Quantifying the relationship between RNA sequence and three-dimensional structure conservation for homology detection

Emidio Capriotti* & Marc A. Marti-Renom[1]

1. Structural Genomics Unit, Bioinformatics and Genomics Department. CIPF, Valencia, Spain.
*now at Department of Bioengineering., Stanford University, Stanford (CA), USA.

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Aim

In recent years, the number of available RNA structures has rapidly grown reflecting the increased interest on RNA biology. Similarly to the studies carried out two decades ago for proteins, which gave the fundamental grounds for developing comparative protein structure prediction methods, we are now able to quantify the relationship between sequence and structure conservation in RNA.

Here we introduce an all-against-all sequence- and three-dimensional (3D) structure-based comparison of a representative set of RNA structures, which have allowed us to quantitatively confirm that: (i) there is a measurable relationship between sequence and structure conservation that weakens for alignments resulting in below 60% sequence identity, (ii) evolution tends to conserve more RNA structure than sequence, and (iii) there is a twilight zone for RNA homology detection.

The computational analysis here presented quantitatively describes the relationship between sequence and structure for RNA molecules and defines a twilight zone region for detecting RNA homology. Our work could represent the theoretical basis and limitations for future developments in comparative RNA 3D structure prediction.

Results

We found that similar to proteins, the structure identity decreased with the decrease of the sequence identity and in agreement with previous work[1] and we observed a higher mean value of percentage of structure identity with respect to the average sequence identity. Using the Infernal program[2], an e-value threshold of ~5x10-4 has been found to be the lower limit of the "twilight zone" for sequence alignment longer than 100 nucleotides. The computational analysis here presented quantitatively describes the relationship between sequence and structure for RNA molecules and defines a twilight zone region for detecting RNA homology. Our work could represent the theoretical basis and limitations for future developments in comparative RNA 3D structure prediction.

Conclusion

Despite the increasing interest on RNA function and the accelerated deposition of RNA structures, there was a gap between the common knowledge and a quantitative analysis of the relationship between sequence and structure conservation in RNA. Here we have addressed this knowledge gap by applying our RNA alignment method[3] to a set of 451 non-identical RNA structures. The relationship we quantified confirms previous studies in ribosomal RNA and could prove useful to assess whether a particular alignment could be reliably used for comparative RNA structure prediction. We have quantitatively shown that: (i) there is an exponential decay relationship between sequence and structure conservation, (ii) evolution tends to conserve more structure than sequence, and (iii) there is a "twilight zone" for RNA homology detection.

Our study provides an initial assessment of the current limits of comparative modeling of RNA structures. We anticipate that our work will aid the development of new methods for RNA structure prediction from sequence. In the near future, it is expected that large-scale comparative modeling of RNA structures will extend opportunities for answering key questions about RNA evolution, such as the origin of RNA as functional molecules. We have estimated that it would be possible to model by comparative modeling segments of approximately one quarter of all RNA sequences in the RFam database. More accurate RNA sequence alignment methods, including those that explicitly use base-pairing information, will be needed to increase the coverage, diversity and accuracy of reliable comparative RNA 3D structure models. Assuming the current growth rate in the number of known RNA structures, comparative modeling will be applicable to a significant number of RNA families within the next few years and thus could play an important bridging role in understanding the atomic mechanisms of RNA folding.

Availability

The datasets of structures and alignments used in this work are available at http://sgu.bioinfo.cipf.es/.

Acknowledgments

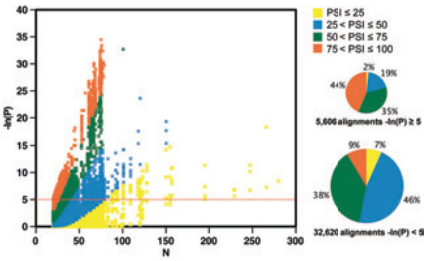
EC acknowledges support from the Marie Curie IOF (PIOF-GA-2009-237225). MAM-R acknowledges support from the Marie Curie IRG (FP6-039722), the Generalitat Valenciana (GV/2007/065), and the Spanish Ministerio de Ciencia e Innovación (BIO2007/66670).

References

1. Abraham et al. (2008). RNA. 14:2274-2289.
2. Nawrocki EP, et al.. (2009). Bioinformatics. 25:1335-1337.
3. Capriotti E, Marti-Renom MA (2008) Current Bioinformatics. 3:32-45.

Structural alignments

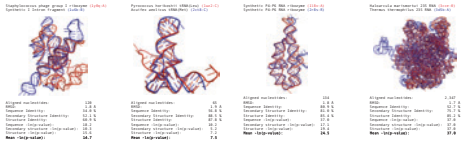
all-against-all comparison of structures in the NR95 set



7

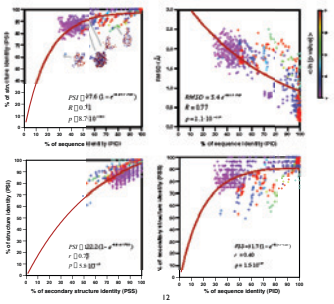
Seq-Str datasets

http://sgu.bioinfo.cipf.es/datasets/



10

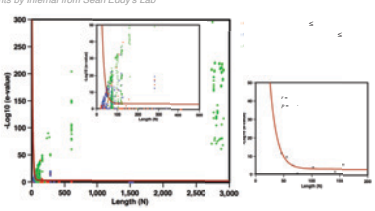
Seq-Str conservation



12

Twilight zone

Alignments by Infernal from Sean Eddy's Lab



CM would result in accurate models for **1/4 of the RFam** database

Infernal e-value < 10^{-4} to a known structure

13