TP1. Génotypage probabiliste

Nicolas Lartillot
nicolas.lartillot@univ-lyon1.fr

1er février 2018

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 10 individus d'une espèce bactérienne *Agrobacterium* ont été échantillonnés dans divers environnements naturels et leur génome a été séquencé. Les données de séquence se présentent sous la forme de lectures très courtes (de l'ordre de la centaine de paires de bases), qui sont ensuite alignées sur un génome de référence.

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 `gencounts.txt`. Ici, on vous donne un petit échantillon de 100 positions (sur un total de plusieurs millions).

Prenez le temps de regarder le fichier `gencounts.txt`. Chaque ligne correspond à une position dans le génome. Par ailleurs, le fichier possède $4 \times 10 = 40$ colonnes, 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, $N = 100$ et $P = 10$). 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 :

- $k_{ij}(a)$, 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 `gencounts.txt` ;
- $k_{ij} = (k_{ij}(a))_{a=A,C,G,T}$ est le vecteur de dimension 4 correspondant aux comptages pour la position i et l'individu j ;
- $n_{ij} = \sum_{a=A,C,G,T} k_{ij}(a)$ est le nombre total de lectures pour l'individu j en la position i ;
- $m_i(a) = \sum_{j=1..P} k_{ij}(a)$ 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 formule exprimant votre estimée des fréquences des 4 bases $\hat{\pi}_i(a)$, pour $a = A, C, G, T$ en la position i et dans la population globale des Agrobacterium.*

Réponse : en ajoutant des pseudo-comptes :

$$\hat{\pi}_i(a) \quad = \quad \frac{m_i(a) + 1}{\sum_{b=A,C,G,T} m_i(b) + 1}.$$

Agrobacterium est un organisme haploide. En chaque position, on note h_{ij} le génotype en la position i pour l'individu j (le génotype est donc l'état caché que l'on cherche à inférer ; cet état peut prendre n'importe quelle valeur $a = A, C, G, T$).

Question 2 *Étant donné les fréquences alléliques $\pi_i(a)$ en la position i, et en supposant que l'individu j a été pris au hasard dans la population, quelle est la probabilité a priori que le génotype en cette position et pour cet individu, soit égal à a : $p(h_{ij} = a \mid \pi_i)$, pour $a = A, C, G, T$?*

Réponse : $p(h_{ij} = a \mid \pi_i) = \pi_i(a)$.

Modèle d'erreurs de séquençage

Les méthodes de séquençage massif actuellement utilisées ont un taux d'erreur assez élevé. 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. 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 ϵ. Dans un premier temps, on fixera $\epsilon = 0.03$. On estimera ce paramètre dans un second temps, par maximum de vraisemblance. La probabilité de lire la base b alors que la vraie base est a, notée E_{ab}, est donc donnée par les formules suivantes :

- $E_{aa} = 1 - \epsilon$.
- $E_{ab} = \epsilon/3$ pour $b \neq a$,

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 $k_{ij} = (k_{ij}(a))_{a=A,C,G,T}$, de dimension 4.

Question 3 *Quelle est la probabilité $p(k_{ij} \mid h_{ij} = a)$ d'obtenir les comptages k_{ij}, à supposer que le vrai génotype de l'individu j en cette position est a ? Exprimer cette probabilité uniquement en fonction de ϵ, n_{ij} et $k_{ij}(a)$. On ignorera les facteurs combinatoires, qui sont constants.*

Réponse :

$$p(k_{ij} \mid h_{ij} = a) \quad = \quad Z \, (1-\epsilon)^{k_{ij}(a)} \left(\frac{\epsilon}{3}\right)^{n_{ij} - k_{ij}(a)}$$

où Z est le facteur combinatoire associé (que l'on ne cherche pas à estimer).

Question 4 *Quelle est la probabilité marginale $p(k_{ij} \mid \pi_i)$ d'obtenir les comptages k_{ij} chez un individu tiré aléatoirement de la population (sachant les fréquences alléliques π_i) ?*

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

$$p(k_{ij} \mid \pi_i) \quad = \quad \sum_{a=A,C,G,T} p(h_{ij} = a \mid \pi_i) \, p(k_{ij} \mid h_{ij} = a)$$

Dans cette équation, le facteur $p(h_{ij} = a \mid \pi_i)$ est donné par la réponse à la question 2, tandis que le facteur $p(k_{ij} \mid h_{ij} = a)$ est donné par la réponse à la question 3.

Question 5 *Exprimer la probabilité a posteriori du génotype* h_{ij} *:* $p(h_{ij} = a \mid k_{ij}, \pi_i)$*, pour* $a = A, C, G, T$*.*

On applique le théorème de Bayes :

$$p(h_{ij} = a \mid k_{ij}, \pi_i) \quad = \quad \frac{p(h_{ij} = a \mid \pi_i)\, p(k_{ij} \mid h_{ij} = a)}{p(k_{ij} \mid \pi_i)}$$

À noter que cette 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.

Question 6 *Écrire la vraisemblance du modèle, ainsi que son logarithme*

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

ou, en logarithme :

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

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

Programmation en R

Lancez R et chargez le fichier `gencounts.txt`

```
data <- read.table("gencounts.txt",header=TRUE)

# get number of positions and number of individuals
npos <- dim(data)[1]
nind <- dim(data)[2]/4

# redimension data as a 3-dim array
countmatrix <- as.vector(as.matrix(data))
dim(countmatrix) <- c(npos,4,nind)
```

Question 7 *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 8 *De même, écrire un petit programme qui calcule le tableau des valeurs de* n_{ij}*, pour* $i = 1..N$ *et* $j = 1..P$

```
#define a npos x 4 array: m_i(a)
basecount <- array(0,dim=c(npos,4))
#define a npos x nind array: n_ij
indcount <- array(0,dim=c(npos,nind))
```

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

pi <- array(0,dim=c(npos,4))
for (i in 1:npos)   {
    pi[i,] <- basecount[i,] / sum(basecount[i,])
}
```

Question 9 *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.*

```
lnL <- 0

for (i in 1:npos)   {
    for (j in 1:nind)   {

            # store prior and likelihood for the 10 genotypes
            prior <- array(0,dim=4)
            likelihood <- array(0,dim=4)

            # loop over all possible genotypes
            for (a in 1:4)  {

                prior[a] <- pi[i,a];
                likelihood[a] <- (1-eps)^countmatrix[i,a,j] * (eps/3)^(indcount[i,j]-countmatri
            }

            # joint probability is product of prior and likelihood
            # thus, sum in log
            jointprob <- prior * likelihood

            # now make the sum
            marginalprob <- sum(jointprob)

            # and normalize, to get posterior probabilities
            posterior <- jointprob / marginalprob
```

```
            # log likelihood
            lnL <- lnL + log(marginalprob)

            # get best genotype
            genotype[i,j] <- which.max(posterior)
            score[i,j] <- max(posterior)
    }
}
```

Question 10 *En sommant le logarithme de la probabilité marginale sur tous les individus et toutes les positions, calculer le log de la vraisemblance. Effectuer ce calcul pour les valeurs suivantes de ϵ : 0.01, 0.02, ... , 0.30, puis visualiser graphiquement la courbe de la log vraisemblance en fonction de ϵ. Quelle valeur de ϵ semble-t-il indiqué d'utiliser ?*

Réponse : voir script `genotyper.R`. Au bout du compte, la valeur maximisant la vraisemblance est aux alentours de $\epsilon = 0.10$.

Question 11 *Pour chaque position et chaque individu, choisir le génotype le plus probable a posteriori, et sortir dans un fichier les génotypes avec le niveau de confiance associé*

Références

Une méthode similaire à celle introduite dans ce TP (mais pour des organismes diploides) est utilisée dans le programme `read2snp`, introduit dans l'article suivant : Tsagkogeorga, G., Cahais, V., and Galtier, N. (2012). The Population Genomics of a Fast Evolver : High Levels of Diversity, Functional Constraint, and Molecular Adaptation in the Tunicate Ciona intestinalis. Genome Biology and Evolution 4 :852.

Une autre variante de cette méthode, intégrant l'incertitude sur les π_i, est implémentée dans `samtools` : Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), 2987 :2993.