



**HAL**  
open science

## Les mots de l'ADN

Sophie S. Schbath

► **To cite this version:**

Sophie S. Schbath. Les mots de l'ADN. Journées des nouveaux arrivants du centre INRA de Jouy-en-Josas, Institut National de Recherche Agronomique (INRA). UR Mathématique, Informatique et Génome (1077)., Dec 2009, Jouy-en-Josas, France. 15 diapos. hal-02813617

**HAL Id: hal-02813617**

**<https://hal.inrae.fr/hal-02813617v1>**

Submitted on 6 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

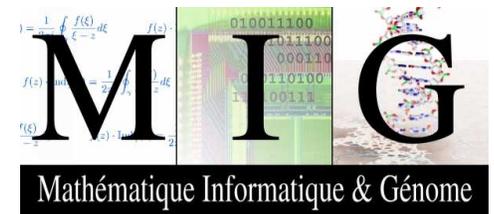
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 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  $w$  sur l'alphabet  $\{a, c, g, t\}$ .

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

$N^{\text{obs}}$  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  $\hat{N}^{\text{att}}$  dépend du modèle pour la séquence.

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

Modèle	Info sur la séquence	$\hat{N}^{\text{att}}$
indép. + équirép.	longueur $n = 100\,000$	$\frac{n}{4^5} = n \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \simeq \mathbf{98}$

# Modèle et comptage inattendu

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

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

Modèle	Info sur la séquence	$\hat{N}^{\text{att}}$
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%	$n\hat{p}(t)\hat{p}(c)\hat{p}(a)\hat{p}(t)\hat{p}(g) \simeq \mathbf{47}$

# Modèle et comptage inattendu

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

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

Modèle	Info sur la séquence	$\hat{N}^{\text{att}}$
indép. + équirép.	longueur $n = 100\,000$	$\frac{n}{4^5} = n \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \simeq \mathbf{98}$
indép. M0	%a, %c, %g, %t 10%, 15%, 50%, 25%	$n\hat{p}(t)\hat{p}(c)\hat{p}(a)\hat{p}(t)\hat{p}(g) \simeq \mathbf{47}$
dépendance 1 lettre M1	%aa, %ac, %ag, ..., %tg, %tt	$n\hat{p}(t)\hat{p}(c \text{ après } t)\hat{p}(a \text{ après } c)\hat{p}(t \text{ après } a) \times \hat{p}(g \text{ après } t) \simeq \mathbf{62}$

# Modèle et comptage inattendu

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

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

Modèle	Info sur la séquence	$\hat{N}^{\text{att}}$
indép. + équirép.	longueur $n = 100\,000$	$\frac{n}{4^5} = n \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \simeq \mathbf{98}$
indép. M0	%a, %c, %g, %t 10%,15%,50%,25%	$n\hat{p}(t)\hat{p}(c)\hat{p}(a)\hat{p}(t)\hat{p}(g) \simeq \mathbf{47}$
dépendance 1 lettre M1	%aa, %ac, %ag, ..., %tg, %tt	$n\hat{p}(t)\hat{p}(c \text{ après } t)\hat{p}(a \text{ après } c)\hat{p}(t \text{ après } a) \times \hat{p}(g \text{ après } t) \simeq \mathbf{62}$
⋮		
dépendance 3 lettres M3	%aaaa, %aaac, ..., %tttg, %tttt	$n\hat{p}(tca)\hat{p}(t \text{ après } tca)\hat{p}(g \text{ après } cat) \simeq \mathbf{55}$

# Mesure d'exceptionnalité



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

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

# Mesure d'exceptionnalité



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

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

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

# Mesure d'exceptionnalité

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

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

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

On peut donc calculer la significativité de  $N^{\text{obs}}$  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  $N^{\text{obs}}$  et  $\hat{N}^{\text{att}}$ , par exemple en faisant

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

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

On peut donc calculer la significativité de  $N^{\text{obs}}$  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  $\hat{N}^{\text{att}}$  est petit, on peut montrer que le comptage suit une loi de type Poisson.

# Contraste significatif ?



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

On observe son comptage le long d'une 1ère séquence d'ADN :  $N_1^{\text{obs}}$ .

On observe son comptage le long d'une 2ème séquence d'ADN :  $N_2^{\text{obs}}$ .

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

# Contraste significatif ?



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

On observe son comptage le long d'une 1ère séquence d'ADN :  $N_1^{\text{obs}}$ .

On observe son comptage le long d'une 2ème séquence d'ADN :  $N_2^{\text{obs}}$ .

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

Pour cela, on quantifie l'exceptionnalité du motif en écrivant  $N^{\text{obs}} = k \times \hat{N}^{\text{att}}$   
et on teste statistiquement l'hypothèse  $H_0 : \{k_1 = k_2\}$  contre  $H_1 : \{k_1 > k_2\}$

# Contraste significatif ?



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

On observe son comptage le long d'une 1ère séquence d'ADN :  $N_1^{\text{obs}}$ .

On observe son comptage le long d'une 2ème séquence d'ADN :  $N_2^{\text{obs}}$ .

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

Pour cela, on quantifie l'exceptionnalité du motif en écrivant  $N^{\text{obs}} = k \times \hat{N}^{\text{att}}$   
et on teste statistiquement l'hypothèse  $H_0 : \{k_1 = k_2\}$  contre  $H_1 : \{k_1 > k_2\}$

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  $w$  sur l'alphabet  $\{a, c, g, t\}$ .

On observe son comptage le long d'une 1ère séquence d'ADN :  $N_1^{\text{obs}}$ .

On observe son comptage le long d'une 2ème séquence d'ADN :  $N_2^{\text{obs}}$ .

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

Pour cela, on quantifie l'exceptionnalité du motif en écrivant  $N^{\text{obs}} = k \times \hat{N}^{\text{att}}$   
et on teste statistiquement l'hypothèse  $H_0 : \{k_1 = k_2\}$  contre  $H_1 : \{k_1 > k_2\}$

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  $H_0$

$\implies p\text{-value} \implies \text{score de contraste}$