BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 4 2008, pages 579–580
doi:10.1093/bioinformatics/btm635

Sequence analysis

Colorstock, SScolor, Ratón: RNA alignment visualization tools

Yuri R. Bendaña and Ian H. Holmes*

Department of Bioengineering, UC Berkeley, Berkeley, CA, USA

Received on October 22, 2007; revised on December 11, 2007; accepted on December 22, 2007

Advance Access publication January 24, 2008

Associate Editor: Limsoon Wong

ABSTRACT

Summary: Interactive examination of RNA multiple alignments for covariant mutations is a useful step in non-coding RNA sequence analysis. We present three parallel implementations of an RNA visualization metaphor: Colorstock, a command-line script using ANSI terminal color; SScolor, a Perl script that generates static HTML pages; and Ratón, an AJAX web application generating dynamic HTML. Each tool can be used to color RNA alignments by secondary structure and to visually highlight compensatory mutations in stems.

Availability: All source code is freely available under the GPL. The source code can be downloaded and a prototype of Ratón can be accessed at http://biowiki.org/RnaAlignmentViewers

Contact: ihh@berkeley.edu

1 INTRODUCTION

Non-coding RNA (ncRNA) is an important part of the *cis*-regulatory picture (Ambros and Chen, 2007; Pheasant and Mattick, 2007) and has a broad chemical repertoire of great potential value (Breaker, 2004). Several 'pipelines' for discovery of novel ncRNA motifs by comparative genomics have been described (Pedersen *et al*., 2006; Rivas *et al*., 2001; Bradley *et al*., 2007; Washietl *et al*., 2005) adding to a comprehensive database of known ncRNA alignments (Griffiths-Jones *et al*., 2003). Visual inspection of alignments is an important quality control step in such pipelines.

Generally, in comparative genomics, alignments of sequences from related species are used to look for evidence of conservation of genomic features through evolution. Non-coding RNAs in particular, however, often do not exhibit conservation at the sequence level but do display conservation at the structural level. Compensatory mutations of the bases in an RNA alignment are a signature of this structural-level conservation. (By 'compensatory mutations' we mean substitutions of both the sites involved in a base-pair, such that canonical Watson-Crick or wobble base-pairing is maintained: although we do not know in which order the two substitutions occurred, we presume that the second substitution 'compensated' for the first.) When eyballing an RNA alignment to determine if it includes a structural ncRNA, it is extremely helpful to have a visual indicator of such mutations.

Using color to denote basepairing patterns in RNA alignments has been common and, workers place for a while. To our knowledge, the idea of specifically highlighting compensatory mutations was introduced by Pedersen in the EVOFOLD track of the UCSC Genome Browser, and by Griffiths-Jones in RALEE, the RNA alignment mode for Emacs (Griffiths-Jones, 2005). In the EVOFOLD track, rendered HTML pages show the colorized alignment for a predicted ncRNA feature on clickthrough from the corresponding browser track (Pedersen *et al*., 2006).

We found this visual paradigm to be useful enough to warrant a standalone implementation. We observe, however, that bioinformatics workflow often goes beyond static HTML. Experienced analysts spend a lot of time at the command line in an X11, VT100 or other terminal window, which (for an expert) is usually the most interactive way to work. However, extra interactivity is also starting to show up in 'Web 2.0' applications via dynamic HTML technologies such as Javascript/AJAX. An example of this migration of command-line workflow to a Javascript-led interface is the heavily Unix-influenced web application 'Yahoo Pipes': `http://pipes.yahoo.com/`

To analyze RNA effectively we needed visualization tools that matched all three parts of our workflow: command-line terminal hackery, static HTML page browsing and smart AJAX components. We also found that a variety of coloring paradigms is effective. Here, we describe three tools—Colorstock, SScolor and Ratón—that grew from these needs.

2 DESIGN

All three scripts take a Stockholm format alignment as input (see `biowiki.org/StockholmFormat`). The alignment should include a line beginning '`#=GC SS_cons`' that describes the consensus secondary structure, as per the Stockholm format spec. A colorized, annotated alignment is produced as output. Optionally, compensatory mutations are shown relative to a designated reference sequence.

The visual outputs of the three tools are compared in Figure 1. The coloring scheme used by Colorstock is slightly different than the scheme used by the other two programs. In Colorstock, coloration is per-column and by stem; the number of compensatory mutations is indicated above each stem. In SScolor and Ratón, coloration is per-basepair and depends on whether the bases (i) are complementary and (ii) display mutations relative to a reference sequence.

2.1 Colorstock

The Perl script `colorstock.pl` renders a colorized RNA alignment in ANSI terminal color. Optionally, a reference sequence can be highlighted. Colorstock also outputs a

*To whom correspondence should be addressed.

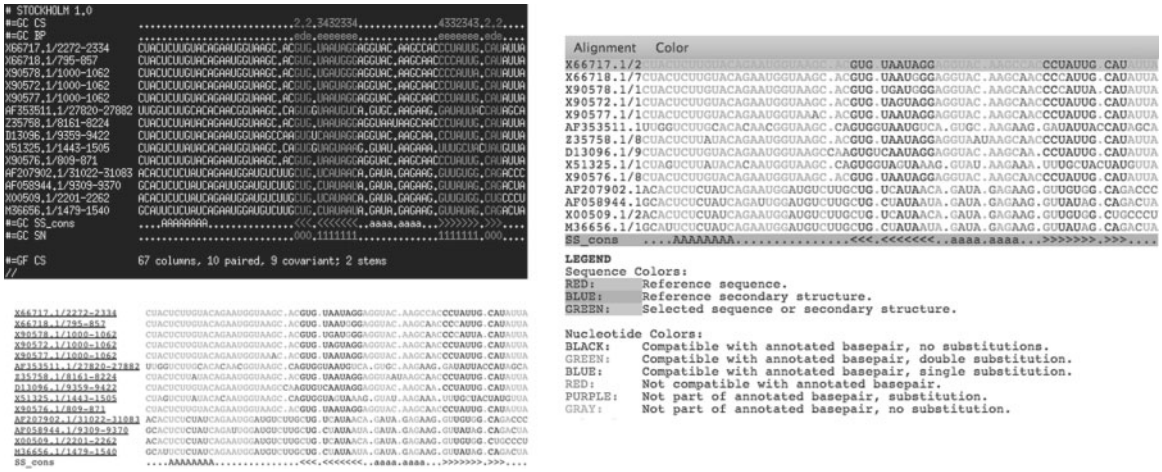

Fig. 1. RFAM alignment RF00165 (coronavirus 3'UTR replication element), as displayed by Colorstock (*top left*), SScolor (*lower left*) and Ratón (*right*). The first sequence is the reference sequence for this particular coloring. The SS_cons line displays the consensus secondary structure previously predicted by another program, such as `xrate` (Klosterman *et al*., 2006), and uploaded with the alignment (matching angle-brackets denote base-paired columns; see `biowiki.org/StockholmFormat` for explanation of this line). Note that the programs currently do *not* highlight basepairs in pseudoknots (some of which exist in this alignment).

summary line counting total paired columns, the number of these paired columns which display compensatory mutations and the number of distinct stems. Output is piped through `less` or another Unix pager. Optionally, the output can be rendered in HTML using the ANSI terminal color scheme. Extensive documentation is available by typing '`perldoc colorstock.pl`'.

2.2 SScolor

The `sscolor.pl` script outputs static HTML without Javascript, with a color scheme similar to Ratón's (described below). The accompanying `sscolorMult.pl` script calls `sscolor.pl` repeatedly, generating a family of interlinked HTML pages where each page has a different row of the alignment designated as the reference sequence. Extensive documentation of both tools is available by typing '`perldoc sscolor.pl`'.

2.3 Ratón

Ratón is an AJAX web application created to make the `sscolor.pl` alignment coloring function more interactive. At the beginning of a session, the user first uploads an RNA alignment in Stockholm format to the server. If the alignment contains a consensus secondary structure, the user is then able to select one of the sequences in the alignment to be the reference sequence. If there is more than one consensus secondary structure, the user can also select which one will be used for coloring.

The Ratón program acts as an interface to the `xrate` phylogrammar engine (Klosterman *et al*., 2006), so that the user can also request that a consensus structure to be computed by an `xrate` server.While this is processing, the user can (asynchronously) perform interactive coloring operations on the alignment, selecting any sequence to be the 'reference' and observing the mutations in the other sequences relative to this reference. With the graphical interface, the user is quickly able to see how basepairing covariation changes with respect to a given sequence in the alignment and consensus secondary structure.

The coloring scheme was inspired by the UCSC Genome Browser Evofold track, which uses color to distinguish Watson-Crick/wobble base-pairs from non-canonical base-pairs and to indicate compensatory mutations. Complementary basepairs are colored black if both are identical to the reference sequence, blue if one is different and green if both are different. Non-complementary basepairs are colored red. Unpaired bases are colored gray if they are identical to the reference sequence or purple if they are different.

ACKNOWLEDGEMENTS

YRB was funded in part by a Berkeley EDGE scholarship. IH was funded in part by NIH/NHGRI grant 1R01GM076705-01. The authors thank Mitch Skinner, Andrew Uzilov and Robert Bradley.

Conflict of Interest: none declared.

REFERENCES

Ambros,V. and Chen,X. (2007) The regulation of genes and genomes by small RNAs. *Development*, **134**, 1635–1641.

Bradley, *et al*. (2007) An XRATE ncRNA pipeline. In preparation.

Breaker,R.R. (2004) Natural and engineered nucleic acids as tools to explore biology. *Nature*, **432**, 838–845.

Griffiths-Jones,S. (2005) RALEE–RNA ALignment editor in Emacs. *Bioinformatics*, **21**, 257–9.

Griffiths-Jones,S. *et al*. (2003) Rfam: an RNA family database. *Nucleic Acids Res*., **31**, 439–441.

Klosterman,P.S. *et al*. (2006) XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics*, **7**, 428.

Pedersen,J.S. *et al*. (2007) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Computational Biology*, **2**, e33.

Pheasant,M. and Mattick,J.S. (2007) Raising the estimate of functional human sequences. *Genome Res*., **17**, 1245–1253.

Rivas,E., Klein,R.J., Jones,T.A. and Eddy,S.R. (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr. Biol*., **11**, 1369–1373.

Washietl,S. *et al*. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Nat. Acad. Sci. USA*, **102**, 2454–2459.