RNA StrAT: RNA Structure Analysis Toolkit

Valentin Guignon[1,*], Cedric Chauve[2,3] and Sylvie Hamel[1]

[1]DIRO, Université de Montréal, Montréal (QC), Canada.

[2]Comparative Genomics Laboratory and LaCIM, Université du Québec à Montréal, Montréal (QC), Canada.

[3]Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada.

Submitted to Bioinformatics. December 12th, 2007.

ABSTRACT

Summary: We present a web server dedicated to the comparison of RNA secondary structures. This server allows to compare pairs of RNA secondary structures and to search, for a given RNA secondary structure, RNA genes with similar structures in a large database of RNA structures. The comparison and search are performed using an edit distance algorithm that considers a wide range of edit operations.

Availability: The website is freely available without registration at http://www-lbit.iro.umontreal.ca/mastrat/

Contact: guignonv@yahoo.fr

1 INTRODUCTION

RNA molecules, and non-coding RNA (ncRNA) in particular, play an important role in several fundamental biological processes (Mattick and Makunin, 2006) and recently hundred of thousand of potential ncRNA genes have been discovered (He *et al.*, 2007). In general, the function of ncRNAs is strongly linked to their three dimensional structure. This 3D structure is hard to determine precisely and many efforts have been devoted to predict an intermediate level of structure: the secondary structure. Computer methods that allow to predict one or several possible secondary structures for a given RNA gene sequence have been developed (Gardner and Giegerich, 2004; Aksay *et al.,* 2007) and are now widely used in high-throughput ncRNA prediction (Pedersen *et al.*, 2006). The number of available RNA secondary structures (either validated or predicted) is increasing dramatically. They are recorded in databases like Rfam (Griffiths-Jones *et al.*, 2005) which contains alignments of RNA sequences with consensus secondary structures for many RNA families.

It then becomes increasingly important to be able, given a new RNA secondary structure, to detect which other known RNA genes have a similar structure. Some servers provide tools to align pairs or sets of RNA secondary structures provided by the user, like RNAForester (Höchsmann *et al.*, 2004), Gardenia (http://bioinfo.lifl.fr/RNA/gardenia/index.php) or Marna (Siebert and Backofen, 2005); as far as we know, Radar (Khaladkar *et al.*, 2007) is the only server that allows to search for structurally similar RNAs in a database of known RNA secondary structures, using the consensus structures of Rfam. Most of these tools perform comparison using an edit-distance approach, but with different sets of edit operations. Among the edit distance models, the most general was defined in (Jiang *et al.,* 2002); it has been shown to be intractable and the most general implementation of an edit distance algorithm that approximates this model (Herrbach *et al.*, 2007*)* is available only in Gardenia.

Our server proposes several tools for the comparison of RNA secondary structures. It includes pairwise comparison of user-provided structures and search for similar structures in the Rfam, database. We use an edit distance model that considers all edit operations defined in (Jiang *et al.,* 2002) but is tractable as we consider it only on stems and stem-loops substructures (see Section 3.1). RNA StrAT is the first server offering both these features (general RNA edit model and search database) together.

2 SERVER OVERVIEW

The website interface is composed of 3 main sections: database, tools, information.

The database section offers a web interface to browse RNA secondary structures, either by RNA family, species, ID (in our database) but also by structural properties ("Subset" subsection). Users can access to structure information including links to its RNA family (in the Rfam classification), its organism taxonomy (EMBL), its sequence (EMBL); structures can also be displayed, as well as RNA families multiple alignments. Structures stored in our database are extracted from the Rfam seed alignments (release 8.1) and for each RNA gene, its specific secondary structure is obtained from both its sequence and the family consensus structure .

The tools section provides several options for comparing RNA secondary structures that will be described in Section 3: comparing pairs of RNA secondary structures, searching a database for similar structures, computing a distance table for a set of RNA secondary structures, several rendering methods for single structures, simple alignments and consensus alignments.

The information section provides documentation and help about the site.

The server was developed using PHP technology to generate pages that are compliant to XHTML 1.0 Transitional, CSS2 and Javascript W3C standards. The site has been tested and should display correctly on the following web browsers: Internet Explorer 7, Firefox 2, Safari 2, Opera 9 and Konqueror 3. The relational database uses a MySQL engine.

3 TOOLS

3.1 Structures comparison

The main feature of RNA StrAT is the pairwise comparison of RNA secondary structures using an edit distance.

The edit distance method we use works in three steps: first the pair of compared structures is broken down into stems and stem-loops, then these basic substructures are compared pairwisely using the edit distance algorithm defined in (Guignon *et al*, 2005), and finally an alignment of the set of stems and stem-loops is performed based on the results of the stems and stem-loops comparisons for the matching scores. Hence structural comparisons are performed only on pairs of stems and/or stem-loops, which allows to consider all edit operations defined in (Jiang *et al.,* 2002) with an efficient algorithm, whose complexity is in practice a little more than quadratic in time and space. This approach is particularly well suited to the comparison of ncRNA structures having only a few stems and stem-loops.

In order to speed-up the database search the search engines first analyze the structural characteristics of the query structure to select a group of candidates in the database that share similar characteristics close to the query ones, eliminating at the same time irrelevant structures. Then, the query structure is compared to these candidates to find which ones have the best similarity scores. The user can modify the parametersthat define the candidates.

RNA StrAT also offers the possibility to compute a pairwise distances table. The user can provide a set of structures to compare and the tool will output a distance table between structures that can then be used to cluster the set of RNA structures.

The cost of edit operations can also be decided by the user. Costs can be specified for each edit operation without taking care of the type of bases involved but it also can be both base-type and operation-type specific.

3.2 Structures rendering

With web browsers that support Javascript, comparison and search results will be sorted by score in a table providing links to the database, to the alignments and also to our rendering engines. The rendering engine (RS^{3}R for **R**NA**S**traT **S**econdary **S**tructure **R**endering engine) can render a single structure, an alignment of 2 structures and also an alignment of a structure against several other structures which we call a “consensus structure alignment”. The engine offers many rendering parameters like stems and stem-loops coloration, various way to represent bases and hydrogen bonds, image output format and resolution. A navigation panel allows the user to zoom in and out and also move displayed region. To our knowledge, no web tool of this kind already exists.

The “consensus structure alignment” rendering is a new way to display statistical informations about a query structure compared to a group of structures. The query structure is rendered with annotation indicating how often a base is deleted or substituted by another and the position of insertions.

3.3 Typical uses

The site can be used in various ways. The full structures and stems and stem-loops search engines can quickly find similar structures or stems and stem-loops in the database. For example, in the design of small interfering RNA, this could be a way to check if other structures share characteristics similar to the one designed and may interact. Database searching can also give hints on what kind of functions could have an unknown RNA of which only the structure is known.

The distance table computation is a useful tool for the clustering and classification of new RNA secondary structures based on structural similarity.

The “consensus structure alignment” rendering allows to point out the specificities of a structure against a group of structures. For example, one could be interested in comparing the secondary structure of a new viral RNA variant against a group of viral RNA of the same kind to find out what is conserved and what is specific to that structure.

4 CONCLUSION AND FUTURE WORK

Our server provides a comprehensive set of tools to compare and study RNA secondary structures through a friendly user interface. Provided tools can help to answer several kinds of problems related to RNA secondary structures, including a database search using an efficient distance algorithm based on a complete set of edit operations.

We plan to add additional features to RNA StrAT in the near future. The first one is the importation of RNA secondary structures from 3D structures stored in PDB files and from the non-seed alignments of Rfam. We also plan to incorporate new tractable edit distance algorithms based on (Jiang *et al.,* 2002) such as (Herrbach *et* al., 2006). These algorithms have a higher complexity but are more precise and could then be useful, especially for comparing small sets of RNA structures.

ACKNOWLEDGEMENTS

This work was supported by a Genome Quebec scholarship to V.G. and NSERC research grants to C.C. and S.H..

REFERENCES

Aksay, C *et al.* (2007) taveRNA: a web suite for RNA algorithms and applications. *Nucleic Acids Res.*, **35**, W325-9.

Gardner, P.P and Giegerich, R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5, 140.

Griffiths-Jones, S *et al.* (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*., **33**, D121-4.

Guignon, V *et al.* (2005) An edit distance between RNA stem-loops. In *SPIRE 2005* vol. 3772 of *Lecture Notes Comput. Sci.*. Springer, Berlin/Heildeberg, pp. 334-45.

He, S (2007) NONCODE v2.0: decoding the non-coding. To appear in *Nucl. Acids Res.*

Herrbach, C *et al.* (2006) Alignment of RNA secondary structures using a full set of operations. LRI Research Report 1451, Université Paris-Sud 11.

Höchsmann, M *et al.* (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 53-62.

Jiang, T *et al.* (2002) A general edit distance between RNA structures. *J. Comput. Biol.*, **9**, 371-88.

Khaladkar, M *et al.* (2007) RADAR: a web server for RNA data analysis and research. *Nucleic Acids Res.*, **35**, W300-4.

Mattick, J.S and Makunin, V (2006) Non-coding RNA. *Hum. Mol. Genet.,* **15**, R17-29.

Pedersen, J.S *et al. (*2006) Identification and classsification of conserved RNA secondary structures in the human genome. *PloS Comput. Biol.*, **2**, e33.

Siebert, S and Backofen, R (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, **21**, 3352-9.