Une histoire de mots innattendus et de génomes

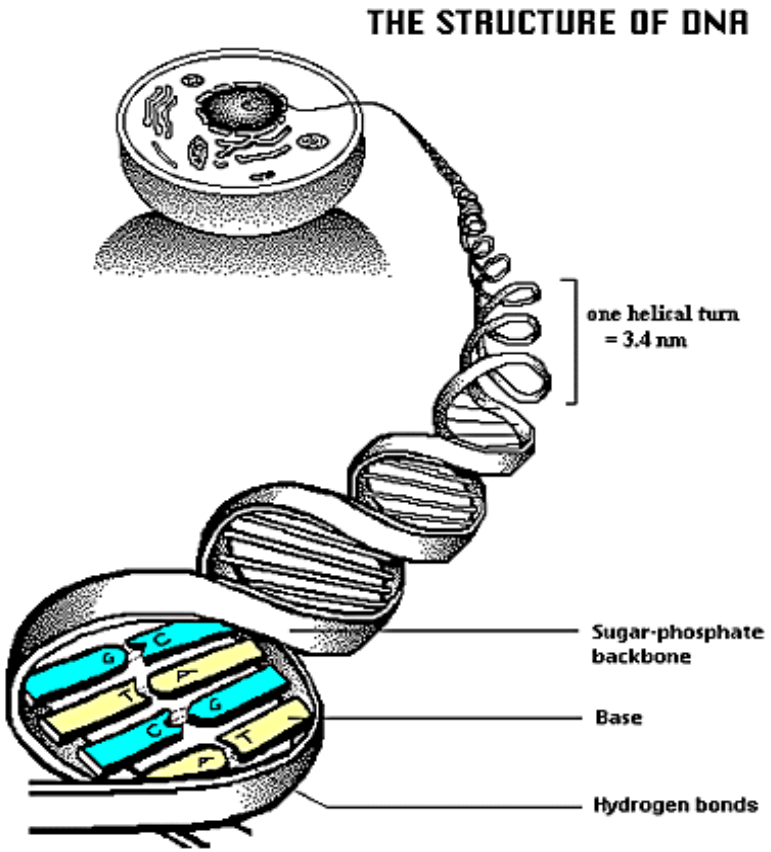
Sophie Schbath

ALEA 2017, Marseille, 22 mars 2017

Introduction

DNA and motifs

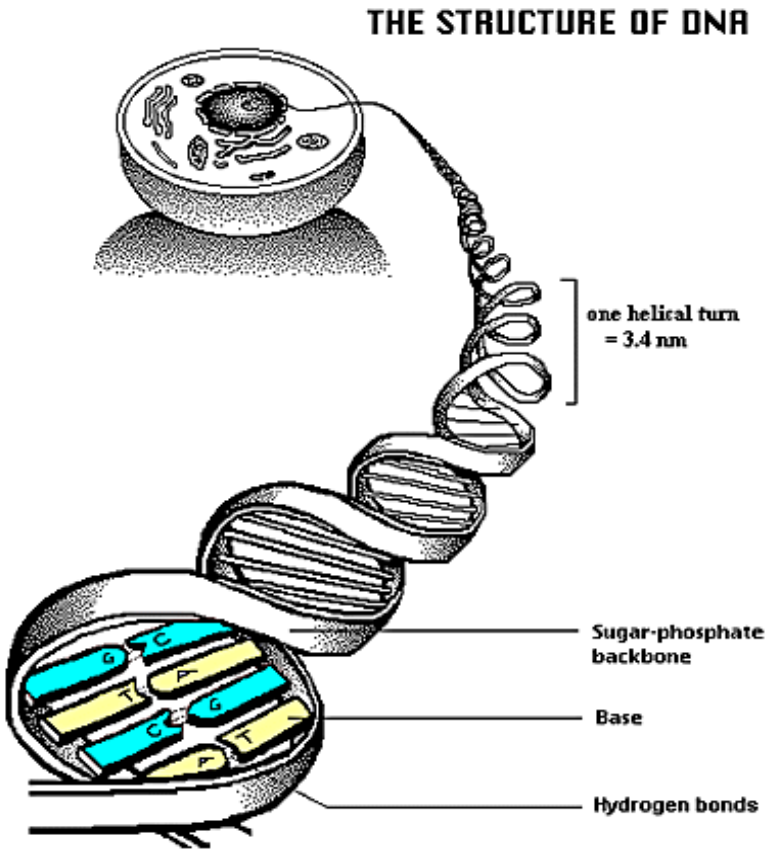
- **DNA** : Long molecule, sequence of nucleotides
- **Nucleotides** : **A**(denine), **C**(ytosine), **G**(uanine), **T**(hymine).



...GTTCAATCGTAGGTAGGTACTGAATGGTAGGTATGTTGA...

DNA and motifs

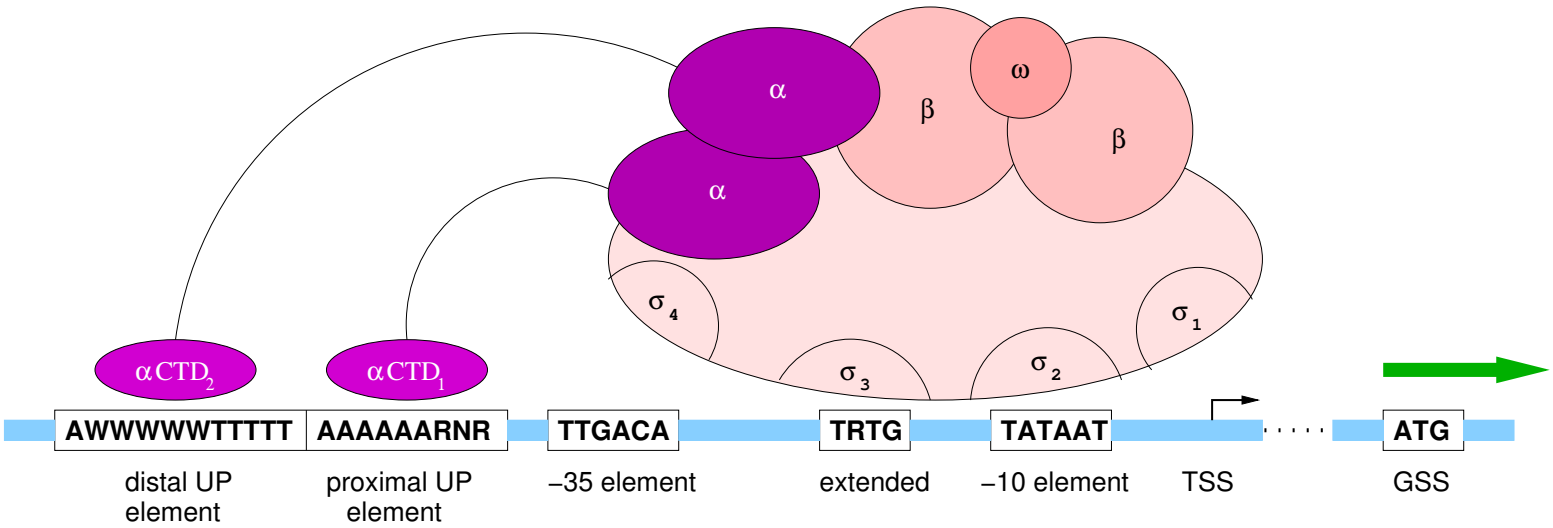
- **DNA** : Long molecule, sequence of nucleotides
- **Nucleotides** : **A**(denine), **C**(ytosine), **G**(uanine), **T**(hymine).
- **Motif** (= oligonucleotides) : short sequence of nucleotides, e.g. AGGTA



...GTTCAATCGT**A**GGT**A**GGTACTGAATGGT**A**GGTATGTTGA...

DNA and binding sites

- **Functional motif** : recognized by proteins or enzymes to initiate a biological process



Some functional motifs

- **Restriction sites** : recognized by specific bacterial restriction enzymes $\Rightarrow$ double-strand DNA break.
  E.g. GAATTC recognized by *Eco*RI

- **Chi motif** : recognized by an enzyme which processes along DNA sequence and degrades it $\Rightarrow$ enzyme degradation activity stopped and DNA repair is stimulated by recombination.
  E.g. GCTGGTGG recognized by *RecBCD* (*E. coli*)

- ***parS*** : recognized by the *Spo0J* protein $\Rightarrow$ organization of *B. subtilis* genome into macro-domains.
  $\underset{c}{\mathrm{T}}\mathrm{GTT}\overset{t}{\underset{c}{\mathrm{A}}}\underset{t}{\mathrm{AC}}\mathrm{ACGTG}\underset{t}{\mathrm{A}}\mathrm{AACA}$

- **promoter** : structured motif recognized by the RNA polymerase to initiate gene transcription.
  E.g. $\boxed{\text{TTGAC}}\overset{(16;18)}{---}\boxed{\text{TATAAT}}$ (*E. coli*).

Some functional motifs

- **Restriction sites** : recognized by specific bacterial restriction enzymes $\Rightarrow$ double-strand DNA break.
  E.g. GAATTC recognized by *Eco*RI
  very rare along bacterial genomes
- **Chi motif** : recognized by an enzyme which processes along DNA sequence and degrades it $\Rightarrow$ enzyme degradation activity stopped and DNA repair is stimulated by recombination.
  E.g. GCTGGTGG recognized by *RecBCD* (*E. coli*)
  very frequent along *E. coli* genome
- ***parS*** : recognized by the *Spo0J* protein $\Rightarrow$ organization of *B. subtilis* genome into macro-domains.
  $\underset{c}{\text{T}}\underset{c}{\text{G}}\underset{t}{\text{T}}\overset{t}{\underset{}{\text{A}}}\text{ACACGTG}\underset{t}{\text{A}}\text{AACA}$
  very frequent into the ORI domain, rare elsewhere
- **promoter** : structured motif recognized by the RNA polymerase to initiate gene transcription.
  E.g. $\boxed{\text{TTGAC}}\overset{(16;18)}{---}\boxed{\text{TATAAT}}$ (*E. coli*).
  particularly located in front of genes

Prediction of functional motifs

Most of the functional motifs are unknown in the different species.

For instance,

- which would be the **Chi motif** of *S. aureus* ? [Halpern *et al.* (07)]
- Is there an equivalent of ***parS*** in *E. coli* ? [Mercier *et al.* (08)]

Statistical approach : to identify candidate motifs based on their statistical properties.

The most over-represented 8-letter words under M1

E. coli ($\ell = 4.6\ 10^6$)

word	obs	exp	score
gctggtgg	762	84.9	73.5
ggcgctgg	828	125.9	62.6
cgctggcg	870	150.8	58.6
gctggcgg	723	125.9	53.3
cgctggtg	619	101.7	51.3

The most over-represented families *an*bcdefg under M1

H. influenzae ($\ell = 1.8\ 10^6$)

motif	obs	exp	score
gntggtgg	223	55.3	22.33
anttcatc	469	180.3	21.59
anatcgcc	288	87.8	21.38
tnatcgcc	279	84.5	21.18
gnagaaga	270	83.6	20.10

Statistical questions on word occurrences

Here are some quantities of interest.

- Number of occurrences (overlapping or not) :
  - Is $N^{\text{obs}}(\mathbf{w})$ significantly high ?
  - Is $N^{\text{obs}}(\mathbf{w})$ significantly higher than $N^{\text{obs}}(\mathbf{w}')$?
  - Is $N_1^{\text{obs}}(\mathbf{w})$ significantly more unexpected than $N_2^{\text{obs}}(\mathbf{w})$?
- Distance between motif occurrences :
  - Are there significantly rich regions with motif $\mathbf{w}$
  - Are two motifs significantly correlated ?
- Waiting time till the first occurrence :
  - Is the presence of a motif $\mathbf{w}$ significant ?

A model to define what to expect

Assessing the significance of an observed value (count, distance, occurrence, etc.) requires to define a **null model** to set what to expect.

A model for random sequences :

- **Markov chain** models : a Markov chain of order m (Mm) fits the h-mers frequencies for $h = 1, \ldots, (m+1)$.
- Hidden Markov models allow to integrate heterogeneity.

A model to define what to expect

Assessing the significance of an observed value (count, distance, occurrence, etc.) requires to define a **null model** to set what to expect.

A model for random sequences :

- **Markov chain** models : a Markov chain of order m (Mm) fits the h-mers frequencies for $h = 1, \ldots, (m+1)$.
- Hidden Markov models allow to integrate heterogeneity.

A model for the occurrence processes :

- **(compound) Poisson processes** allow to fit the number of occurrences and then to study the significance of inter-arrival times ([Robin (02)], or to compare the exceptionality of a word in two sequences ([Robin et al. (07)]).
- **Hawkes processes** allow to estimate the dependence between occurrence processes ([Gusto and S. (05)], [Reynaud and S. (10)])

Markov chains of order m : model Mm

Let $X_1X_2X_3\cdots X_\ell\cdots$ be a stationary Markov chain of order m on $\mathcal{A}=\{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}$, i.e.

$$\mathbb{P}(X_i=b\,|\,X_1,X_2,\ldots,X_{i-1})=\mathbb{P}(X_i=b\,|\,X_{i-m},\ldots,X_{i-1}).$$

Transition probabilities are denoted by

$$\pi(a_1\cdots a_m,b)=\mathbb{P}(X_i=b\,|\,X_{i-m}\cdots X_{i-1}=a_1\cdots a_m),$$

whereas the stationary distribution is given by

$$\mu(a_1a_2\cdots a_m):=\mathbb{P}(X_i=a_1,\ldots,X_{i+m-1}=a_m),\quad\forall i.$$

Markov chains of order m : model Mm

Let $X_1 X_2 X_3 \cdots X_\ell \cdots$ be a stationary Markov chain of order m on $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, i.e.

$$\mathbb{P}(X_i = b \,|\, X_1, X_2, \ldots, X_{i-1}) = \mathbb{P}(X_i = b \,|\, X_{i-m}, \ldots, X_{i-1}).$$

Transition probabilities are denoted by

$$\pi(a_1 \cdots a_m, b) = \mathbb{P}(X_i = b \,|\, X_{i-m} \cdots X_{i-1} = a_1 \cdots a_m),$$

whereas the stationary distribution is given by

$$\mu(a_1 a_2 \cdots a_m) := \mathbb{P}(X_i = a_1, \ldots, X_{i+m-1} = a_m), \quad \forall i.$$

The MLE are

$$\widehat{\pi}(a_1 \cdots a_m, a_{m+1}) = \frac{N^{\text{obs}}(a_1 \cdots a_m a_{m+1})}{N^{\text{obs}}(a_1 a_2 \cdots a_m +)}, \quad \widehat{\mu}(a_1 \cdots a_m) = \frac{N^{\text{obs}}(a_1 \cdots a_m)}{\ell - m + 1}$$

$$\rightarrow \widehat{\mathbb{E}} N(a_1 \cdots a_m a_{m+1}) \simeq N^{\text{obs}}(a_1 \cdots a_m a_{m+1})$$

Overlapping occurrences

Occurrences of words may overlap in DNA sequences (no space between words).
$\Rightarrow$ occurrences are not independent.

- Occurrences of overlapping words will tend to occur in clumps. For instance, they are 3 overlapping occurrences of CAGCAG below :

  TAGACAGATAGACGAT **C**AG**C**AG**C**AGCAG ACAGTAGGCATGA...

- On the contrary, occurrences of non-overlapping words will never overlap.

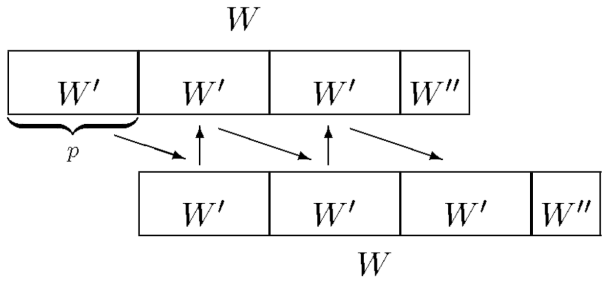
Overlapping occurrences (2)

All results on word occurrences will depend on the overlapping structure of the words.

Classically, this structure is described thanks to the periods of a word :

$$p \text{ is a period of } \mathbf{w} := w_1 w_2 \cdots w_h \text{ iff } w_i = w_{i+p}, \quad \forall i$$

meaning that 2 occurrences of $\mathbf{w}$ can overlap on $h - p$ letters.



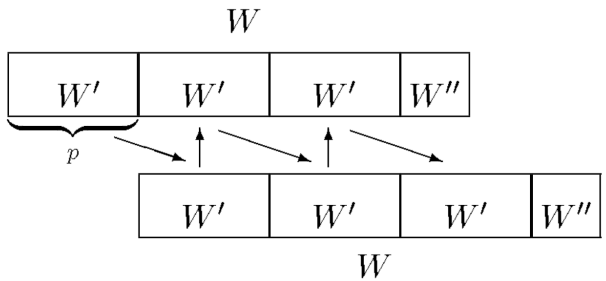
Overlapping occurrences (2)

All results on word occurrences will depend on the overlapping structure of the words.

Classically, this structure is described thanks to the periods of a word :

$$p \text{ is a period of } \mathbf{w} := w_1 w_2 \cdots w_h \text{ iff } w_i = w_{i+p}, \quad \forall i$$

meaning that 2 occurrences of $\mathbf{w}$ can overlap on $h - p$ letters.



We also define the overlapping indicator :

$$\varepsilon_{h-p}(\mathbf{w}) = 1 \text{ if } p \text{ is a period of } \mathbf{w}, \text{ and 0 otherwise}$$

Detecting words with significanly unexpected counts

Problem

Let $N(\mathbf{w})$ be the number of occurrences of the word $\mathbf{w} := w_1 w_2 \cdots w_h$ in the sequence $X_1 X_2 X_3 \cdots X_\ell$ (model M1) :

$$N(\mathbf{w}) = \sum_{i=1}^{\ell-h+1} Y_i$$

where

$$Y_i = \mathbf{1}\{\mathbf{w} \text{ starts at position } i\} \sim \mathcal{B}(\mu(\mathbf{w}))$$

and

$$\mu(\mathbf{w}) = \mu(w_1) \prod_{j=1}^{h-1} \pi(w_j, w_{j+1}).$$

Problem

Let $N(\mathbf{w})$ be the number of occurrences of the word $\mathbf{w} := w_1 w_2 \cdots w_h$ in the sequence $X_1 X_2 X_3 \cdots X_\ell$ (model M1) :

$$N(\mathbf{w}) = \sum_{i=1}^{\ell-h+1} Y_i$$

where

$$Y_i = \mathbf{1}\{\mathbf{w} \text{ starts at position } i\} \sim \mathcal{B}(\mu(\mathbf{w}))$$

and

$$\mu(\mathbf{w}) = \mu(w_1) \prod_{j=1}^{h-1} \pi(w_j, w_{j+1}).$$

Question : how to decide if $N^{\text{obs}}(\mathbf{w})$ is significantly unexpected (under model M1) ?

Ideally : one should compute the p-value $\mathbb{P}(N(\mathbf{w}) \geq N^{\text{obs}}(\mathbf{w}))$ or at least compare $N^{\text{obs}}(\mathbf{w})$ with the expected count $\mathbb{E}N(\mathbf{w})$

Hint

If the Y_i's were independent, we would have

$$\sum_{i=1}^{\ell} Y_i \sim \mathcal{B}(\ell, \mu(\mathbf{w})) \text{ approx by } \begin{cases} \mathcal{P}(\ell\mu(\mathbf{w})) & \text{if } \ell\mu(\mathbf{w}) \text{ small,} \\ \mathcal{N}(\ell\mu(\mathbf{w}), \ell\mu(\mathbf{w})(1-\mu(\mathbf{w}))) & \text{if } \ell\mu(\mathbf{w}) \sim \infty. \end{cases}$$

But **the Y_i's are not independent** (overlaps) :

- For non-overlapping words, such as ATGAC, $Y_i = 1 \Rightarrow Y_{i+1} = 0$.
- For overlapping words, such as ATGAT, $\mathbb{P}(Y_{i+3} = 1 \mid Y_i = 1) > \mathbb{P}(Y_{i+3} = 1)$.

Scores of exceptionality

- In the 80's, the ratio $\frac{N^{\text{obs}}(\mathbf{w})}{\mathbb{E}N(\mathbf{w})}$ was used with

$$\mathbb{E}N(\mathbf{w}) = (\ell - h + 1)\mu(\mathbf{w}) = (\ell - h + 1)\mu(w_1)\prod_{j=1}^{h-1}\pi(w_j, w_{j+1})$$

$\rightarrow$ problem with the variability around 1 : Var(N) ?

Scores of exceptionality

- In the 80's, the ratio $\frac{N^{\text{obs}}(\mathbf{w})}{\mathbb{E}N(\mathbf{w})}$ was used with

$$\mathbb{E}N(\mathbf{w}) = (\ell - h + 1)\mu(\mathbf{w}) = (\ell - h + 1)\mu(w_1)\prod_{j=1}^{h-1}\pi(w_j, w_{j+1})$$

  $\rightarrow$ problem with the variability around 1 : Var(N) ?

- Normalization by $\mathbb{E}N(\mathbf{w})$ like for a Poisson variable [Brendel et al. (86)]

$$\frac{N^{\text{obs}}(\mathbf{w}) - \mathbb{E}N(\mathbf{w})}{\sqrt{\mathbb{E}N(\mathbf{w})}}.$$

  $\rightarrow$ problem with the variability around 0 : Var(N) $\neq \mathbb{E}(N)$.

Scores (2)

- The variance formula was published in 1992 by Kleffe & Borodowsky. The overlapping structure of the word clearly appears.

$$\begin{aligned}\mathrm{Var}[N(\mathbf{w})] \quad = \quad & (\ell - h + 1)\mu(\mathbf{w})[1 - \mu(\mathbf{w})] \\ & + 2\mu(\mathbf{w}) \sum_{d=1}^{h-1} (\ell - h - d + 1) \left[\varepsilon_{h-d}(\mathbf{w}) \prod_{j=h-d+1}^{h} \pi(w_{j-1}, w_j) - \mu(\mathbf{w}) \right] \\ & + 2\mu^2(\mathbf{w}) \sum_{t=1}^{\ell-2h+1} (\ell - 2h - t + 2) \left[\frac{1}{\mu(w_1)} \pi^t(w_h, w_1) - 1 \right]\end{aligned}$$

One then uses the z-score and the Central Limit Theorem :

$$\frac{N(\mathbf{w}) - \mathbb{E}N(\mathbf{w})}{\sqrt{\mathrm{Var}(N(\mathbf{w}))}} \longrightarrow \mathcal{N}(0,1) \text{ as } \ell \to \infty.$$

Scores (2)

- The variance formula was published in 1992 by Kleffe & Borodowsky. The overlapping structure of the word clearly appears.

  One then uses the z-score and the Central Limit Theorem :

$$\frac{N(\mathbf{w}) - \mathbb{E}N(\mathbf{w})}{\sqrt{\mathrm{Var}(N(\mathbf{w}))}} \longrightarrow \mathcal{N}(0,1) \text{ as } \ell \to \infty.$$

  $\to$ problem when parameters (π, μ) are unknown and have to be estimated by their MLE $(\widehat{\pi}, \widehat{\mu})$

  Indeed,

$$\mathrm{Var}(N - \widehat{\mathbb{E}(N)}) \neq \widehat{\mathrm{Var}(N)}$$

Scores (3)

- Prum, Rodolphe, de Turckheim (95) proposed an appropriate normalizing factor $\widehat{\sigma}$ for $N - \widehat{\mathbb{E}N}$ which depends on the overlapping structure of the word.

  It leads to the following score

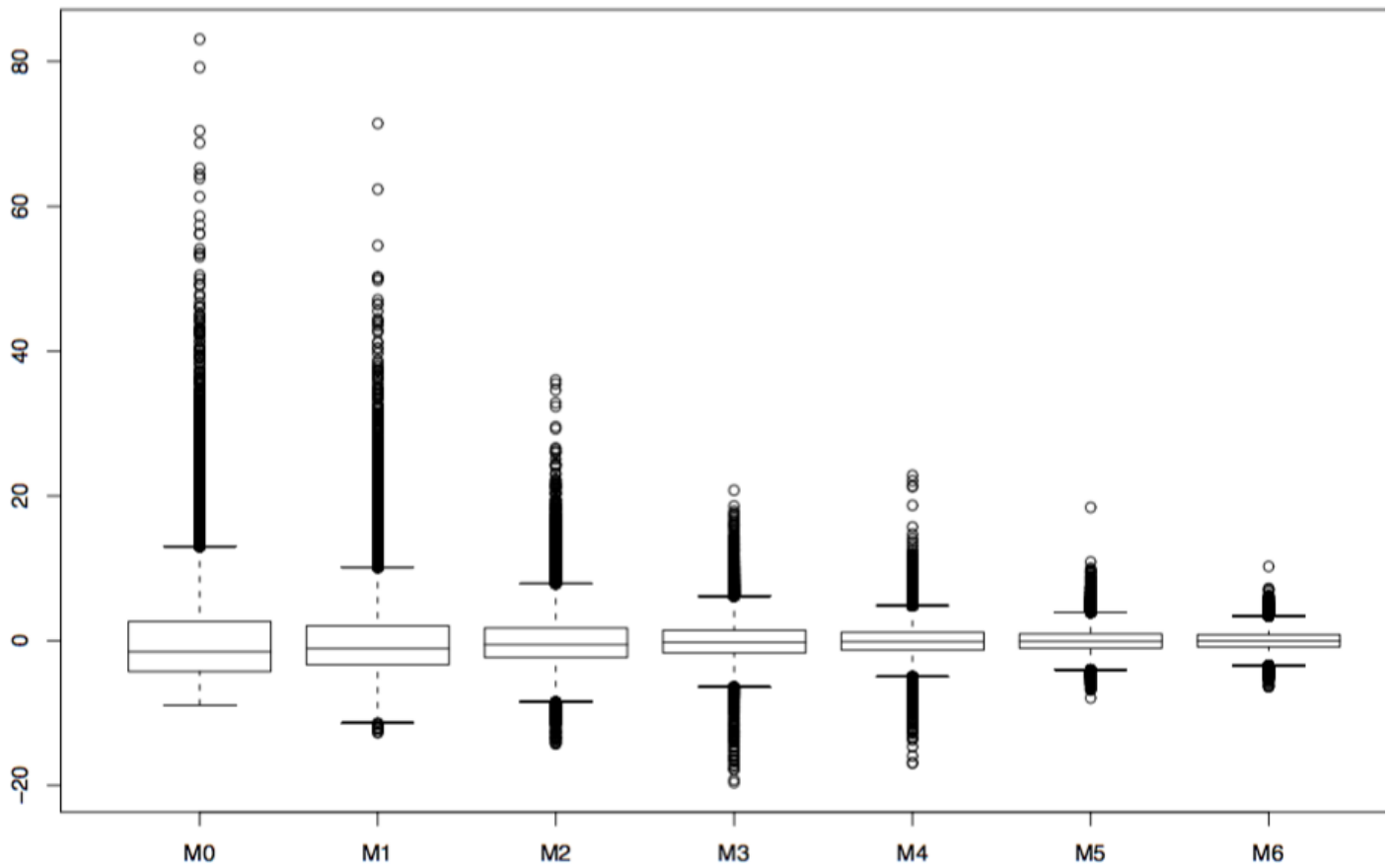
$$\frac{N(\mathbf{w}) - \widehat{\mathbb{E}N(\mathbf{w})}}{\widehat{\sigma(\mathbf{w})}} \longrightarrow \mathcal{N}(0,1) \text{ as } \ell \to \infty.$$

  and an approximation of the p-value :

$$\mathbb{P}(N \geq N^{\text{obs}}) \simeq \mathbb{P}\left(\mathcal{N}(0,1) \geq \frac{N^{\text{obs}} - \widehat{\mathbb{E}N}}{\widehat{\sigma}}\right)$$

A similar score has been derived under model Mm.

Which model to use ?



Scores of exceptionality for the 65,536 8-letter words in the *E.coli* backbone.

Influence of the model

Example : **gctggtgg** occurs 762 times in the *E. coli*'s genome (leading strands, $\ell = 4.610^6$).

model	fit	expected	score	p-value	rank
M00	length	70.783			
M0	bases	85.944	72.9	$< 10^{-323}$	3
M1	dinucl.	84.943	73.5	$< 10^{-323}$	1
M2	trinucl.	206.791	38.8	$< 10^{-323}$	1
M3	tetranucl.	355.508	22.0	$1.4\,10^{-107}$	5
M4	pentanucl.	355.312	22.9	$2.3\,10^{-116}$	2
M5	hexanucl.	420.867	19.7	$1.0\,10^{-86}$	1
M6	heptanucl.	610.114	10.6	$1.5\,10^{-26}$	3

Influence of the model (2)

	`gctggtgg` 762 occ.		`ggcgctgg` 828 occ.		`ccggccta` 71 occ.	
M0	85.944	$< 10^{-323}$ (3)	85.524	$< 10^{-323}$ (2)	70.445	0.47 (25608)
M1	84.943	$< 10^{-323}$ (1)	125.919	$< 10^{-323}$ (2)	48.173	10^{-3} (13081)
M2	206.791	$< 10^{-323}$ (1)	255.638	10^{-283} (3)	35.830	10^{-8} (4436)
M3	355.508	$1.4\,10^{-107}$ (5)	441.226	10^{-78} (15)	14.697	10^{-49} (47)
M4	355.312	$2.3\,10^{-116}$ (2)	392.252	10^{-120} (1)	15.341	10^{-46} (21)
M5	420.867	$1.0\,10^{-86}$ (1)	633.453	10^{-22} (24)	27.761	10^{-18} (36)
M6	610.114	$1.5\,10^{-26}$ (3)	812.339	0.16 (14686)	25.777	10^{-26} (4)

Expected counts and *p*-values (rank) under models M*m*,
$m = 0, 1, \ldots, 6$,
estimated from the *E. coli*'s genome (4 638 858 bps, leading strands).

Another approximation for rare words

The Gaussian approximation appeared to be not accurate for expectedly rare words ($\mathbb{E}(N(\mathbf{w})) = O(1)$ as $\ell \to +\infty$).

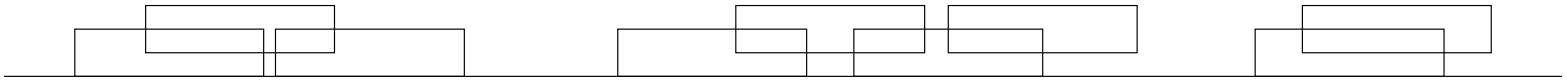
Here "**w**" is rare along the sequence.

- If **w** is not self-overlapping : $N(\mathbf{w}) \sim Pois(\mathbb{E}[N(\mathbf{w})])$ (Chen-Stein method).
- In the general case, $N(\mathbf{w})$ is approximated by a Geometric-Poisson distribution with parameter $((1 - a(\mathbf{w}))\mathbb{E}[N(\mathbf{w})]; a(\mathbf{w}))$ (S. (95)).

Both (compound) Poisson approximations are still valid when plugging the estimated parameters of the Markov model.

Compound Poisson approximation ($\mathbb{E}[N(\mathbf{w})] = O(1)$)

In the general case : clump decomposition



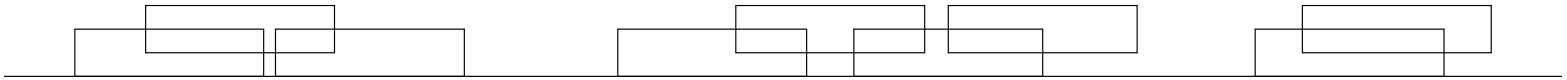
$$N(\mathbf{w}) = \sum_{c=1}^{\widetilde{N}(\mathbf{w})} K_c$$

where

- $\widetilde{N}(\mathbf{w})$ is the number of clumps
- K_c is the size of the c-th clump

Compound Poisson approximation ($\mathbb{E}[N(\mathbf{w})] = O(1)$)

In the general case : clump decomposition



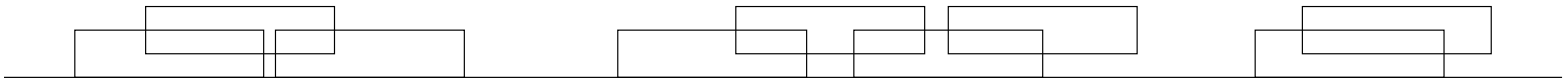
$$N(\mathbf{w}) = \sum_{c=1}^{\widetilde{N}(\mathbf{w})} K_c \qquad a(\mathbf{w}) = \sum_{\text{main periods } p} \prod_{j=1}^{p} \pi(w_j, w_{j+1}).$$

where is the overlapping proba.

- $\widetilde{N}(\mathbf{w})$ is the number of clumps
  can be approximated by $Pois((1 - a(\mathbf{w}))\mathbb{E}[N(\mathbf{w})])$ (Chen-Stein)
- K_c is the size of the c-th clump

Compound Poisson approximation ($\mathbb{E}[N(\mathbf{w})] = O(1)$)

In the general case : clump decomposition



$$N(\mathbf{w}) = \sum_{c=1}^{\widetilde{N}(\mathbf{w})} K_c \qquad a(\mathbf{w}) = \sum_{\text{main periods } p} \prod_{j=1}^{p} \pi(w_j, w_{j+1}).$$

where is the overlapping proba.

- $\widetilde{N}(\mathbf{w})$ is the number of clumps
  can be approximated by $Pois((1 - a(\mathbf{w}))\mathbb{E}[N(\mathbf{w})])$ (Chen-Stein)
- K_c is the size of the c-th clump
  follows a geometric distribution $\mathcal{G}(a(\mathbf{w}))$

Chen-Stein method

Chen(75), Stein(71), Arratia et al. (89)

$$Y_i \sim \mathcal{B}(p_i) \qquad N = \sum_{i \in I} Y_i$$

$$Z_i \sim \mathcal{P}\mathrm{o}(p_i) \text{ indep.} \qquad Z = \sum_{i \in I} Z_i$$

$$d_{\mathrm{TV}}(\mathcal{L}(N) - \mathcal{L}(Z)) \leq 2(b_1 + b_2 + b_3)$$

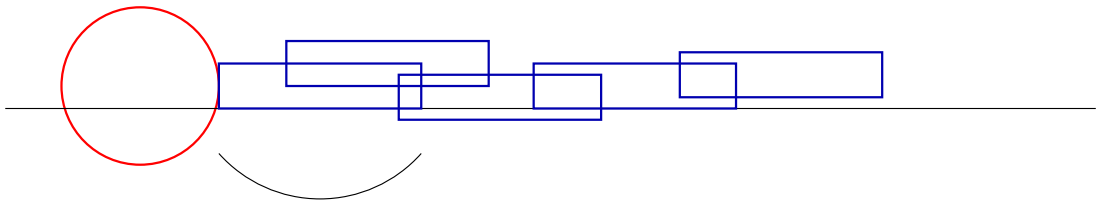
where
$$b_1 = \sum_{i \in I} \sum_{j \in B_i} \mathbb{E} Y_i \, \mathbb{E} Y_j, \quad B_i \text{ is any neighborhood of } i \text{ in } I$$

$$b_2 = \sum_{i \in I} \sum_{j \in B_i \setminus \{i\}} \mathbb{E}(Y_i Y_j)$$

$$b_3 = \sum_{i \in I} \mathbb{E} | \, \mathbb{E}(Y_i - p_i \mid \sigma(Y_j, j \notin B_i)) \, |.$$

Geometric distribution for the clump size

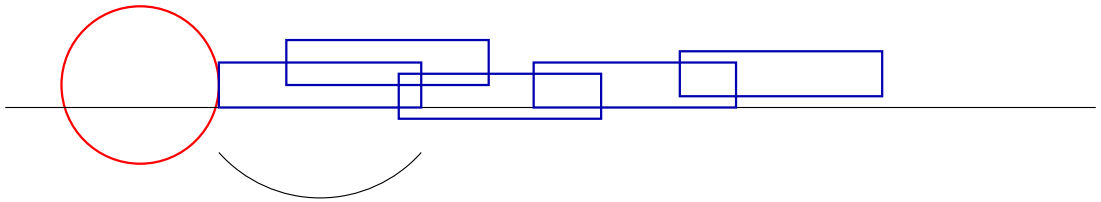
- Probability for a clump of **w** to start at a given position



$$(1-a) \quad \mu(\mathbf{w})$$

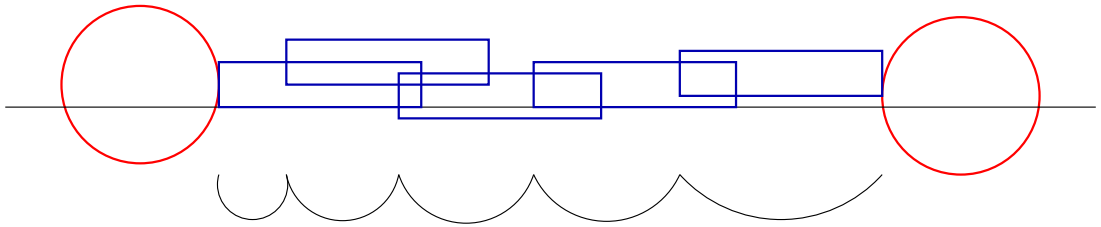
Geometric distribution for the clump size

- Probability for a clump of **w** to start at a given position



$$(1-a) \quad \mu(\mathbf{w})$$

- Probability for a k-clump of **w** to start at a given position



$$(1-a)\underbrace{a \quad a \quad a \quad a}_{a^{k-1}} \quad \mu(\mathbf{w}) \quad (1-a)$$

Other approaches

Under model M1 (with known parameters)

- the exact distribution of the count $N(\mathbf{w})$ can be computed
  - via its generating function [Régnier (00)],
  - via the duality equation $\mathbb{P}(N(\mathbf{w}) \geq x) = \mathbb{P}(T_x \leq \ell)$ where T_x is the position of the x-th occurrence ; The distribution of T_x can be obtained by recursion or via its generating function (Robin & Daudin (99), Stefanov (03)).
- large deviation technique can be used to directly approximate the p-value [Nuel (04)] :

$$\mathbb{P}(N \geq N^{\text{obs}}) \simeq \exp(-\ell\, I(N^{\text{obs}})).$$

It is a very accurate (but numerically costly) method for very exceptional words.

Prediction and identification of functional DNA motifs

Chi motifs in bacterial genomes

- Motif involved in the repair of double-strand DNA breaks.
  Chi needs to be frequent along bacterial genomes.
- Chi motifs have been identified for few bacterial species. They are not conserved through species.
- Known Chi motifs are 5 to 8 nucleotides long and can be degenerated.
- Moreover, Chi activity is strongly orientation-dependent (direction of DNA replication).
  It is present preferentially on the leading strands (high skew).

The skew of a motif $\mathbf{w}$ *is defined by* $N^{obs}(\mathbf{w})/N^{obs}(\overline{\mathbf{w}})$ *where* $\overline{\mathbf{w}}$ *is the reverse complementary of* $\mathbf{w}$*.*

E. coli as a learning case

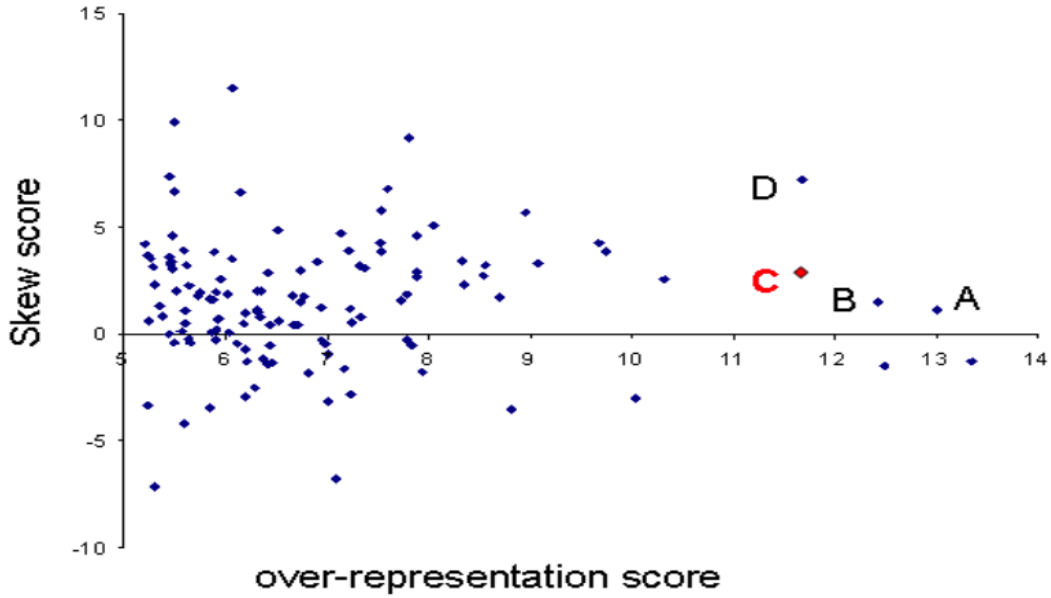
- 8-letter word GCTGGTGG
- 762 occurrences on the leading strands ($\ell = 4.6\,10^6$)
- Among the most over-represented 8-letter words (whatever the model M*m*)
  $\Rightarrow$ its frequency cannot be explained by the genome composition.
- Its rank is improved if one analyzes only the backbone genome (genome conserved in several strains of the species).
- Its skew equals 3.20 (*p*-value of 3.310^{-11}).

The skew significance can be evaluated thanks to the Gaussian approximation of word counts.

Identification of Chi motif in *S. aureus*

Halpern *et al.* (07)

- Analysis of the *S. aureus* backbone ($\ell = 2.44\,10^6$).
- 8-letter words : none of the most over-represented and skewed motifs were frequent enough.
- 7-letter words :



A=gaaaatg (1067), B=ggattag (266), C=gaagcgg (272), D=gaattag (614)

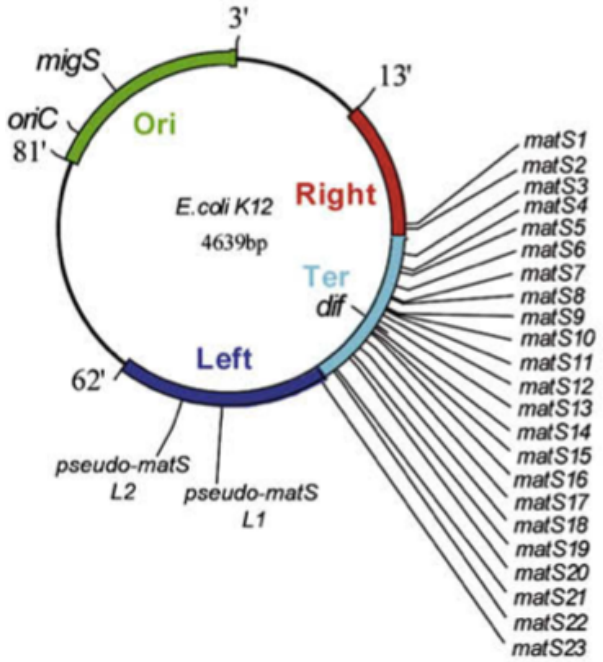
Toward more complex motifs

Signature Motif of the Ter Macrodomain of *E. coli*

The MatP/*matS* Site-Specific System Organizes the Terminus Region of the *E. coli* Chromosome into a Macrodomain

Romain Mercier,[1] Marie-Agnès Petit,[2] Sophie Schbath,[3] Stéphane Robin,[4] Meriem El Karoui,[2] Frédéric Boccard,[1,*] and Olivier Espéli[1,*]

Cell (2008)



Use of R'MES software :

- exceptional frequency
- exceptional contrast

Degenerated motifs

Definition : a word whose one or more positions may tolerate different nucleotides. The IUPAC alphabet maybe used (R=A or G, Y= C or T, N=A or C or G or T, etc.)

Examples :

- the Chi motif of *H. influenzae* is gNtggtgg
- the matS motif of *E. coli* is gtgacRNYgtcac

$\rightarrow$ one will consider them like a family $\mathcal{W}$ of words :

$$N(\mathcal{W}) = \sum_{\mathbf{w} \in \mathcal{W}} N(\mathbf{w})$$

Degenerated motifs (2)

Gaussian approximation [S. (95)] :

- $\mathbb{E}(N(\mathcal{W})) = \sum_{\mathbf{w}\in\mathcal{W}} \mathbb{E}(N(\mathbf{w}))$
- $\mathrm{Var}(N(\mathcal{W}))$: need for $\mathrm{Cov}(N(\mathbf{w}), N(\mathbf{w}'))$
  $\rightarrow$ one needs to know all possible overlaps between $\mathbf{w}$ and $\mathbf{w}'$

Degenerated motifs (2)

Gaussian approximation [S. (95)] :

- $\mathbb{E}(N(\mathcal{W})) = \sum_{\mathbf{w}\in\mathcal{W}} \mathbb{E}(N(\mathbf{w}))$
- $\mathrm{Var}(N(\mathcal{W}))$: need for $\mathrm{Cov}(N(\mathbf{w}), N(\mathbf{w}'))$
  $\rightarrow$ one needs to know all possible overlaps between $\mathbf{w}$ and $\mathbf{w}'$

Compound Poisson approximation [Roquain and S. (07)] :

- mixed clumps need to be considered (again it requires all possible overlaps between any $\mathbf{w}$ and $\mathbf{w}'$ in the $\mathcal{W}$ family)
- the clump size is no more geometric, the overlap probability $a(\mathbf{w})$ is replaced by a matrix $A = (a(\mathbf{w}, \mathbf{w}'))$
- we still get a compound Poisson distribution for $N(\mathcal{W})$

Position Weight Matrix

Here is an example of a PWM of length $h = 5$:

$$\mathbf{m} = \begin{pmatrix} 1 & 0.25 & 0 & 0.25 & 0.3 \\ 0 & 0 & 0 & 0.25 & 0.1 \\ 0 & 0.75 & 1 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.6 \end{pmatrix} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

$m_{a,j} =$ probability of letter a at motif position j

Such representation induces a set of “compatible” words having different probabilities (or “weights”)

How to count occurrences of a PWM ?

$$\mathbf{m} = \begin{pmatrix} 1 & 0.25 & 0 & 0.25 & 0.3 \\ 0 & 0 & 0 & 0.25 & 0.1 \\ 0 & 0.75 & 1 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.6 \end{pmatrix} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

Weights :

$$\nu_i = \prod_{j=1}^{h} m(X_{i+j-1}, j).$$

...GTTCGT**AGGTA**CGG**TACTG**ATGGT**AAGTA**TG**AGGCT**...

weights	0.05	0	0.02	0.1

How to count occurrences of a PWM ?

$$\mathbf{m} = \begin{pmatrix} 1 & 0.25 & 0 & 0.25 & 0.3 \\ 0 & 0 & 0 & 0.25 & 0.1 \\ 0 & 0.75 & 1 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.6 \end{pmatrix} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

Weights :

$$\nu_i = \prod_{j=1}^{h} m(X_{i+j-1}, j).$$

...GTTCGT**AGGTA**CGG**TACTG**ATGGT**AAGTA**TG**AGGCT**...

weights	0.05	0	0.02	0.1

Classical approach : to count the number of "hits" i.e. $\sum_i 1\{\nu_i \geq \alpha\}$

$\rightarrow$ If the set of words $\mathcal{W} = \{w, \nu(w) \geq \alpha\}$ is not too large, one can use previous results for a word family

How to count occurrences of a PWM ?

$$\mathbf{m} = \begin{pmatrix} 1 & 0.25 & 0 & 0.25 & 0.3 \\ 0 & 0 & 0 & 0.25 & 0.1 \\ 0 & 0.75 & 1 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.6 \end{pmatrix} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

Weights :

$$\nu_i = \prod_{j=1}^{h} m(X_{i+j-1}, j).$$

...GTTCGT**AGGTA**CGG**TACTG**ATGGT**AAGTA**TG**AGGCT**...

weights	0.05	0	0.02	0.1

Classical approach : to count the number of "hits" i.e. $\sum_i 1\{\nu_i \geq \alpha\}$

$\rightarrow$ If the set of words $\mathcal{W} = \{w, \nu(w) \geq \alpha\}$ is not too large, one can use previous results for a word family

Othewise, there exists dedicated results [Touzet & Varré (07)], [Pape et al. (08)], [Turatsinze et al. (08)].

Another approach : weighted count

Drawbacks of the classical approach :

- choice of the threshold α
- the hits are not weighted anymore

$\rightarrow$ New approach : to directly study the distribution of the weighted count defined by

$$T(\mathbf{m}) = \sum_{i=1}^{\ell-h+1} \nu_i.$$

Another approach : weighted count

Drawbacks of the classical approach :

- choice of the threshold α
- the hits are not weighted anymore

$\rightarrow$ New approach : to directly study the distribution of the weighted count defined by

$$T(\mathbf{m}) = \sum_{i=1}^{\ell-h+1} \nu_i.$$

Note : if **m** is a word, there is a unique compatible word and both counts are equal to the total word count

Ongoing results for PWM

Expectation and variance of $T(\mathbf{m})$ can be analytically derived

A Gaussian approximation can be performed

A compound Poisson approximation has been derived for $T = \sum_{c=1}^{C} K_c$:

- the number C of clumps of compatible words can be approximated by a Poisson variable with explicit parameter (Chen-Stein method)
- the distribution of K_c, the total weight of the cth clump can be simulated

A compound Poisson approximation is better than a Gaussian approximation as soon as occurrences of compatible words are rare (h large enough and card(compatible words)$<<4^h$).

NBS motif

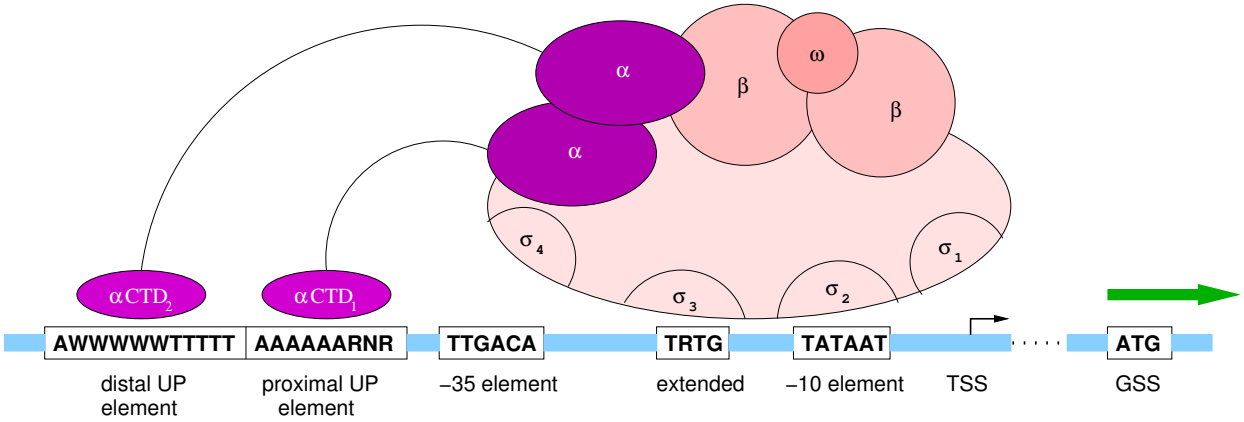
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	.88	0	0	0	0	0	0	0	0	0	0	0	0	.88
C	0	0	0	.11	1	.88	.73	.73	.88	1	.11	0	0	0
G	0	0	0	.11	0	0	.12	.12	0	0	.11	0	0	0
T	.12	1	1	.78	0	.12	.15	.15	.12	0	.78	1	1	.12

Ongoing results for PWM (2)

To be done :

- check the Gaussian approximation is good when few 0's in the PWM
- study the influence of the parameter estimation
- generalize to Markovian sequences

Structured motifs



What is the probability for a structured motif to occur in a given sequence ?

Difficulty : even for 2 boxes, previous results on word counts cannot be used because the overlapping structure is too complicated. The structure of the motif need to be considered.

Structured motifs (2)

The 2 next approaches rely on the exact distribution of the following intersite distances in Markovian sequences (recursive formula or probability generating function) [Robin and Daudin (01)], [Stefanov (03)] :

- $T_{\alpha,\mathbf{w}}$, the waiting time to reach pattern $\mathbf{w}$ from state α
- $T_{\mathbf{w},\mathbf{w}'}$, the waiting time to reach pattern $\mathbf{w}'$ from pattern $\mathbf{w}$

Structured motifs (3)

A first order approximation ([Robin *et al.* (02)])

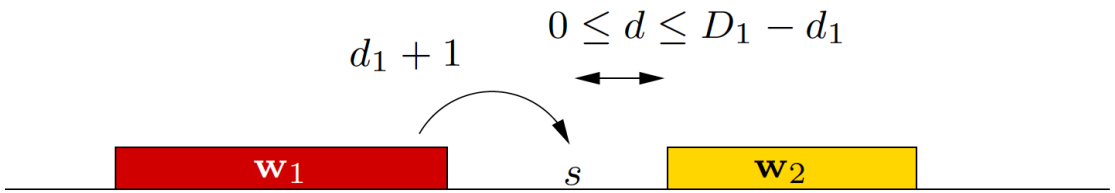
- The probability $\mathbb{P}(N(\mathbf{m}) = 0) = 1 - \mathbb{P}(N(\mathbf{m}) \geq 1)$ is approximated by

$$(1 - \mu(\mathbf{m}))\left(1 - \gamma(\mathbf{m})\right)^{\ell - |\mathbf{m}|}$$

where $\mu(\mathbf{m}) = \mathbb{P}(\mathbf{m} \text{ occurs at position } i)$

$\gamma(\mathbf{m}) = \mathbb{P}(\mathbf{m} \text{ at } i \mid \mathbf{m} \text{ not at } i-1)$

- The occurrence probability of $\mathbf{m}$ is calculated like



$$\mu(\mathbf{m}) = \mu(\mathbf{w}_1) \sum_{s \in \mathcal{A}} \pi_{u,s}^{(d_1+1)} \mathbb{P}(T_{s,\mathbf{w}_2} \leq D_1 - d_1)$$

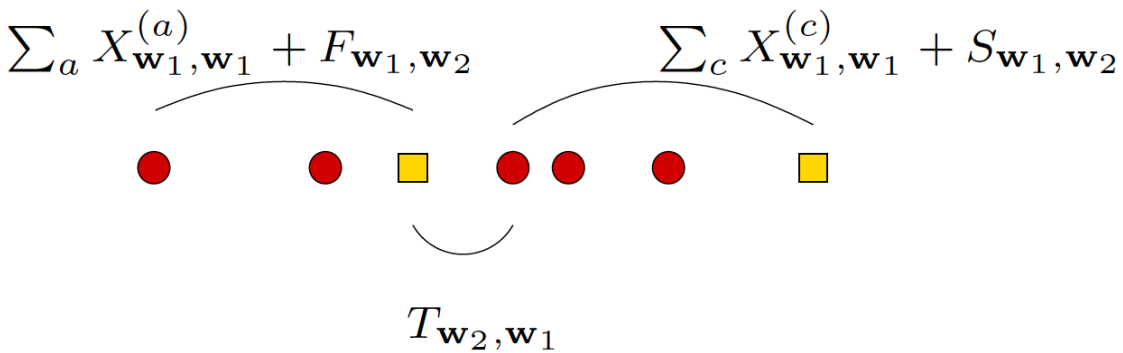
where $T_{s,\mathbf{w}_2}$ is the waiting time to reach pattern $\mathbf{w}_2$ from state s

Structured motifs (4)

An exact approach via random sums ([Stefanov *et al.* (06)])

- Assumption : $\mathbf{w}_2$ should not occur in between the 2 boxes.
- The explicit formula for the pgf of $\tau_{\mathbf{m}}$ is given thanks to the following decomposition :

$$\tau_{\mathbf{m}} \stackrel{\mathcal{D}}{=} T_{\alpha,\mathbf{w}_1} + \sum_{b=1}^{L'} \left(\underbrace{\sum_{a=1}^{L_1} X^{(ab)}_{\mathbf{w}_1,\mathbf{w}_1} + F^{(b)}_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{failure}} + T^{(b)}_{\mathbf{w}_2,\mathbf{w}_1} \right) + \underbrace{\sum_{c=1}^{L_2} X^{(c)}_{\mathbf{w}_1,\mathbf{w}_1} + S_{\mathbf{w}_1,\mathbf{w}_2}}_{\stackrel{\mathcal{D}}{=} T_{\mathbf{w}_1,\mathbf{w}_2} \mid \text{success}},$$



L_1, L_2 and L' are independent geometric variables.

Bibliography

Overviews :

- REINERT, G., SCHBATH, S. and WATERMAN, M. (2005). *Applied Combinatorics on Words*. volume 105 of *Encyclopedia of Mathematics and its Applications*, chapter Statistics on Words with Applications to Biological Sequences. Cambridge University Press.
- REINERT, G., SCHBATH, S. and WATERMAN, M. (2000). Probabilistic and statistical properties of words : an overview. *J. Comp. Biol.* **7** 1–46.
- ROBIN, S., RODOLPHE, F. and SCHBATH, S. (2005). *DNA, Words and Models*. Cambridge University Press.
- ROBIN, S., RODOLPHE, F. and SCHBATH, S. (2003). *ADN, mots et modèles*. BELIN.
- SCHBATH, S. and ROBIN, S. (2008). *How pattern statistics can be useful for DNA motif discovery ?*. To appear in *Scan Statistics - Methods and Applications*, Glaz, J., Pozdnyakov, I. and Wallenstein , S. Eds., Statistics for Industry and Technology series, Birkhauser.

Bibliography

Word count (Poisson approximations) :

- ERHARDSSON, T. (2000). Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains *Ann. Appl. Prob.*, **10**, 573–591.
- GESKE, M. X., GODBOLE, A. P., SCHAFFNER, A. A., SKOLNICK, A. M. and WALLSTROM, G. L. (1995). Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Prob.*, **32**, 877–892.
- REINERT, G. and SCHBATH, S. (1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol.*, **5**, 223–253.
- ROQUAIN, E. and SCHBATH, S. (2007). Efficient compound Poisson approximation for the number of occurrences of multiple words in Markov chains. *Adv. Appl. Prob.*, **39**, 128–140.
- SCHBATH, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM : Probability and Statistics*. **1** 1–16.

Bibliography

Word count (others) :

- KLEFFE, J. and BORODOVSKY, M. (1992). First and second moment of counts of words in random texts generated by Markov chains *Comp. Applic. Biosci.*, **8**, 433–441.
- NUEL, G. (2004). LD-SPatt : Large Deviations Statistics for Patterns on Markov chains. *J. Comp. Biol.*
- PRUM, B. RODOLPHE, F., TURCKHEIM, É. (1995). Finding words with unexpected frequencies in DNA sequences *J. R. Statist. Soc. B*, **57**, 205–220.
- PUDLO, P. (2004). Estimations précises de grandes déviations et applications à la statistique des séquences biologiques. PhD thesis, Université Lyon I.
- RÉGNIER, M. (2000). A unified approach to word occurrence probabilities *Discrete Applied Mathematics*, **104**, 259–280.
- RÉGNIER, M. and DENISE, A. (2004). Rare events and conditional events on random strings. *Discrete Mathematics and Theoretical Computer Science* **6** 191–214.
- ROBIN, S. and SCHBATH, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.* **8** 349–359.
- ROBIN, S., SCHBATH, S. and VANDEWALLE, V. (2007). Statistical tests to compare motif count exceptionalities. *BMC Bioinformatics* **8 :84** 1–20.

Bibliography

Distances and waiting times :

- ROBIN, S. (2002). A compound Poisson model for words occurrences in DNA sequences *J. Royal Statist. Soc., C series*, **51**, 437–451.
- ROBIN, S. and DAUDIN, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters *J. Appl. Prob.*, **36**, 179–193.
- ROBIN, S. and DAUDIN, J.-J. (2001). Exact distribution of the distances between any occurrences of a set of words *Ann. Inst. Statist. Math.*, **36**, 895–905.
- ROBIN, S., DAUDIN, J.-J., RICHARD, H., SAGOT, M.-F. and SCHBATH, S. (2002). Occurrence probability of structured motifs in random sequences. *J. Comp. Biol.* **9** 761–773.
- STEFANOV, V. (2003). The intersite distances between pattern occurrences in strings generated by general discrete - and continuous- time models : an algorithmic approach *J. Appl. Prob.*, *40*.
- STEFANOV, V., ROBIN, S. and SCHBATH, S. (2007). Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Applied Mathematics*. **155** 868–880.

Bibliography

Prediction of functional motifs :

- 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*.
- TOUZAIN, F., SCHBATH, S., DEBLED-RENNESSON, I., AIGLE, B., LEBLOND, P. and KUCHEROV, G. (2008). SIGffRid : a tool to search for σ factor binding sites in bacterial genomes using comparative approach and biologically driven statistics. *BMC Bioinformatics.*. **9 :73** 1–23.

Others :

- GUSTO, G. and SCHBATH, S. (2005). FADO : a statistical method to detect favored or avoided distances between motif occurrences using the Hawkes' model. *Statistical Applications in Genetics and Molecular Biology*.
- REYNAUD-BOURET, P. and SCHBATH, S. (2010). Adaptive estimation for Hawkes' processes ; Application to genome analysis. *Annals of Statistics.* **38 (5)** 2781–2822.