNA alignments provided by the methods.

Tools:

RNAforester [1] is an ordered trees local/global alignment algorithm. It uses a special tree encoding that allows to break nucleotide pairings under certain conditions.

MiGaL [2] uses a multi-level representation of the secondary structure composed by four layers coded by rooted ordered trees. The layers model different structural levels from multiloop network to the sequence of nucleotides composing the RNA. The algorithm is an adapted edit distance successively applied to each layer. *(options: -M --hairpin-strict --indel-once)*

TreeMatching [3] is based on a quotiented tree representation of the secondary structure which is a similar structure made of two rooted ordered trees at 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 [4] and **NestedAlign** [5] 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 while NestedAlign implements an original local alignment algorithm.

RNAStrAT[6] performs 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.

RNAdistance[7] implements a classical edit distance on a tree representation of the structure. A particularity of RNAdistance is that it does not take into account the RNA sequence.

We also compute the score using **blast** [8] (bl2seq -t blastn -W 4).

[1] M. Höchsmann, T. Töller, R. Giegerich, S. Kurtz
Local Similarity in RNA Secondary Structures,
Proceedings of the IEEE Bioinformatics Conference 2003

[2] J. Allali and M-F. Sagot
A multiple layer model to compare RNA secondary structures
Software: Practice and Experience 2007 (online)

[3] A. Ouangraoua, P. Ferraro, L. Tichit, S. Dulucq
Local similarity between quotiented ordered trees,
Journal of Discrete Algorithms 2007

[4] G. Blin and H. Touzet
How to compare arc-annotated sequences: The alignment hierarchy. *SPIRE* 2006

[5]C. Herrbach
Etude algorithmique et statistique de la comparaison de structures secondaires d'ARN. *Thesis* 2007

[6] V. Guignon, C. Chauve , S. Hamel
An edit distance between RNA stem-loops. *SPIRE* 2005

[7] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster
Fast Folding and Comparison of RNA Secondary Structures
Monatshefte f. Chemie 1994

[8] S.F. Altschul, W. Gish, W. Miller, E. W. Myers,D. J. Lipman
Basic local alignment search tool
Journal of Molecular Biologie 1990.