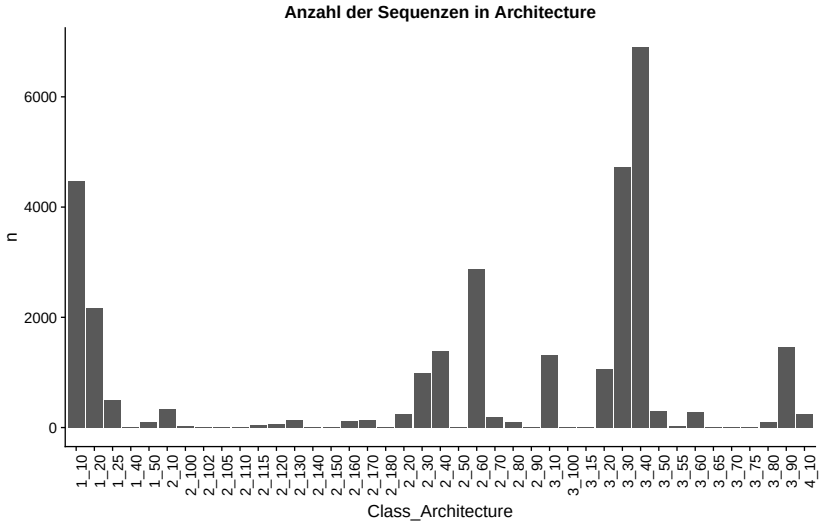
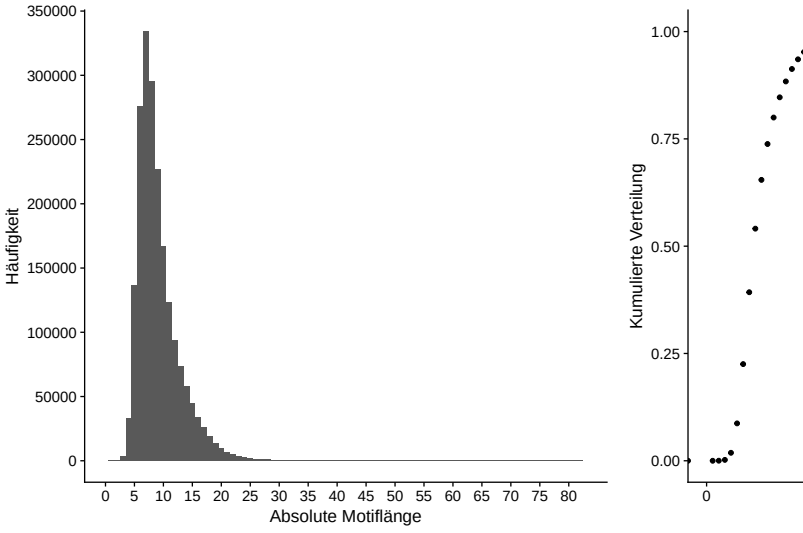
R Notebook

Wie balanciert ist der Datensatz zwischen den Klassen und zwischen den Architekturen?



Was lässt sich über die Länge der Motife auf den ersten Blick sagen?



Etwa 25% der Motife sind mehr als 20 Buchstaben lang.

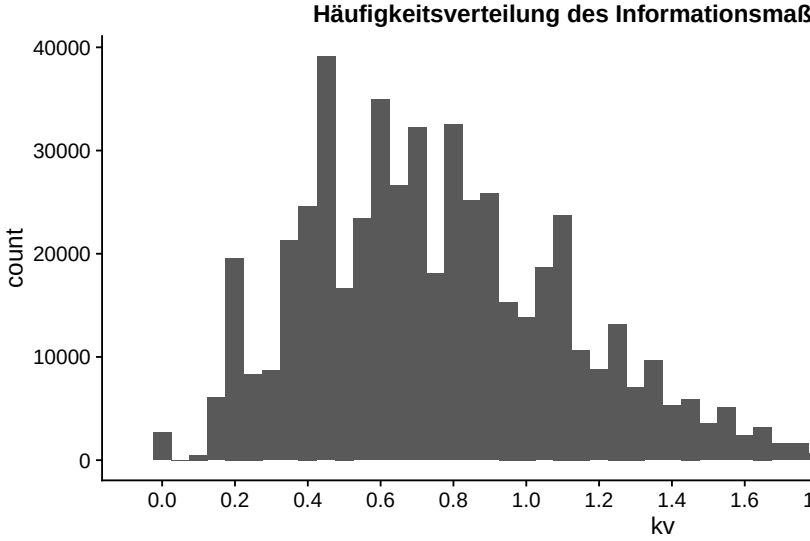
Welche Motife mit mehr als 15 zeichenketten kommen sowohl in Alpha-Klasse als auch in Beta-Klasse vor?

```
## # A tibble: 6 x 5
## # Groups:   motif [6]
##   motif             alpha  beta alpha.beta other
##   <chr>             <int> <int>      <int> <int>
## 1 agahhhhhhhhhhhhh     12     1         19    NA
## 2 agahhhhhhhhhhhhi      2     1          5    NA
## 3 agahhhhhhhhhhhhik     1     1          1    NA
## 4 agcaiihhhhhhhhh       2     1          2    NA
```

```
## 5 agcpmihhhhhhhhhh      1     1          NA    NA
## 6 aglajhhhhhhhhhh       1     1           1    NA
```

Berechne Informationsmaß für Motife bzw wie verzerrt sind die Motife?

Es lassen sich viele lange Sequenzen beobachten mit einem stark verzerrten Aufkommen weniger Buchstaben, z.B hhhhhhhhhhhhhhhhhhhhhhhhhhhh Wobei die Buchstaben I, J, H, K alpha-spezifischen Strukturen und F,



E, C, D sheet-spezifischen Strukturen zugeordnet sind.

Wie sehen "informative" Motife aus

```
## # A tibble: 6 x 5
##   motif                C     A     TO       kv
##   <chr>                <chr> <chr> <chr> <dbl>
## 1 efeeecaiimkhhhhh     3     40    190    1.84
## 2 ajhhhhikagcahhhiij   1     10    10     1.69
## 3 hhhhhhhhhhjogplede   3     20    20     1.53
## 4 mgnmlcageedajhhhhhiih 3    40    50     2.20
## 5 eeecajmpihhhhhhhh    3     40    630    1.66
## 6 hhhhhhhhhjiogcaji    1     20    58     1.51
```