Logolas - Guided Tutorial

Section 1: Why Logolas?

How Logo plots have been popularly used in the field of genomics for revealing TF patterns and amino acid compositions (cite the name of the packages)

Popular Logo building softwares are constrained in their limited size of the library of symbols that can be used to plot (seqLogo etc).

Logolas provides a much wider library of symbols comprising of all English alphabets, numbers, punctuations etc.

But most importantly, Logolas lets the user plot alphanumeric strings as logos, which extends the scope of the visualization massively, beyond the TFs or protein sequences. We show examples of how this string representation of logos is effective in visualizing mutation signature patterns, ecological species abundance patterns etc.

Additionally with Logolas, besides the position weights of logos at each site, the user can also plot enrichments and depletions of symbols in terms of their occurrence likelihood at each site.

We use the enrichment and depletion of the logos/symbols for symbol calling at each position. This Logolas based nomenclature is a more generic alternative to the IUPAC and the Prosite nomenclatures used for calling nucleotide and amino acid respectively.

Most logo plotting softwares take position weight matrix as input for plotting the logo plots. But this approach neglects the frequency scale underlying the position weight matrix. For example, a position weight computed for a TF data based on just 10 fragments mapping to that position is less reliable compared to one based on 100 fragments. In such a case, the user would want to shrink the composition probability to the pre-defined background or prior much more in the first case compared to the second.

In Logolas, we provide a Dirichlet Adaptive Shrinkage method (**dash**), in similar lines to the adaptive shrinkage **ashr** approach due to Stephens 2016, to adaptively shrink the positional weights based on the positional frequency scale.

Besides all the above benefits, Logolas allows for new and flexible stylizations, textures and color patterns of the logos , allows the user to create her own logos and add them to the library, and also provides an easy interface to combine Logolas plots with R base graphics and ggplot2 graphs in multipanel visualizations.

Section 2: First Look at Logolas

Figure 1 : Take a demo data with A, C, G, T - may be the same one as of now. Plot the logomaker plot.

Explain how the information content was calculated

Figure 2: Do a multiple panel plot - changing colors, changing stylization, with yscale_change, changing viewport settings, different background. The plot will show firstly how multiple panel plot can be used + different ways to plot a logo plot in Logolas

Figure 3: Do another Multiple panel logomaker plot with Renyi entropy with different alpha values. (try two different alpha parameters). Also mention how Renyi entropy works.

Figure 4: Introduce nlogomaker plot for the same demo example - a multiple panel plot with log and log odds. State how the median adjustment is done. Mention how enrichment and depletion can be identified better using this plot.

Figure 5: A multiple panel plot with different stylizations (border and fill) for the logos, for both positive and negative Logolas plots.

Figure 6: Depletion weight adjustment - a multiple panel plot to show how the depletion weight adjustment can affect the nlogomaker plots making the depletion look more prominent as the depletion weight is increased.

Section 3: Extended features of Logolas

Figure 7: A protein logo plot using symbols beyond the A, C, G and T. This would provide an example of how other alphabets can be used in Logolas besides the A, C, G, T. Use the BLOSUM background probability to automatically point to the use of different background being important when dealing with proteins data.

Figure 8: A PSSM plot (logopssm) for plotting PSSM data of the proteins.

Figure 9: Mutation signature plot. A multiple panel plot showing how for strings, coloring can be done in two different ways - per row and per symbol.

Figure 10: Introduce full string logos. First example can be the ecological abundance data or the histone modifications data. I think just one of these two would be enough to drive home the point. The other can be dropped.

Figure 11: aRxiv example can be placed here. It shows how punctuation marks like . and - can be used in strings as well besides alphabets and numbers.

Figure 12 : A multi panel plot where we have a Logolas plot combined with a R base graphics and/or ggplot2 graph. Could be as simple as hist() or qplot(). This is just to illustrate that the user can combine Logolas plots with other plots.

Section 4: Symbol calling

IUPAC Nomenclature

Prosite Nomenclature

Logolas Nomenclature

Take the help of figures as you see fit to discuss how you define the Logolas nomenclature and how connected it is to the IUPAC and Prosite nomenclature

Discuss examples to demonstrate why the Logolas nomenclature is an improvement on either nomenclature. For proteins, I think Prosite is as good, but we already saw that Logolas type nomenclature is better than getIUPAC().

You can also mention about the action space nomenclature as an alternative nomenclature specifically applicable for the TF data. But focus more on the generic Logolas nomenclature based on the heights of the symbols in the nlogomaker plot.

Section 5: Dirichlet Adaptive Shrinkage (dash)

The Model formulation of dash

The intuition behind different parameters and the choice of defaults for them.

Figure 13 : Application of dash to low frequency, medium frequency and high frequency data. Separate panels for logomaker and nlogomaker in a single multi-

panel plot.

Figure 14: Application of dash when the background is not uniform. Comparison of without dash without bg, with dash without bg, without dash with bg and with dash with bg.

Section 6 : Applications of Logolas

TF databases - Encode, HOCOMOCO, Plant TFDB, Manollis Kellis webpage (I think this has a overlap to a great extent with Encode). Data can come from Chip-seq, Chip-chip, HT-SELEX etc.

Figure 15: Application of Logolas on HT-SELEX data and Chip-seq data. Talk about how HT-SELEX TFBS are shorter compared to Chip-seq ones.

Figure 16: Demonstrate how n-Logolas plots capture the symmetry in dimerized TFBS.

Figure 17: Plant TFDB, how different backgrounds can be used in real life application for plotting the logo plots (may be show plots for a couple of species and with and without dash).

Figure 18: Do a combined dash versus each signature wise dash, and compare the results. The combined dash shrinks less than the signature wise dash.

Protein database: webpage, explanation how the PWM was obtained and how a PFM was estimated from number of mapped reads from PWM.

Figure 19: General application of Logolas and n-Logolas for protein database.

Figure 20: Use of dash for some protein family PWM with very few sequences mapped.

Mutation signature data. How the data were obtained. Reference to Yuichi's paper and the Alexandrov et al (2013) paper.

Figure 21: The Logolas and n-Logolas plots for some of the clusters and comparsion with the visPMsignature plots from Yuichi's paper.

Figure 22: The n-Logolas plots for some of the tissues in Alexandrov et al (refer to the workflow page for all the figures).

UMI data: refer to Tung et al (2016). How FastQ files were processed.

Figure 23 : How the barcode composition shows bias in the initial part of the read. The enrichment of purines (A and G) among the first five bases. The enrichment of Ts otherwise in the data. The plots to be done for cells across individuals to show that the bias is consistent across individuals and is most possibly a result of the Fluidigm experiment.

aRxiv data: the process of getting the aRxiv field categories of professors. The code is already provided.

Figure 24: Logolas and n-Logolas plots with and without dash for the aRviv data. This will be an example of using dash on string logo based data.