[Re] Local alignment statistics - Amino acid frequencies

Nathan Brouwer

11/6/2019

Amino acid frequencies (Robinson & Robinson 1991)

Amino acid frequencies from Robinson and Robinson (1991) "Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins" PNAS. (R & R 1991). These frequencies are used by Altschul and Gish (1996) for their simulated polypeptides.

Make a vector of all letters that represent amino acid

```
# 1 letter codes
aa1 <- c("A", "C", "D", "E", "F", "G", "H",
         "I", "K", "L", "M", "N", "P", "Q",
         "R", "S", "T", "V", "W", "Y")


aa3 <- c("Ala","Cys","Asp","Glu","Phe","Gly","His",
         "Ile","Lys","Leu","Met","Asn","Pro","Gln",
         "Arg","Ser","Thr","Val","Trp","Tyr")
```

Number of each amino acid reported in (R & R 1991):

```
aa.count <- c(35155,8669,24161,28354,17367,
              33229,9906,23161,25872,40625,
              10101,20212,23435,19208,23105,
              32070,26311,29012,5990,14488)
```

Check frequency; should sum to 450431

```
aa.total <- sum(aa.count)
aa.total
```

```
## [1] 450431
```

Setting up amino acid relative frequencies

Create dataframe of amino acid frequencies (Note: stringsAsFactors = F prevents a default R behavior which is annoying)

```
robinson.aafreq <- data.frame(aa3,
                              aa1,
                              aa.count = aa.count,
                              stringsAsFactors = F)
```

Calculate **relative frequency** of each amino acid:

```
robinson.aafreq$aa.freq <- robinson.aafreq$aa.count/aa.total
```

Check that relative frequency sums to 1

```
sum(robinson.aafreq$aa.freq)
```

```
## [1] 1
```

Save

```
write.csv(robinson.aafreq, file = "robinson_aafreq.csv")
```