Les mots de l'ADN

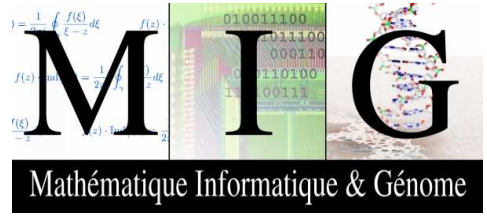
Sophie S. Schbath

Les mots de l'ADN

M.-A. Petit et S. Schbath

INRA, Jouy-en-Josas, France

http://genome.jouy.inra.fr/ssb/



Modélisation statistique

Comptage inattendu ?

On considère un motif $\mathbf{w}$ sur l'alphabet $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

On observe son comptage le long d'une séquence d'ADN : N^{obs}.

N^{obs} est-il anormalement grand ? anormalement petit ? ou normal ($=$ attendu) ?

"*attendu*" fait référence à un a priori que nous avons compte tenu de l'information que nous avons sur la séquence $\Rightarrow$ le modèle.

Modèle et comptage inattendu

Le comptage attendu $\widehat{N}^{\text{att}}$ dépend du modèle pour la séquence.

Exemple : `tcatg` observé 100 fois dans une séquence de longueur $n = 100\,000$.

Modèle	Info sur la séquence	$\widehat{N}^{\text{att}}$
indép. + équirép.	longueur $n = 100000$	$\frac{n}{4^5} = n \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \simeq \mathbf{98}$

Modèle et comptage inattendu

Le comptage attendu $\widehat{N}^{\text{att}}$ dépend du modèle pour la séquence.

Exemple : `tcatg` observé 100 fois dans une séquence de longueur $n = 100\,000$.

Modèle	Info sur la séquence	$\widehat{N}^{\text{att}}$
indép. + équirép.	longueur $n = 100000$	$\frac{n}{4^5} = n \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \simeq \mathbf{98}$
indép. M0	%a, %c, %g, %t 10%,15%,50%,25%	$n\widehat{p}(\mathtt{t})\widehat{p}(\mathtt{c})\widehat{p}(\mathtt{a})\widehat{p}(\mathtt{t})\widehat{p}(\mathtt{g}) \simeq \mathbf{47}$

Modèle et comptage inattendu

Le comptage attendu $\widehat{N}^{\text{att}}$ dépend du modèle pour la séquence.

Exemple : `tcatg` observé 100 fois dans une séquence de longueur $n = 100\,000$.

Modèle	Info sur la séquence	$\widehat{N}^{\text{att}}$
indép. + équirép.	longueur $n = 100000$	$\frac{n}{4^5} = n \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \simeq \mathbf{98}$
indép. M0	%a, %c, %g, %t 10%,15%,50%,25%	$n\widehat{p}(\mathtt{t})\widehat{p}(\mathtt{c})\widehat{p}(\mathtt{a})\widehat{p}(\mathtt{t})\widehat{p}(\mathtt{g}) \simeq \mathbf{47}$
dépendance 1 lettre M1	%aa, %ac, %ag, ..., %tg, %tt	$n\widehat{p}(\mathtt{t})\widehat{p}(\mathtt{c}\text{ après }\mathtt{t})\widehat{p}(\mathtt{a}\text{ après }\mathtt{c})\widehat{p}(\mathtt{t}\text{ après }\mathtt{a}) \times \widehat{p}(\mathtt{g}\text{ après }\mathtt{t}) \simeq \mathbf{62}$

Modèle et comptage inattendu

Le comptage attendu $\widehat{N}^{\text{att}}$ dépend du modèle pour la séquence.

Exemple : `tcatg` observé 100 fois dans une séquence de longueur $n = 100\,000$.

Modèle	Info sur la séquence	$\widehat{N}^{\text{att}}$
indép. + équirép.	longueur $n = 100000$	$\frac{n}{4^5} = n \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \simeq \mathbf{98}$
indép. M0	%a, %c, %g, %t 10%,15%,50%,25%	$n\widehat{p}(\mathtt{t})\widehat{p}(\mathtt{c})\widehat{p}(\mathtt{a})\widehat{p}(\mathtt{t})\widehat{p}(\mathtt{g}) \simeq \mathbf{47}$
dépendance 1 lettre M1	%aa, %ac, %ag, ..., %tg, %tt	$n\widehat{p}(\mathtt{t})\widehat{p}(\mathtt{c}\text{ après }\mathtt{t})\widehat{p}(\mathtt{a}\text{ après }\mathtt{c})\widehat{p}(\mathtt{t}\text{ après }\mathtt{a}) \times \widehat{p}(\mathtt{g}\text{ après }\mathtt{t}) \simeq \mathbf{62}$
⋮		
dépendance 3 lettres M3	%aaaa, %aaac, ..., %tttg, %tttt	$n\widehat{p}(\mathtt{tca})\widehat{p}(\mathtt{t}\text{ après }\mathtt{tca})\widehat{p}(\mathtt{g}\text{ après }\mathtt{cat}) \simeq \mathbf{55}$

Mesure d'exceptionnalité

Intuitivement on veut comparer N^{obs} et $\widehat{N}^{\text{att}}$, par exemple en faisant

$$\text{score}^{\text{obs}} = \frac{N^{\text{obs}} - \widehat{N}^{\text{att}}}{\widehat{\sigma}}$$

Mesure d'exceptionnalité

Intuitivement on veut comparer N^{obs} et $\widehat{N}^{\text{att}}$, par exemple en faisant

$$\text{score}^{\text{obs}} = \frac{N^{\text{obs}} - \widehat{N}^{\text{att}}}{\widehat{\sigma}}$$

On peut montrer que si $\widehat{N}^{\text{att}}$ est assez grand alors le score $\sim \mathcal{N}(0, 1)$.

Mesure d'exceptionnalité

Intuitivement on veut comparer N^{obs} et $\widehat{N}^{\text{att}}$, par exemple en faisant

$$\text{score}^{\text{obs}} = \frac{N^{\text{obs}} - \widehat{N}^{\text{att}}}{\widehat{\sigma}}$$

On peut montrer que si $\widehat{N}^{\text{att}}$ est assez grand alors le score $\sim \mathcal{N}(0,1)$.

On peut donc calculer la significativité de N^{obs} grâce à la "p-value" :

$$\mathbb{P}(\text{comptage} \geq N^{\text{obs}}) = \mathbb{P}(\mathcal{N}(0,1) \geq \text{score}^{\text{obs}}).$$

Une p-value proche de 0 $\leftrightarrow$ score proche de $+\infty$ $\leftrightarrow$ motif exceptionnellement fréquent
Une p-value proche de 0.5 $\leftrightarrow$ score proche de 0 $\leftrightarrow$ motif attendu
Une p-value proche de 1 $\leftrightarrow$ score proche de $-\infty$ $\leftrightarrow$ motif exceptionnellement rare

Mesure d'exceptionnalité

Intuitivement on veut comparer N^{obs} et $\widehat{N}^{\text{att}}$, par exemple en faisant

$$\text{score}^{\text{obs}} = \frac{N^{\text{obs}} - \widehat{N}^{\text{att}}}{\widehat{\sigma}}$$

On peut montrer que si $\widehat{N}^{\text{att}}$ est assez grand alors le score $\sim \mathcal{N}(0,1)$.

On peut donc calculer la significativité de N^{obs} grâce à la "p-value" :

$$\mathbb{P}(\text{comptage} \geq N^{\text{obs}}) = \mathbb{P}(\mathcal{N}(0,1) \geq \text{score}^{\text{obs}}).$$

Une p-value proche de 0 $\leftrightarrow$ score proche de $+\infty$ $\leftrightarrow$ motif exceptionnellement fréquent
Une p-value proche de 0.5 $\leftrightarrow$ score proche de 0 $\leftrightarrow$ motif attendu
Une p-value proche de 1 $\leftrightarrow$ score proche de $-\infty$ $\leftrightarrow$ motif exceptionnellement rare

Si $\widehat{N}^{\text{att}}$ est petit, on peut montrer que le comptage suit une loi de type Poisson.

Contraste significatif ?

On considère un motif $\mathbf{w}$ sur l'alphabet $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

On observe son comptage le long d'une 1ère séquence d'ADN : N_1^{obs}.
On observe son comptage le long d'une 2ème séquence d'ADN : N_2^{obs}.

Le motif $\mathbf{w}$ est-il davantage inattendu dans la 1ère séquence que dans la 2ème ?

Contraste significatif ?

On considère un motif $\mathbf{w}$ sur l'alphabet $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

On observe son comptage le long d'une 1ère séquence d'ADN : N_1^{obs}.
On observe son comptage le long d'une 2ème séquence d'ADN : N_2^{obs}.

Le motif $\mathbf{w}$ est-il davantage inattendu dans la 1ère séquence que dans la 2ème ?

Pour cela, on quantifie l'exceptionnalité du motif en écrivant $N^{\text{obs}} = k \times \widehat{N}^{\text{att}}$
et on teste statistiquement l'hypothèse H0 : $\{k_1 = k_2\}$ contre H1 : $\{k_1 > k_2\}$

Contraste significatif ?

On considère un motif $\mathbf{w}$ sur l'alphabet $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

On observe son comptage le long d'une 1ère séquence d'ADN : N_1^{obs}.
On observe son comptage le long d'une 2ème séquence d'ADN : N_2^{obs}.

Le motif $\mathbf{w}$ est-il davantage inattendu dans la 1ère séquence que dans la 2ème ?

Pour cela, on quantifie l'exceptionnalité du motif en écrivant $N^{\text{obs}} = k \times \widehat{N}^{\text{att}}$
et on teste statistiquement l'hypothèse H0 : $\{k_1 = k_2\}$ contre H1 : $\{k_1 > k_2\}$

Ici la modélisation va porter sur le processus des occurrences du motif le long de chacune des séquences (processus de Poisson),

Contraste significatif ?

On considère un motif $\mathbf{w}$ sur l'alphabet $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

On observe son comptage le long d'une 1ère séquence d'ADN : N_1^{obs}.
On observe son comptage le long d'une 2ème séquence d'ADN : N_2^{obs}.

Le motif $\mathbf{w}$ est-il davantage inattendu dans la 1ère séquence que dans la 2ème ?

Pour cela, on quantifie l'exceptionnalité du motif en écrivant $N^{\text{obs}} = k \times \widehat{N}^{\text{att}}$
et on teste statistiquement l'hypothèse H0 : $\{k_1 = k_2\}$ contre H1 : $\{k_1 > k_2\}$

Ici la modélisation va porter sur le processus des occurrences du motif le long de chacune des séquences (processus de Poisson),

et on utilise que $N_1 \mid N_1 + N_2$ suit une loi binomiale dont les paramètres ne dépendent ni de k_1 ni de k_2 sous l'hypothèse H0

$\Longrightarrow$ p-value $\Longrightarrow$ score de contraste