BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 13 2013, pages 1690–1692
doi:10.1093/bioinformatics/btt264

Sequence analysis

Advance Access publication May 8, 2013

Twine: display and analysis of *cis*-regulatory modules

Joseph C. Pearson[1,2] and Stephen T. Crews[1,2,*]

[1]Department of Biochemistry and Biophysics and [2]Program in Molecular Biology and Biotechnology, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3280, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Many algorithms analyze enhancers for overrepresentation of known and novel motifs, with the goal of identifying binding sites for direct regulators of gene expression. Twine is a Java GUI with multiple graphical representations ('Views') of enhancer alignments that displays motifs, as IUPAC consensus sequences or position frequency matrices, in the context of phylogenetic conservation to facilitate *cis*-regulatory element discovery. Thresholds of phylogenetic conservation and motif stringency can be altered dynamically to facilitate detailed analysis of enhancer architecture. Views can be exported to vector graphics programs to generate high-quality figures for publication. Twine can be extended via Java plugins to manipulate alignments and analyze sequences.

Availability: Twine is freely available as a compiled Java .jar package or Java source code at http://labs.bio.unc.edu/crews/twine/.

Contact: steve_crews@unc.edu

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

Received on December 4, 2012; revised on April 10, 2013; accepted on May 5, 2013

1 INTRODUCTION

Transcription is controlled by binding of sequence-specific transcription factors to DNA sequences near a gene. These binding sites are organized into modules called *cis*-regulatory modules or enhancers (Ong and Corces, 2011). Understanding how specific combinations of binding sites lead to the precise control of gene expression in developmental patterns or in response to physiological regulation is a major goal of modern biology.

Many software tools have been devised to aid in deciphering *cis*-regulatory logic, by predicting binding sites for known transcriptional regulators or by identifying putative motifs for novel regulators. Databases of *in vitro* and *in vivo* binding specificities for thousands of transcription factors enable predictions of possible regulators for an enhancer of interest (Das and Dai, 2007, Matys *et al.*, 2003, Portales-Casamar *et al.*, 2010, Zhu *et al.*, 2011). Various algorithms search for enrichment of short sequences (motifs) above statistical 'noise' (reviewed in Das and Dai, 2007).

However, most program outputs are either text-based or output graphical representations as raster images (Thomas-Chollier *et al.*, 2011), and it is tedious to manually annotate enhancers with binding site predictions. Genome browsers (Homann and Johnson, 2010, Kent *et al.*, 2002, Nicol *et al.*, 2009, Stein *et al.*, 2002) and gene browsers (Rebeiz and Posakony, 2004) map binding sites and functional genomic information on the whole-genome and single-gene scales in a dynamic fashion that allow greater user control. Twine complements these programs as an interactive graphical tool to analyze and compare enhancers. Twine displays the most common information used by researchers (motif locations and evolutionary conservation) in several intuitive 'Views' to help analysis and prediction of regulatory information, and the Views can be exported as vector graphics files to generate figures with scaled representations of conservation and motif locations.

2 FEATURES

Using FASTA-format sequence alignments as input, Twine generates multiple displays ('Views') of each alignment to allow visualization of sequence conservation (Fig. 1): Comparison View, Conservation View and Sequence View. The Comparison View represents each alignment with blocks of conservation and motif matches to the 'reference sequence' (the first sequence in each alignment). The threshold for conservation (0–100% of aligned sequences) and Blur (number of nucleotides used in each window for calculating conservation, 1–20 bp) can be adjusted to alter the number and span of conservation blocks. The selected alignment from the Comparison View is also displayed in Conservation and Sequence Views.

The Conservation View contains three sub-Views. The Aligned Species View is a graphical representation of aligned sequences in the selected alignment, with positions of nucleotides represented as black boxes, and motif matches indicated in each sequence by colored boxes. The Conservation Plot View displays conservation along the reference sequence, using the Blur factor to smooth the plot by averaging the conservation level of adjacent nucleotides. The Unaligned Species View represents each sequence from the alignment, with all gaps removed to reveal dramatic variations in species sequence length, indicating a possible problem in sequence assembly. The Sequence View displays the DNA sequence for the selected alignment, allowing direct visualization of the alignment and organization of motifs.

Matches to sequence motifs, entered as IUPAC strings or Position Frequency Matrices (PFMs), are displayed on the sequence views as colored blocks. IUPAC motifs are input as one or more binding sites, or as a consensus sequence with degenerate nucleotides (A, C, G, T, M, R, W, S, Y, K, V, H, D, B and N). The number of mismatches allowed can be user-specified for each motif. Matrices are imported as horizontal counts or frequencies (vertical matrices can be rotated to horizontal matrices in the input window). Motif thresholds can be

*To whom correspondence should be addressed.

Fig. 1. Twine main window. FASTA alignments of four enhancers active in the *Drosophila* neurogenic ectoderm (NEEs) were opened in Twine. Consensus sites for Suppressor of Hairless [Su(H)] and an unidentified motif (Motif-1), and position frequency matrices for Snail, Twist and Dorsal (from FlyFactorSurvey) were added (Markstein *et al*., 2002; Markstein *et al*., 2004). The Comparison View displays all four enhancers, with dark blocks representing regions conserved in at least 75% of species (specified by Threshold). The *single-minded* (*sim*) mesectoderm enhancer is displayed in the Conservation View and Sequence View. Organization and conservation of motif matches can be visualized at multiple scales and thresholds and exported to vector graphics programs

independently set, allowing control over match density. Strength of each match is indicated by opacity of each block; the range of opacity and threshold are user-adjusted.

Clicking on graphical representations of the alignments (Comparison View or Conservation Views) automatically moves the Sequence View to the appropriate location. Once a matrix is added, thresholds can be adjusted 'on-the-fly' using a slider to adjust the similarity threshold [the threshold score is the negative log of the product of each position's frequency in the matrix; therefore, zero is the most stringent possible score (Sung, 2010)]. To identify conserved motifs in non-optimal alignments and compensatory binding site shifts, the 'drift' of matches between species from linear alignment can be increased so that these will be considered 'conserved' (Supplementary Fig. S1). Matches can also be filtered to display only matches conserved at the current threshold.

Motif libraries can be saved to organize collections, and they can be imported *en masse* from text files containing motif matrices in commonly used formats (e.g. JASPAR). Motif libraries can be filtered using strings (literal or regular expression) matching motif descriptors (Supplementary Fig. S2). Using zero- to third-order Markov Chain background models (Liu *et al*., 2001), enrichment of each motif over the background model are calculated using binomial (Papatsenko, 2007) and Poisson distributions.

After opening enhancer alignments and adding motifs, the organization and conservation of binding sites within each enhancer, as well as patterns of motif clustering between enhancers, can be identified in the different Views. When threshold parameters have been adjusted to user specification, each View can be saved as Scalable Vector Graphics (SVG) files, where each element can be independently manipulated by programs, such as Adobe Illustrator or Inkscape. In addition to SVG outputs of each alignment view, sub-alignments can be saved in FASTA format by drag-selecting a region in the Aligned Species or Conservation Plot Views and saving the selection as FASTA alignments. Thus, potential 'minimal' enhancers from larger fragments can be extracted by analyzing the conservation patterns and clustered binding sites of likely regulators. Alignment data, motif statistics and the set of all motif matches can be exported as a tab-delimited file for further analysis.

To generate representations of all deletions or binding site mutations of a given enhancer (such as from *in vitro* mutagenesis experiments), an alignment of the wild-type sequence to sequences with each variant can be opened in Twine. Inputting motifs for tested binding sites generates an Aligned Sequence view indicating presence or absence (or deletion) of all tested variants (Supplementary Fig. S3).

Using a plugin interface implementing the Java Simple Plugin Framework, AlignedSequence objects (a custom Java class containing all alignments and motifs) can be sent to user-written Java plugins, modified (e.g. aligned, analyzed and manipulated), then returned to Twine for display. Several example plugins, as well as a template, are included. Future work includes expanding the suite of plugins and supporting manual adjustments to alignments.

ACKNOWLEDGEMENTS

The authors thank William McGinnis for support on a precursor to this program, Neil Tedeschi for programming advice and Joseph Watson, Joseph Fontana and Brian Busser for software testing and suggesting features. The authors also thank the creators of Batik SVG library, Apache Commons Mathematics Library and Java Simple Plugin Framework.

Funding: The National Institutes of Health (NICHD F32 HD061175 to J.C.P., NINDS R01 NS64264 to S.T.C.).

REFERENCES

Das,M.K. and Dai,H.K. (2007) A survey of DNA motif finding algorithms. *BMC Bioinformatics*, **8** (**Suppl. 7**), S21.

Homann,O.R. and Johnson,A.D. (2010) MochiView: versatile software for genome browsing and DNA motif analysis. *BMC Biol*., **8**, 49.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.

Liu,X. *et al*. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput*., **6**, 127–138.

Markstein,M. *et al*. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA*, **99**, 763–768.

Markstein,M. *et al*. (2004) A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development*, **131**, 2387–2394.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.

Nicol,J.W. *et al*. (2009) The integrated genome browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, **25**, 2730–2731.

Ong,C.T. and Corces,V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet*., **12**, 283–293.

Papatsenko,D. (2007) ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. *Bioinformatics*, **23**, 1032–1034.

Portales-Casamar,E. *et al.* (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.

Rebeiz,M. and Posakony,J.W. (2004) GenePalette: a universal software tool for genome sequence visualization and analysis. *Dev. Biol.*, **271**, 431–438.

Stein,L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Sung,W.K. (2010) *Algorithms in Bioinformatics: A Practical Introduction*. Chapman & Hall/CRC, Boca Raton, FL, pp. 253–254.

Thomas-Chollier,M. *et al.* (2011) RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res.*, **39**, W86–W91.

Zhu,L.J. *et al.* (2011) FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.*, **39**, D111–D117.