Solutions des exercices: probabilités et statistique

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2017-12-25

Probabilité des longueurs d'ORF

Enoncé

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 (plus précisément, la probabilité qu'à cette position précise commence un cadre ouvert de lecture 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	0.039	478708
ATG	0.018	221902
TAA	0.022	272041
TAG	0.013	156668
TGA	0.020	244627

Approche

Cet exercice repose sur les concepts de probabilités combinées entre événements.

En particulier, nous mobiliserons les règles suivantes:

1. Des événements (A_1, A_2)sont mutuellement exclusifs si leur probabilité jointe est nulle ($P(A_1 \wedge A_2) = 0$).
2. La probabilité de l'union d'événements mutuellement exclusifs est la somme de leurs probabilités;

$$\mathrm{P}(A_1 \wedge A_2) = 0 \iff \mathrm{P}(A_1 \vee A_2) = \mathrm{P}(A_1) + \mathrm{P}(A_2)$$

3. Des événements sont **complémentaires** s'ils sont mutuellement exclusifs et exhaustifs (ils couvrent l'ensemble des possibilités). La somme des probabilité d'événements complémentaires vaut 1.

$$A_1, A_2, \ldots A_m \text{ complémentaires } \Rightarrow \mathrm{P}(A_1 \vee A_2 \vee \ldots \vee A_m) = \mathrm{P}(A_1) + \mathrm{P}(A_1) + \ldots + \mathrm{P}(A_m) = 1$$

4. Deux événements sont **stochastiquement indépendants** si la réalisation de l'un n'affecte pas la probabilité de réalisation de l'autre.

La **probabilité jointe** d'une série d'événements indépendants est le produit de leurs probabilités.

$$A_1, A_2, \ldots A_m \text{ indépendants } \Rightarrow \mathrm{P}(A_1 \wedge A_2 \wedge \ldots \wedge A_m) = \mathrm{P}(A_1)\,\mathrm{P}(A_2)\ldots\mathrm{P}(A_m)$$

Solution: probabilité des longueurs d'ORF

On sélectionne aléatoirement une position du génome (i). Nous allons raisonner en terme de codons (triplets de nucléotides): pour chaque position i on observe le triplet couvrant les nucléotides de i à $i+1$, et on avance par pas de 3 ($i, i+3, +6, \ldots$).

Pour y trouver le début d'un cadre ouvert de lecture d'au moins 100 codons,il faut remplir les conditions suivantes.

1. Présence d'un codon start (trinucléotide ATG) à la position i. La probabilité associée est directement fournie dans le tableau: $P(\text{ATG}) = 0.018$.
2. Absence de codon stop (trinucléotides TAA, TAG ou TGA) pour les 99 triplets suivants (positions).

Probabilité d'absence du codon stop à une position donnée

L'absence d'un codon stop est l'événement *complémentaire* de la présence d'un condon stop (on observe exclusivement l'un ou l'autre). On en déduit que la somme des probabilités (absence ou présence) vaut 1.

$P(\neg\text{STOP}) = 1 - P(\text{STOP})$

Le symbole $\neg$ représente la négation logique (logical NOT).

Calculons la probabilité de **présence d'un codon stop** à une position donnée. Pour cela, il faut observer à cette position l'un des trois triplets suivants: TAG, TGA, TAA. Ces événements sont *mutuellement exclusifs*: à une position donnée, on en peut pas observer *à la fois* un TAA et un TGA. La probabilité d'observer l'un d'entre eux (probabilité de leur union) est donc la somme des probabilités individuelles.

$$\begin{aligned} P(\text{STOP}) &= P(\text{TAA}) + P(\text{TAG}) + P(\text{TGA}) \\ &= 0.022 + 0.013 + 0.02 = 0.055 \end{aligned}$$

On en détuit la probabilité d'**absence d'un codon stop** à une position donnée.

$$P(\neg\text{STOP}) = 1 = P(\text{STOP} = 1 - 0.06 = 0.94$$

Probabilité de présence d'ORF

Nous pouvons maintenant calculer la probabilité d'avoir un cadre ouvert de lecture d'au moins 100 codons qui commence à une position arbitraire i du génome.

Ceci correspond à la *probabilité jointe* (le ET logique) de l'ensemble des conditions requises.

- présence d'un codon start à la position i,
- ET absence de codon stop en position $i+3$,
- ET absence de codon stop en position $i+6$,
- ET ...
- ET absence de codon stop en position $i+197$.

On peut considérer que ces événements sont *indépendants* (le codon observé en position $i+3$ ne dépend pas de celui observé en position i). Leur probabilité jointe est donc le produit des probabilités individuelles.

$$\begin{aligned} P(\text{ORF}_{100}) &= P(\text{START}) \cdot \overbrace{P(\neg\text{STOP}) \cdot \ldots \cdot P(\neg\text{STOP})}^{\text{99 times}} \\ &= P(\text{START}) \cdot P(\neg\text{STOP})^{99} \\ &= 0.018 \cdot 0.945^{99} = 6.653 \times 10^{-5} \end{aligned}$$

A priori cette probabilité n'a pas l'air énorme. Cependant, il faut tenir compte du fait qu'en annotant un génome on considère successivement toutes les positions possibles pour évaluer si on y trouve un cadre ouvert de lecture.

Nombre attendu d'ORF d'au moins 100 codons

Pour calculer le nombre d'ORF attendus au hasard, il faut multiplier la probabilité d'observer un ORF à une position donnée par le nombre de positions considérées.

Pour un génome de 12 Mb ($G = 1.2 \times 10^7$), on va considérer toutes les positions sur chacun des deux brins (on multiple donc par deux la taille du génome).

$$E(\mathrm{ORF}_{100}) = P(\mathrm{ORF}_{100}) \cdot 2 \cdot G = 6.653 \times 10^{-5} \cdot 2 \cdot 1.2 \times 10^7 = 1596.7$$

Interprétation du résultat

En 1996, lors de la première vague d'annotation du génome de la levure du boulanger, la stratégie d'annotation consistait à predire un gène codant chaque fois qu'on détectait un cadre ouvert de lecture d'au moins 100 codons. Les chercheurs étaient conscients du fait que cette stratégie était susceptible de produire un nombre important de fausses prédictions, mais il s'agissait du tout premier génome eucaryote séquencé, et il fallait bien commencr par quelque chose pour se donner une chance d'y découvrir de nouveaux gènes.

Nous venons de calculer le nombre attendu de faux-positifs dans ce processus: si on avait généré une séquence aléatoire de la même taille selon un modèle de Markov respectant les mêmes fréquences de trinucléotides, on s'attendrait à y trouver 1596.7 cadres ouverts de lecture d'au moins 100 codons. Ceci indique que les gènes de taille relativement courte (300 nucléotides) devaient être considérés avec circonspection, et soumis à analyse ultérieure avant d'être annotés comme des gènes fiables.

En 2003, une stratégie complémentaire a été mise à contribution pour évaluer la fiabilité des ORF prédits dans le génome de *Saccharomyces cerevisiae*, en s'appuyant sur la génomique comparative: après avoir séquençé les génomes de quelques autres espèces du même genre (*Saccharomyces*), on a testé la conservation des ORF initialement prédits, en supposant que si on trouvait des ORF similaires dans les 4 espèces, cela suggérait une pression sélective positive pour l'absence de codon stop dans ces régions, et renforçait la prédiction d'un gène codant correspondant à ce cadre ouvert de lecture.

Probabilité d'un motif avec erreurs

On recherche dans un génome les occurrences du motif GATAAG 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 ?

Cadre théorique

On modélise le problème comme un **schéma de Bernoulli**: chaque position de la séquence correspond à un essai, qui peut donner soit un *succès* (correspondance avec motif) soit un *échec* (différence avec le motif), et on suppose que la probabilité de succès est constante (nucléotides indépendants et équiprobables).

Le motif GATAAG fait 6 nucléotides, on aura donc 6 essais successifs, qui viseront à tester successivement l'identité entre le premier, second, ... sixième nucléotide de la séquence et le nucléotide à la position correspondante du motif. Pour chaque position, on a une probabilité de succès $p = 1/4$ et d'échec $q = 1 - p = 3/4$.

Probabilité d'nstance exacte du motif (aucune substitution)

Pour avoir une instance exacte du motif, il faut une succession de 6 correspondances entre nucléotides.

$$P(\text{perfect match}) = p^6 = 0.25^6 = 2.4 \times 10^{-4}$$

Probabilité d'une séquence ne présentant aucune correspondance avec le motif (6 substitutions)

$$P(\text{full mismatch}) = (1-p)^6 = 0.75^6 = 0.178$$

Probabilité d'une instance avec exactement 1 substitution

On peut trouver la réponse par un raisonnement assez simple. Les instances avec 1 substitutions combinent - 5 "succès"" (correspondances) ayant une probabilité $p = 0.25$, - 1 "échec" (substitition) $q = 1 - p = 0.75$.

La substitution peut se trouver à n'importe quelle position du motif. Si l'on représente respectivement par 1 et 0 les succès et échecs, on acceptera les disposition suivantes de succès et d'échecs.

Disposition	Probabilité
111110	$p \cdot p \cdot p \cdot p \cdot p \cdot (1-p) = p^5(1-p)$
111101	$p \cdot p \cdot p \cdot p \cdot (1-p) \cdot p = p^5(1-p)$
111011	$p \cdot p \cdot p \cdot (1-p) \cdot p \cdot p = p^5(1-p)$
110111	$p \cdot p \cdot (1-p) \cdot p \cdot p \cdot p = p^5(1-p)$
101111	$p \cdot (1-p) \cdot p \cdot p \cdot p \cdot p = p^5(1-p)$
011111	$(1-p) \cdot p \cdot p \cdot p \cdot p \cdot p = p^5(1-p)$

Ces dispositions sont mutuellement exclusive (si on a exactement une substitution, elle peut se trouver soit à la première soit à la deuxième position, mais pas aux deux en même temps). On peut donc calculer la prbabilité de l'union des 6 dispositions comme la somme de leurs probabilités.

$$P(\text{match withb 1 subst}) = 6 \cdot p^5 \cdot (1-p) = 6 \cdot 0.25^5 \cdot 0.75 = 0.0044$$

Un raisonnement plus rapide: il s'agit d'un schéma de Bernoulli avec 6 essais indépendant, ayant la même probabilité de succès $p = 0.25$. La probabilité d'observer exactement $x = 5$ succès parmi $n = 6$ essais est donnée par la **distribution binomiale**.

$$\begin{aligned} P(x = 5 | n = 6, p = 0.25) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{6}{5! \cdot 1!} 0.25^5 \cdot 0.75 = 0.0044 \end{aligned}$$

Probabilité d'une instance avec au plus 2 substitutions

Pour avoir au plus deux substitutions, il faut avoir soit 0, soir 1, soit 2 subsitution. Ces possibilités sont mutuellement exclusives (on ne peut avoir à la fois 1 et 2 substitutions), la probabilité de leur union est donc la somme de leurs probabilités.

Nous avons calculé ci-dessus la probabilité d'avoir 0 et 1 substitution. Par le même raisonnement, on peut calculer la prbabilité de 2 substitutions avec le distribution binomiale.

$$\begin{aligned} P(x = 4|n = 6, p = 0.25) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{6}{4! \cdot 2!} 0.25^4 \cdot 0.75^2 = 0.033 \end{aligned}$$

La solution peut s'écrire sous la forme de 3 probabilités binomiales correspondant respectivement à 0, 1 et 2 substitutions, ou, de façon équivalente, à 4, 5 ou 6 succès ($x \geq 4$).

Note: afin de vous familiariser avec les deux formulations alternatives du coefficient binomial, nous utilisons ici la formule "choose" C_n^x, qui est équivalente à $\binom{n}{x}$ (notez l'inversion des positions de x et n).

$$\begin{aligned} P(i \geq 4|n = 6, p = 0.25) &= \sum_{i=4}^{n} C_n^i p^i (1-p)^{n-i} \\ &= \sum_{i=4}^{6} \frac{6}{i! \cdot (n-i)!} 0.25^i \cdot 0.75^{n-i} \\ &= 0.033 + 0.0044 + 2.4 \times 10^{-4} \\ &= 0.038 \end{aligned}$$

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).

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 ?

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 ?

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.

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.

Méthode préconisée pour tester l'égalité des moyennes

Pour choisir un test d'égalité des moyennes, la première question est de savoir si l'on peut ou non appliquer un test paramétrique.

A priori nnous ne disposons pas d'information concernant la distribution des taux d'IgE dans les populations résistantes et susceptibles. Nous ne pouvons donc pas présupposer que les données suivent une distribution normale.

Cependant, les deux échantillons sont de taille suffisante ($n_1 = n_2 = 32$) pour qu'on puisse s'affranchir de cette condition. En effet, même si les données ne suivent pas une distribution normale, la moyenne d'échantillons tend vers la normalité quand l'effectif augmente, et, de façon pragmatique, on considère généralement qu'au-delà de 30 éléments par échantillon on peut appliquer un test paramétrique.

Parmi les tests paramétriques, on choisira soit un test t de Student si les deux populations sont de même variance (*homoscédasticité*), soit un test t de Welch si leurs variances diffèrent (*hétéroscédasticité*). Comme dans notre cas les écarts-types des deux échantillons sont identiques, on peut présupposer que les populations dont ils sont extraits ont la même variance (en tout état de cause, si l'on réalisait un test d'égalité des variances, on serait amené à accepter l'hypothèse nulle).

Hypothèse nulle

On effectue un **test unilatéral** puisqu'on teste specifiquement l'association de la résistance à un taux élevé d'anticorps.

$$H_0 : \mu_R \leq \mu_S H_1 : \mu_R > \mu_S$$

où μ_R et μ_S sont respectivement les taux moyens d'anticorps chez les individus résistants et sensibles.

Calcul du score t de Student

$$\begin{aligned} t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} &= \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{10 - 7}{\sqrt{\frac{32 \cdot 2.8^2 + 32 \cdot 2.8^2}{32 + 32 - 2}\left(\frac{1}{32} + \frac{1}{32}\right)}} \\ &= \frac{10 - 7}{2.8 \cdot \sqrt{\frac{64}{62}}\sqrt{\frac{1}{32} + \frac{1}{32}}} \\ &= \frac{3}{2.8 \cdot 1.016 \cdot \frac{1}{4}} = 4.22 \end{aligned}$$

Estimation de la probabilité critique (p valeur)

On se base sur le tableau de la distribution t de Student (disponible pour l'examen), en considérant le test unilatéral (*one-tail*).

On sélectionne la ligne correspondant aux degrés de liberté $\mathrm{df} = n_1 + n_2 - 2 = 62$). Comme le tableau ne mentionne pas cette valeur, on choisit la ligne la plus proche ($\mathrm{df} = 60$).

La valeur t observée ($t_{obs} = 4.22$) est supérieure à la plus grande valeur indiquée dans la table ($t = 3.460$, pour $p = 0.0005$). On peut donc conclure que la probabilité critique est inférieure à 0.0005.

Décision et jusification

Pour un test isolé, une probabilité critique de $p = 0.0005$ signifie que sous hypothèse nulle – c'est-à-dire si la résistance n'était pas associée à un taux élevé d'anticorps – la probabilité d'obsrver une différence de moyennes aussi grande serait de 0.0005. Cette probabilité est très faible, on peut donc rejeter l'hypothèse nulle, et conclure que la différence observée est significative.

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 ?

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.

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.