TP1. Génotypage probabiliste – Corrigé

Nicolas Lartillot
nicolas.lartillot@univ-lyon1.fr

25 janvier 2017

L'objectif de ce premier TP est d'implémenter une méthode simple de génotypage d'individus, à partir de données de séquençage massif. Un ensemble de 14 individus de l'espèce *Silene vulgaris* (Silène commun, plante herbacée vivace), ont été échantillonnés dans divers environnements naturels, et leur transcriptome a été séquencé. Les données de séquence se présentent sous la forme de lectures très courtes (de l'ordre de la centaine à quelques centaines de paires de bases), qui sont ensuite alignées sur un génome de référence.

Données de séquences alignées

Le fichier `contig1262.ali` vous montre un exemple de données qui sont typiquement obtenues à l'issue de ces étapes préliminaires. Prenez le temps d'examiner ce fichier, qui contient un alignement pouvant être visualisé, par exemple, à l'aide du programme `seaview`.

La séquence correspondant à une région particulière du génome de référence (un *contig*) est donnée (première séquence dans le fichier). Sur cette séquence sont alignées un ensemble de lectures provenant des 14 individus. Chaque individu possède un identifiant (par exemple, `SV17`), qui sert de préfixe à l'ensemble des lectures qui lui correspondent. Comme vous pouvez le voir, le nombre de lectures par individu est variable (tous les individus ne sont d'ailleurs pas représentés, le nombre de lectures pouvant être égal à 0 pour certains).

Données de comptage

En chaque position de cet alignement, on peut compter, pour chaque individu, le nombre de lectures présentant les bases, A, C, G, ou T. Ce comptage a été effectué pour un grand nombre de contigs, et les données ont été rassemblées dans le fichier `data10000.txt`. À noter que, afin de se concentrer sur des données de meilleures qualités, on n'a conservé dans ce fichier que les positions pour lesquelles au moins 3 individus possèdent chacun au moins 10 lectures. Ici, on vous donne un échantillon de 10000 positions (sur un total de plusieurs millions).

Prenez le temps de regarder le fichier `data10000.txt`. Chaque ligne correspond à une position dans le génome. Par ailleurs, le fichier possède $4 \times 14 = 56$ colonnes (57 en comptant la première indiquant la position), donnant, pour chaque individu successivement, les comptages pour les bases A, C, G et T en cette position.

Dans la suite, on note N le nombre total de positions, et P le nombre d'individus (ici, $N = 10000$ et $P = 14$). L'indice $i = 1..N$ sera utilisé pour les positions, et l'indice $j = 1..P$ pour les individus. On utilisera les indices a, b, ou $c = A, C, G, T$ pour les 4 nucléotides. On note enfin :

- $n_{ij}(a)$, pour $a = A, C, G, T$, le nombre de lectures en la position i et pour l'individu j qui ont la base a. Ces comptages sont donc exactement les valeurs tabulées dans le fichier `data10000.txt` ;
- $n_{ij} = (n_{ij}(a))_{a=A,C,G,T}$ est le vecteur de dimension 4 correspondant aux comptages pour la position i et l'individu j ;
- $n_{ij}^{tot} = \sum_{a=A,C,G,T} n_{ij}(a)$ est le nombre total de lectures pour l'individu j en la position i ;
- $m_i(a) = \sum_{j=1..P} n_{ij}(a)$ est le nombre total d'occurrences de la base $a = A, C, G, T$ en la position i, à travers l'ensemble des individus.

Question 1 *Sur la base des données de comptage, donnez une estimée des fréquences des 4 bases $\hat{\pi}_a$, pour $a = A, C, G, T$ en la position i et dans la population globale des Silènes.*

Réponse : en ajoutant des pseudo-comptes :

$$\hat{\pi}_a \quad = \quad \frac{m_a + 1}{\sum_b m_b + 1}.$$

La Silène est un organisme diploide. En chaque position, on énumère l'ensemble des génotypes possibles comme suit : $(AA, AC, AG, AT, CC, CG, CT, GG, GT, TT)$. Il existe donc 10 génotypes. Dans la suite, on notera ab le génotype formé des deux bases a et b (avec $a \leq b$). Pour effectuer les calculs, on utilisera également une indexation par $g = 1..10$, dans l'ordre spécifié ci-dessus. Par ailleurs, on fait l'hypothèse que la population est à l'équilibre de Hardy et Weinberg.

Question 2 *Étant donné les fréquences alléliques $\pi(a)$, $a = A, C, G, T$ en une position, quelle est la probabilité du génotype homozygote aa ? Quelle est la probabilité d'un génotype hétérozygote de type ab, avec $b \neq a$?*

homozygote : $p(aa) = \pi_a^2$ **;**
hétérozygote : $p(ab) = 2\pi_a\pi_b$**, pour** $b \neq a$**.**

Modèle d'erreurs de séquençage

Les méthodes de séquençage massif actuellement utilisées ont un taux d'erreur qui se situe entre 0.1% et 1% (selon la méthode utilisée). On considère ici un modèle d'erreur très simple : en chaque position, et indépendamment du contexte, la probabilité qu'une lecture comporte une erreur est de ϵ (avec ϵ compris entre 0.001 et 0.01). Si une erreur a été commise, alors n'importe laquelle des 3 autres bases a été lue à la place de la vraie base, chacune avec probabilité 1/3. Le modèle ne comporte donc que le paramètre inconnu ϵ. Dans un premier temps, on fixera $\epsilon = 0.01$. On estimera ce paramètre dans un second temps. La probabilité de lire la base b alors que la vraie base est a est donc donnée par les formules suivantes :

- $p(b \mid a) = \epsilon/3$ pour $b \neq a$,
- $p(b \mid a) = 1 - \epsilon$ si $b = a$.

Considérons maintenant un individu donné, en une position du génome, et considérons une lecture qui s'aligne sur cette région du génome. En la position, la lecture possède la base c.

Question 3 *Étant donné le modèle d'erreur présenté ci-dessus, quelle est la probabilité de lire la base c, sachant que le vrai génotype est homozytgoe aa ? (Considérer séparément les cas $c = a$ et $c \neq a$.)*

De façon générale, (pour a, b et c quelconques), on fait l'hypothèse que la lecture s'est faite sur le chromosome portant l'allèle a ou b avec probabilité $1/2$. On fait donc la moyenne des deux probabilités d'émission, sachant que l'on est en train de lire a ou b :

$$p(c \mid ab) = \frac{1}{2}p(c \mid a) + \frac{1}{2}p(c \mid b)$$

On peut ensuite spécialiser cette formule au cas homozygote $a = b$:

$$p(c \mid aa) = p(c \mid a)$$

Et donc, si $c = a$:

$$p(a \mid aa) = p(a \mid a) = 1 - \epsilon$$

et si $c \neq a$:

$$p(c \mid aa) = p(c \mid a) = \epsilon/3$$

Question 4 *Étant donné le modèle d'erreur présenté ci-dessus, quelle est la probabilité de lire la base c, sachant que le vrai génotype est homozytgoe ab, avec $b \neq a$? (Considérer séparément les cas $c = a$, $c = b$, puis $c \neq a, b$.)*

On spécialise la formule donnée plus haut au cas hétérozygote. Dans le cas $c = a$:

$$p(a \mid ab) = \frac{1}{2}p(a \mid a) + \frac{1}{2}p(a \mid b) = \frac{1}{2}(1 - \epsilon) + \frac{1}{2}\frac{\epsilon}{3}$$

Le cas $c = b$ donne le même résultat (par symétrie). Enfin, le cas $c \neq a$ et $c \neq b$:

$$p(c \mid ab) = \frac{1}{2}p(c \mid a) + \frac{1}{2}p(c \mid b) = \frac{1}{2}\frac{\epsilon}{3} + \frac{1}{2}\frac{\epsilon}{3} = \frac{\epsilon}{3}$$

Calcul de la vraisemblance et de la probabilité a posteriori

Revenons maintenant à nos données de comptage. On rappelle qu'en la position i et pour l'individu j, on observe le vecteur de comptage n_{ij}, de dimension 4.

Question 5 *Quelle est la probabilité $p(n_{ij} \mid aa)$ d'obtenir les comptages n_{ij}, à supposer que le vrai génotype de l'individu j en cette position est aa ? Exprimer cette probabilité uniquement en fonction de ϵ, n_{ij}^{tot} et $n_{ij}(a)$.*

Réponse :

$$p(n_{ij} \mid aa) \quad = \quad (1-\epsilon)^{n_{ij}(a)} \left(\frac{\epsilon}{3}\right)^{n_{ij}^{tot}-n_{ij}(a)}$$

Question 6 *Quelle est la probabilité $p(n_{ij} \mid ab)$ d'obtenir les comptages n_{ij}, à supposer que le vrai génotype de l'individu j en cette position est ab, avec $b \neq a$? Exprimer cette probabilité uniquement en fonction de ϵ, n_{ij}^{tot}, $n_{ij}(a)$ et $n_{ij}(b)$.*

Réponse :

$$p(n_{ij} \mid ab) \quad = \quad \left(\frac{1}{2}(1-\epsilon) + \frac{1}{2}\frac{\epsilon}{3}\right)^{n_{ij}(a)+n_{ij}(b)} \left(\frac{\epsilon}{3}\right)^{n_{ij}^{tot}-n_{ij}(a)-n_{ij}(b)}$$

Question 7 *Quelle est la probabilité marginale $p(n_{ij} \mid \pi_i)$ d'obtenir les comptages n_{ij} chez un individu tiré aléatoirement de la population (sachant les fréquences alléliques $\pi_i = (\pi_i(a))_{a=A,C,G,T}$) ?*

Cette probabilité est la somme sur tous les génotype possibles de la probabilité a priori du génotype (sachant les fréquences alléliques π), multipliée par la probabilité d'obtenir les comptages n_{ij} sachant le génotype :

$$p(n_{ij} \mid \pi) \quad = \quad \sum_{a \leq b} p(n_{ij} \mid ab) p(ab)$$

Dans cette équation, le facteur $p(ab)$ est donné par la réponse à la question 2. De même, le facteur $p(n_{ij} \mid ab)$ est donné par les réponses aux questions 5 et 6 (selon que $a = b$ ou $a \neq b$).

Question 8 *Exprimez la probabilité a posteriori du génotype ab (homozgote ou hétérozygote) en la position i et pour l'individu j : $p(ab \mid n_{ij}, \pi_i)$.*

On applique le théorème de Bayes :

$$p(ab \mid n_{ij}) \quad = \quad \frac{p(n_{ij} \mid ab) p(ab)}{\sum_{a \leq b} p(n_{ij} \mid ab) p(ab)}$$

Que l'on peut écrire, en utilisant la question 7 :

$$p(ab \mid n_{ij}) \quad = \quad \frac{p(n_{ij} \mid ab) p(ab)}{p(n_{ij} \mid \pi)}$$

À noter que la probabilité marginale dépend implicitement de ϵ. En prenant le produit de la probabilité marginale sur tous les individus et toutes les positions, on obtient la vraisemblance totale du modèle :

$$L(\epsilon) \quad = \quad \prod_{i=1..N} \prod_{j=1..P} p(n_{ij} \mid \pi_i)$$

ou, en logarithme :

$$\ln L(\epsilon) \quad = \quad \sum_{i=1..N} \prod_{j=1..P} \ln p(n_{ij} \mid \pi_i)$$

Cette vraisemblance sera utilisée dans la suite pour estimer ϵ (par maximum de vraisemblance).

Programmation en R

Lancez R et chargez le fichier data10000.txt

```
# read data: gives a data frame
data <- read.table("data10000.txt", header=TRUE)
# transform into an array
y <- as.vector(data)
# number of columns is 4*number of individuals + 1 (here, 57)
l <- length(y[1,])
# define number of individuals
nind <- (l-1)/4
#define number of positions
npos <- length(y[,1])
#define the count matrix (the n_ij vectors)
countmatrix <- y[,2:l]
```

À noter que, pour aller plus vite, dans un premier temps, il peut être utile de fixer N (ou npos) à 1000. Une fois le programme opérationnel, on pourra alors utiliser la valeur de N correspondant au nombre total de positions présentes dans les données.

Question 9 *Écrire un petit programme qui calcule le tableau des valeurs de $m_i(a)$, pour $i = 1..N$ et $a = A, C, G, T$, ainsi que les estimées des fréquences alléliques en la position i, $\pi_i(a)$.*

Question 10 *De même, écrire un petit programme qui calcule le tableau des valeurs de n_{ij}^{tot}, pour $i = 1..N$ et $j = 1..P$*

```
# m_i(a)
basecount <- array(1,dim=c(npos,4))

# n_ij^tot
indcount <- array(0,dim=c(npos,nind))

for (i in 1:npos)       {
        index <- 1
        for (j in 1:nind)       {
                for (a in 1:4)  {
                        basecount[i,a] <- basecount[i,a] + countmatrix[i,index]
                        indcount[i,j] <- indcount[i,j] + countmatrix[i,index]
                        index <- index+1
                }
        }
}

pi <- array(0,dim=c(npos,4))
for (i in 1:npos)       {
```

```
        pi[i,] <- basecount[i,] / sum(basecount[i,])
}
```

Question 11 *Pour chaque position et chaque individu, calculer les probabilités a priori et les vraisemblances de chaque génotype. Calculer la probabilité marginale puis les probabilités a posteriori de chaque génotype.*

Question 12 *Pour chaque position et chaque individu, choisir le génotype le plus probable a posteriori*

Réponse : voir script `genotyper1.R`. À noter que le calcul de la probabilité a posteriori occasionne parfois des erreurs numériques. Une version plus robuste est fournie dans le script `genotyper2.R`.

Question 13 *En sommant le logarithme de la probabilité marginale sur tous les individus et toutes les positions, calculer le log de la vraisemblance. Effectuez ce calcul pour toutes les valeurs de ϵ : 0.001, 0.002, ... 0.009, 0.01. Quelle valeur de ϵ semble-t-il indiqué d'utiliser ?*

Réponse : voir script `genotyper3.R`. Au bout du compte, la valeur maximisant la vraisemblance est aux alentours de $\epsilon = 0.008$.