



HAL
open science

À la recherche de mots exceptionnels dans les génomes

Sophie Schbath

► **To cite this version:**

Sophie Schbath. À la recherche de mots exceptionnels dans les génomes. K'fêt des sciences du collège de La Gyonnerie, Mar 2018, Bures-sur-Yvette, France. 22 diapos. hal-01953364

HAL Id: hal-01953364

<https://hal.science/hal-01953364>

Submitted on 4 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.



Distributed under a Creative Commons Attribution 4.0 International License



À la recherche de mots exceptionnels dans les génomes



K'fêt des sciences - Guyonnerie – 9 mars 2018

Mon parcours

Un bac scientifique (Les Ulis)

Une licence de mathématique (Orsay)

Un master de probabilités et statistique (Orsay)

Un doctorat à l'Inra (Jouy-en-Josas)

Une année post-doctorale en Californie

Chercheure à l'Inra depuis 1996

Directrice de laboratoire (MIG, puis MaIAGE) depuis 2012.

INRA

Un organisme de recherche public avec trois grands domaines de recherche

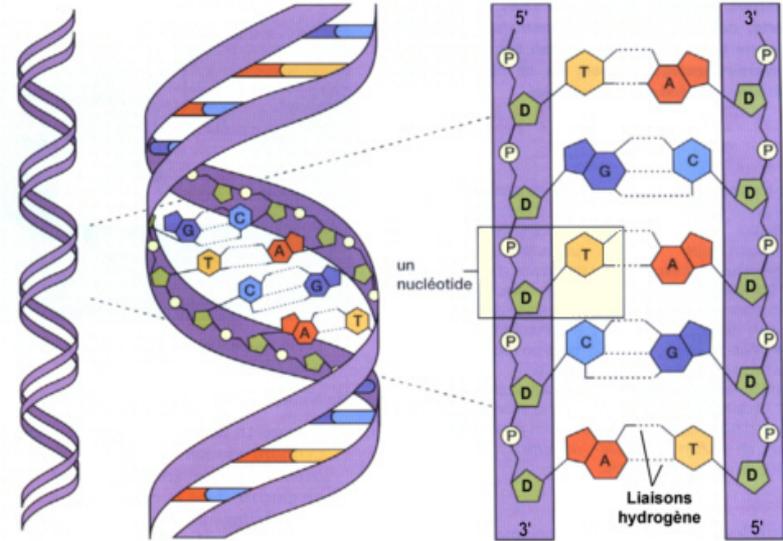
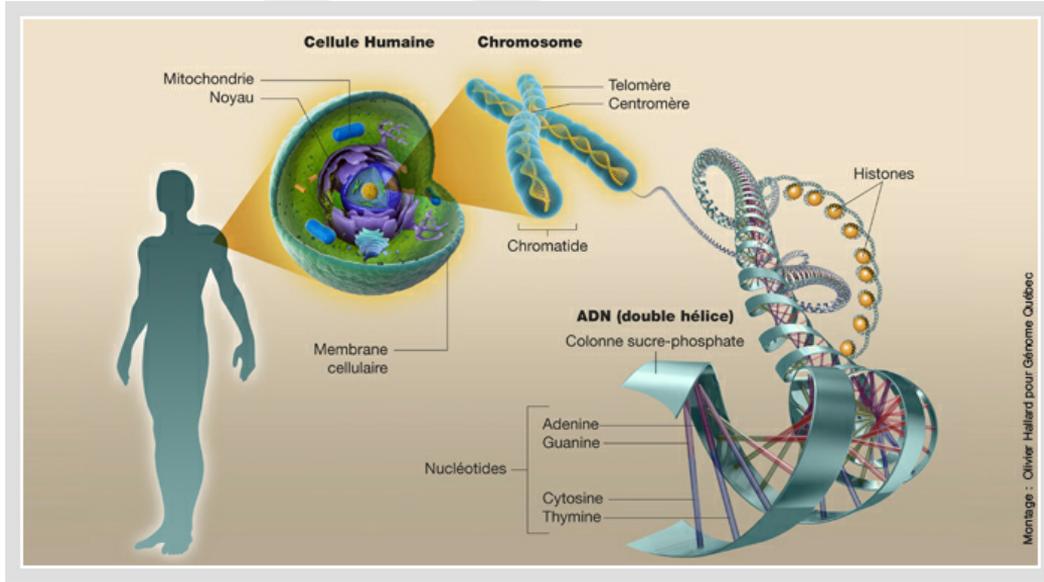
ALIMENTATION AGRICULTURE ENVIRONNEMENT



17 centres de recherche
8200 agents titulaires dont 1800 chercheurs
13 départements de recherche
dont « Math Info Appliquées » (60 chercheurs)



ADN, génome, chromosomes, ...



A G C C A T G T A A T G C A G T T C T G A A C C G G T

Séquence d'ARN transcrite

Séquence d'ADN

Machinerie transcriptionnelle

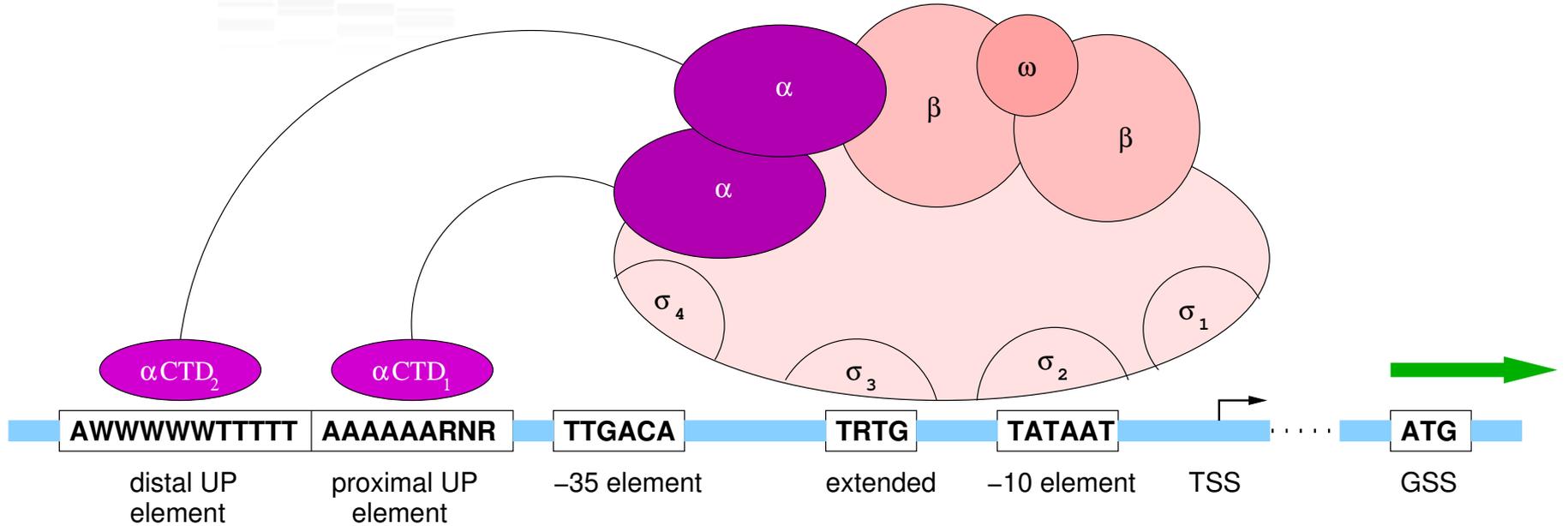
DNAPK

Machinerie de réparation

Facteur provocant des cassures

Cassure double brin

ADN et mots



À la recherche des mots exceptionnels

Si on était capable d'identifier tous les mots **anormalement très fréquents** (« exceptionnels ») d'un génome donné, alors on aurait des nouveaux mots candidats à avoir un rôle important, qu'il faudra étudier.

Le mathématicien va s'attaquer à la question « comment savoir si un mot est anormalement très fréquent dans une séquence d'ADN ».

Le biologiste s'attaquera à la question de découvrir le rôle biologique du mot.



Exercice de mathématique !



La bactérie *E. coli* a un génome de longueur 4 638 850.

Le mot de 8 lettres **GCTGGTGG** est présent **762 fois** le long de ce génome.

Est-ce normal ? Est-ce anormalement beaucoup ou anormalement faible ?

Pour répondre à cette question (« **est-ce normal ?** »), il faut savoir « **normal, par rapport à quoi ?** »

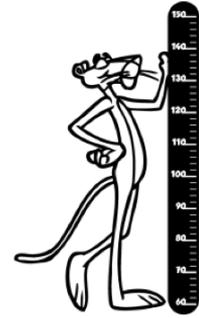
Par exemple, la réponse ne sera pas la même si le génome de *E. coli* est riche en **G** ou au contraire est pauvre en **G**

Une idée de la démarche (1)

Prenons un autre exercice similaire :

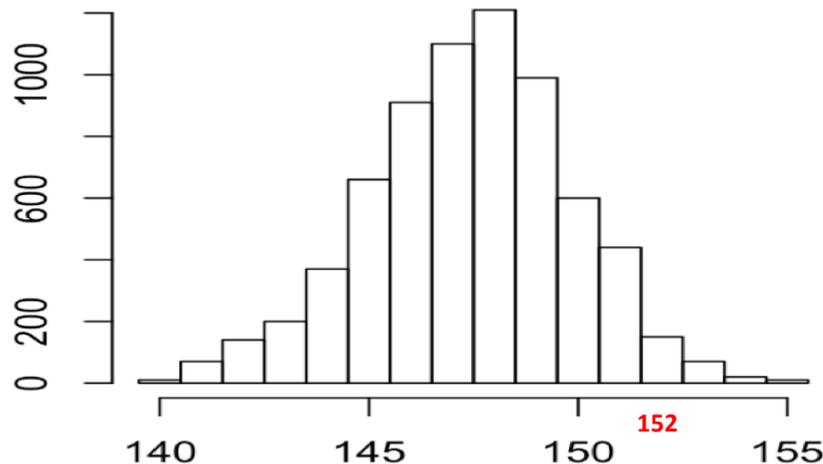
La taille moyenne des collégiens de la Guyonnerie est de 1m52.

Peut-on affirmer que ces collégiens sont particulièrement grands par rapport aux collégiens français ?



Une idée de la démarche (2)

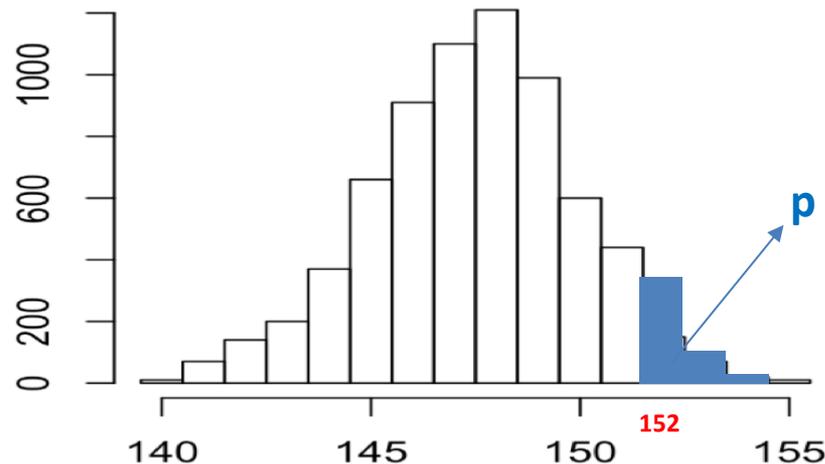
→ Collecter la taille moyenne pour chaque collège (6950)
et comparer 1m52 à la
« **distribution** » de tailles
obtenue :



Ici, 1m52 est vraiment très à droite de la distribution, donc on a envie de dire que
« **oui, les collégiens de Bures sont particulièrement grands !** »

Une idée de la démarche (2)

→ Collecter la taille moyenne pour chaque collège (6950)
et comparer 1m52 à la
« **distribution** » de tailles
obtenue :

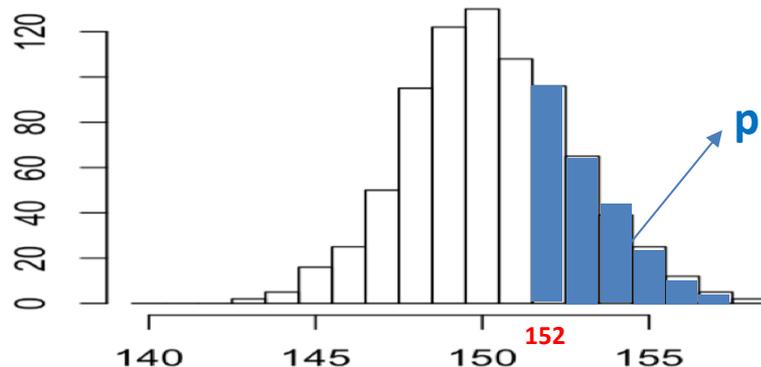


Ici, 1m52 est vraiment très à droite de la distribution, donc on a envie de dire que
« **oui, les collégiens de Bures sont particulièrement grands !** »

Le « très à droite » est mathématiquement quantifié par la proportion « **p** ».

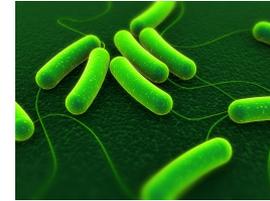
Une idée de la démarche (3)

MAIS, il serait probablement plus pertinent de tenir compte du fait que « *les garçons sont généralement plus grands que les filles* » et refaire l'analyse avec seulement les collèves (797) ayant la même proportion de garçons qu'à La Guyonnerie :



1m52 devient dans ce cas « plus normal », la proportion p étant assez grande
→ « **Les collégiens de Bures sont particulièrement grands mais si l'on tient compte de la proportion de garçons, cela n'est pas particulièrement surprenant !** »

Revenons à nos mots sur l'ADN



La bactérie *E. coli* a un génome de longueur 4 638 850.

Le mot de 8 lettres **GCTGGTGG** est présent **762 fois** le long de ce génome. Normal ?

On va comparer le génome de *E. coli* à toutes les suites de 4 638 850 lettres dans l'alphabet {**A**, **G**, **C**, **T**} - **MAIS** - qui ont autant de **A**, **G**, **C**, **T** que le génome de *E. coli* et représenter la distribution des comptages obtenus.

Voyons sur un exemple plus simple que la tâche s'avère longue et fastidieuse !



Exemple « jouet »



Prenons le mot **AT** dans la séquence de 5 lettres **ATATC** (il apparaît 2 fois)
Un calcul mathématique nous dit qu'il y a $5!/2!2!1!=30$ suites possibles :

AATTC	AACTT	CAATT	AATCT	ACATT
ATATC	ATCAT	CATAT	ATACT	ACTAT
ATTAC	ATCTA	CATTA	ATTCA	ACTTA
TAATC	TACAT	CTAAT	TAACT	TCAAT
TATAC	TACTA	CTATA	TATCA	TCATA
TTAAC	TTCAA	CTTAA	TTACA	TCTAA

Exemple « jouet »



Prenons le mot **AT** dans la séquence de 5 lettres **ATATC** (il apparaît 2 fois)
Un calcul mathématique nous dit qu'il y a $5!/2!2!1! = 30$ séquences possibles :

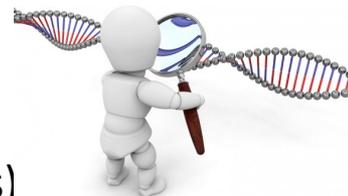
AATTC	AACTT	CAATT	AATCT	ACATT
ATA TC	ATCAT	CATAT	ATACT	ACTAT
ATTAC	ATCTA	CATTA	ATTCA	ACTTA
TAATC	TACAT	CTAAT	TAACT	TCAAT
TATAC	TACTA	CTATA	TATCA	TCATA
TTAAC	TTCAA	CTTAA	TTACA	TCTAA

$N(\mathbf{AT}) = 0$ dans 9 cas sur 30

$N(\mathbf{AT}) = 1$ dans 18 cas sur 30

$N(\mathbf{AT}) = 2$ dans 3 cas sur 30

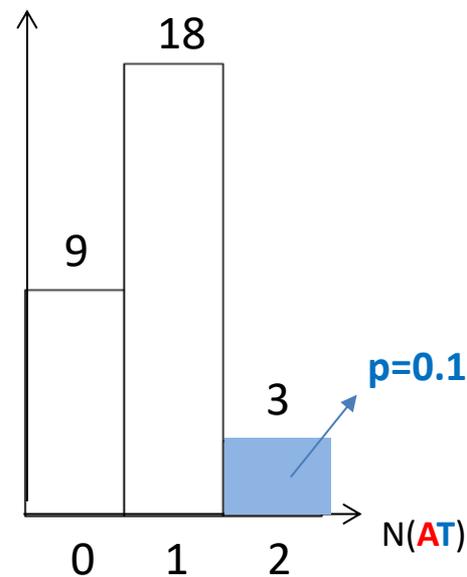
Exemple « jouet »



Prenons le mot **AT** dans la séquence de 5 lettres **ATATC** (il apparaît 2 fois)
Un calcul mathématique nous dit qu'il y a $5!/2!2!1!=30$ séquences possibles :

A AT TC	A A CTT	C AAT	A AT CT	A CA TT
AT ATC	AT CA T	C AT AT	AT ACT	ACT AT
AT TAC	AT CTA	C AT TA	AT TCA	ACTTA
T AAT C	T ACA T	C TAA T	T AA CT	T CAAT
T ATA C	T ACTA	C TATA	T AT CA	T CATA
T TAA C	T TCAA	C TTAA	T TACA	T CTAA

$N(\mathbf{AT}) = 0$ dans 9 cas sur 30
 $N(\mathbf{AT}) = 1$ dans 18 cas sur 30
 $N(\mathbf{AT}) = 2$ dans 3 cas sur 30
→ La moyenne est 0.8 occurrences



Les mathématiques à la rescousse

Pour une séquence de 5 lettres : 30 séquences possibles

Pour une séquence de 12 lettres : environ 20 millions séquences possibles

Pour une séquence de 100 lettres : environ 10^{132} séquences possibles

Pour une séquence de 4 638 850 lettres... impossible.

Les mathématiques à la rescousse

Pour une séquence de 5 lettres : 30 séquences possibles

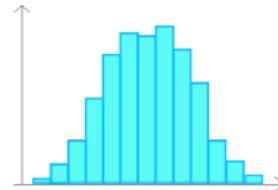
Pour une séquence de 12 lettres : environ 20 millions séquences possibles

Pour une séquence de 100 lettres : environ 10^{132} séquences possibles

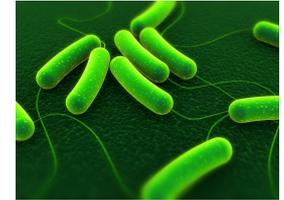
Pour une séquence de 4 638 850 lettres... impossible.



Grâce à des calculs de probabilité compliqués, une formule mathématique a été démontrée pour calculer la distribution du nombre d'occurrences de n'importe quel mot dans n'importe quelle séquence et donc de connaître la valeur de p .



Et alors pour *E. coli* ?



La bactérie *E. coli* a un génome de longueur 4 638 850.

Le mot de 8 lettres **GCTGGTGG** est présent **762 fois** le long de ce génome.

Sachant $N(\mathbf{A})$, $N(\mathbf{G})$, $N(\mathbf{C})$, $N(\mathbf{T})$, $\rightarrow p < 10^{-323} \simeq 0$

Sachant $N(\mathbf{GC})$, $N(\mathbf{CT})$, $N(\mathbf{TG})$, $N(\mathbf{GG})$, etc. $\rightarrow p < 10^{-323} \simeq 0$

etc.

Sachant $N(\mathbf{GCTGGTGG})$, $N(\mathbf{CTGGTGG})$, etc, $p=1.5 \cdot 10^{-26}$

c'est le champion des 8-mer !

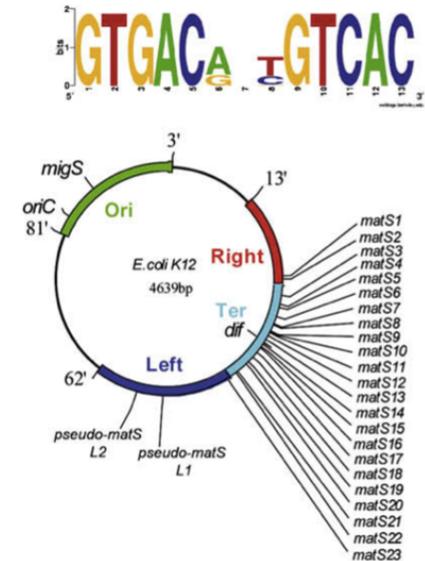


Conclusion : GCTGGTGG, connu sous le nom de « chi », est le mot de 8 lettres le plus exceptionnel dans le génome de *E. coli*.

Conclusion

Grâce à ce type de calcul mathématique, de nombreux mots ont été découverts, par exemple :

- Halpern, D., Chiapello, H., Schbath, S., Robin, S., Hennequet-Antier, C., Gruss, A. and El Karoui, M. (2007). Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modelling. *PLoS Genetics*. **3(9)** e153.
- Mercier, R., Petit, M.-A., Schbath, S., Robin, S., El Karoui, M., Boccard, F. and Espeli, O. (2008). The MatP/matS site specific system organizes the Terminus region of the *E. coli* chromosome into a Macrodomain. *Cell*. **135** 475-485.





Un même morceau d'ADN chez différentes espèces

```
cagaaactgcagattagcgtgtatatttatctgtttatgct  
cagaaactgcagattagcgtgtatatttatctgtttgct  
cagaaactgcagatttatgtgtatatttatctgtttatgct  
cagaaactgcagattttgtgtatatttatctgtttatgct  
cagaaactggcgggtgtatgtgtatatttatctgtttatgca
```

boeuf

cheval

porc

chèvre

poulet