Bioinformatics Advance Access published August 13, 2016

Application Notes

Gene Slider: Sequence Logo Interactive Data-visualization for Education and Research

Jamie Waese, Asher Pasha, Tingting Wang, Anna van Weringh, David S. Guttman, and Nicholas J. Provart*

Department of Cell and Systems Biology / Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, Ontario, M5S 3B2

Associate Editor: Dr. John Hancock

***ABSTRACT**

Summary: Gene Slider helps visualize the conservation and entropy of orthologous DNA and protein sequences by presenting them as one long sequence logo that can be zoomed in and out of, from an overview of the entire sequence down to just a few residues at a time. A search function enables users to find motifs such as *cis*-elements in promoter regions by simply "drawing" a sequence logo representation of the desired motif as a query. In addition to displaying user-supplied FASTA files, our demonstration version of Gene Slider loads and displays a rich database of 90,000+ conserved non-coding regions across the Brassicaceae indexed to the TAIR10 Col-0 *Arabidopsis thaliana* sequence. It also displays transcription factor binding sites, enabling easy identification of regions that are both conserved across multiple species and may contain transcription factor binding sites.

Availability and Implementation: Freely available on the web at: http://www.bar.utoronto.ca/GeneSlider and also as an app on http://araport.org. Website implemented in JavaScript and Processing.js with all major browsers supported. Source code available under GNU GPLv2 at SourceForge: https://sourceforge.net/projects/geneslider/.

Contact: nicholas.provart@utoronto.ca

1 INTRODUCTION

Sequence logos are a well-known bioinformatic data visualization paradigm for displaying conservation and entropy in aligned sequences (Schneider and Stephens, 1990). Each of the residues in a sequence is represented by a series of stacked characters in which the height of the stack signifies the degree of conservation for that position, and the height of the characters signifies the frequency of that residue. Traditionally, sequence logos are generated with tools such as WebLogo (Crooks et al., 2004) and Seq2Logo (Thomsen and Nielsen, 2012) in which any sequence beyond 30 or so bases must be broken into separate rows. This makes it difficult to identify motifs that span multiple rows. Skylign (Wheeler et al., 2014) addresses this problem by generating long sequence logos that can be zoomed in and out of, however it is still difficult to identify amino acid motifs when there is "wobble" in one or more of the bases.

Gene Slider addresses these challenges by combining the ability to make long sequence logos with a search panel that automatically highlights motifs of interest, even when there is a high degree of wobble. Slider controls assign certainty thresholds for each position of a search query to control whether the search function will recognize partially conserved matches. Up to six different motifs can be searched for, with each motif containing up to eight different bases. A navigation slider at the bottom of the screen indicates which region of the sequence is currently being displayed. Coloured markers along this slider indicate where query motifs have been identified, making it easy to see where multiple matches have been found. Figure 1 shows a portion of an alignment in which there is only one occurrence of the motif AAACA. Without the search function it would be challenging to find this motif because of the entropy in positions 3 and 4 of the match.

Fig. 1. The motif AAACA is highlighted according to the query in the search panel. The horizontal bar superimposed over each column represents the number of sequences with a residue at that position. The six residues on the right side of the figure have high bit scores suggesting a high degree of conservation, however the horizontal bar indicates that only one of the nineteen species in the alignment has a base in those positions.

2 METHODS

Gene Slider loads FASTA formatted aligned sequences into a large matrix and then counts the frequency of each base per column. Bit scores for each base are calculated according to Schneider & Ste-

*To whom correspondence should be addressed.

phens's (1990) application of the Shannon entropy formula. Individual bit scores are summed to make a total bit score for the column. Symbols are then scaled according to the frequency of residues found in each column and drawn to the screen according to the start and end points of the view settings.

The usual Shannon logo can be misleading because column heights do not take insertion and deletion events into account. A base that is only found in one sequence while each of the other sequences show a gap will have the same bit score and column height as a base that is conserved across every sequence. Gene Slider provides two modes for visualizing how many sequences are contributing to the bit score for any given base. The first mode uses a horizontal bar indicator, as can be seen in Figure 1 where positions 534-539 only have one species with a base in that column. The second mode manipulates the saturation level of each column, which takes advantage of the preattentive ability in our visual cortex to quickly recognize subtle changes in luminance (Healey and Enns, 2012).

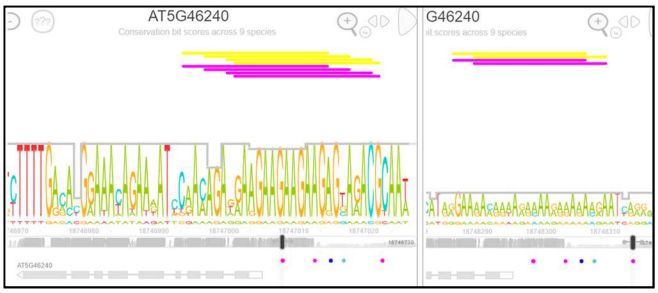

Fig. 2. Pre-mapped JASPAR and Weirauch et al. (2014) motifs in the promoter of *KAT1* (At5g46240). It has been shown that the transcription factor SOC1 affects stomatal opening, which is regulated by KAT1 (Kimura et al., 2015). The left panel shows a well-conserved SOC1 binding site (yellow) immediately upstream of the KAT1 promoter, while the right panel shows a SOC1 site that is not well conserved. The hypothesis would be that the former is the functional SOC1 binding site. The panels are displayed with the same vertical weighted bit score scale.

Gene Slider provides another method for overcoming this problem by introducing a new logotype we call a "Weighted Shannon". This approach factors the number of sequences contributing to the bit score into the calculation of that position's height by multiplying the bit score with the number of residues in that column divided by the total number of sequences in the alignment. In contrast to the various Kullback-Leibler modifications in Seq2Logo (Thomsen and Nielsen, 2012), this produces a logotype that is similar to the convention, but highly conserved regions are not conflated with ones having just one or two residues in an otherwise gapped region.

Work sessions can be saved and reloaded via a dynamically generated URL accessed through the 'Share' button. The 'Screen Grab' button creates high-resolution images suitable for publication. Gene Slider can display DNA or protein FASTA files, as well as GFF files that are uploaded by the user.

Gene Slider was written in Processing.js and runs in most web browsers. It can display a large number of sequences of any length, however performance decreases as the sequence length and sequence number goes up. Contemporary laptops can display sequences up to 3000 bases long without noticeable lag. Data can be loaded directly from the users' hard drive, however, conserved noncoding sequence data can be loaded from RESTful web services running on the BAR (Toufighi et al., 2005).

3 RESULT

To demonstrate Gene Slider, we have pre-processed 90,000+ conserved non-coding regions across the Brassicaceae (Haudry et al., 2013), indexed to the TAIR10 Col-0 *Arabidopsis thaliana* sequence, for easy identification of regions that are conserved across multiple species. Further, we have mapped transcription factor binding sites from the JASPAR database (Mathelier et al., 2014) and from Weirauch et al. (2014) to the TAIR10 genome sequence using FIMO (Grant et al., 2011) to permit the identification of regions that are both conserved and may contain transcription factor binding sites (Figure 2). Upon visiting the web site, simply enter the name of the Arabidopsis gene and how many bases upstream and downstream of the gene you wish to load, or select a chromosomal region by entering the chromosome number, start and end point. This data can also be accessed directly as a web service. A link for this can be found on the website's About page.

Although FIMO-mapped motifs are found in many places in the genome, a strong positional disequilibrium for a subset of these is found approximately -200 bp upstream of the transcriptional start site in Arabidopsis promoters if a conservation criterion is applied. Such a disequilibrium is consistent with functionally active binding sites. In Arabidopsis, 79% of the TFBSs mapped from the Weirauch et al. (2014) data set and some other smaller data sets have identical positions in the alignment of the promoter region with three other closely related Brassicaceae species (Yu et al., 2016). This subset shows a marked positional disequilibrium. In the same study, the authors also examined DNAse I hypersensitivity data from Sullivan et al. (2014), and found a similar disequilibrium in the -200 bp upstream region. For this reason, we have also included a "Link to Regulome" for a set of DNAse I data from Sullivan et al. (2014).

REFERENCES

Crooks, G.E. et al. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Grant, C.E. et al. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Haudry, A. et al. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. *45*, 891–898.

Healey, C.G., and Enns, J.T. (2012). Attention and visual memory in visualization and computer graphics. IEEE Trans. Vis. Comput. Graph. *18*, 1170–1188.

Kimura, Y. et al. (2015). A Flowering Integrator, SOC1, Affects Stomatal Opening in Arabidopsis thaliana. Plant Cell Physiol. *56*, 640–649.

Mathelier, A. et al. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. *42*, D142–D147.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Sullivan, A.M. et al. (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. Cell Rep. *8*, 2015–2030.

Thomsen, M.C.F., and Nielsen, M. (2012). Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res. *40*, W281–W287.

Toufighi, K. et al. (2005). The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. Plant J. *43*, 153–163.

Weirauch, M.T. et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

Wheeler, T.J. et al. (2014). Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics *15*, 7.

Yu, C.-P. et al. (2016). Positional distribution of transcription factor binding sites in Arabidopsis thaliana. Sci. Rep. *6*, 25164.