Exercices: probabilités et statistique

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2017-10-05

Exercice: 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

Exercice: probabilité d'un motif avec erreurs

On recherche dans un génome les occurrences du motif GATAAG (où W signifie "A ou T") en admettant un certain nombre de substitutions. En supposant que les nucléotides sont indépendants et équiprobables, quelle est la probabilité de trouver à une position du génome:

a. Une instance exacte du motif (aucune substitution) ?
b. Une séquence ne présentant aucune correspondance avec le motif (6 substitutions) ?
c. Une instance avec exactement 1 substitution ?
d. Une instance avec au plus 2 substitutions ?

Exercice: alignement de lectures NGS sans erreur

Au terme d'un séquençage de type "Next Generation Sequencing" (NGS), on dispose d'une librairie de $N = 10^6$ lectures courtes. On aligne la librairie sur le génome de référence, dont la somme des chromosomes fait $G = 10^9$ paires de bases, en utilisant un algorithme d'alignement sans gap.

Calculez la distribution de probabilité du nombre de correspondances en fonction de la longueur des lectures (k).

Exercice: alignement de lecture NGS avec erreurs

Un biologiste a fait séquencer un échantillon et a obtenu un fichier comportant 50 millions de lectures (« short reads ») de 35 paires de base, qu'il aligne sur le génome humain (3 gigabases répartis sur 23 chromosomes). Durant l'alignement, il choisit d'accepter au maximum 3 substitutions par lecture.

a. En supposant un modèle de fond basé sur des nucléotides équiprobables et distribués de façon indépendante, comment calculeriez-vous la probabilité pour qu'un read s'aligne complètement à une position arbitraire du génome, avec au plus 3 substitutions (sans indel). Indiquez la formule et justifiez votre choix.

b. Sous ces mêmes conditions, quel serait le nombre de faux-positifs attendus si l'on aligne l'ensemble de la librairie de séquences sur l'ensemble du génome ?

Exercice: sites de restriction

Dans un génome bactérien de 4 Mb avec une composition de 50% de G+C, on observe 130 occurrences de l'hexanucléotide GGCGCC. On suppose un schéma de Bernoulli et une composition équiprobable de nucléotides.

(a) Quelle est la probabilité d'observer une occurrence de GGCGCC à une position donnée du génome ?
(b) Combien d'occurrences s'attend-on à trouver dans l'ensemble du génome ?
(c) Quelle serait la probabilité d'observer un nombre aussi faible d'occurrences (130 ou moins) si l'on générait une séquence aléatoire selon le modèle de Bernoulli avec nucléotides équiprobables ?
(d) Comment peut-on interpréter cette sous-représentation de l'hexanucléotide GGCGCC du point de vue biologique ?

Exercice: détection de différence d'expression

Un chercheur a analysé, à l'aide de biopuces, le niveau d'expression de l'ensemble des gènes à partir d'échantillons sanguins prélevés chez 50 patients (np=50) et chez 50 sujets témoins (nt=50). Il s'intéresse particulièrement à un gène qui semble montrer une différence entre les 2 groupes. Ainsi, il ré-analyse l'expression du même gène dans les mêmes échantillons en utilisant une autre technique, la qPCR. Il obtient

- pour les patients, une moyenne $m_p = 21$
- pour les contrôles, une moyenne $m_t = 10$
- des écarts-types identiques pour les 2 groupes $s_p = s_t = s = 15$

Afin de tester si la différence observée entre les moyennes est significative, le chercheur décide d'effectuer un test de Student.

a. Le choix du test de Student vous semble-t-il approprié ? Justifiez le choix du chercheur. Quelles auraient été les alternatives envisageables ?

b. Sachant qu'a priori on ne sait pas dans quel sens la maladie pourrait affecter le niveau d'expression du gène, préconisez-vous un test uni- ou bilatéral ?

c. Formulez l'hypothèse nulle et expliquez-la en une phrase.

d. Sur base de la formule ci-dessous, calculez la statistique t de Student.

e. Indiquez, en vous basant sur les tables fournies, la p-valeur correspondante.

f. Interprétez la p-valeur, et aidez le chercheur à tirer les conclusions de son étude.

Exercice: taux d'anticorps

Un groupe de chercheurs a détecté l'association, avec la résistance à la bilharziose, de taux élevés d'IgE spécifiques, une classe particulière d'anticorps. D'autres chercheurs ont cherché à répliquer ce résultat dans une population indépendante exposée à la bilharziose. Les résultats obtenus sont indiqués ci-dessous.

- Pour les sujets résistants ($n_r = 32$), la moyenne $m_r = 10$.
- Pour les sujets susceptibles ($n_s = 32$), la moyenne $m_s = 7$.

- Les écarts-types des deux groupes sont égaux : $s_r = s_s = s = 2.8$.

a. Quelle méthode préconisez-vous pour tester l'égalité des moyennes (justifiez) ? Quelles sont les hypothèses de travail de ce test ?

b. En partant du principe que ces conditions sont remplies dans le cas présent, formulez l'hypothèse nulle et calculez le score t de Student (formule ci-dessous). Enfin, estimez P valeur à partir de la table fournie.

c. A l'issue du test, quelle décision prenez-vous ? Justifiez votre réponse.

Exercice: enrichissement fonctionnel

Dans le génome de la levure, 40 gènes ont été assignés à la classe fonctionnelle "Biological Process: Methionin Biosynthesis". Une expérience de transcriptome rapporte 80 gènes différentiellement exprimés, dont 10 appartiennent à cette classe fonctionnelle. Sachant que le génome comporte 6000 gènes, peut-on consiérer ce résultat comme significatif ?

Exercice: 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 d'indiquer la formule, en indiquant les nombres correspondant aux différents symboles mathématiques. Justifiez votre réponse en expliquant votre raisonnement.

Exercice: concepts de probabilité

En quoi consiste le modèle de Bernoulli ? Ce modèle est-il généralement adapté à l'analyse des séquences biologiques ? Justifiez en quelques phrases.