K-mer matching probabilities

Jacques van Helden

2015-10-22

Contents

Parameters

```
## Nucleotide matching probability
## (assuming equiprobability and independence !)
p <- 1/4  ## prior residue probability
G <- 3e9 ## Genome size
k <- 26  ## Read length
```

Concepts

Bernoulli process

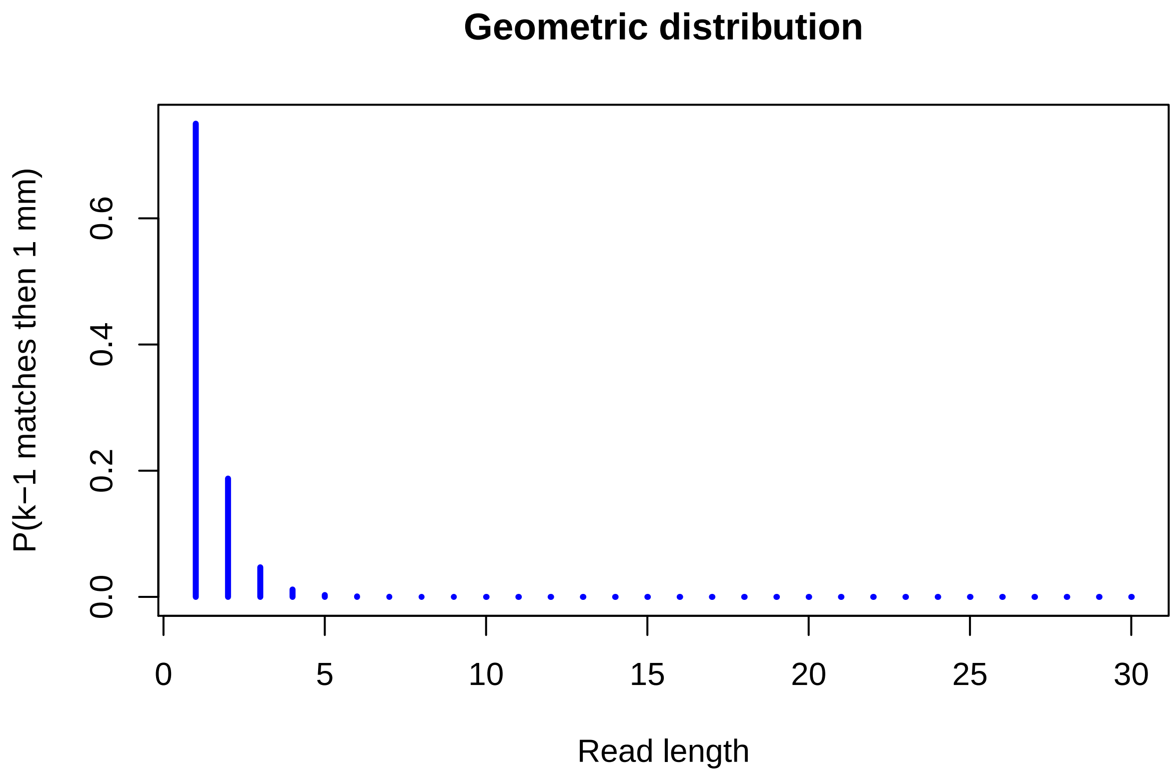
A Bernoulli process is defined as a succession of trials, where

1. each trial can result in two possible (and exclusive) outomes: success or failure;
2. successive trials are independent from each oter;
3. the probability of success (p) is constant.

Probability to observe a succession of $k-1$ matches followed by 1 mismatch

The geometric distribution describes the probability to observe $k-1$ successes followed by 1 failure, in a Bernoulli process.

```
x <- 1:30
dens.geo <- p^(x-1) * (1-p)
plot(x, dens.geo, type="h", lwd=3, col="blue",
     main="Geometric distribution",
     xlab="Read length", ylab="P(k-1 matches then 1 mm)")
```



Etape XX

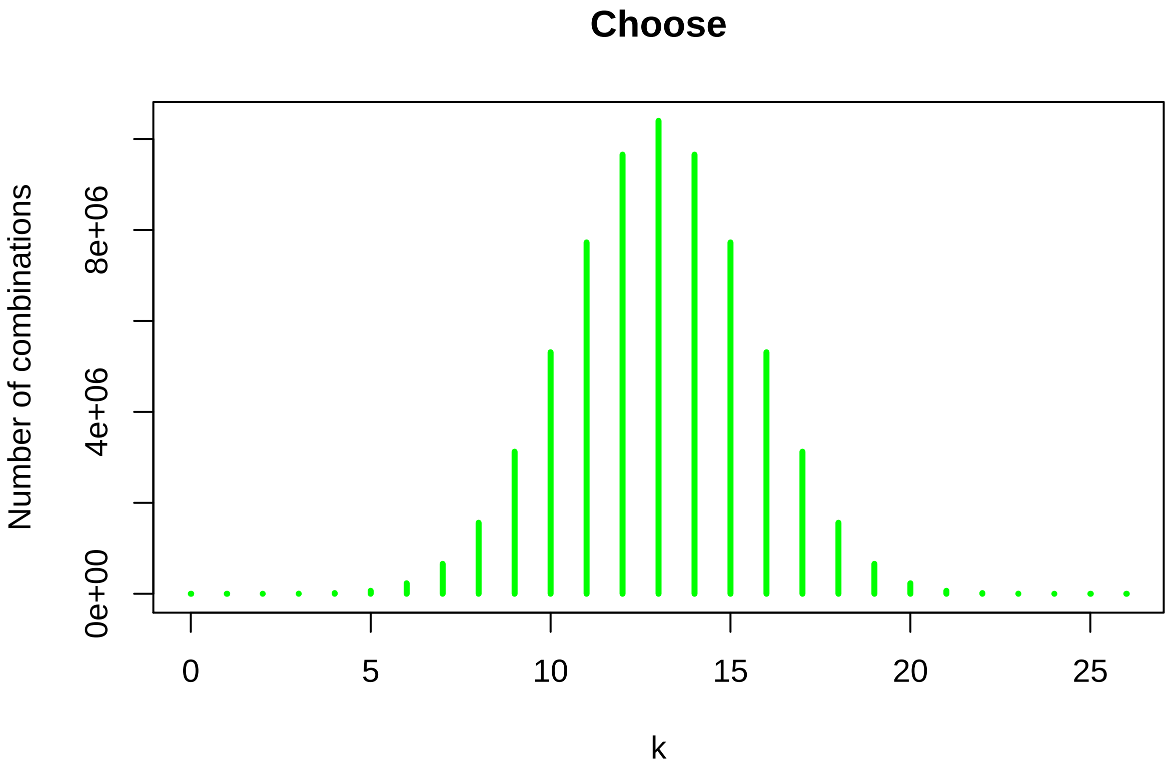
Etape XX

Etape XX

Etape XX

Etape XX

```
r plot(0:k, choose(k=0:k, n=k), lwd=3, col="green", type="h", main="Choose", xlab="k",
ylab="Number of combinations")
```



Etape 6: probability to observer exactly x matches and $k-x$ mismatches, at any position

$P(X=x) = C_k^x \cdot p^x \cdot (1-p)^{k-x}$