RNAspace: an integrated environment
for the prediction, annotation and analysis of non-coding RNA

Marie-Josée Cros[1], Antoine de Monte[2], Jérôme Mariette[3], Philippe Bardou[4],
Daniel Gautheret[5], Hélène Touzet[2] and Christine Gaspin[1,3]

[1] INRA, Unité de Biométrie et Intelligence Artificielle, UR 875, F-31320 Castanet, France
[2] LIFL, UMR CNRS 8022 Université Lille 1 and INRIA Lille Nord Europe, France
[3] INRA, Plateforme bioinformatique, F-31320 Castanet, France
[4] INRA, SIGENAE, UMR 444, F-31320 Castanet, France
[5] IGM UMR 8621 CNRS-U Paris sud, France
contact@rnaspace.org

Abstract *RNAspace is an environment that allows to create web sites dedicated to non-protein-coding RNA (ncRNA) prediction, annotation and analysis. The web sites allow users to run a variety of tools in an integrated and flexible way. RNAspace is focused on the integration of complementary ncRNA gene finders. It also offers a set of tools for the comparison, visualization, edition and export of ncRNAs candidates. Predictions can be filtered according to a large set of characteristics.*
A public web site http://rnaspace.org *has been created that allows for on line annotation of a complete bacterial genome or a small eukaryotic chromosome.*

Keywords non-protein-coding RNA, genome annotation, ncRNA gene finder.

The availability of complete genome sequences and the development of high throughput technologies have led to the accumulation of raw biological data at an unprecedented scale. Whereas structural and functional protein annotation is now considered as a task which is relatively well solved, ncRNA genes are not (or at a weak level) integrated in these environments. This fact can be explained by a few reasons which are respectively a recent interest for ncRNA, the absence of general ncRNA prediction methods and the difficulty to analyze these molecules with regard to their sequence and structure conservation. The latter task generally requires an expertise level not widespread and the need to use analysis and edition tools more sophisticated than pure similarity search. The increasing number of ncRNA discovered and the lack of user friendly tools for finding and annotating them, led us to propose to biologists an *in silico* environment allowing structural and functional annotations of these molecules. For this purpose, an environment called RNAspace was developed that allows to install dedicated web sites just by adjusting various global parameters (gene finders to consider, maximal size for input genomic sequences ...).

A web site allows to:

- run a variety of ncRNA gene finders in an integrated environment,
- explore computed results with dedicated tools for comparison, visualization, alignment and edition,
- and export them in various formats (FASTA, GFF, RNAML).

Gene finders are organized into three categories containing respectively :

1] known ncRNA based gene finders including (i) sequence homology search tools: BLAST [1], YASS [2] on ncRNA databases: Rfam [3], fRNAdb [4], miRBase [5], (ii) general purpose ncRNA motif search tools: Infernal [6], Darn [7], Erpin [8], (iii) specialized search tools: RNAmmer for ribosomal RNAs [9], tRNAscan_SE [10] for transfer RNAs;

2] a comparative analysis gene finder: an *ad hoc* pipeline [11] has been implemented based on BLAST or YASS for similarities search and caRNAc [12] or RNAz [13] for consensus structure inference;

3] an *ab initio* gene finder based on detection of atypical GC% regions.

All gene finders can be run with default parameters values. However it is also possible for users, through a dedicated interface, to set some of these parameters to specific values according to the level of knowledge of biological data and user expertize.

Once the execution of selected gene finders is achieved, combination of predictions is possible on demand. For example, predictions that have only tiny differences in positions on the input genomic sequence are merged into a single prediction. This avoids having a lot of redundant predictions for ncRNA families (*e.g.* tRNA) predicted by several gene finders. An overview of all putative ncRNAs found on the genomic sequence is provided. Their main characteristics are displayed in a list that can be dynamically explored by sorting and filtering its content. For each putative ncRNA or a selection of them, more details are computed on line (*e.g.,* compute and visualize a secondary structure, align a selection of predictions ...). Any putative ncRNA could be edited and deleted. It is also possible to visualize putative ncRNAs on the input genomic sequence with several genome browsers: JBrowse, CGview, ApolloRNA. Finally, a functionality allows the export of candidate ncRNAs in several formats. Thus, it is possible to save them in one or several of the proposed formats for a future usage in others contexts (literature …).

The environment relies on collaboratively-developed code using the Python language and the HTTP framework CherryPy (see http://cherrypy.org for more detail). The code is open source under GPL license and is available on Source Forge (https://sourceforge.net/projects/rnaspace/). It has been conceived to be as parameterizable and extensible as possible. This allows to configure web sites for special uses: A site dedicated to a species with limited accesses, a site shared by a group of biologists ... Parametrization of a site includes declaration of available gene-finders, limits for process time execution, disk space, storage duration, execution on a connected computer cluster via a job scheduler. It is also worth to note that the environment could be used in command line and thus inserted in a pipeline.

Using the RNAspace environment, a public site http://rnaspace.org has been created. It accepts genomic sequences up to 5Mb, which permits on line annotation of a complete bacterial genome or a small eukaryotic chromosome. Computations are executed on the computer cluster of the Genotoul bioinformatic platform. For example, it is possible to get an annotation for the *E. coli* genome (4.9M nucleotides) using a wide selection of gene finders and recovering the majority of known RNA genes in less than one hour.

In the near future, we plan to incorporate supplementary prediction approaches, to provide more advanced methods to eliminate redundant results, to include information on the genomic context, to define and compute a common normalized prediction score (indeed some gene finders now provide a score but these scores are not comparable). Furthermore with the huge quantity of high-throughput sequencing data obtained by transcriptome studies, it is also highly desirable to consider RNAseq and sRNAseq data. We will consider handling such large sequence sets of NGS data for the annotation and quantification of non-coding transcripts and the search for potential targets of regulatory RNA acting through RNA-RNA interactions.

Acknowledgements

This work was supported by IBiSA (Infrastructures en Biologie, Santé et Agronomie).

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J Mol Biol 215 (3): 403–410, 1990.

[2] L. Noe, G. Kucherov, YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Research, 33(2), 2005.

[3] P.P. Gardner, J. Daub, J.G. Tate, E.P. Nawrocki, D.L. Kolbe, S. Lindgreen, A.C. Wilkinson, R.D. Finn, S. Griffiths-Jones, S.R. Eddy and A. Bateman, Rfam: updates to the RNA families database. Nucleic Acids Research, 37(Database Issue), 2009.

[4] T. Kin, K. Yamada, G. Terai, H. Okida, Y. Yoshinari, Y. Ono, A. Kojima, Y. Kimura, T. Komori, K. Asai, fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. Nucleic Acids Research, 35 (Database Issue), 2007.

[5] S. Griffiths-Jones, H.K. Saini, S. van Dongen, A.J. Enright, miRBase: tools for microRNA genomics. Nucleic Acids Research, 36 (Database Issue), 2008.

[6] E.P. Nawrocki, D.L. Kolbe, S.R. Eddy, Infernal 1.0: Inference of RNA alignments. Bioinformatics 25(10), 2009.

[7] M. Zytnicki, C. Gaspin, T. Schiex, DARN! A Weighted Constraint Solver for RNA Motif Localization. Constraints, Vol. 13, 2008.

[8] D. Gautheret, A. Lambert, Direct RNA Motif Definition and Identification from Multiple Sequence Alignments using Secondary Structure Profiles. J Mol Biol. 313:1003-11, 2001.

[9] K. Lagesen, P. Hallin, E.A. Rødland, H.-H. Stærfeldt, T. Rognes, D.W. Ussery, RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research, 35(9), 2003.

[10] T.M. Lowe, S.R. Eddy, tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research, 25(5), 1997.

[11] B. Grenier-Boley, A. de Monte, H. Touzet, CG-seq: a toolbox for automatic annotation of genomes by comparative analysis. INRIA Research Report N°7428, available on HAL, Oct. 2010.

[12] H. Touzet, O. Perriquet, CARNAC: folding families of non coding RNAs. Nucleic Acids Research, 142(Web Server Issue), 2004.

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