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

Jamie Waese Asher Pasha Nicholas Provart

Cell and Systems Biology, University of Toronto

Abstract

Gene Slider is a web-based interactive data visualization tool that enables users to analyze up to 10,000 aligned DNA and protein sequences, and manipulate the display from a broad overview of the entire sequence to a very close view of just a few residues at a time. In addition, a search function enables users to enter up to six different query motifs by using sliders in the search panel to “draw” a SeqLogo representation of the desired query. Gene Slider then automatically finds, highlights and labels matching regions. Gene Slider assists comprehension and hypothesis generation by presenting the information visually and with an intuitive user interface.

Background

While there is no doubt that sequence logos provide an elegant approach to visualizing the conservation / entropy of aligned DNA and protein sequences, any sequence beyond 30 or so bases must be broken into separate rows. This makes it difficult to identify motifs that span across multiple rows because the visual disconnect of a line break interferes with our precognitive abilities to recognize patterns and clusters based on proximity. Also, it is difficult to recognize motifs when there is "wobble" in several of the bases because it takes additional mental processing to filter out the symbols we're not looking for.

Objective

To develop a web-based tool for identifying conserved bases and motifs in aligned DNA and protein sequences. Also, to facilitate comprehension and hypothesis generation by providing educators and researchers with an easy to use tool that incorporates current thinking on effective data visualization.

Methods

Gene Slider was written with Processing and Process.js and is embedded as a canvas element on an HTML page.

Upon confirming the validity of a FASTA data file, the program loads all the sequences into a large matrix, counts the frequency of each base per column, and computes the bit scores for each base. Bit scores are calculated according to Schneider & Stephens’s (1990) application of the Shannon entropy formula, where:

$$H(l) = -\sum_{b=a}^{t} f(b,l)\log_2 f(b,l)$$

Individual bit scores are then summed to make a total bit score for the column. Objects that contain the individual bases according to their relative height are then created, stored in a new array, and drawn to the screen according to the start and end points of the user’s current view settings. Depending on the user’s computer processing speed as well as the number of objects that need to be displayed each cycle, the program reacts to mouse events and redraws the screen an average of 20-40 times per second.

When the user adjusts a search panel slider, Gene Slider scans the matrix comparing the new search query with each position in the sequence, background colors are updated accordingly and the draw cycle begins again.

GUI controls enable quick manipulation of the display in and out points. Because the SeqLogo letters are based on TrueType fonts that can be distorted, it is possible to zoom in and out without distortion as well as change fonts and color schemes on the fly.

Save and load functions are handled through a separate JavaScript program that sends session data from the Gene Slider program to the Bio-Analytic Resource (BAR) server. The server formats the data into a JSON object, returns it to the user, and presents it as a downloadable file. The load function parses FASTA files and Gene Slider session files then sends the data to the main program, which displays the data in the Gene Slider window.

A new contribution to the SeqLogo visual paradigm is the addition of a horizontal bar that runs across the display to indicate how many sequences are contributing to each residue’s bit score. Without such an indicator, it would be difficult to differentiate between residues that are highly conserved across multiple species and residues that are only conserved in the few sequences containing an insertion at that position. When the alignment indicator is turned off, this information is applied to the saturation level of each base.

Future Modifications

Gene Slider will soon be able to load GFF files and display gene feature data such as introns, exons and known promoter regions. It will also output a list of which samples contain a specific search query and where it is located, and generate high resolution SeqLogo images suitable for publication.

Gene Slider was developed in the Provart Laboratory and is available at GeneSlider.org.

BioVis 2013 Redesign Contest

Using data from the BioVis 2013 Redesign Contest “to develop a practical replacement to the long-standing convention of sequence logos”, we present a figure in which we have used Gene Slider’s search feature to rapidly identify and highlight residues that are somewhat conserved in all organisms, and have juxtaposed this search with two searches for residues that are well conserved in just Gram-negative or just Gram-positive bacteria. In this manner, residues which exhibit distinctly different requirements in the two subsets of bacteria are easily identified (in Figure 1 for the second stabilizing residue, S/C).

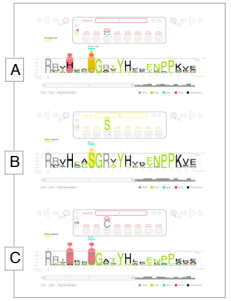

Figure 1. Sequence logos showing the amino acid usage in the adenylate kinase lid (AKL) domain. (A) Across all organisms. (B) from Gram-negative bacteria. (C) from Gram-positive bacteria.

The ADK lid domain structure is universally conserved, but is stabilized in the Gram-negatives by a hydrogen bonding network between residues 4, 7, 9, 24, 27, and 29 (and several other residues in some organisms), while the Gram-positives are stabilized by a bound metal ion, tetrahedrally coordinated by the Cysteines at 4, 7, 24 and 27. The identities of several other positions (eg 5, 8, 30, 32) are differentially constrained in each subfamily as well, apparently due to steric requirements of the stabilizing residues.

Conclusion

Gene Slider leverages the many good qualities of sequence logos. It was developed with reference to recent thinking about precognitive processing of visual information and data visualization. As such it represents an intuitive, familiar way for exploring sequence alignments.

References

Schneider, TD & Stephens, RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res., 1990 Oct 25;18*20:6097-100

C. E. Shannon, ``A mathematical theory of communication,'' Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.