

Une méthode d'échantillonnage pour la reconstitution de séquences de variants de gènes d'antibiorésistance depuis des données métagénomiques

Daniel Bonnéry

Séminaire, 3 juin 2024



Daniel Bonnéry* (Ensaе) Guillaume Kon Kam King (Inrae)
Anne-Laure Abraham (Inrae)
Ouleye Sidibe (Inrae)
Sebastien Leclercq (Inrae)
Nicolas Chopin (Crest-Ensaе)

Contents

- 1 Le modèle paramétrique et la question statistique
- 2 Desman
- 3 Stratégies
- 4 Travaux en cours

1 Le modèle paramétrique et la question statistique

Observations

Variants

Proportions

Erreur de mesure

Distribution du tableau de comptages

Approche bayésienne

2 Desman

3 Stratégies

4 Travaux en cours

1 Le modèle paramétrique et la question statistique

Observations

Variants

Proportions

Erreur de mesure

Distribution du tableau de comptages

Approche bayésienne

2 Desman

3 Stratégies

4 Travaux en cours

Observations

We observe a 3 dimensional array of counts.

$$(n_{v,s,a}) \quad \begin{aligned} v &\in \{1, \dots, V\} \\ s &\in \{1, \dots, S\} \\ a &\in \{1, \dots, 4\} \end{aligned}$$

- $(n_{v,s,a})$: counts of nucleotides
 - at position v
 - in sample s
 - of type a :
 - $a = 1 = A = (1, 0, 0, 0)$
 - $a = 2 = C = (0, 1, 0, 0)$
 - $a = 3 = G = (0, 0, 1, 0)$
 - $a = 4 = T = (0, 0, 0, 1)$

$$n_{.,s=1,.} = \begin{matrix} & a = A & a = C & a = G & a = T \\ \begin{matrix} v=1 \\ v=2 \end{matrix} & \begin{pmatrix} 600 & 400 & 0 & 0 \\ 400 & 600 & 1 & 0 \end{pmatrix} \end{matrix}$$

$$n_{.,s=2,.} = \begin{matrix} & a = A & a = C & a = G & a = T \\ \begin{matrix} v=1 \\ v=2 \end{matrix} & \begin{pmatrix} 200 & 800 & 0 & 1 \\ 800 & 200 & 0 & 0 \end{pmatrix} \end{matrix}$$

1 Le modèle paramétrique et la question statistique

Observations

Variants

Proportions

Erreur de mesure

Distribution du tableau de comptages

Approche bayésienne

2 Desman

3 Stratégies

4 Travaux en cours

Il semblerait qu'il y a deux variants dans chacun des deux échantillons

$$a = 12^{34}$$

$$\tau = \begin{matrix} & \begin{matrix} g=1 & g=2 \end{matrix} \\ \begin{matrix} v=1 \\ v=2 \end{matrix} & \begin{pmatrix} A & C \\ C & A \end{pmatrix} \end{matrix} = \begin{matrix} & \begin{matrix} g=1 & g=2 \end{matrix} \\ \begin{matrix} v=1 \\ v=2 \end{matrix} & \begin{pmatrix} \mathbf{1}^{00} & \mathbf{01}^{00} \\ \mathbf{01}^{00} & \mathbf{1}^{00} \end{pmatrix} \end{matrix}$$

- $\tau_{v,g,a}$ indique si
 - à la position v
 - le nucléotide du variant g
 - est a .

1 Le modèle paramétrique et la question statistique

Observations

Variants

Proportions

Erreur de mesure

Distribution du tableau de comptages

Approche bayésienne

2 Desman

3 Stratégies

4 Travaux en cours

$$\pi = \begin{matrix} & \begin{matrix} s=1 & s=2 \end{matrix} \\ \begin{matrix} g=1 \\ g=2 \end{matrix} & \begin{pmatrix} 0.6 & 0.2 \\ .04 & 0.8 \end{pmatrix} \end{matrix}$$

$\pi_{g,s}$ est la proportion de

- variant g
- dans l'échantillon s .

1 Le modèle paramétrique et la question statistique

Observations

Variants

Proportions

Erreur de mesure

Distribution du tableau de comptages

Approche bayésienne

2 Desman

3 Stratégies

4 Travaux en cours

Erreur de mesure

La variable $\epsilon_{b,a}$ est la probabilité que la mesure d'un nucleotide de type b donne a

$$\epsilon = \begin{matrix} & \begin{matrix} a=1 & a=2 & a=3 & a=4 \end{matrix} \\ \begin{matrix} b=1 \\ b=2 \\ b=3 \\ b=4 \end{matrix} & \begin{pmatrix} \mathbf{0.91} & 0.03 & 0.03 & 0.03 \\ 0.03 & \mathbf{0.91} & 0.03 & 0.03 \\ 0.02 & 0.02 & \mathbf{0.94} & 0.02 \\ 0.05 & 0.04 & 0.01 & \mathbf{0.90} \end{pmatrix} \end{matrix}$$

1 Le modèle paramétrique et la question statistique

Observations

Variants

Proportions

Erreur de mesure

Distribution du tableau de comptages

Approche bayésienne

2 Desman

3 Stratégies

4 Travaux en cours

Distribution du tableau de comptages

$$\mathcal{L}(n|\pi, \tau, \epsilon)$$

$$= \prod_{v=1}^V \prod_{s=1}^S (n_{v,s,+})! \times \frac{\prod_{a=1}^4 \left(\sum_{g=1}^G \sum_{b=1}^4 \tau_{v,g,b} \epsilon_{b,a} \pi_{g,s} \right)^{n_{v,s,a}}}{\prod_{a=1}^4 n_{v,s,a}!}$$

1 Le modèle paramétrique et la question statistique

Observations

Variants

Proportions

Erreur de mesure

Distribution du tableau de comptages

Approche bayésienne

2 Desman

3 Stratégies

4 Travaux en cours

Quelle est la loi de probabilité du résultat du lancé d'un dé sachant que le résultat est pair ? **A priori**, le dé est non pipé, et la loi η du résultat est uniforme.

Règle de Bayes:

$$\eta(\{x\}|\{2, 4, 6\}) = \frac{\eta(\{x\} \cap \{2, 4, 6\})}{\eta(\{2\}) + \eta(\{4\}) + \eta(\{6\})}$$

Notre problème est similaire. Une loi sur le vecteur de variables aléatoires (n, π, τ, ϵ) est construit à partir de lois a priori sur τ, π et ϵ , et à partir de la loi de n sachant ces paramètres.

$$\eta(n, \pi, \tau, \epsilon) = \eta(n \mid \tau, \pi, \epsilon) \times \eta(\tau) \times \eta(\pi) \times \eta(\epsilon)$$

La solution de notre problème est:

$$\eta(\pi, \tau, \epsilon \mid n) = \frac{\eta(n, \pi, \tau, \epsilon)}{\int \eta(n, \pi, \tau, \epsilon) d\tau d\pi d\epsilon}$$

1 Le modèle paramétrique et la question statistique

2 Desman

Présentation

Algorithme de Gibbs

Les étapes de Gibbs dans Desman

Identification d'un problème

Diagnostics

3 Stratégies

4 Travaux en cours

1 Le modèle paramétrique et la question statistique

2 Desman

Présentation

Algorithme de Gibbs

Les étapes de Gibbs dans Desman

Identification d'un problème

Diagnostics

3 Stratégies

4 Travaux en cours

Quince, C., Delmont, T. O., Raguideau, S., Alneberg, J., Darling, A. E., Collins, G., and Eren, A. M. (2017).

Desman: a new tool for de novo extraction of strains from metagenomes.

Genome biology, 18:1–22.

- Desman est un algorithme d'échantillonnage qui permet d'obtenir un échantillon dont les propriétés sont proches d'un échantillon iid de la loi a posteriori de (τ, π, ϵ) .
- Il est basé sur des méthodes de Monte Carlo par chaînes de Markov
- obtenues par échantillonnage de Gibbs
- après introduction de variables latentes pour obtenir un maximum de lois conjuguées et éviter de devoir tirer avec Metropolis Hastings

1 Le modèle paramétrique et la question statistique

2 Desman

Présentation

Algorithme de Gibbs

Les étapes de Gibbs dans Desman

Identification d'un problème

Diagnostics

3 Stratégies

4 Travaux en cours

Algorithme de Gibbs

Algorithm 1 Algorithme de Gibbs

- 1: Initialiser $\Theta^{(0)} = (\Theta_1^{(0)}, \Theta_2^{(0)}, \dots, \Theta_l^{(0)})$
 - 2: **for** $t = 1$ to T **do**
 - 3: **for** $i = 1$ to l **do**
 - 4: Échantillonner
 - 5: $\Theta_i^{(t)} \sim \eta \left(\Theta_i \mid \Theta_1^{(t)}, \dots, \Theta_{i-1}^{(t)}, \Theta_{i+1}^{(t-1)}, \dots, \Theta_n^{(t-1)} \right)$
 - 6: **end for**
 - 7: **end for**
-

Explications

- L'algorithme commence par une initialisation de $\Theta^{(0)}$.
- À chaque étape t , chaque variable Θ_i est mise à jour en échantillonnant de sa distribution conditionnelle.
- Le processus est répété pour un nombre d'itérations T .
- En fin de compte, les échantillons $\Theta^{(T)}$ sont utilisés pour estimer la distribution cible.

$$\theta_i^{(t)} \sim \eta \left(\theta_i \mid \theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \dots, \theta_n^{(t-1)} \right)$$

Comment faire ?

- 1 Lois conjuguées
- 2 Metropolis Hasting

1 Le modèle paramétrique et la question statistique

2 Desman

Présentation

Algorithme de Gibbs

Les étapes de Gibbs dans Desman

Identification d'un problème

Diagnostics

3 Stratégies

4 Travaux en cours

Algorithm 2 Noyau MCMC

```
procedure  $M(n, \Theta = (\tau, \pi, \epsilon); \alpha_\pi, \alpha_\epsilon, e)$   
   $(\nu, \mu) \leftarrow \text{sample}_{(\nu, \mu)}(n, \Theta)$   
   $\pi \leftarrow \text{sample}_\pi(\mu; \alpha_\pi)$   
   $\tau \leftarrow \text{sample}_\tau(n, \tau, \pi, \epsilon)$   
   $\epsilon \leftarrow \text{sample}_\epsilon(\nu; \alpha_\epsilon)$   
  Return  $(\tau, \pi, \epsilon)$   
end procedure
```

1 Le modèle paramétrique et la question statistique

2 Desman

Présentation

Algorithme de Gibbs

Les étapes de Gibbs dans Desman

Identification d'un problème

Diagnostics

3 Stratégies

4 Travaux en cours

L'étape de Gibbs correspondante à τ semble problématique: τ ne varie pas avec t .

1 Le modèle paramétrique et la question statistique

2 Desman

Présentation

Algorithme de Gibbs

Les étapes de Gibbs dans Desman

Identification d'un problème

Diagnostics

3 Stratégies

4 Travaux en cours

$$n_{.,s=1,.} = \begin{matrix} & a=A & a=C & a=G & a=T \\ \begin{matrix} v=1 \\ v=2 \end{matrix} & \begin{pmatrix} 600 & 400 & 0 & 0 \\ 400 & 600 & 1 & 0 \end{pmatrix} \end{matrix}$$

$$n_{.,s=2,.} = \begin{matrix} & a=A & a=C & a=G & a=T \\ \begin{matrix} v=1 \\ v=2 \end{matrix} & \begin{pmatrix} 200 & 800 & 0 & 1 \\ 800 & 200 & 0 & 0 \end{pmatrix} \end{matrix}$$

t=1

t=2

t=3

$$\begin{matrix} g=1 & g=2 \\ v=1 & v=2 \end{matrix} \begin{pmatrix} A & C \\ A & C \end{pmatrix} \rightarrow \begin{matrix} g=1 & g=2 \\ v=1 & v=2 \end{matrix} \begin{pmatrix} A & C \\ C & C \end{pmatrix} \rightarrow \begin{matrix} g=1 & g=2 \\ v=1 & v=2 \end{matrix} \begin{pmatrix} A & C \\ C & A \end{pmatrix}$$

1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies

Block sampling

Fixed variants

Cibler ρ

SMC and Tempering

4 Travaux en cours

1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies

Block sampling

Fixed variants

Cibler ρ

SMC and Tempering

4 Travaux en cours

Il s'agit d'échantillonner en block $\tau_{v,..}$:

$$\begin{array}{cc}
 t=1 & t=2 \\
 \\
 \begin{array}{cc}
 g=1 & g=2 \\
 v=1 \left(\begin{array}{cc} A & C \end{array} \right) \\
 v=2 \left(\begin{array}{cc} A & C \end{array} \right)
 \end{array}
 \rightarrow
 \begin{array}{cc}
 g=1 & g=2 \\
 v=1 \left(\begin{array}{cc} A & C \end{array} \right) \\
 v=2 \left(\begin{array}{cc} C & A \end{array} \right)
 \end{array}
 \end{array}$$

Problème:

4^1	4
4^2	16
4^3	64
4^4	256
4^5	1 024
4^{10}	1 048 576
4^{15}	1 073 741 824
4^{20}	1 099 511 627 776

1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies

Block sampling

Fixed variants

Cibler ρ

SMC and Tempering

4 Travaux en cours

Il s'agit de remplacer la distribution a priori de τ par $\text{Uniform}(\mathcal{T})$

- Metropolis Hasting est évité.
- Satisfaisant lorsque les variants sont connus
- Ne répond pas à l'objectif initial de détection.

1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies

Block sampling

Fixed variants

Cibler ρ

SMC and Tempering

4 Travaux en cours

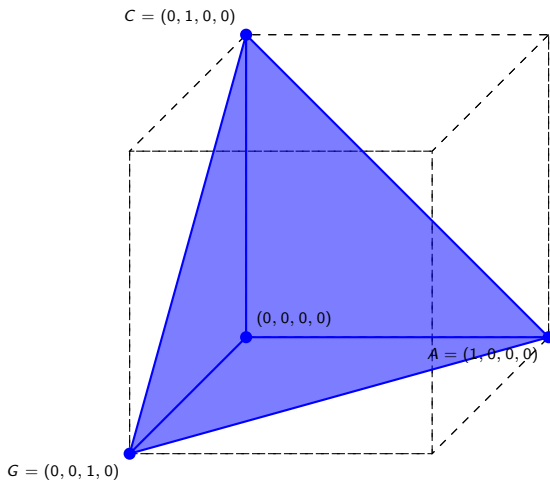
$$\begin{aligned}
& \mathcal{L}(n|\pi, \tau, \epsilon) \\
&= \prod_{v=1}^V \prod_{s=1}^S (n_{v,s,+})! \times \frac{\prod_{a=1}^4 \left(\sum_{g=1}^G \sum_{b=1}^4 \tau_{v,g,b} \epsilon_{b,a} \pi_{g,s} \right)^{n_{v,s,a}}}{\prod_{a=1}^4 n_{v,s,a}!} \\
&= \prod_{v=1}^V \prod_{s=1}^S (n_{v,s,+})! \times \frac{\prod_{a=1}^4 \left(\sum_{g=1}^G \rho_{v,g,a} \pi_{g,s} \right)^{n_{v,s,a}}}{\prod_{a=1}^4 n_{v,s,a}!}
\end{aligned}$$

avec

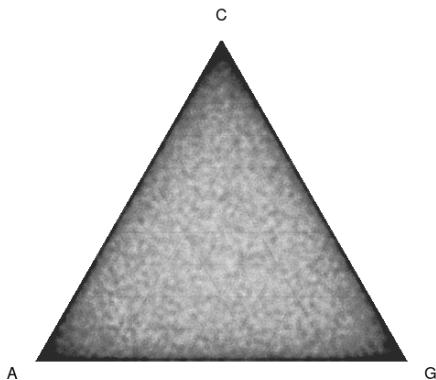
$$\rho_{v,g,a} = \sum_{b=1}^4 \tau_{vgb} \epsilon_{ba}$$

$$\rho_{\cdot, g, \cdot} = \begin{matrix} & A & C & G & T \\ \begin{matrix} v=1 \\ \vdots \\ v=V \end{matrix} & \begin{pmatrix} 0.03 & \mathbf{0.91} & 0.03 & 0.03 \\ \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.04 & \mathbf{0.90} & 0.01 \end{pmatrix} \end{matrix}$$

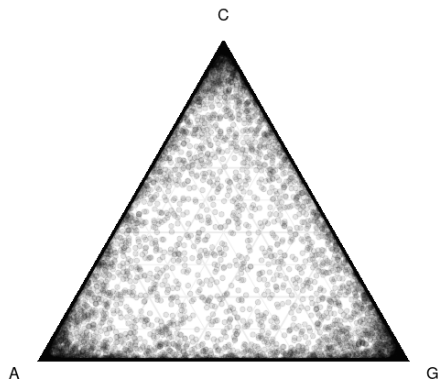
- On utilise un hyperparamètre de forme très faible pour la loi de Dirichlet, ce qui assure que les $\rho_{v, g, \cdot}$ soient concentrés autour de A, C, G, ou T.
- On garde des lois conjuguées pour ρ .
- La mise à jour se fait naturellement par block.
- Rajout d'un a priori non conjugué sur l'hyperparamètre de forme.



Dirichlet Distribution on a 3-Simplex



Dirichlet Distribution on a 3-Simplex



1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies

Block sampling

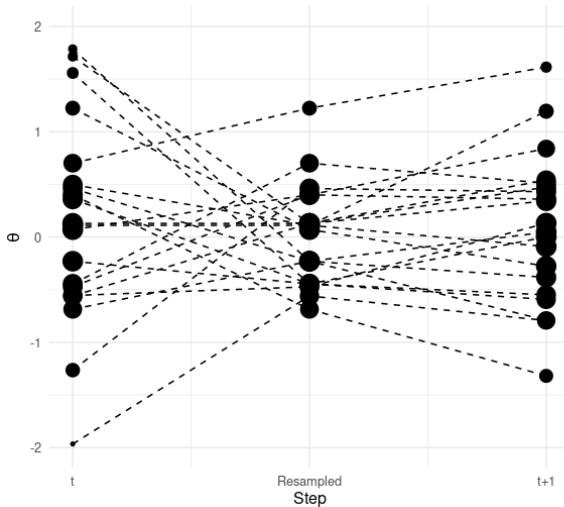
Fixed variants

Cibler ρ

SMC and Tempering

4 Travaux en cours

Sequential Monte Carlo Step

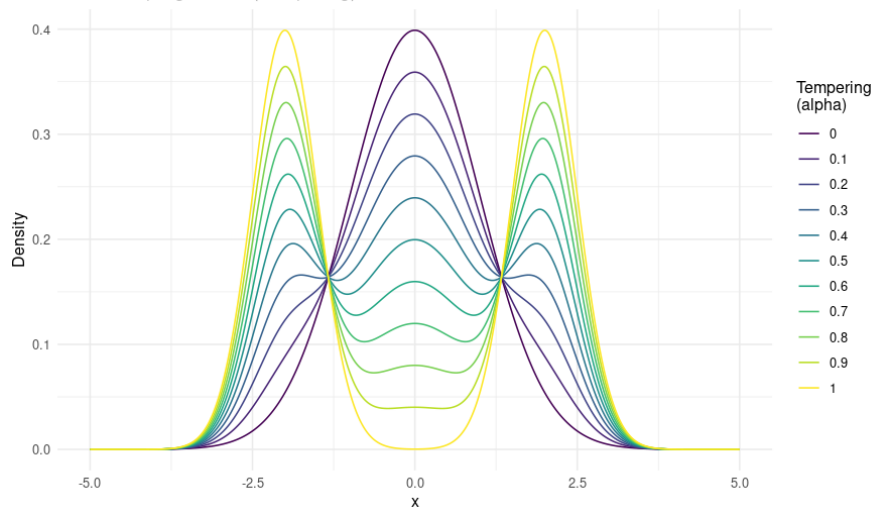


weight

- 0.1
- 0.2
- 0.3



Transition progressive (Tempering)



Stratégies de tempering:

- Puissance de la vraisemblance
- Data tempering
- Paramètre de tempering naturel.

Stratégie retenue

- Cibler τ , ϵ et π
- Relaxer τ
- Le paramètre de forme sur τ devient le paramètre de tempering naturel.

- ① Le modèle paramétrique et la question statistique
- ② Desman
- ③ Stratégies
- ④ Travaux en cours

- Développement d'algorithmes en R basés sur jags.
- SMC
- tempering.
- Choix de modèle (G)
- ESS
- Exploitation des trajectoires de particules obtenues.

- Dau, H.-D. and Chopin, N. (2022). Waste-free sequential monte carlo. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(1):114–148.
- Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). Identifying mixtures of mixtures using bayesian estimation. Journal of Computational and Graphical Statistics, 26(2):285–295.
- Quince, C., Delmont, T. O., Raguideau, S., Alneberg, J., Darling, A. E., Collins, G., and Eren, A. M. (2017). Desman: a new tool for de novo extraction of strains from metagenomes. Genome biology, 18:1–22.