Sublogo dendrograms: visualizing correlation in biological sequence motifs

Toby Dylan Hocking[1,*]

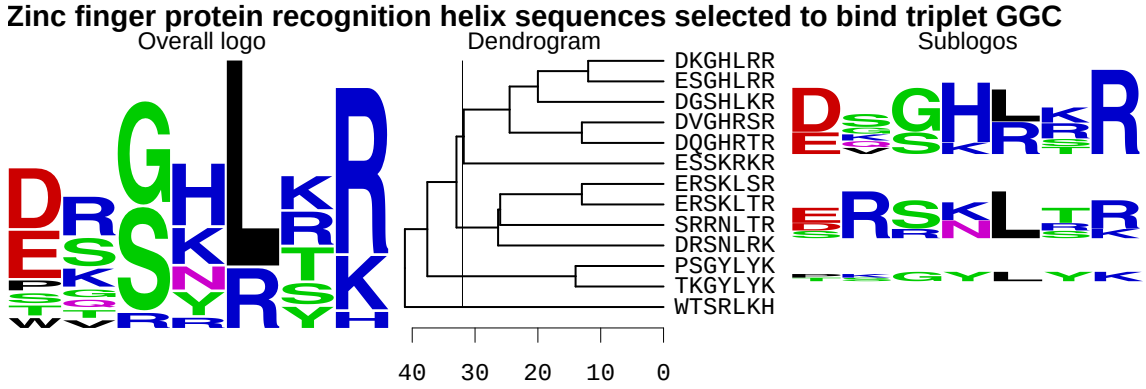
1. LSTA; Université Paris 6, 175 Rue du Chevaleret, 75013 Paris, France

* Contact author: tdhock@ocf.berkeley.edu

Keywords: Sequence logo, statistical graphics, biological sequence correlation, clustering

DNA and protein sequence motifs are usually visualized using graphical sequence logo plots (Schneider and Stephens, 1990). Though sequence logos excel at communicating the information at each position in the motif, one problem with their use is that they make no attempt to show the joint distribution between positions. I propose the sublogo dendrogram as a new type of statistical plot that uses logos and dendrograms to show this joint distribution. In addition, sublogo dendrograms reveal significant details in subfamilies of sequences that can go unnoticed using a standard logo.

In this work, sublogo dendrograms have been implemented using the WebLogo program (Crooks *et al.*, 2004) in conjunction with R's `grImport` package (Murrell, in press). First, R is used to perform a hierarchical clustering on the aligned input sequences. The tree that results from the clustering is cut, yielding several subfamilies of sequences. Next, WebLogo creates an overall logo image for the entire set of sequences as well as a "sublogo" for each subfamily. The dendrogram resulting from the hierarchical clustering is drawn in the center of the plot, with the overall logo on the left, and the sublogos on the right, next to the relevant leaves of the dendrogram.



For user convenience, a web server has been set up to make sublogo dendrograms. Furthermore, the R software and code for the web interface is freely available for download, usage and modification.

`http://www.ocf.berkeley.edu/~tdhock/sublogo_dendrogram/form.php`

References

Schneider TD Stephens RM (1990). Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.*, 18, 6097–6100.

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004). WebLogo: A sequence logo generator. *Genome Research*, 14, 1188–1190,
`http://weblogo.berkeley.edu`

Paul Murrell (in press). Importing Vector Graphics: the GrImport package for R. *Journal of Statistical Software*,
`http://www.stat.auckland.ac.nz/~paul/R/grImport/import.pdf`.