[Re] Local alignment statistics - simulating polypeptides

Nathan Brouwer

11/18/2019

Polypeptide simulation

Simulation of polypeptides is done using R's sample() function. This requires a vector of possible amino acids letters and their probabilities of occurring. The following code expalins the general principles.

Possible amino acids

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")
```

AA Frequencies

The frequencies below were derived from Robinson and Robinson. For further details see the associated script.

```
robinson.aafreq <- c(0.07804747,0.01924601,0.05363974,0.06294860,0.03855640,
                     0.07377157,0.02199227,0.05141964,0.05743832,0.09019139,
                     0.02242519,0.04487258,0.05202795,0.04264360,0.05129531,
                     0.07119847,0.05841294,0.06440942,0.01329837,0.03216475)
```

Select a single random amino acid

Example: Randomly select a single amino acid with equal frequency

```
sample(x = aa1, size = 1, replace = T)
```

```
## [1] "A"
```

Create a random polypeptide

Make a vector of 20 amino acids (eg, a simulated polypeptide), asumming equal frequencies

```
pp.length <- 20
sample(x = aa1, size = pp.length, replace = T)
```

```
##  [1] "A" "I" "F" "Y" "Q" "K" "K" "S" "I" "N" "A" "E" "I" "H" "D" "D" "W"
## [18] "F" "N" "A"
```

Make a vector of 20 amino acids (eg, a simulated polypeptide), asumming frequencies from Robinson and Robinson

```
pp.length <- 20
sample(x = aa1, size = pp.length, prob = robinson.aafreq, replace = T)
```

```
##  [1] "S" "A" "A" "Y" "F" "S" "Q" "H" "Y" "G" "G" "D" "D" "P" "C" "G" "S"
## [18] "P" "V" "T"
```