Discrete distributions, with applications in bioinformatics

Jacques van Helden

First version: 2016-12-10; Last update: 2016-12-10

Introduction

Perfect match probability

Binomial: global alignment with m mismatches

Negative binomial: local alignment with m mismatches

Introduction

Discrete probabilities and NGS

The advent of Next Generation Sequencing (**NGS**) technologies revived the importance of discrete distributions of probabilities for biologists.
This tutorial aims at providing a rapid overview of some discrete distributions commonly used to analyse NGS data, and highlight the relationship between them.

Overview

Distribution	Applications
Geometric	Local read mapping without mismatche (read extension until first mismatch)
Binomial	Global read mapping with a given number of mismatches
Negative binomial	Local read mapping with m mismatches (waiting time for $(m+1)^{th}$ mismatch); Detection of differentially expressed genes from RNA-seq data
Poisson	ChIP-seq peak calling
Hypergeometric	Enrichment of a set of differentially expressed genes for functional classes

Perfect match probability

Perfect match probability

We align a library of 50 million short reads of 25 base pairs onto a genome that comprises 23 chromosomes totalling 3 Gigabases. For the sake of simplicity, we assume that nucleotides are equiprobable and independently distributed in the genome. What is the probability to observe the following events by chance?

1. A perfect match for a given read at a given genomic position.
2. A perfect match for a given read anywhere in the genome (searched on two strands).
3. A perfect match for any read of the library at any position of the genome.
4. How many matches do we expect by chance if the whole library is aligned onto the whole genome?

Perfect match - parameters

Let us define the variables of our problem. Since we assume equiprobable and independent nucleotides we can define p as probability to observe a match by chance for a given nucleotide.

$$p = P(A) = P(C) = P(G) = P(T) = 0.25$$

```
k <- 25     # Read length
L <- 50e6   # Library size
C <- 23     # Number of chromosomes
G <- 3e9    # Genome size
p <- 1/4    # Matching probability for a nucleotide
```

Exercise: use these parameters to compute the matching probability for a read (*solution is on next slide*).

Perfect match for a given read at a given genomic position

Since we assume independence, the joint probability (probability to match all the nucleotides) is the product of the individual matching probabilities for each nucleotide.

```r
# Matching probabilty for a given read
# at a given genomic position
P.read <- p^k
```

$$P_{\text{read}} = P(n_1 \wedge n_2 \wedge \ldots \wedge n_k) = p^k = 0.25^{25} = 8.9e - 16$$

This looks a rather small probability. However we need to take into account that this risk will be challenged many times:

- the size of the genome (3 000 000 000)
- the size of the sequencing library (50 000 000)

Number of genomic alignments

The read will be aligned to each genomic position, but we should keep in mind the following facts.

1. For each chromosome, we will skip the last 24 positions, since a 25 bp read cannot be fully aligned there.
2. We double the number of alignments since we try to map the read on two strands.

$$N = 2\sum_{i=1}^{C}(L_i - k + 1) = 2\,(G - C(k-1))$$

```r
N <- 2 * (G - C * (k - 1))
```

In total, we will thus try to align each read on 5 999 998 896 genomic positions.

Genome-wise matching probability for one read

We reason in 3 steps, by computing the following probabilities.

Formula	Rationale
$1 - P_{\text{read}}$	no match at a given genomic position
$(1 - P_{\text{read}})^N$	not a single match in the genome
$1 - (1 - P_{\text{read}})^N$	at least one match in the genome

```
P.genomic <- 1 - (1 - P.read)^N
```

This gives $P_{\text{genomic}} = 0.00000533$.

Library-wise probability

We can apply the same reasoning for the library-wise probability.

Formula	Rationale
$1 - P_{\text{genomic}} = (1 - P_{\text{read}})^G$	no genomic match for a given read
$(1 - P_{\text{read}})^{GL}$	not a single genomic match in the library
$1 - (1 - P_{\text{read}})^{GL}$	at least one genomic match in the library

```r
P.library <- 1 - (1 - P.read)^(G*L)
```

This gives $P_{\text{library}} = 1$, which should however not be literally interpreted as a certainty, but as a probability so close to 1 that it cannot be distiguished from it.

Expected number of matches

The expected number of matches is the read matching probability mutliplie by the number of matching trials, i.e. $G \cdot L$ since each read will be matched against each genomic position.

$$E(X) = P_{read} \cdot N \cdot L$$

```
E <- P.read * N * L
```

In total, we expect 266 perfect matches by chance for the whole library against the whole genome.

Binomial: global alignment with m mismatches

Global alignment with mismatches

What is the probability to observe a global alignment with at most $m = 3$ mismatches for a given read of 25bp aligned on a particular genomic position?

This question can be formulated as a Bernoulli schema, where each nucleotide is a trial, which can result in either a success (nucleotide match between the read and the genome) or a failure (mismatch).

We can label each position of the alignment with a Boolean value indicating whether it maches (1) or not (0), as examplified below.

```
   ATGCG ACTAG CGTAC GACTG ACTAA
   10000 10000 11000 00000 00011
...AGCTC AGCTA CGACT ACGAC TACAA....
```

At each position, we have a probability of success $p = 0.25$, and a probability of failure $q = 1 - p = 0.75$.

Probability to observe exactly k matches

```
n <- 25      # Number of trials, i.e. the length of the ali
m <- 3       # Maximal number of accepted mismatches
p <- 1/4     # Matching probability for one nucleotide
```

Let us denote by k the number of matching residues. The probability to observe k successes in a Bernoulli schema with n trials and

$$P(X=k)=\mathcal{B}(k;n,p)=\binom{n}{k}p^k(1-p)^{n-k}=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

Remark: the perfect match probability seen above is a particular case of the binomial.

$$P(X=n)=\frac{n!}{n!0!}p^n(1-p)^{n-n}=p^n$$

Probability of hit with at least m mismatches

We can sum the probabilities for all possible values of matches from $k = n - m$ (m mismatches) to $k = n$ (no mismatch).

$$P(M \leq m) = \sum_{k=n-m}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial density

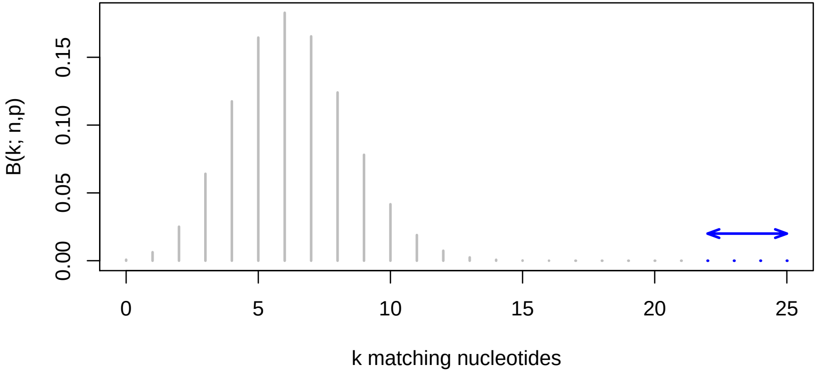

Figure 1: Binomial density function. Alignemnts with less than *m* mismatches are highlighted in blue.

Binomial distribution

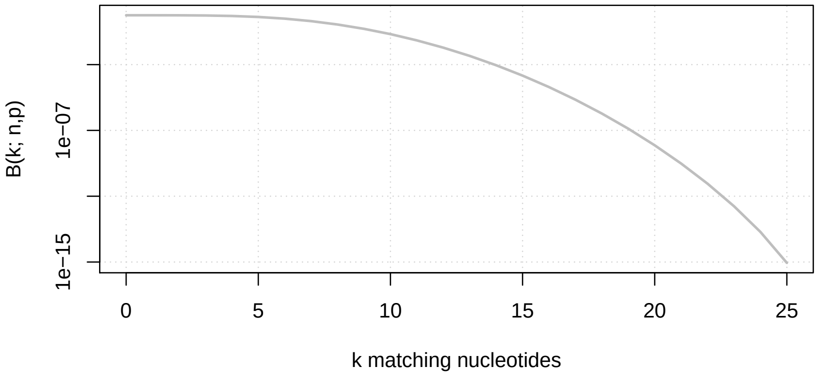

Figure 2: Binomial p-value. The X axis indicates the probability to obtain at least X matches by chance.

Negative binomial: local alignment with m mismatches