BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 2 2013, pages 271–272
doi:10.1093/bioinformatics/bts667

Structural bioinformatics

Advance Access publication November 21, 2012

DOOSS: a tool for visual analysis of data overlaid on secondary structures

Michael Golden* and Darren Martin

Computational Biology Group, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town 4579, South Africa

Associate Editor: Ivo Hofacter

ABSTRACT

Motivation: DOOSS (Data Overlaid On Secondary Structures) is a tool for visualizing annotated secondary structures of large single-stranded nucleotide sequences (such as full-length virus genomes). The purpose of this tool is to assist investigators in evaluating the biological relevance of secondary structures within particular sequences.

Availability and implementation: DOOSS is written in Java and is available from: http://dooss.computingforbiology.org

Contact: michaelgolden0@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 4, 2012; revised on November 4, 2012; accepted on November 11, 2012

1 INTRODUCTION

Large single-stranded nucleotide sequences, such as cellular RNA molecules and single-stranded DNA and RNA virus genomes, apparently form complex secondary structures through Watson–Crick base-pairing between bases on the same DNA/RNA strand. Identified using either computational folding (Markham and Zuker, 2008) or experimental techniques (Watts et al., 2009), only a few of these structures have any known functional importance. Although techniques used to identify secondary structures typically give little or no information regarding the biological significance of identified structures, comparative structural and evolutionary analyses of related sequences can be illuminating in this regard (Hofacker et al., 2002). Specifically, the evolutionary processes operating both to conserve these structures and to maintain the coding potential of the nucleic acid sequences from which they are composed, could potentially be used to identify structures with likely biological relevance (Tuplin et al., 2002).

DOOSS (Data Overlaid On Secondary Structures) permits visualization of individual secondary structures simultaneously annotated with large amounts of relevant data including, for example, gene locations, codon usage statistics, computational or biochemically determined nucleotide pairing probabilities, apparently co-evolving nucleotide pairs and the degrees to which individual nucleotides, codons or base pairings are evolutionarily conserved. The intended purpose of such visualizations is ultimately to assist researchers in focusing the attention of their wet-lab and computational research onto the computationally predicted or experimentally demonstrated nucleic acid secondary structures that are most likely to have some unknown biological relevance.

*To whom correspondence should be addressed.

2 IMPLEMENTATION

2.1 Secondary structure and data visualization

At its core DOOSS uses the VARNA package (Darty *et al.*, 2009) to generate geometric coordinates for plotting nucleotide positions within structured nucleotide sequences, DOOSS then displays these structures together with various user-specifiable graphical data-overlays.

Secondary structures are displayed by DOOSS at two scales with a full-scale view (Supplementary Fig. S1) of the whole sequence facilitating ease of navigation across the entire sequence and a sub-structure view (Fig. 1) providing close-up details of individual structures. DOOSS will then annotate these structures with three different categories of data-overlay:

The first category is a nucleotide composition overlay that can be derived from sequence alignments, where the columns of the alignment are used to determine nucleotide composition at each nucleotide position within a structure. The nucleotide composition data are displayed in the form of a sequence logo at each nucleotide position (depicted by a circle) of the structure.

The second category of data-overlay is for the visualization of one-dimensional single nucleotide or single codon resolution quantitative data such as degrees of evolutionary conservation or biochemically determined base-pairing probabilities where each nucleotide corresponds with a single numeric value. These data values are visualized by way of a user-defined colour gradient.

The third category of data-overlay is for visualization of two-dimensional quantitative data relating to pairs of nucleotides such as the distance in ångströms between the nucleotides within the predicted or experimentally determined tertiary structure, the degree to which pairing between the nucleotides is conserved or probabilities that the nucleotides are co-evolving (see Supplementary Fig. S2). Wherever a pair of nucleotides in the sequence has a corresponding numeric value within a certain range it is depicted as a coloured line (also with a user-definable colour gradient) between the nucleotides.

The resulting graphically intense representation of the secondary structures can then be saved as publication quality images in various formats (PNG, SVG and EMF).

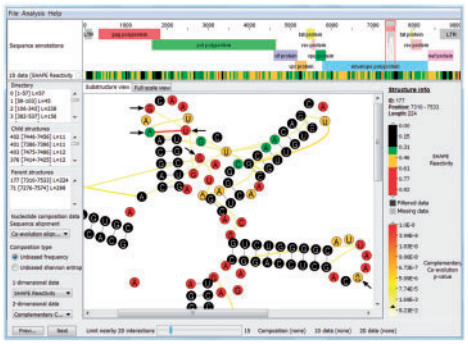

Fig. 1. A portion of the Rev response element with SHAPE reactivities at individual nucleotides as a one-dimensional data-overlay, complementarily co-evolving nucleotide pairs as a two-dimensional data-overlay (lines linking nucleotides) and the sequence alignment that was used to infer complementary co-evolution as a nucleotide composition overlay. Sequence logos with significant nucleotide variation are indicated using arrows

2.2 The dataset creation wizard

DOOSS provides a dataset creation wizard that steps the user through the process of creating a dataset. The wizard requires the user to specify a structure file (in any of various formats), a reference alignment file (also in any of various formats but which must contain sequences detectably homologous to that in the structure file) and a set of mapping alignments and their associated data files (in comma separated values format) containing data values to be visualized on the structure. Whereas the reference alignment must contain one or more sequences, where the columns of the alignment correspond precisely to nucleotide positions within the structure, each of the mapping alignments must contain the sequence datasets used to produce the nucleotide- or codon-specific values in their associated data files. The purpose of the mapping alignments is to allow matching up of values corresponding to individual nucleotides/codons or nucleotide pairs that were generated with diverse nucleotide sequence alignments to the appropriate nucleotide positions in the structure. Specifically, the wizard uses the reference alignment corresponding to the structure and the mapping alignments corresponding to each data source to map data points within each data source to the structure. The mapping of data sources to the structure is done in an automated fashion via MUSCLE (Edgar, 2004) generated pairwise alignments of reference and mapping alignment sequences. To provide a visual representation of the data values corresponding to each data source, the wizard asks the user to specify a 'data legend', which is a colour gradient that specifies a range of colours that represent the range of numeric values contained within the data source; these colours are then mapped to the corresponding visual elements on the structure. Finally, the wizard can be used to specify annotations that are displayed at the top of application, either by selecting a Genbank file or by manually entering the coordinates of the annotations.

2.3 Ranking of structures and sequence searching

To assist the investigator in identifying structures with potential biological significance, a ranking tool is provided that can be used to rank individual structures using one- or two-dimensional data sources. The ranking is done using a heuristic that compares the distribution of data values within an individual structure against the distribution of values in all other structures for a particular data source. This allows the user to mine out what are potentially the most biologically interesting structures amongst those present.

A search tool allows the user to find all structures containing a specified motif. This is useful for finding a structure that has been identified in the literature or elsewhere but might otherwise be difficult to find in a dataset using only coordinate information.

2.4 The HIV-1 SHAPE dataset as an example

To demonstrate the ease of visualizing a sequence with heterogeneous data sources, we used the dataset creation wizard to create a dataset from the HIV-1 SHAPE experimental data (Watts *et al.*, 2009). Data sources generated in the original study (the secondary structure itself, SHAPE reactivities and nucleotide pairing probabilities) along with an additional data source of our own (evidence of complementary co-evolution) were used to create a dataset viewable in DOOSS (Fig. 1).

The ranking tool was used to rank structures by the SHAPE reactivities (a chemical proxy for accessibility) of nucleotides within them. As expected, the Rev response element (RRE; a well-characterized secondary structure in the HIV genome) ranked first (Supplementary Table S1). The ranking indicated that the SHAPE reactivities of nucleotides within the RRE were much lower than could be accounted for by chance given the distribution of SHAPE reactivities for all other nucleotides within the HIV-1 genome.

3 CONCLUSION

By providing a simple way to graphically view nucleic acid secondary structures in the context of sequence characteristics such as degrees of evolutionary conservation or probabilities of nucleotide co-evolution and a mechanism to rank structures, DOOSS can assist tremendously in the identification and characterization of secondary structures that are likely to have biological significance.

Funding: National Research Foundation (NRF), Poliomyletis Research Foundation (PRF).

Conflict of Interest: none declared.

REFERENCES

Darty,K. *et al.* (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Hofacker,I. *et al.* (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.

Tuplin,A. *et al.* (2002) Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. *RNA*, **8**, 824–841.

Watts,J.M. *et al.* (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–716.