Sequence logos for DNA sequence alignments

Oliver Bembom

Division of Biostatistics, University of California, Berkeley

October 30, 2018

1   Introduction

An alignment of DNA or amino acid sequences is commonly represented in the form of a position weight matrix (PWM), a $J \times W$ matrix in which position (j, w) gives the probability of observing nucleotide j in position w of an alignment of length W. Here J denotes the number of letters in the alphabet from which the sequences were derived. An important summary measure of a given position weight matrix is its information content profile (Schneider et al., 1986). The information content at position w of the motif is given by

$$IC(w) = \log_2(J) + \sum_{j=1}^{J} p_{wj} \log_2(p_{wj}) = \log_2(J) - entropy(w).$$

The information content is measured in bits and, in the case of DNA sequences, ranges from 0 to 2 bits. A position in the motif at which all nucleotides occur with equal probability has an information content of 0 bits, while a position at which only a single nucleotide can occur has an information content of 2 bits. The information content at a given position can therefore be thought of as giving a measure of the tolerance for substitutions in that position: Positions that are highly conserved and thus have a low tolerance for substitutions correspond to high information content, while positions with a high tolerance for substitutions correspond to low information content.

Sequence logos are a graphical representation of sequence alignments developed by Schneider and Stephens (1990). Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack is proportional to the information content at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

2 Software implementation

The `seqLogo` package provides an R implementation for plotting such sequence logos for alignments consisting of DNA sequences. Before being able to access this functionality, the user is required to load the package using the `library()` command:

```
> library(seqLogo)
```

2.1 The `pwm` class

The `seqLogo` package defines the class `pwm` which can be used to represent position weight matrices. An instance of this class can be constructed from a simple matrix or a data frame using the function `makePWM()`:

```
> mFile <- system.file("Exfiles/pwm1", package="seqLogo")
> m <- read.table(mFile)
> m

   V1  V2  V3  V4  V5  V6  V7  V8
1 0.0 0.0 0.0 0.3 0.2 0.0 0.0 0.0
2 0.8 0.2 0.8 0.3 0.4 0.2 0.8 0.2
3 0.2 0.8 0.2 0.4 0.3 0.8 0.2 0.8
4 0.0 0.0 0.0 0.0 0.1 0.0 0.0 0.0

> p <- makePWM(m)
```

`makePWM()` checks that all column probabilities add up to 1.0 and also obtains the information content profile and consensus sequence for the position weight matrix. These can then be accessed through the corresponding slots of the created object:

```
> slotNames(p)

[1] "pwm"       "consensus" "ic"        "width"     "alphabet"

> p@pwm

    1   2   3   4   5   6   7   8
A 0.0 0.0 0.0 0.3 0.2 0.0 0.0 0.0
C 0.8 0.2 0.8 0.3 0.4 0.2 0.8 0.2
G 0.2 0.8 0.2 0.4 0.3 0.8 0.2 0.8
T 0.0 0.0 0.0 0.0 0.1 0.0 0.0 0.0

> p@ic

[1] 1.2780719 1.2780719 1.2780719 0.4290494 0.1535607 1.2780719 1.2780719
[8] 1.2780719

> p@consensus

[1] "CGCGCGCG"
```

2.2 Plotting sequence logos

The `seqLogo()` function plots sequence logos.

INPUT.

```
> args(seqLogo)

function (pwm, ic.scale = TRUE, xaxis = TRUE, yaxis = TRUE, xfontsize = 15,
    yfontsize = 15)
NULL
```

1. The position weight matrix for which the sequence logo is to be plotted, `pwm`. This may be either an instance of class `pwm`, as defined by the package `seqLogo`, a `matrix`, or a `data.frame`.

2. A `logical ic.scale` indicating whether the height of each column is to be proportional to its information content, as originally proposed by Schneider et al. (1986). If `ic.scale=FALSE`, all columns have the same height.

EXAMPLE.

The call `seqLogo(p)` produces the sequence logo shown in figure 1. Alternatively, we can use `seqLogo(p, ic.scale=FALSE)` to obtain the sequence logo shown in figure 2 in which all columns have the same height.

2.3 Software Design

The following features of the programming approach employed in `seqLogo` may be of interest to users.

Class/method object-oriented programming. Like many other Bioconductor packages, `seqLogo` has adopted the *S4 class/method objected-oriented programming approach* presented in Chambers (1998). In particular, a new class, `pwm`, is defined to represent a position weight matrix. The plot method for this class is set to produce the sequence logo corresponding to this class.

Use of the `grid` package. The `grid` package is used to draw the sequence letters from graphical primitives. We note that this should make it easy to extend the package to amino acid sequences.

References

J.M. Chambers *Programming with Data: A Guide to the S Language*. Springer Verlag, New York, 1998.

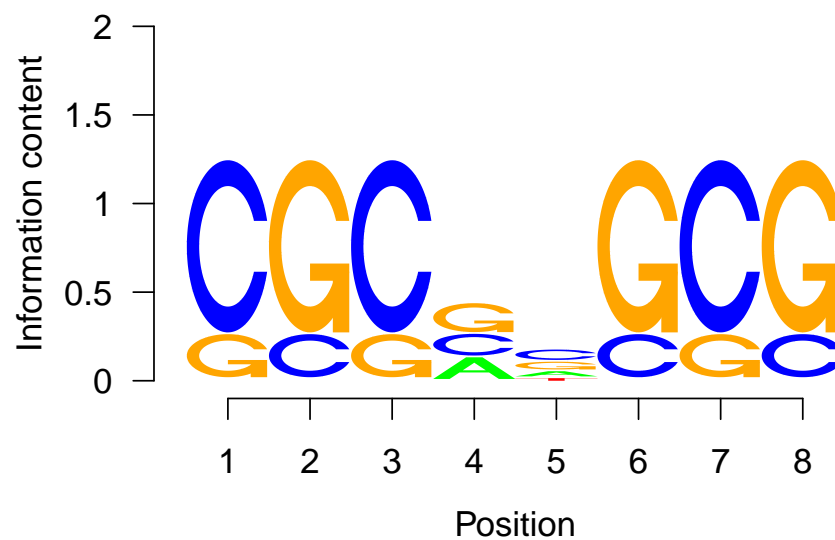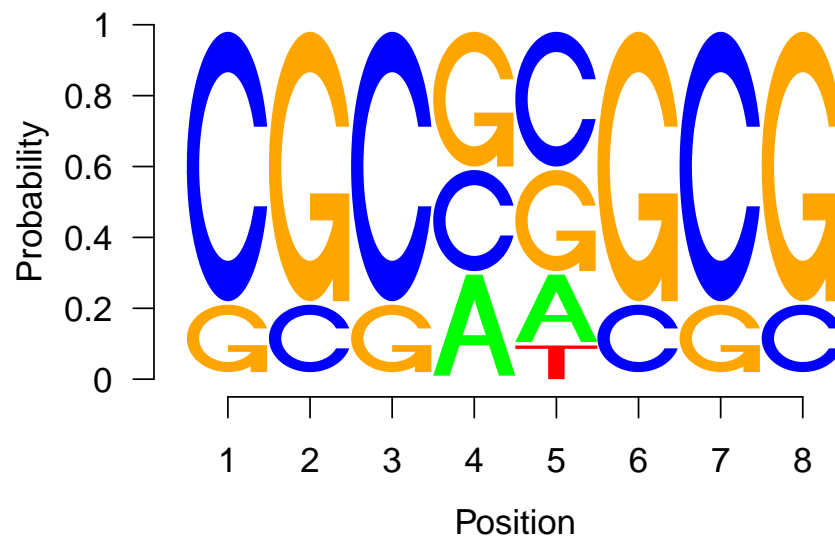Figure 1: Sequence logo with column heights proportional to information content.

Figure 2: Sequence logo with uniform column heights.

4

T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188:415–431, 1986.

T. D. Schneider, and R. R. Stephens. Sequence Logos: A New Way to Display Consensus Sequences *Nucleic Acid Research*, 18:6097–6100, 1990.