Exercises: discrete distributions

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2019-09-20

Exercises: discrete distributions

Exercise 04.1: probability of a motif with mismatches

We scan a genome to search all occurrences of the motif *GATAAG*, and we admit a given numnber of mismatches. Assuming equiprobable and independent nucleotides, what would be the probability to find, at a given genomic position:

a. An perfect match with the motif (not a signle mismatch)?
b. An alignement with not a single matching nucleotide (6 mismatches)?
c. An instance with exactly 1 substitution?
d. An instance with at most 2 substitutions?

Exercise 04.2: alignment of NGS sequences

After an experiment of *Next Generation Sequencing* (*NGS*), we dispose of a sequence library containing $N = 10^6$ short reads. We run a read mapping software to find the position of these sequences on a reference genome that contain $C = 20$ chromosomes totaling $G = 10^9$ base pairs. We use a gapless algorithm and we don't accept any mismatch.

We would like to compute the probability of a perfect alignment (not a single mismatch) for a particular short read at a given position of the genome, as a function of the read length (k).

- Which theoretical distribution would you use to model this problem? Justify your choice.
- Write the formula of the corresponding probability. ‘

Note: during the practicals we will draw this distribution with *R*.

Exercise 04.3: restriction sites

In a bacterial genome of 4Mb containing 50% $G + C$ we observe 130 occurrences of the hexanucleotide *GGCGCC*. We assume an equiprobable and independent distribution of nucleotides.

(a) What is the probability to observe one occurrence of *GGCGCC* at a given genomic position?

(b) How many occurrences would we expect to find in the whole genome?

(c) What is the probability to observe such a number of occurrences at least as weak as what we observe (at most 130 poccurrences) in a random sequence of the same size?

(d) Is there a biological explanation for this under-representation of the hexanucleotide *GGCGCC* in this bacterial genome?

Exercise 04.4: Jeu de roulette

La roulette comporte 37 nombres allant du 0 au 36. Un joueur a décidé de miser systématiquement 1 euro sur le nombre 17 jusqu'à ce que ce nombre sorte, et de s'arrêter ensuite.

Sachant que quand on mise sur un seul nombre, le gain vaut 36 fois la mise, quelle est la probabilité pour que le joueur sorte du casino en ayant gagné de l'argent?

Il n'est pas nécessaire de fournir une réponse numérique, vous pouvez vous contenter de fournir la formule, en indiquant les nombres correspondant aux différents symboles mathématiques. Justifiez votre réponse en expliquant votre raisonnement.

Exercise 04.5: probabilité des longueurs d'ORF

On détecte les cadres ouverts de lecture (*open reading frames*, *ORF*) d'un génome en identifiant toutes les séquences de taille multiple de 3 comprises entre un start (ATG) et un stop (TAA, TAG ou TGA).

a. Sur base des fréquences génomiques de trinucléotides, calculer la probabilité de trouver à une position donnée du génome un ORF d'au moins 100 codons.

b. Sachant que le génome fait 12 Mb, quel est le nombre attendu d'ORF d'au moins 100 codons?

sequence	frequency	occurrences
AAA	0.0394	478708
ATG	0.0183	221902
TAA	0.0224	272041
TAG	0.0129	156668
TGA	0.0201	244627

Exercise 04.6: mutagenèse

On soumet une librairie de molécules d'ADN de 1 kilobase à un traitement mutagène qui provoque un nombre moyen de 5 mutations ponctuelles (substitutions) par molécule.

a. Quelle est la probabilité d'avoir exactement 5 mutations pour une molécule donnée?
b. Quelle est la probabilité pour une molécule d'ADN de n'avoir subi aucune mutation au cours du traitement?
c. Quelle est la probabilité d'obtenir au moins 10 mutations ?

Formulation attendue pour la réponse.

- Expliquez le raisonnement qui vous permet de modéliser ce problème.
- Justifiez vos choix des hypothèses de travail.
- Ecrivez les formules avec les symboles, puis remplacez le symboles par les valeurs numériques correspondantes. Il n'est pas nécessaire de calculer la valeur finale (ceci nécessite un