How to detect non-coding RNAs ?

Arnaud Fontaine[1] and Hélène Touzet[1]

Laboratoire d'Informatique Fondamentale de Lille - UMR CNRS 8022 - USTL, France
fontaina, touzet@lifl.fr

Structured non-coding RNAs have a very important functional role in the cell. They are involved in a variety of processes such as transcriptional and post-transcriptional regulation, chromosome replication, RNA processing and modification, protein degradation and translocation. Current research for annotation of eukaryotic genomes show that there is a need for novel tools that are able to identify functional RNA structures in long genomic sequences.

From a computational viewpoint, non-coding RNAs lack a simple statistical signal in their primary sequence, such as codon bias for protein encoding genes. It is not clear whether the measure of the theoretical thermodynamic stability alone is a general distinctive feature [8,1]. This makes it a difficult task to detect novel non-coding RNAs in a genomic sequence. Comparative methods provide a way to cut through the abundance of potential structures: only structures that are conserved along evolution are likely to be biologically significant. This problem of *structure detection* is related to structure inference, but the philosophy is different. Structure prediction requires sufficient sensitivity to predict almost all base pairs. On the contrary, detection of structure in a full genome scan requires much greater specificity, but it can accomodate with lower base pair sensitivity.

We perform a systematic evaluation of four recent programs that are devised for the detection of conserved structural motifs in a set of aligned sequences. All approaches combine comparative analysis, structure prediction and statistical measures of significance of the structure. But the detection strategies are divergent and produce substantially different results. Our goal is to furnish a guideline for potential users.

- `QRNA` [5] is a supervised learning method whose key idea is to test the pattern of substitutions observed in the alignment. A conserved coding region tends to show a pattern of synonymous substitutions, whereas a conserved structural RNA tends to show a pattern of compensatory mutations consistent with some consensus secondary structure. This intuition is implanted with a SCFG for RNA structures and two HMM, the first one modelling sequences constrained by coding sequence evolution, and the latter one modelling a null hypothesis of position-independent evolution. `QRNA` is restrained to pairwise alignments.
- `RNAz` [7] consists of two basic components. It constructs a consensus secondary structure with `RNAalifold` [4], incorporating covariance information into the energy model. This structure is then assigned a measure for thermodynamic stability, which takes into account the minimal free energy of single sequences on one hand, and the distribution of energies for equivalent randomized alignments on the other hand.
- `ddbRNA` [3] uses the same two-step scheme as `RNAz`. The main difference is that the stems of the consensus structure are selected in a greedy fashion. So the computational time is proportional to the square of the sequence length, instead of being cubic, which makes it more appropriate of long sequences scan. The assessment of the significance of the conserved structure is based on shuffled alignments.
- The outline of `MSARI` [2] is similar to `ddbRNA`. The statistical evidence for conservation of common stems employs a distribution-mixture method. It also allows for small variations between positions of complementary base pairs in the alignment. The current version is restrained to alignments composed of at least 10 sequences.

Methodology. We compiled positive data sets, constituted from non-coding RNAs, as well a negative data sets, constituted of similar sequences that are assumed not to share a conserved global structure. The length of the

sequences ranges from 70 nt to 300 nt. The conservation of the primary structure is variable: from 50% to more than 95%. For positive data sets, sequences were retrieved from public databases. We exclude structures that contain thermodynamically stable pseudoknots, since all methods equally fail on such sequences. This makes 20 families of at least 10 sequences. For negative data sets, two kinds of sequences are used. The first ones are generated from positive sequences with random shuffles of multiple alignments, and then re-aligned (as described in [6]). The latter ones are fragments of homologous coding sequences, both eukariotic and prokariotic: 15 families of sequences.

We construct a large number of test sets combining various sizes (2, 3, 5 and 10 sequences) and various alignment programs (Blast and Needleman&Wunsh for pairwise alignments, ClustalW, Dialign2 and T-coffee for all alignments). This gives more than 80,000 alignments. The accuracy of each algorithm is measured through its sensitivity (ratio of positive sequences that are classified as non-coding RNAs) and selectivity (ratio of negative sequences that are not classified as non-coding RNAs).

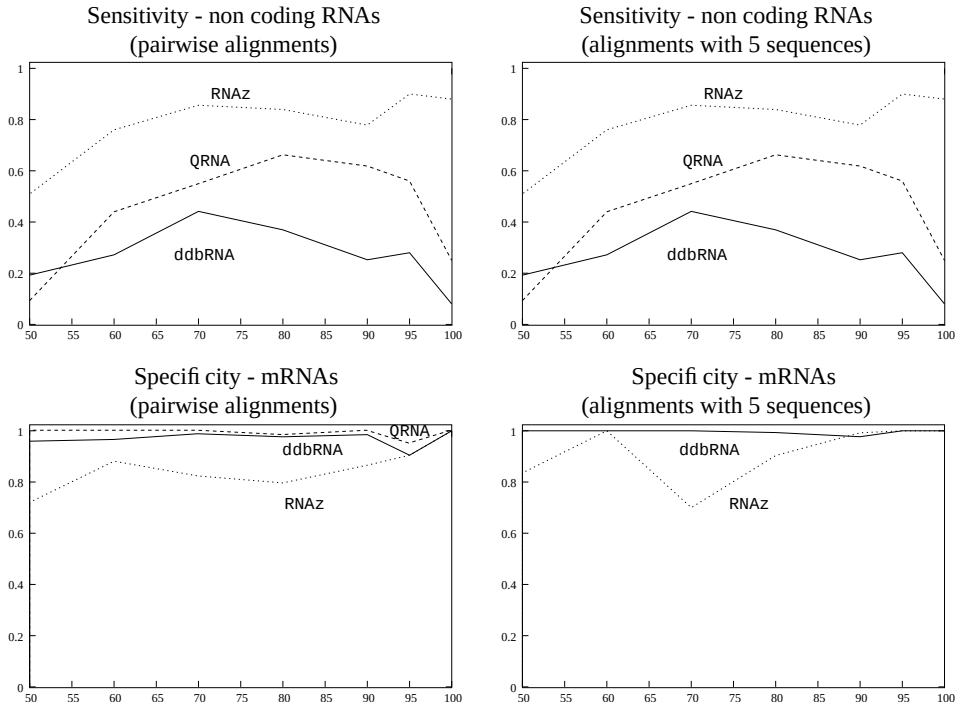

Figure 1. Sensitivity and specifi city for RNAZ, QRNA and ddbRNA.

Conclusions. The first lesson is that the behaviour of the four methods is strongly dependent of the quality of the input alignment. The results are much better when the alignment is generated by ClustalW, which produces few gaps. Another general observation is that sentivity is a main limitation. RNAz outperforms clearly ddbRNA and QRNA. But accuracy may be very low (less than 0.45) when the identity percentage is poor, even for thermodynamically-stable well-conserved structures such as Human microRNAs. Finally, abundance of data may be a major cause of error: neither ddbRNA, nor RNAz are able to detect non-coding RNAs with alignments of 10 sequences. Only MSARI, that is more permissive to search for common stems, manage to detect common

motifs in such large data sets, but the overall accuracy is very low. Concerning the specificity, all methods, except `MSARI`, prove to be excellent with shuffled sequences: the rate of rejection is near 100%. Tests on fragments of coding sequences lead to relatively mediocre results for `RNAz`: this is the price to pay for its higher sensitivity. `QRNA` turns out to be the most selective algorithm, demonstrating that including an auxiliary evolutionary model for coding sequences is a good choice.

Figure 1 shows sensitivity for non-coding RNAs and specificity for coding regions of messenger RNAs. Data sets are composed of two and five sequences, and all alignments are produced by ClustalW and curated so as to remove gaps in the 5' and 3' regions. Results are classified according to the evolutionary distance: the horizontal axis indicates the average identity percentage (from $< 50\%$ to $> 95\%$). Full results and analyses are available at `http://bioinfo.lifl.fr/rna`.

References

[1] P. Clote, E. Ferre, E. Kranakis, and D. Krizanc. Structural RNA has lower folding energy than random rna of the same dinucleotide frequency. *RNA*, 11 (5), 2005.

[2] A. Coventry, D.J. Kleitman, and B. Berger. MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, 101(33):12102–12107, 2004.

[3] D. di Bernardo D, T. Down, and T. Hubbard. ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics*, 19(13):1606–11, 2003.

[4] I.L. Hofacker, M. Fekete, and P.F. Stadler. Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*, pages 1059–1066, 2002.

[5] E. Rivas and S.R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(8), 2001.

[6] S. Washietl and I L. Hofacker. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *Journal of Molecular Biology*, 342:19–30, 2004.

[7] S. Washietl, I.L. Hofacker, and P.F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, 102:2454–2459, 2005.

[8] C. Workman and A. Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Research*, 27(24), 1999.