À la recherche de mots exceptionnels dans les génomes

Sophie Schbath

INRA
SCIENCE & IMPACT

MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES
MaIAGE
DU GÉNOME À L'ENVIRONNEMENT

À la recherche de mots exceptionnels dans les génomes

K'fêt des sciences - Guyonnerie – 9 mars 2018

Sophie SCHBATH

26/ 06 / 2017

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

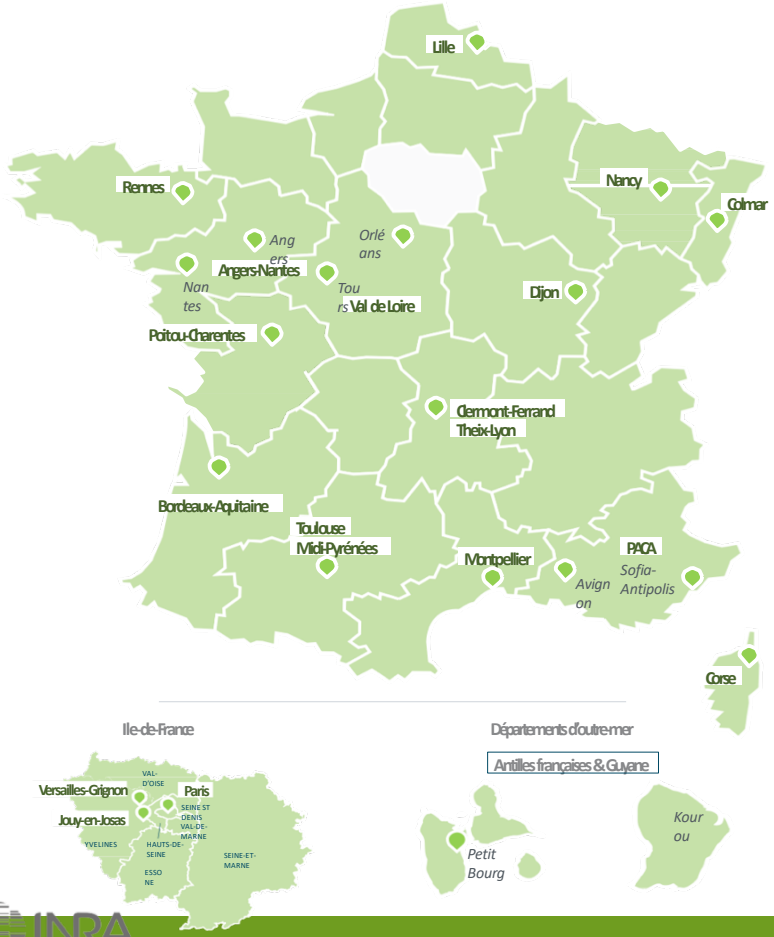
Directrice de laboratoire (MIG, puis MaIAGE) depuis 2012.

INRA

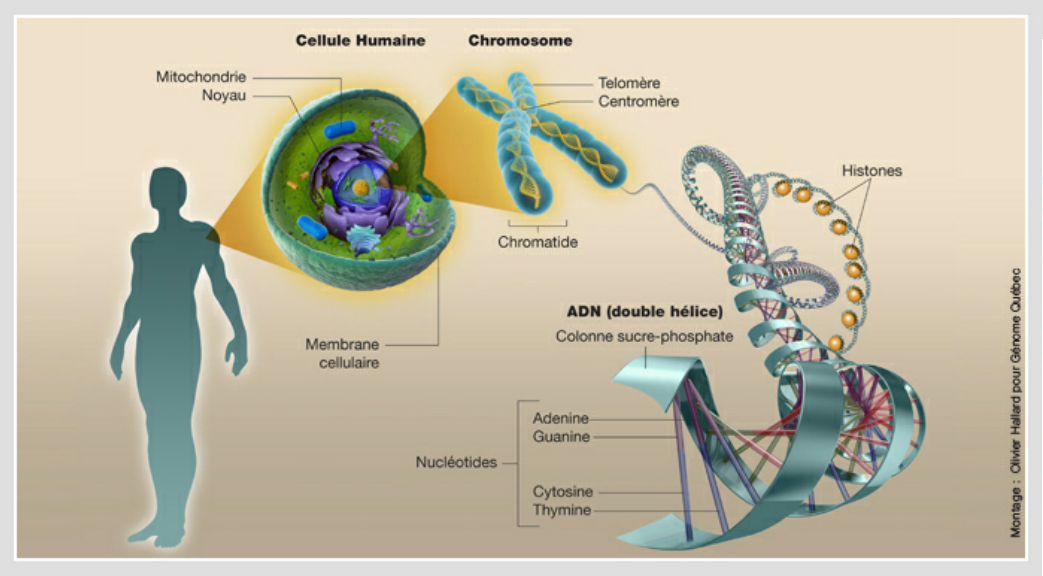
Un organisme de recherche public avec trois grands domaines de recherche

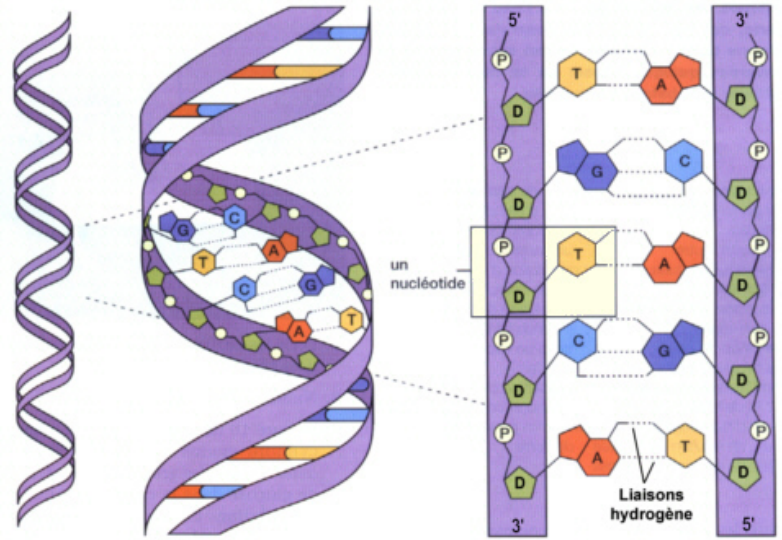
ALIMENTATION AGRICULTURE ENVIRONNEMENT

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

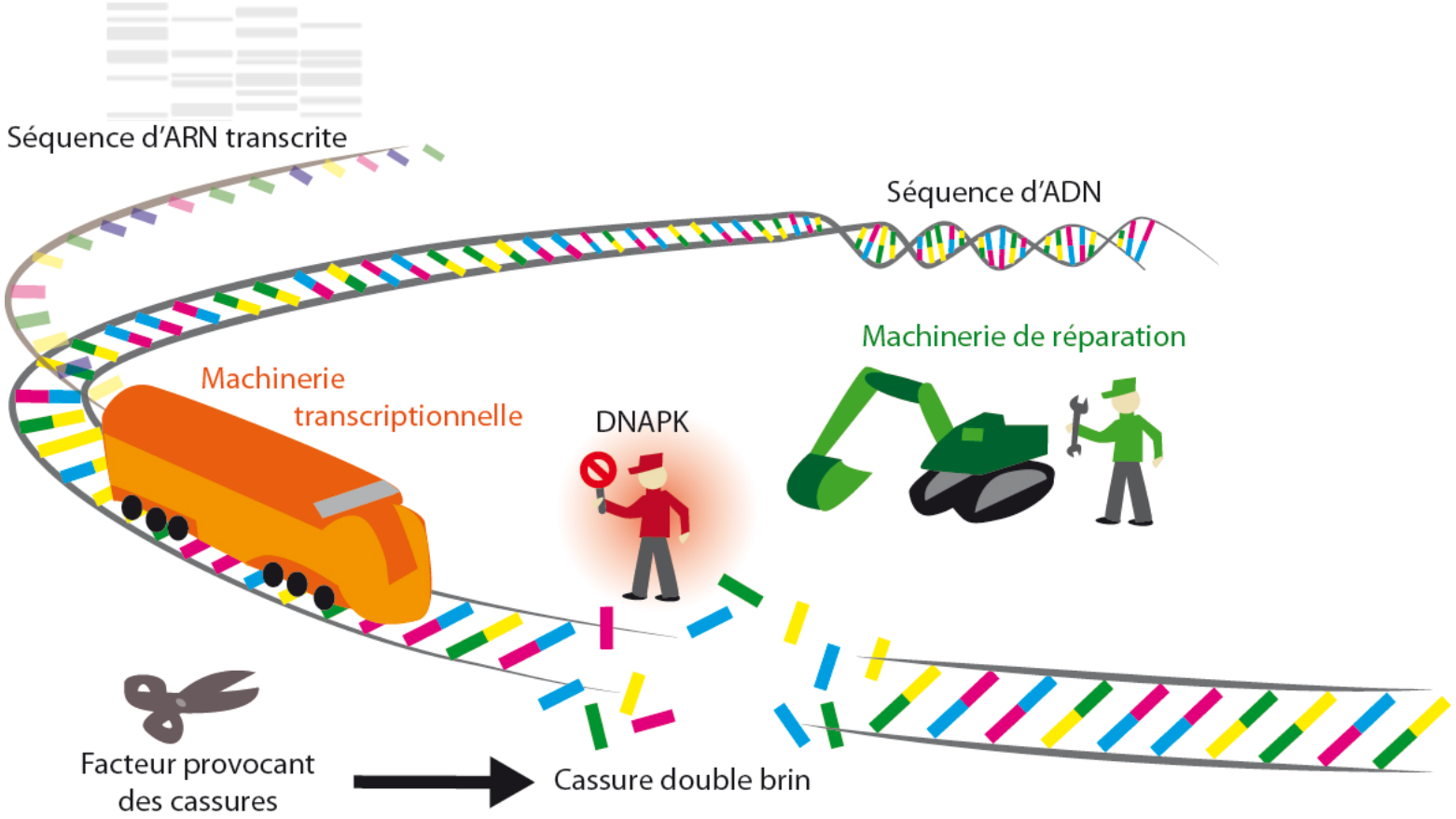


ADN, génome, chromosomes, …

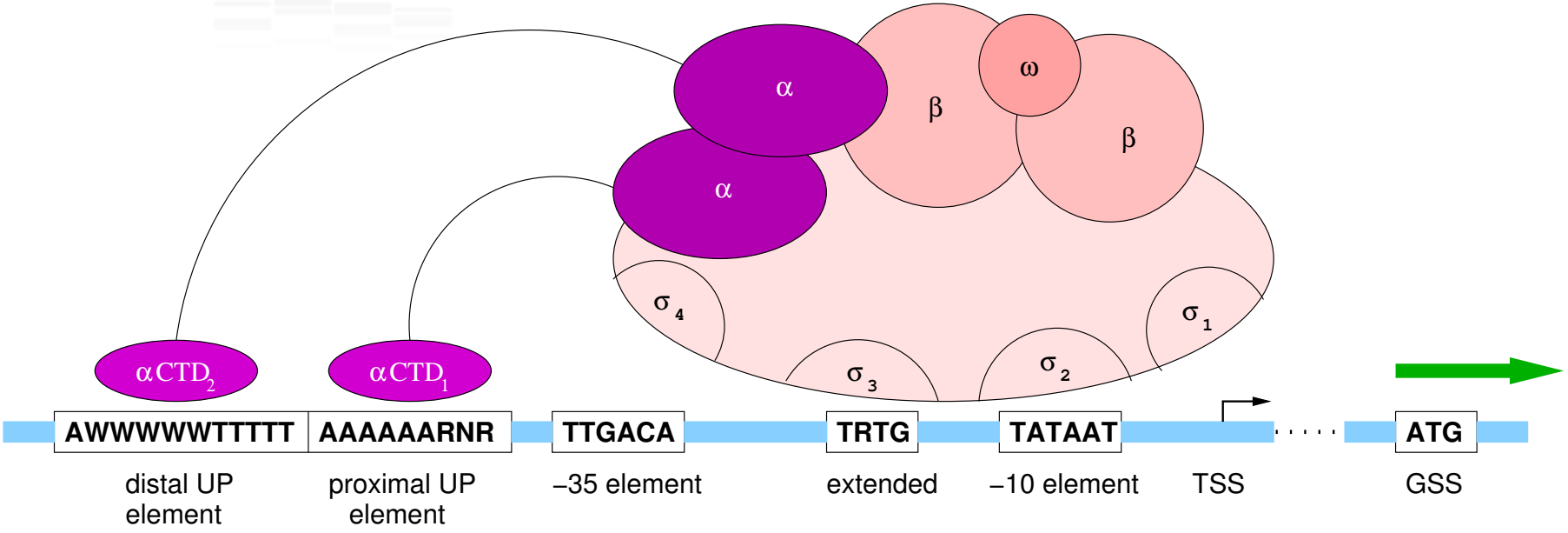




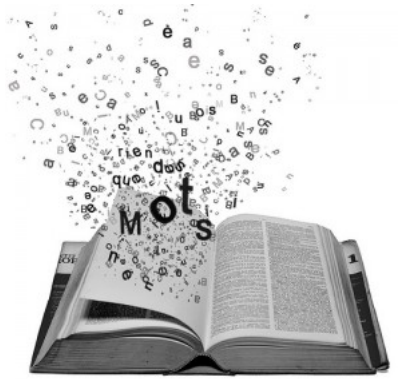
A G C C A T G T A A T G C A G T T C T G A A C C G G T


Séquence d'ARN transcrite
Séquence d'ADN
Machinerie de réparation
Machinerie transcriptionnelle
DNAPK
Facteur provocant des cassures
Cassure double brin

ADN et mots


α
α
β
β
ω
σ 4
σ 3
σ 2
σ 1
αCTD2
αCTD1
AWWWWWTTTTT
AAAAAARNR
TTGACA
TRTG
TATAAT
ATG
distal UP element
proximal UP element
–35 element
extended
–10 element
TSS
GSS

Mots exceptionnels



Chaque protéine reconnaît son mot spécifique.

Ces mots sont différents d'une espèce vivante à l'autre.

Nous en connaissons beaucoup, pour certaines protéines et certaines espèces, mais **la majorité reste à découvrir.**

Comment ?

Ces mots ont souvent des propriétés très particulières en terme de nombre d'occurrences le long du génome, par exemple ils vont être **anormalement très fréquents**.

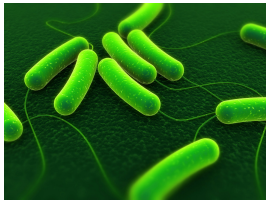
À la recherche des mots exceptionnels

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 !

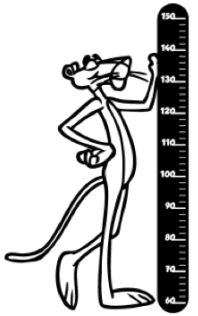


La bactérie *E. coli* a un génome de longueur 4 638 850.
Le mot de 8 lettres **GCTGGTGG** est présent 762 fois le long de ce génome.
Est-ce normal ? Est-ce anormalement beaucoup ou anormalement faible ?

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**

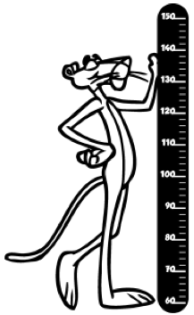
Une idée de la démarche (1)



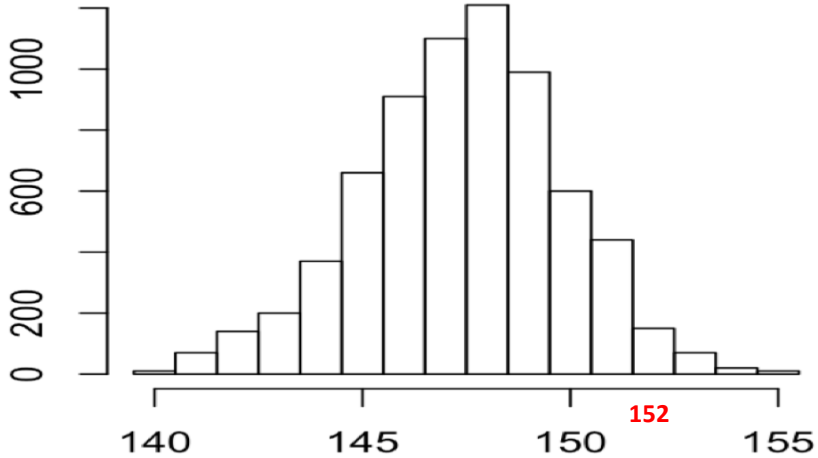
Prenons un autre exercice similaire :

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)



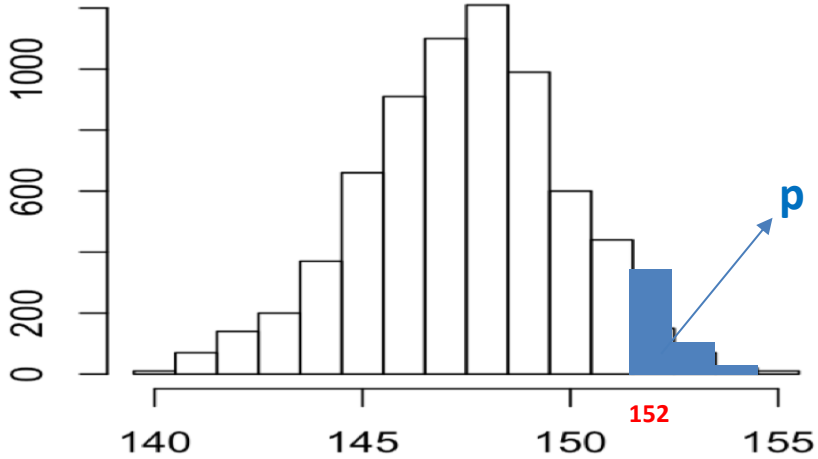
→ Collecter la taille moyenne pour chaque collège (6950) et comparer 1m52 à la « distribution » de tailles obtenue :



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 :

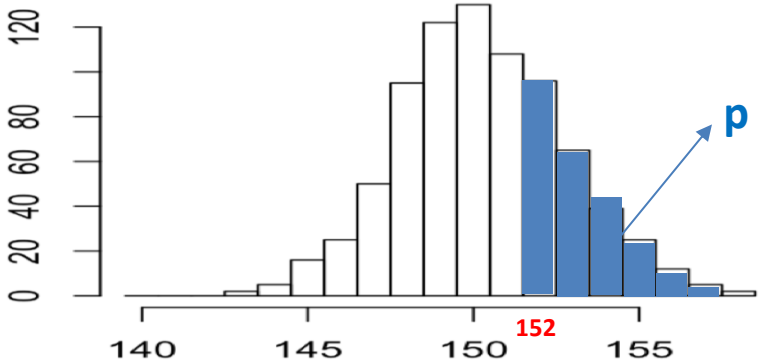


Ici, 1m52 est vraiment très à droite de la distribution, donc on a envie de dire que « **oui, les collégiens de Bures sont particulièrement grands** ! »

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

Une idée de la démarche (3)

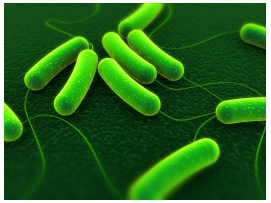
MAIS, il serait probablement plus pertinent de tenir compte du fait que « *les garçons sont généralement plus grands que les filles* » et refaire l'analyse avec seulement les collèges (797) ayant la même proportion de garçons qu'à La Guyonnerie :



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 !** »

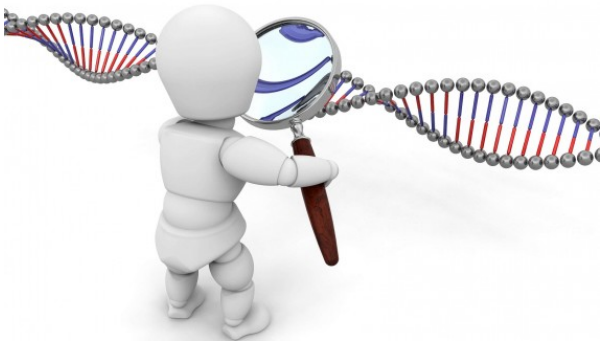
Revenons à nos mots sur l'ADN



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
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

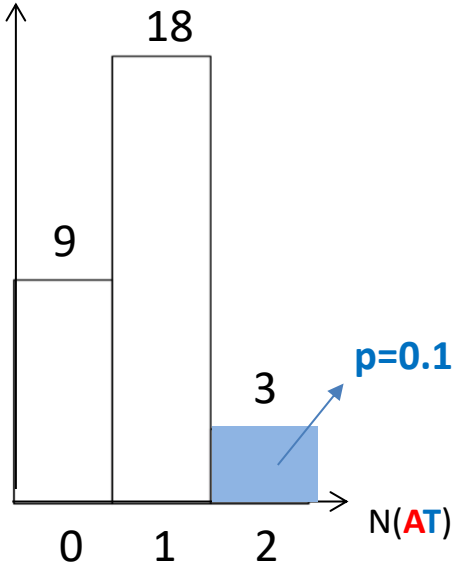
N(**AT**) = 0 dans 9 cas sur 30
N(**AT**) = 1 dans 18 cas sur 30
N(**AT**) = 2 dans 3 cas sur 30

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 :

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

N(**AT**) = 0 dans 9 cas sur 30
N(**AT**) = 1 dans 18 cas sur 30
N(**AT**) = 2 dans 3 cas sur 30
→ La moyenne est 0.8 occurrences



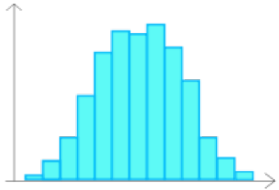
Les mathématiques à la rescousse

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.

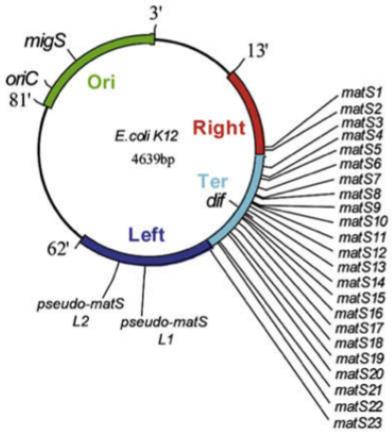
Sachant N(A), N(G), N(C), N(T), → $p < 10^{-323} \simeq 0$
Sachant N(GC), N(CT), N(TG), N(GG), etc. → $p < 10^{-323} \simeq 0$
etc.
Sachant N(GCTGGTG), N(CTGGTGG), etc, $p=1.5\ 10^{-26}$

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