BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 12 2011, pages 1706–1707
doi:10.1093/bioinformatics/btr258

Sequence analysis

Advance Access publication May 5, 2011

Boulder ALignment Editor (ALE): a web-based RNA alignment tool

Jesse Stombaugh[1], Jeremy Widmann[1], Daniel McDonald[1] and Rob Knight[1,2,*]

[1]Department of Chemistry and Biochemistry, University of Colorado and [2]Howard Hughes Medical Institute, Boulder, CO USA

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: The explosion of interest in non-coding RNAs, together with improvements in RNA X-ray crystallography, has led to a rapid increase in RNA structures at atomic resolution from 847 in 2005 to 1900 in 2010. The success of whole-genome sequencing has led to an explosive growth of unaligned homologous sequences. Consequently, there is a compelling and urgent need for user-friendly tools for producing structure-informed RNA alignments. Most alignment software considers the primary sequence alone; some specialized alignment software can also include Watson–Crick base pairs, but none adequately addresses the needs introduced by the rapid influx of both sequence and structural data. Therefore, we have developed the Boulder ALignment Editor (ALE), which is a web-based RNA alignment editor, designed for editing and assessing alignments using structural information. Some features of BoulderALE include the annotation and evaluation of an alignment based on isostericity of Watson–Crick and non-Watson–Crick base pairs, along with the collapsing (horizontally and vertically) of the alignment, while maintaining the ability to edit the alignment.

Availability: http://www.microbio.me/boulderale.

Contact: jesse.stombaugh@colorado.edu

Received on November 20, 2010; revised on March 19, 2011; accepted on April 13, 2011

1 INTRODUCTION

The RNA Alignment Ontology (Brown *et al.*, 2009) provides several key recommendations that are essential for the development of a user-friendly editor of alignments of a few dozen to a few hundred sequences, consisting of a few hundred base pairs. These are: (i) the incorporation of concepts introduced by the RNA Ontology, in particular the Leontis–Westhof classification system for non-Watson–Crick base pairs (Leontis *et al.*, 2002; Leontis and Westhof, 2001; Stombaugh *et al.*, 2009), which are the building blocks of tertiary motifs (Nasalean *et al.*, 2009); (ii) the annotation of specific regions within the structures (e.g. the P4 helix of RNase P), which can be used to support alternative notions of correspondence (sequence level versus structure level), including homology; and (iii) the ability to perform both horizontal and vertical collapsing of the alignment, allowing the user to focus on specific sequences or on specific regions of the alignment. Several additional considerations that are especially useful for curating large databases of structured RNAs such as Rfam (Griffiths-Jones *et al.*, 2003), the RNase P database (Brown, 1999) and the tRNA database (Juhling *et al.*, 2009) are (i) the functionality to dynamically score an alignment based on its ability to preserve features of a known structure, including non-Watson–Crick pairing and to highlight mismatches in a context, where the user can edit the alignment to resolve these mismatches; (ii) to visualize the secondary structure of any sequence within the RNA family based on the consensus secondary structure using standard tools that can be embedded in a web context (Darty *et al.*, 2009); and (iii) to exploit recently discovered compositional preferences in RNA structural regions (Smit *et al.*, 2009) to indicate when an alignment is a plausible representative of a putative secondary structure. Additional desiderata include a high level of interactivity, for example, the ability to dynamically rearrange rows of the alignments to juxtapose relevant groups, and the ability to stretch, rotate and otherwise manipulate the picture of the secondary structure, while keeping the bases aligned.

2 THE BOULDERALE SOFTWARE

BoulderALE is built on the PyCogent toolkit (Knight *et al.*, 2007) and combines these features into a single web application that will greatly assist both in the curation of RNA family databases and in the understanding of novel RNA structures. BoulderALE is available at http://www.microbio.me/boulderale, and source code and unit-tests can be obtained from sourceforge under the GPL (http://sourceforge.net/projects/boulderale). The availability of the source code will allow the developers of other RNA resources to integrate BoulderALE in their own web sites. BoulderALE fully implements several of the ROC recommendations, including the ability to display and evaluate non-Watson–Crick base pairs, annotate structural regions within the RNA interactively [including automatic inference of these annotations from infernal (Nawrocki *et al.*, 2009) covariance models], horizontal and vertical collapsing based on manual choices of sequences or regions, and display and evaluation of secondary structures. Some other considerations include automatically deciding which sequences or regions to collapse, and fully implementing the RNA Alignment Ontology correspondence concepts.

A typical workflow is as follows: first, the alignment is input as a Stockholm or FASTA file. Then, a list of valid base pairs, including non-Watson–Crick base pairs, associated with one reference sequence, is uploaded. A tab-delimited file including locations of regions or motifs can also be uploaded. Alternatively, the list of valid base pairs and features can be stored in the Stockholm file. Using the secondary structure (including non-Watson–Crick base pairs), it is then possible to highlight base pairs in the secondary structure that do or do not match in the alignment, and the user can then edit the alignment to optimize this matching. Base composition metrics can also be produced, and the secondary structure can be

*To whom correspondence should be addressed.

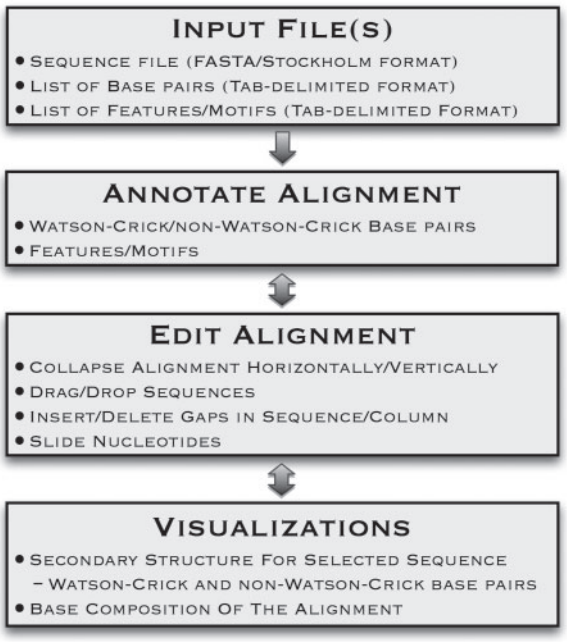

Fig. 1. Illustration of an alignment analysis with BoulderALE. Users upload a Stockholm or FASTA-formatted sequence file along with tab-delimited base pair and motif files. Next, users have options to annotate and edit their alignment, and can produce visualizations to aid in their analysis.

plotted. This workflow is illustrated in Figure 1. Finally, features can be mapped onto the alignment, and BoulderALE allows the alignment to be vertically or horizontally collapsed to focus the user's attention on specific taxa or regions of the sequence. In practice, it is often useful to do this iteratively, cleaning up a particular region of the alignment in each of several closely related groups of sequences, then aligning the groups of sequences to each other to reveal higher-level correspondences that rely more on structure than on sequence.

3 COMPARISONS WITH OTHER SOFTWARE

There are other software packages that offer some overlapping functionality with BoulderALE, but they are targeted for different alignment problems. Jalview (Clamp *et al.*, 2004), although web-embeddable, lacks the ability to incorporate structural data. BioEdit (Hall, 1999), although user-friendly and allowing for Watson–Crick pairing, is restricted to the Windows platform and does not allow for horizontal collapsing. S2S (Jossinet and Westhof, 2005) allows for non-Watson–Crick base pairs; however, many users find its interface conventions counterintuitive, since it was primarily designed for modeling RNA, and it cannot annotate and collapse structural motifs. MultiSeq (Roberts *et al.*, 2006) can do filtering and grouping of redundant sequences, but lacks a representation of non-Watson–Crick base pairs. SARSE (Andersen *et al.*, 2007) and RALEE (Griffiths-Jones, 2005) allow for feature coloring, however; they both lack the ability to annotate non-Watson–Crick basepairing and horizontal collapsing. These examples are intended to be illustrative rather than exhaustive, since there are several sequence alignment editors to choose from, many of which are optimized for specific tasks other than those addressed here.

4 CONCLUSIONS

In conclusion, BoulderALE provides a user-friendly package that allows rapid visualization of RNA sequence alignments that have previously been inaccessible, especially through the collapsing of features that rapidly focus the user's attention on specific parts of the alignment, while highlighting features allow users to identify specific sequences or regions that require manual cleanup. We believe BoulderALE will thus assist users in dealing with the flood of structural and sequence data now becoming available.

ACKNOWLEDGEMENTS

We would like to thank Jim Brown, Yann Ponty, Craig Zirbel, Neocles Leontis for their correspondence, the Howard Hughes Medical Institute and the RNA Ontology Consortium—(NSF 0443508). We would also like to thank Greg Caporaso, Justin Kuczynski and Jose Clemente Litran for editing the manuscript.

Funding: National Institutes of Health (Grant HG4872, to R.K.); NASA Astrobiology (Grant NNX08AP60G, to R.K.).

Conflict of Interest: none declared.

REFERENCES

Andersen,E.S. *et al.* (2007) Semiautomated improvement of RNA alignments. *RNA*, **13**, 1850–1859.
Brown,J.W. (1999) The ribonuclease P database. *Nucleic Acids Res.*, **27**, 314.
Brown,J.W. *et al.* (2009) The RNA structure alignment ontology. *RNA*, **15**, 1623–1631.
Clamp,M. *et al.* (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
Darty, K. *et al.* (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
Griffiths-Jones,S. (2005) RALEE–RNA ALignment editor in Emacs. *Bioinformatics*, **21**, 257–259.
Griffiths-Jones,S. *et al.* (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
Hall,T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.*, **41**, 95–98.
Jossinet,F. and Westhof,E. (2005) Sequence to structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics*, **21**, 3320–3321.
Juhling,F. *et al.* (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes, *Nucleic Acids Res.*, **37**, D159–D162.
Knight,R. *et al.* (2007) PyCogent: a toolkit for making sense from sequence, *Genome Biol.*, **8**, R171.
Leontis,N.B. *et al.* (2002) The non-Watson–Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
Nasalean,L. *et al.* (2009) RNA 3D structural motifs: definition, identification, annotation, and database searching. In Walter,N.G. *et al.* (eds) *Non-Protein Coding RNAs*. Springer, Berlin Heidelberg, pp. 1–26.
Nawrocki,E.P. *et al.* (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
Roberts,E. *et al.* (2006) MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics*, **7**, 382.
Smit,S. *et al.* (2009) RNA structure prediction from evolutionary patterns of nucleotide composition. *Nucleic Acids Res.*, **37**, 1378–1386.
Stombaugh,J. *et al.* (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.*, **37**, 2294–2312.