Phylo seq logo

In collaboration with David Liberles.
Sequence logos (seqlogos) is a popular way to visualize patterns in biological sequences, both for DNA and peptides (small protein parts). A seqlogo has a bar for each position in a pattern and the height of each bar corresponds to the relative entropy of the probability distribution for the symbols estimated in that position compared to the uniform distribution. Within each bar, the actual symbols are drawn with the size of each symbol being proportional to its frequency. As an example: if a DNA position contains all A's, then the bar has height 2 because the relative entropy is 1*lg(1/0.25) = 2. Within that bar, a single A is shown. If a position has A and G at equal frequencies, then the relative entropy is 0.5*lg(0.5/0.25) + 0.5*lg(0.5/0.25) = 1 and equal-sized A and G are drawn atop of each other.
There are web services that makes it easy to retrieve seqlogos and several open source projects for making seqlogos on your own computer. For example

- https://services.healthtech.dtu.dk/services/Seq2Logo-2.0/
- http://weblogo.berkeley.edu/logo.cgi

are two commonly used services that have plenty of examples and links to the literature.
A complaint with these systems, that has become emphasized in recent years due to the massive amounts of sequencing of populations rather than model species, is that they do not account for the evolutionary history of the sequences. The problem is that some species or groups of species might be sampled more than others and that skews the distribution. For example, if one is interested in a peptide pattern in mouse, dog and primates, then it is likely that the primates have the same or very similar versions of the pattern while mouse and dog probably has more differences. But the pattern found in mouse and dog is drowned in the many samples from the primates. The seqlogo is drawn to visualize a pattern in some context and the question is whether the samples are representative to the context.
It has been proposed to deal with this problem by weighing the samples using a phylogenetic tree.
Felsenstein defined phylogenetic independent contrast (PIC), which could be adapted for "sequence logos", but no software for that is available.

- https://www.jstor.org/stable/pdf/2461605.pdf

There is a tutorial for PIC by Luke Harmon. I am not a fan, but maybe it helps:

- Estimating Rates using Independent Contrasts

The presentation given by Felsenstein and Harmon is for scalar variables modeled as Brownian motion over a rooted tree, with one variable per edge. The value on one edge

is a weighted average of the 'child edges', using variance as weights as can be justified when using Brownian motion as a model
We would want the variables to be symbol distributions, one for each position in a pattern. The weighs would be from branchlengths in the input phylogeny.

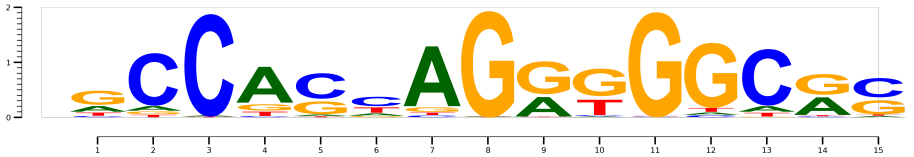
Graphics could (should?) be handled by this code:

- https://github.com/saketkc/pyseqlogo

Its benefit is that it takes low-level logo description as input:

```
matrix = [[('C', 0.02247014831444764),
('T', 0.057903843733384308),
('A', 0.10370837683591219),
('G', 0.24803586793255664)],
...
```

And the output can be:



One is free to provide both symbols and symbol heights.
The goal would be to write a unix app that reads a MSA and a tree for a set of sequences, DNA or proteins/peptides, and outputs an image with a seqlogo.
The scientific question is: how does the visualization change?

Untitled Attachment

Untitled Attachment