1

How Pattern Statistics Can Be Useful For DNA Motif Discovery?

Sophie Schbath[1] **and Stéphane Robin**[2]

[1] *INRA, UR1077 Mathématique, Informatique et Génome, F-78350 Jouy-en-Josas, France*
[2] *AgroParisTech/INRA, UMR518 Mathématique et Informatique Appliquées, F-75231, Paris, France*

Abstract: Statistics of motifs have been widely revisited in the last 15 years due to the increasing availability of genomic sequences. The identification of DNA motifs with biological functions is still a huge challenge of genome analysis. Many functional and essential motifs have the particularity to be very frequent all along the chromosome or to be concentrated in some particular regions (e.g. in front of genes) or to be co-oriented with the replication direction. The prediction of functional motifs is then mostly based on statistical properties of pattern occurrences in Markovian sequences. This paper will be mostly devoted to such properties with a special focus on pattern frequency. How to compute or approximate the count distribution to assess motif exceptionality? How to test if a motif is significantly unbalanced between two (sets of) sequences? How to deal with degenerated patterns? How to model occurrences to find regions significantly enriched with a given pattern? etc. Examples of functional motifs will illustrate all these questions and we will see how the Chi motif has been identified in *Staphylococcus aureus* thanks to its statistical properties.

Keywords and phrases: Pattern statistics, word count, Markov chain, DNA sequence, exceptional words, unexpected frequency, compound Poisson process

1.1 Introduction

Genomic sequence analysis is probably the applied domain which has been offering for the last 15 years the widest variety of problems on pattern statistics. This variety is due to the huge length of the sequences, to their heterogeneous composition and structure, but also to the complexity of the functional motifs. These motifs take place in fundamental molecular processes like chromosome

maintenance or gene transcription but few of them have been completely identified (i.e. whose sequence of letters is known). Moreover, they are rarely conserved through species leading to a very challenging area of DNA motif discovery. This chapter is related to the statistical approach used to predict candidate functional motifs. Indeed, many known functional motifs are characterized by an exceptional behavior of their occurrences. Some of them are extremely frequent along the entire genome (or along a particular DNA strand), others are avoided because their occurrences are lethal for the chromosome, some are preferred in particular genomic regions. Two main quantities have then been widely studied from a probabilistic and statistical point of view: the number of occurrences of a motif in a random sequence and the distances (cumulated or not) between occurrences of a motif. To avoid a huge list of references, we just point to chapter 6 and 7 from Lothaire (2005) for technical expositions and to Robin *et al.* (2005) for a more applied exposition. In this chapter, we have chosen to present the main statistical results that are really used in practice to help identifying functional DNA motifs. Many biological examples will then be given to illustrate the usefulness of the approaches. The biggest part will be devoted to the question of detecting words with an exceptional frequency in a given sequence; Distribution of a word count in Markovian sequences will then be studied (Section 1.2). We will also consider the related problem of comparing the exceptionality of a word frequency between two independent sequences. Functional motifs can indeed be specific from known parts of the chromosome (or from some particular chromosomes); In this case, the word occurrences themselves are modeled and a statistical test is derived from the two count processes (Section 1.3). However, when one look for regions significantly enriched (or devoid) of a given word, the quantity of interest becomes the distance between occurrences. Section 1.3 also presents results on the distance distribution when the occurrences are modeled by a compound Poisson process. Other results on distances and waiting times can be found in Stefanov's chapter when the sequence is Markovian. Section 1.4 addresses the generalization to more complex patterns, namely degenerated patterns and structured motifs. Finally, we end with some ongoing works and open problems.

1.2 Words With Exceptional Frequency

Lots of functional DNA motifs are extremely over-represented in complete genomes, or in specific genomic regions, whatever the composition level of the biological sequence one takes into account. This statistical property reveals a strong constraint on the DNA sequence. For instance, if we look for the two most over-represented 9-letter words in the complete genome of the bacteria

Haemophilus influenzae (1830140 letters long), we find the two reverse complementary oligo-nucleotides `aagtgcggt` and `accgcactt` which occur respectively 740 and 731 times. As an illustration, Table 1.1 just gives the expected count of these two words when fitting the sequence composition of smaller words. These

Markov model	fitted composition	expected count of `aagtgcggt`	expected count of `accgcactt`
M0	letters	4.694	3.779
M1	2-letter words	6.279	4.847
M2	3-letter words	8.603	6.208
M3	4-letter words	18.601	15.080
M4	5-letter words	55.704	48.658
M5	6-letter words	219.081	220.284
M6	7-letter words	549.815	574.734
M7	8-letter words	719.440	722.366

Table 1.1: Expected counts of`aagtgcggt` and `accgcactt` in random sequences having in average the same composition than the *H. influenzae* complete genome.

two 9-letter words are in fact very well known from the biologists: they are the two DNA *uptake* sequences involved in discriminating self from foreign entering DNA during competence in the bacteria.
Another example is the word `gctggtgg` which is the "crossover hotspot instigator" (*Chi*) motif in the bacteria *Escherichia coli* and is involved in chromosome maintenance. Chi is among the 5 most over-represented 8-letter words in the *E. coli* genome (4638858 letters long). This example will be detailed in Section 1.2.5.
On the contrary, many restriction sites (generally 6-letter words) are strongly under-represented along bacterial genomes, which is not surprising because they induce a double-strand break of the bacterial DNA.
The aim of this section is precisely to show how to assess the significance of over- and under-representations.

When we want to analyze the distribution of a word along a sequence or when we want to know if a word occurs significantly more often in one sequence compared to another one (Section 1.3), it is relevant to model the occurrences themselves in order to fit the observed frequencies of this word. However, if the problem is precisely to know if a given word occurs in a DNA sequence with a frequency that seems either too low or too high, one needs to compare it to an expected frequency. Usually, one compares the observation with what one would expect in random sequences sharing common properties with the DNA sequence. Under classical sequence models (Section 1.2.1), we can analytically

calculate the moments of the count (Section 1.2.2) and sometimes get its distribution or some approximations (Section 1.2.3), leading to p-values (Section 1.2.4). We will end this section by presenting how the Chi motif of *Staphylococcus aureus* has been predicted, thanks to its exceptional frequency, before being experimentally validated [Halpern *et al.* (2007)].

1.2.1 Sequence models

The commonly used sequence models have the property to fit the letter composition of the observed sequence and more generally its composition in small words of a given length. For instance, it is usual to fit the 3-letter word composition of coding DNA sequences because the letters of these sequences are read 3 by 3 by the ribosome which translates each disjoint triplets into amino acids to form a protein. The most intuitive model is therefore the permutation model (or shuffling model) consisting in shuffling the letters of the observed sequence so that the composition remains exactly the same. Preserving exactly the letter composition is an easy task but it is more difficult for 2-letter words or longer words, both from algorithmic and probabilistic points of view. In that respect, stationary Markov chains are particularly interesting if one accepts to fit the composition in average rather than exactly. Moreover, if one wants to take some periodicity or an heterogeneous composition along the sequence into account, permutation models become very complicated to manipulate.

In the remainder, we will consider a random sequence $\mathbf{S} = X_1 X_2 \cdots X_n$ on the four-letter DNA alphabet, i.e. $X_i \in \mathcal{A} := \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

Permutation models These models assume that random sequences are uniformly drawn from the set $\mathcal{S}_m$ of sequences having exactly the same counts of words of length 1 up to m than the observed DNA sequence, for a given integer $m \geq 1$. The probability of a sequence $\mathbf{S}$ is then $1/|\mathcal{S}_m|$. For $m = 1$ or $m = 2$, for instance, we have:

$$\begin{aligned} |\mathcal{S}_1| &= \frac{n!}{N_{\text{obs}}(\mathtt{a})! \times N_{\text{obs}}(\mathtt{c})! \times N_{\text{obs}}(\mathtt{g})! \times N_{\text{obs}}(\mathtt{t})!} \\ |\mathcal{S}_2| &= \prod_{a \in \mathcal{A}} \frac{N_{\text{obs}}(a+)!}{\prod_{b \in \mathcal{A}} N_{\text{obs}}(ab)!} \times H_{X_n, X_1}(\mathcal{S}) \end{aligned}$$

where $N_{\text{obs}}(\cdot)$ denotes the count in the observed sequence $\mathbf{S}_{\text{obs}}$, $N_{\text{obs}}(a+) := \sum_b N_{\text{obs}}(ab)$ and $H_{X_n,X_1}(\mathcal{S})$ is the cofactor corresponding to row X_n and column X_1 of the matrix $\big(\mathbb{1}\{a = b\} - N(ab)/N(a+)\big)_{a,b \in \mathcal{A}}$ [Whittle (1955)]. Note that the constraint for $\mathbf{S} \in \mathcal{S}_2$ to have the same letter composition than $\mathbf{S}_{\text{obs}}$ is equivalent to start (resp. to end) with the first (resp. last) letter of $\mathbf{S}_{\text{obs}}$. Indeed, we have $N_{\text{obs}}(a+) = N_{\text{obs}}(a)$ for all $a \in \mathcal{A}$ but the last nucleotide of $\mathbf{S}_{\text{obs}}$; The counts then differ from 1. Knowing the letter composition, additionally

to the dinucleotide composition, determines the last letter X_n of the sequences $\mathbf{S} \in \mathcal{S}$. It is the same for the first letter X_1 by using the numbers $N_{\text{obs}}(+b)$ of dinucleotides that end with b.
Working with these permutation models require lots of combinatorics.

Stationary Markov chains Let us consider the first order stationary Markov model, denoted by M1; It means that the random letters X_i's are not independent and satisfy the following Markov property: $\mathbb{P}(X_i = b \mid X_1, X_2, \ldots, X_{i-1}) = \mathbb{P}(X_i = b \mid X_{i-1})$, $\forall b \in \mathcal{A}$. The transition probabilities will be denoted as follows:

$$\pi(a, b) = \mathbb{P}(X_i = a \mid X_{i-1} = b), \forall a, b \in \mathcal{A};$$

$\Pi = (\pi(a, b))_{a,b}$ will denote the transition matrix. Moreover, all X_i's have the same distribution, namely the stationary distribution μ which satisfies the relation $\mu = \mu\Pi$.

The transition probabilities are estimated by their maximum likelihood estimators, i.e.

$$\widehat{\pi}(a, b) = \frac{N(ab)}{N(a+)}, \quad a, b, \in \mathcal{A}, \tag{1.1}$$

where $N(\cdot)$ denotes the number of occurrences in the sequence $\mathbf{S} = X_1 X_2 \cdots X_n$. Moreover, the letter probability $\mu(a)$ is usually estimated by $\widehat{\mu}(a) = \frac{N(a)}{n}$.

An important consequence of such estimation is that the plug-in estimator of the expected number of ab in model M1 is approximately equal to the observed count of ab in the DNA sequence. Indeed, we will see in Section 1.2.2 than $\mathbb{E}[N(ab)] = (n-1)\mu(a)\pi(a, b)$ which leads to

$$\widehat{\mathbb{E}}[N(ab)] := (n-1)\widehat{\mu}(a)\widehat{\pi}(a, b) \simeq N(ab).$$

In other words, model M1 fits in average the 2-letter word composition of the observed sequence.

Similarly, the stationary m-th order Markov chain model (Mm) fits in average the $(m+1)$-letter word composition of the observed sequence. In practice, the choice of the order m of the model Mm is important because it defines the set of reference sequences and, as we will see in Section 1.2.5, this choice often has a strong influence on the statistical results. This influence can already be observed on Table 1.1: expected counts vary a lots with respect to the chosen model.
Since model Mm on the $\mathcal{A}$ alphabet can be considered like a model M1 on the larger alphabet $\mathcal{A}^m$, we will focus on first order Markov chains in this chapter.

Phased Markov chains for coding sequences The interest of considering phased Markov chains came from the analysis of coding DNA sequences. Such sequences are split into adjacent 3-letter words called codons, each of them

being translated into an amino acid to form a protein. The succession of codons ensures the reading frame for the translation. The nucleotides of a coding DNA sequence are then alternatively the first letter of a codon, the second letter of a codon, the third letter of a codon, and so on. The phase of a nucleotide is its position with respect to the codons; A letter can then be in 3 different phases in a coding sequence. The three positions of a codon do not have the same importance. First of all, an amino acid is often determined by the two first letters of a codon according to the genetic code. Moreover, the 3D structure of the protein usually implies constraints on the succession of amino acids. It is therefore important to take the phase of the nucleotides into account when modeling coding DNA sequences.

In a phased Markov chain of order 1, the transition probability from letter a to letter b depends on the phase $\phi \in \{1,2,3\}$ the nucleotide b will be. We then have the three following transition probabilities

$$\pi_\phi(a,b) = \mathbb{P}(X_{3i+\phi} = b \mid X_{3i+\phi-1} = a), a, b \in \mathcal{A}.$$

We can also define the distributions $\boldsymbol{\mu}_\phi$ of letters on each phase $\phi \in \{1,2,3\}$; They satisfy $\mu_1 = \mu_3\Pi_1$, $\mu_2 = \mu_1\Pi_2$ and $\mu_3 = \mu_2\Pi_3$.

When estimating these parameters by the maximum likelihood method, it allows to fit in average the composition of the coding DNA sequence in ab's on phase 1, in ab's on phase 2 and ab's on phase 3, for all $a, b \in \mathcal{A}$.

Thanks to an appropriate change of alphabet, the phased Markov model on the $\mathcal{A}$ alphabet can be considered like a model M1 on $\mathcal{A} \times \{1,2,3\}$. It suffices to rewrite the sequence $\mathbf{S}$ over the alphabet $\mathcal{A} \times \{1,2,3\}$ by defining $X_i^\star = (X_i, i \text{ modulo } 3)$. The transition probability from (a,ϕ') to (b,ϕ) is then equal to $\pi_\phi(a,b)$ if $\phi = \phi' + 1$ modulo 3, and 0 otherwise.

Heterogeneous Markov models Some entire chromosomes have been now completely sequenced for several years, and it has been quickly noticed that their composition is more or less heterogeneous. Many reasons may explain this heterogeneity: genes are more constrained than intergenic regions because they have to code for functional proteins, bacterias can exchange genomic regions (so-called horizontal transfers) but they all have their own signature in terms of composition, etc. It is then natural to use heterogeneous Markov models. Usually the heterogeneity is considered like a piecewise homogeneity, i.e. homogeneous regions alternate along the genome. If the heterogeneity is known in advance (for instance genes/intergenic regions), one may then use piecewise homogeneous Markov models. When the aim is precisely to recover the heterogeneous structure then the most popular models in genome analysis are hidden Markov models. Note that a hidden Markov chain with a hidden state space $\mathcal{Q}$ and an observation space $\mathcal{A}$ can be considered like a Markov chain on $\mathcal{A} \times \mathcal{Q}$.

1.2.2 Mean and variance for the count

The derivation of the expectation and the variance of a word count under the permutation model based on $\mathcal{S}_2$ can be found in Cowan (1991) and Prum *et al.* (1995) (see Schbath (1995b) and Robin *et al.* (2005) for the letter permutation model).

In this section, we assume that the sequence $\mathbf{S} = X_1 X_2 \cdots X_n$ is a first-order stationary Markov chain (model M1) with non zero transition probabilities.

The number of occurrences $N(\mathbf{w})$ of a h-letter word $\mathbf{w} = w_1 w_2 \cdots w_h$ in the sequence $\mathbf{S} = X_1 X_2 \cdots X_n$ can be simply defined by

$$N(\mathbf{w}) = \sum_{i=1}^{n-h+1} Y_i(\mathbf{w}), \tag{1.2}$$

where $Y_i(\mathbf{w})$ equals 1 if and only if an occurrence of $\mathbf{w}$ starts at position i in the sequence and 0 otherwise. Therefore, to get the mean and variance of the count, we need to study the distribution of the random indicators $Y_i(\mathbf{w})$'s, namely their expectation, variance and covariances.

Random indicator of an occurrence The position of an occurrence of $\mathbf{w}$ is defined by the position of its first letter w_1. We define the random indicator $Y_i(\mathbf{w})$ of an occurrence of $\mathbf{w}$ at position i, $1 \leq i \leq n-h+1$, in $\mathbf{S}$ by:

$$Y_i(\mathbf{w}) = \begin{cases} 1 & \text{if } (X_i, X_{i+1}, \ldots, X_{i+h-1}) = (w_1, w_2, \ldots w_h), \\ 0 & \text{otherwise.} \end{cases}$$

It is a random Bernoulli variable with parameter $\mathbb{P}(Y_i(\mathbf{w}) = 1)$ given by

$$\begin{aligned} \mathbb{P}(Y_i(\mathbf{w}) = 1) &= \mathbb{P}(X_i = w_1, \ldots, X_{i+h-1} = w_h) \\ &= \mu(w_1) \times \pi(w_1, w_2) \times \cdots \times \pi(w_{h-1}, w_h). \end{aligned}$$

For convenience, $\mu(\mathbf{w})$ will denote the probability for the word $\mathbf{w}$ to appear at a given position in the sequence. The $Y_i(\mathbf{w})$'s are then Bernoulli variables with expectation $\mu(\mathbf{w})$ and variance $\mu(\mathbf{w})[1-\mu(\mathbf{w})]$, with

$$\mu(\mathbf{w}) = \mu(w_1) \times \prod_{j=2}^{h} \pi(w_{j-1}, w_j). \tag{1.3}$$

However, these random indicators $Y_i(\mathbf{w})$ are not independent, not only because the sequence is Markovian but most importantly because occurrences of a given word may overlap in a sequence. Consequently, their sum over the positions $i = \{1, \ldots, n-h+1\}$ (namely the number of occurrences – or count – of the word) is not distributed according to a binomial distribution.

Overlaps Occurrences of a given word may overlap in a sequence. For instance, $\mathbf{w}$ = `aataa` occurs 4 times in the sequence given in Figure 1.1, at positions $i = 2$, 11, 15 and 18. The third occurrence overlaps both the second and the fourth occurrences leading to a clump of 3 overlapping occurrences of `aataa` starting at position 11.

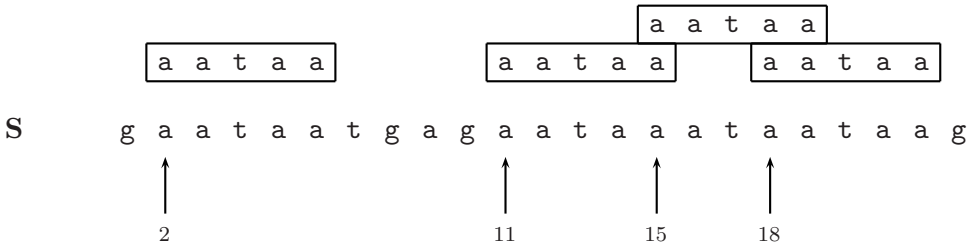


Figure 1.1: Four occurrences of `aataa` in sequence $\mathbf{S}$ leading to two clumps of `aataa`, the first one of size 1 and the second one of size 3.

The overlapping structure of a word can be described by two equivalent quantities: the overlapping indicators or the periods.

Overlapping indicators The overlapping indicator $\varepsilon_u(\mathbf{w})$, for $1 \leq u \leq h$, is equal to 1 if two occurrences of $\mathbf{w}$ can overlap on u letters, meaning the last u letters of $\mathbf{w}$ are identical to its first u letters, and 0 otherwise:

$$\varepsilon_u(\mathbf{w}) = \begin{cases} 1 & \text{if } (w_{h-u+1}, w_{h-u+2}, \ldots, w_h) = (w_1, w_2, \ldots, w_u), \\ 0 & \text{otherwise.} \end{cases}$$

By definition, $\varepsilon_h(\mathbf{w}) = 1$. A *non-overlapping* word $\mathbf{w}$ is such that $\varepsilon_u(\mathbf{w}) = 0$ for all $1 \leq u \leq h-1$.

Periods of a word An integer $p \in \{1, \ldots, h-1\}$ is said to be a period of $\mathbf{w}$ if and only if two occurrences of $\mathbf{w}$ can start at a distance p apart ($\varepsilon_{h-p}(\mathbf{w}) = 1$). It implies the following periodicity: $w_j = w_{j+p}$ for all $j \in \{1, \ldots, h-p\}$.
We denote by $\mathcal{P}(\mathbf{w})$ the set of periods of the word $\mathbf{w}$; For instance $\mathcal{P}(\texttt{aataataa}) = \{3, 6, 7\}$. Periods that are not a strict multiple of the smallest period are said to be *principal* since they will be more important, as we will see later. $\mathcal{P}'(\mathbf{w})$ denotes the set of the principal periods of $\mathbf{w}$; For instance $\mathcal{P}'(\texttt{aataataa}) = \{3, 7\}$.

In the remainder, we will use the periods rather than the overlapping indicators because this simplifies formulas. We will denote by $\mathbf{w}^p\mathbf{w}$ the word composed of two overlapping occurrences of $\mathbf{w}$ starting at a distance p apart:

$$\mathbf{w}^p\mathbf{w} = w_1 \cdots w_p w_1 \cdots w_h.$$

Dependence between occurrences The variables $Y_i(\mathbf{w})$ and $Y_{i+d}(\mathbf{w})$, $d > 0$, are not independent. Their covariance is defined by

$$\begin{aligned}\mathbb{C}[Y_i(\mathbf{w}), Y_{i+d}(\mathbf{w})] &= \mathbb{E}[Y_i(\mathbf{w}) \times Y_{i+d}(\mathbf{w})] - \mathbb{E}[Y_i(\mathbf{w})] \times \mathbb{E}[Y_{i+d}(\mathbf{w})] \\ &= \mathbb{P}(Y_i(\mathbf{w}) = 1,\ Y_{i+d}(\mathbf{w}) = 1) - [\mu(\mathbf{w})]^2. \qquad (1.4)\end{aligned}$$

To calculate the probability $\mathbb{P}(Y_i(\mathbf{w}) = 1,\ Y_{i+d}(\mathbf{w}) = 1)$, we distinguish two cases: $1 \leq d < h$ (two overlapping occurrences) and $d \geq h$ (two disjoint occurrences).

- The probability that $\mathbf{w}$ occurs both at positions i and $i + d$, $1 \leq d < h$, is different from 0 only if d is a period of $\mathbf{w}$. In this case, it is equal to $\mu(\mathbf{w}^d\mathbf{w})$.
- The probability that two disjoints occurrences of $\mathbf{w}$ are separated by $d-h$ letters ($d \geq h$) is given by $\mu(\mathbf{w})\pi^{d-h+1}(w_h, w_1)\mu(\mathbf{w})/\mu(w_1)$, where $\pi^\ell(\cdot,\cdot)$ denotes ℓ-step transition probabilities in $\mathbf{S}$.

The covariance between two random indicators of occurrence is thus:

$$\mathbb{C}[Y_i(\mathbf{w}), Y_{i+d}(\mathbf{w})] = \begin{cases} -[\mu(\mathbf{w})]^2 & \text{if } 0 < d < h,\ d \notin \mathcal{P}(\mathbf{w}), \\ \mu(\mathbf{w}^d\mathbf{w}) - [\mu(\mathbf{w})]^2 & \text{if } d \in \mathcal{P}(\mathbf{w}), \\ [\mu(\mathbf{w})]^2 \left[\dfrac{\pi^{d-h+1}(w_h, w_1)}{\mu(w_1)} - 1\right] & \text{if } d \geq h. \end{cases} \qquad (1.5)$$

Mean and variance of the count Finally, we get the following expression for the expectation and the variance of $N(\mathbf{w})$:

$$\mathbb{E}[N(\mathbf{w})] = \sum_{i=1}^{n-h+1} \mathbb{E}[Y_i(\mathbf{w})] = (n-h+1)\mu(\mathbf{w}) \qquad (1.6)$$

$$\begin{aligned}\mathbb{V}[N(\mathbf{w})] &= \sum_{i=1}^{n-h+1} \mathbb{V}[Y_i(\mathbf{w})] + 2\sum_{i=1}^{n-h+1}\sum_{j=i+1}^{n-h+1} \mathbb{C}[Y_i(\mathbf{w}), Y_j(\mathbf{w})]. \qquad (1.7) \\ &= (n-h+1)\mu(\mathbf{w})\big(1-\mu(\mathbf{w})\big) + 2\sum_{i=1}^{n-h+1}\sum_{d=1}^{n-h-i+1} \mathbb{C}[Y_i(\mathbf{w}), Y_{i+d}(\mathbf{w})]\end{aligned}$$

where $\mu(\mathbf{w})$ is given by Eq. (1.3), p. 7 and the covariance term is given by Eq. (1.5).

1.2.3 Word count distribution

We will now focus on the statistical distribution of the count $N(\mathbf{w})$. Several methods have been proposed to derive the exact distribution of $N(\mathbf{w})$ in a

sequence of independent letters (model M0) or in model M1. Most of them use pattern matching principles or language theory (see for instance chapter 7 from Lothaire (2005)). The most probabilistic approach is probably the one consisting in using the following duality principle: $\mathbb{P}(N(\mathbf{w}) \geq j) = \mathbb{P}(T_j \leq n)$ where T_j denotes the position of the j-th occurrence of the word $\mathbf{w}$ along a random sequence $\mathbf{S}$ of length n. The distribution of T_j can be obtained via the distribution of the distance between two successive occurrences of $\mathbf{w}$ (see Robin and Daudin (1999)). However, all these methods are fastidious to implement, with many technical limitations as soon as the sequence is long, or as the order of the Markov model is greater than 1, or the motif is complex. In practice, approximate distributions are used. In this section, we will present two approximations of the word count distribution that have been theoretically proved under some asymptotic framework: the Gaussian approximation which is valid if the expected count is far enough from zero (Section 1.2.3) and a compound Poisson approximation which is adapted for the count of rare and clumping events (Section 1.2.3). The quality of these approximations have been studied in Robin and Schbath (2001) and Nuel (2006). No theoretical result exists so far on the binomial approximation that would result from neglecting the dependence between the occurrences.

Gaussian approximation

Recall that $N(\mathbf{w})$ is a sum of $(n-h+1)$ random Bernoulli variables $Y_i(\mathbf{w})$ with mean $\mu(\mathbf{w})$ and variance $\mu(\mathbf{w})[1-\mu(\mathbf{w})]$.

Asymptotic normality If the Bernoulli variables $Y_i(\mathbf{w})$'s were independent, then the classical Central Limit Theorem would ensure that the count converges in distribution to a Gaussian variable. But the $Y_i(\mathbf{w})$'s are not independent for two reasons: the occurrences of $\mathbf{w}$ can overlap and the letters of the sequence are not independent. Nonetheless, by using a Central Limit Theorem for Markov chains, the asymptotic normality of the count can be established:

$$\frac{N(\mathbf{w}) - \mathbb{E}[N(\mathbf{w})]}{\sqrt{\mathbb{V}[N(\mathbf{w})]}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ as } n \to +\infty. \tag{1.8}$$

Estimating the parameters In the previous convergence, both expectation and variance of the count depend on the model parameters which are not known in practice. Let us estimate the expected count by its plug-in estimator, i.e. by replacing the transition probabilities $\pi(a,b)$ by their MLE $\widehat{\pi}(a,b) = N(ab)/N(a+)$ and the probability $\mu(w_1)$ by $\widehat{\mu}(w_1) = N(w_1)/n$ in equation (1.6). We then consider the following estimator:

$$\widehat{\mathbb{E}}[N(\mathbf{w})] = \frac{N(w_1w_2) \times \cdots \times N(w_{h-1}w_h)}{N(w_2) \times \cdots \times N(w_{h-1})}. \tag{1.9}$$

Because the estimator $\widehat{\mathbb{E}}_1[N(\mathbf{w})]$ is expressed like a function of several asymptotically Gaussian counts, the δ-method ensures that there exists a constant $v^2(\mathbf{w})$ such that

$$\frac{N(\mathbf{w}) - \widehat{\mathbb{E}}[N(\mathbf{w})]}{\sqrt{(n-h+1)v^2(\mathbf{w})}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ as } n \to +\infty. \tag{1.10}$$

However, since $\widehat{\mathbb{E}}[N(\mathbf{w})]$ is random, the variance of $\{N(\mathbf{w}) - \widehat{\mathbb{E}}[N(\mathbf{w})]\}$ is different from $\mathbb{V}[N(\mathbf{w})]$ and $(n-h+1)v^2(\mathbf{w})$ is therefore not related to $\mathbb{V}[N(\mathbf{w})]$.

Asymptotic variance Several approaches have been used to derive the asymptotic variance $(n-h+1)v^2(\mathbf{w})$. The first one is the δ-method in Lundstrom (1990): it uses the fact that $n^{-1/2}\{N(\mathbf{w}) - \widehat{\mathbb{E}}[N(\mathbf{w})]\}$ is a function of the asymptotically Gaussian vector $\big(N(\mathbf{w}), N(w_1w_2), \ldots, N(w_{h-1}w_h), N(w_2), \ldots, N(w_{h-1})\big)$ from (1.8). However, the function and the size of this vector depends both on the length and on the 2-letter composition of $\mathbf{w}$, so it does not give a unified formula for the asymptotic variance.
Prum *et al.* (1995) proposed a second method: they showed that the estimator $\widehat{\mathbb{E}}[N(\mathbf{w})]$ is asymptotically equivalent to $\mathbb{E}[N(\mathbf{w}) \mid \mathcal{S}_2]$, the expected count of $N(\mathbf{w})$ under the 2-letter word permutation model, and that $v^2(\mathbf{w})$ is the limit of $n^{-1}\mathbb{V}[N(\mathbf{w}) \mid \mathcal{S}_2]$. They obtained:

$$\begin{aligned} v^2(\mathbf{w}) = \mu(\mathbf{w}) + 2 \sum_{p \in \mathcal{P}(\mathbf{w}),\, p<h-1} \mu(\mathbf{w}^p\mathbf{w}) \\ + [\mu(\mathbf{w})]^2 \left[\sum_a \frac{[N_{\mathbf{w}}(a+)]^2}{\mu(a)} - \sum_{a,b} \frac{[N_{\mathbf{w}}(ab)]^2}{\mu(ab)} + \frac{1 - 2N_{\mathbf{w}}(w_1+)}{\mu(w_1)} \right], \end{aligned} \tag{1.11}$$

where $N_{\mathbf{w}}(\cdot)$ stands for the count inside the word $\mathbf{w}$. The overlaps of $\mathbf{w}$ on two or more letters explicitly appear in this formula ($p < h-1$); The overlap on a unique letter is taken into account in the $[\mu(\mathbf{w})]^2$ term.

Since model M1 allows more variability than the corresponding permutation model, one expects the variance $(n-h+1)v^2(\mathbf{w})$ to be smaller than the variance $\mathbb{V}[N(\mathbf{w})]$. This is not difficult to show it in the Bernoulli model ($m = 0$); For higher models, it has been numerically verified.

Generalizations to $m > 1$ and to phased models can be found in Schbath *et al.* (1995) and Schbath (1995b). When $m = h-2$, i.e. in the Markov chain model fitting the counts of all the $(h-1)$-letter words (we call this model the maximal model regarding the analysis of h-letter words), a third approach can be used to derived the asymptotic variance. This approach is based on martingale theory and provides a simpler expression for the asymptotic variance (see Prum *et al.* (1995) or Reinert *et al.* (2000)).

Compound Poisson approximation

Poisson approximations can also be used for the count of rare events, i.e. when $\mathbb{E}[N(\mathbf{w})] = \mathrm{O}(1)$. Note that this condition implies that $\log n = \mathrm{O}(h)$ (long enough words). In this paragraph, we will assume the rare event condition but also that $h = \mathrm{o}(n)$.

A nice method to establish Poisson approximations of counts is the Chen-Stein method (see Arratia *et al.* (1990) for an introduction and Barbour *et al.* (1992b) for a more general presentation). This method gives a bound on the total variation distance between the distribution of a sum of dependent Bernoulli variables and the Poisson distribution with same expectation. Lower the dependence, better the Poisson approximation quality. Unfortunately, the local dependence between occurrences of an overlapping word $\mathbf{w}$ is too important and a Poisson approximation of the distribution of $N(\mathbf{w})$ generally does not hold. One can clearly show that the bound provided by the Chen-Stein method does not converge to zero (it is of order $\mu(\mathbf{w}^{p_0}\mathbf{w})$ with p_0 the minimal period of $\mathbf{w}$, see Schbath (1995a)). But one can also show that a geometric distribution (discrete version of the exponential distribution) does not fit the distribution of the distance between two successive occurrences of an overlapping word [Robin and Daudin (1999)].

The solution is to take advantage of the clump structure (clumps do not overlap) and to use the following relations between the number of occurrences $N(\mathbf{w})$ and the clumps (size and count). Indeed we have

$$N(\mathbf{w}) = \sum_{i=1}^{\widetilde{N}(\mathbf{w})} K_i(\mathbf{w}) \tag{1.12}$$

where $\widetilde{N}(\mathbf{w})$ is the number of clumps of $\mathbf{w}$ and $K_i(\mathbf{w})$ is the size of the i-th clump, but we also have

$$N(\mathbf{w}) = \sum_{k>0} k\widetilde{N}_k(\mathbf{w}) \tag{1.13}$$

where $\widetilde{N}_k(\mathbf{w})$ is the number of clumps of $\mathbf{w}$ of size k in $\mathbf{S}$. Since a compound Poisson variable is defined like $\sum_{k>0} k\, Z_k$ with Z_k's independent Poisson variables, or like $\sum_{i=1}^{Z} C_i$ with Z a Poisson variable and C_i's i.i.d. variables, the Poisson approximation of the number of clumps (of any size or of size k) is the core of the compound Poisson approximation of the word count. In the remainder of this section we will then explicitly define the clumps and give some of their probabilistic properties.

Random indicator of a clump occurrence A clump of a word $\mathbf{w}$ in a sequence $\mathbf{S}$ is a maximal succession of overlapping occurrences of $\mathbf{w}$. The size of a clump is the number of occurrences of $\mathbf{w}$ the clump is composed of. For

instance, in Figure 1.1, there are two clumps of `aataa`: one of size 1 starting at position 2, the other one of size 3 starting at position 11. The position of a clump of $\mathbf{w}$ in the sequence is defined by the position (start) of the first occurrence of $\mathbf{w}$ in the clump. Let us define $\widetilde{Y}_i(\mathbf{w})$ the random indicator that an occurrence of a clump of $\mathbf{w}$ starts at position i in $\mathbf{S}$. A clump of $\mathbf{w}$ occurs at position i if and only if an occurrence of $\mathbf{w}$ occurs at position i without overlapping a previous occurrences of $\mathbf{w}$. Therefore, if we neglect end effects (i.e. when $i < h$), we can write

$$\widetilde{Y}_i(\mathbf{w}) = Y_i(w)[1 - Y_{i-1}(w)] \times \cdots \times [1 - Y_{i-h+1}(w)]. \tag{1.14}$$

(End effects are corrected by considering an infinite sequence). Now an occurrence of $\mathbf{w}$ which overlaps a previous occurrence of $\mathbf{w}$ is necessarily preceded by a prefix $w_1 \cdots w_p$ of $\mathbf{w}$, where p is a period of $\mathbf{w}$. If we restrict ourselves to principal periods, this is a necessary and sufficient condition [Schbath (1995a)]. For instance, an occurrence of `aataataa` overlaps a previous occurrence of `aataataa` if and only if it is preceded either by `aat` (prefix of size 3) or by `aataata` (prefix of size 7). If it was preceded by `aataat` (prefix of size 6), it would also be preceded by `aat`.

Therefore, we have

$$\widetilde{Y}_i(\mathbf{w}) = \sum_{p \in \mathcal{P}'(\mathbf{w})} [1 - Y_{i-p}(w_1 \cdots w_p)] \times Y_i(w).$$

Clump probability Let denote by $\widetilde{\mu}(\mathbf{w})$ the probability that a clump of $\mathbf{w}$ occurs at a given position, i.e. $\widetilde{\mu}(\mathbf{w}) = \mathbb{E}[\widetilde{Y}_i(\mathbf{w})]$. The previous equation gives

$$\widetilde{\mu}(\mathbf{w}) = [1 - a(\mathbf{w})] \times \mu(\mathbf{w}), \tag{1.15}$$

where $a(\mathbf{w})$ is the probability that an occurrence of $\mathbf{w}$ overlaps a previous occurrence of $\mathbf{w}$ and is given by

$$a(\mathbf{w}) = \sum_{p \in \mathcal{P}'(\mathbf{w})} \prod_{j=1}^{p} \pi(w_j, w_{j+1}). \tag{1.16}$$

Symmetrically, the probability that an occurrence of $\mathbf{w}$ overlaps a next occurrence of $\mathbf{w}$ is also equal to $a(\mathbf{w})$; Therefore, $a(\mathbf{w})$ will be simply called the probability of self-overlap of $\mathbf{w}$. Note that $a(\mathbf{w}) = 0$ if and only if $\mathbf{w}$ is a non-overlapping word (we assumed that all transition probabilities were non zero). In that case we also have $\widetilde{Y}_i(\mathbf{w}) = Y_i(\mathbf{w})$ and $\widetilde{\mu}(\mathbf{w}) = \mu(\mathbf{w})$.

Poisson approximation for the number of clumps Let define the number of clumps of $\mathbf{w}$ by $\widetilde{N}(\mathbf{w}) := \sum_{i=1}^{n-h+1} \widetilde{Y}_i(\mathbf{w})$. The mean number of clumps is

then equal to $(n-h+1)\widetilde{\mu}(\mathbf{w}) = [1-a(\mathbf{w})]\mathbb{E}[N(\mathbf{w})]$ from (1.15). The Poisson approximation of $\widetilde{N}(\mathbf{w})$ follows from a direct application of the Chen-Stein method to the Bernoulli variables $\widetilde{Y}_i(\mathbf{w})$ [Schbath (1995a)]. The error bound is indeed of order $(\rho^h + h\mu(\mathbf{w}))$ where $0 < \rho < 1$ is the second largest eigenvalue (in modulus) of the transition matrix Π. Recall that $n\mu(\mathbf{w}) = \mathrm{O}(1)$ from the rare event condition and that $h = \mathrm{o}(n)$.

The exact distribution of the number of clumps of $\mathbf{w}$ in model M1 has been recently derived through its generating function [Stefanov *et al.* (2007)] and compared to the Poisson distribution; The conclusion was that the Poisson approximation is as better than the expected count of the word is small.

Size of a clump A clump is of size k if and only if the first occurrence of $\mathbf{w}$ in the clump overlaps from the right a second occurrence (probability $a(\mathbf{w})$), the second occurrence of $\mathbf{w}$ in the clump overlaps a third occurrence (probability $a(\mathbf{w})$), ..., the $(k-1)$-th occurrence overlaps a k-th occurrence of $\mathbf{w}$ (probability $a(\mathbf{w})$), and this k-th occurrence of $\mathbf{w}$ does not overlap a next occurrence (probability $1-a(\mathbf{w})$). Thus, if we denote by $K_i(\mathbf{w})$ the size of the i-th clump of $\mathbf{w}$ in the sequence, the random variable $K_i(\mathbf{w})$ is geometrically distributed:

$$\mathbb{P}(K_i(\mathbf{w}) = k) = [1-a(\mathbf{w})] \times [a(\mathbf{w})]^{(k-1)}. \tag{1.17}$$

Compound Poisson approximation for rare word counts As previously said, the Poisson approximations of the number of clumps of any size and more particularly of size k for $k \geq 1$ are the key ingredients for the compound Poisson approximation of $N(\mathbf{w})$. Indeed, let denote by $\mathcal{CP}(\lambda_k, k \geq >)$ the compound Poisson distribution of $\sum_{k>0} kZ_k$ with $Z_k \sim \mathcal{P}(\lambda_k)$. Since $N(\mathbf{w}) = \sum_{k>0} k\widetilde{N}_k(\mathbf{w})$, the total variation distance properties give

$$d_{\mathrm{TV}}(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}(\mathbb{E}[\widetilde{N}_k(\mathbf{w})], k \geq 1)) \leq d_{\mathrm{TV}}(\mathcal{L}(\widetilde{N}_k(\mathbf{w}), k \geq 1), \otimes\mathcal{P}(\mathbb{E}[\widetilde{N}_k(\mathbf{w})])).$$

The joint Poisson approximation of $(\widetilde{N}_k(\mathbf{w}), k \geq 1)$ is a little more involved to get than the one for $\widetilde{N}(\mathbf{w})$ [Schbath (1995a)] but the error bound is of the same order and

$$\mathbb{E}[\widetilde{N}_k(\mathbf{w})] = [1-a(\mathbf{w})]^2[a(\mathbf{w})]^{(k-1)}\mathbb{E}[N(\mathbf{w})].$$

The above formula means that the limiting compound Poisson distribution $\mathcal{CP}(\mathbb{E}[\widetilde{N}_k(\mathbf{w})], k \geq 1)$ is in fact a Pólya-Aeppli distribution (also called Geometric-Poisson distribution) with parameter $(\mathbb{E}[\widetilde{N}(\mathbf{w}), a(\mathbf{w}))$ [Johnson *et al.* (1992)].

Direct compound Poisson approximation methods exist and can be alternatively applied to the word count [Erhardsson (1999), Erhardsson (2000)]. Their advantage is to provide better error bounds but it gives the same limiting compound Poisson distribution as above (see Lothaire (2005), chapter 6).

Generalization to Mm and phased models Like for the Gaussian approximation, the generalization to the phased Markov model of order 1 is done by rewriting the sequence with the new alphabet $\mathcal{A} \times \{1, 2, 3\}$ (see page 5). However, note that the occurrence of a single word $\mathbf{w}$ in sequence $\mathbf{S}$ corresponds to the occurrence of a word family composed of three phased words in the new sequence. Therefore, one has to use the compound Poisson approximation for the count of a set of words in M1 presented in Section 1.4.1.

When one changes the alphabet (see page 5) to generalize the compound Poisson approximation in model M1 to model Mm, $m > 1$, one has to be very careful with the word overlaps. Indeed, there is no one-to-one transformation between clumps of $\mathbf{w}$ in $\mathbf{S}$ and clumps of $\mathbf{w}^\star$ (word $\mathbf{w}$ written on $\mathcal{A}^m$) in the new sequence $\mathbf{S}^\star$. Let us take an example with $m = 2$. Put $\mathbf{w} = \texttt{aataa}$ and let $\mathbf{S}$ be the following sequence on the $\mathcal{A}$ alphabet:

$$\mathbf{S} = \underline{\texttt{gaataa}}\texttt{tgag}\underline{\texttt{aataaataataag}}.$$

$\mathbf{S}$ contains 4 occurrences of $\mathbf{w}$ and two clumps of $\mathbf{w}$ (one of size 1, the other one of size 3). Now, we write the word and the sequence in the new alphabet $\mathcal{A}^2$. For this, we put $\texttt{ga} = \gamma$, $\texttt{aa} = \alpha$, $\texttt{at} = \beta$, $\texttt{ta} = \tau$, $\texttt{tg} = \delta$, $\texttt{ag} = \kappa$. We have

$$\mathbf{w}^\star = \alpha\beta\tau\alpha \quad \text{and} \quad \mathbf{S}^\star = \gamma\underline{\alpha\beta\tau\alpha}\beta\delta\gamma\kappa\gamma\underline{\alpha\beta\tau\alpha}\,\underline{\alpha\beta\tau\alpha\beta\tau\alpha}\kappa.$$

We can see that the word $\mathbf{w}^\star$ still appear four times in the sequence $\mathbf{S}^\star$ ($N(\mathbf{w})$ is equal to the count of $\mathbf{w}^\star$ in $\mathbf{S}^\star$) but there are now three clumps of $\mathbf{w}^\star$ in $\mathbf{S}^\star$ (two of size 1 and one of size 2). This is due to the fact that $\mathbf{w}^\star$ has just a unique period ($\mathcal{P}(\alpha\beta\tau\alpha) = \{3\}$) whereas $\mathbf{w}$ has two periods ($\mathcal{P}(\texttt{aataa}) = \{3, 4\}$). Therefore, when the results for the word $\mathbf{w}^\star$ in M1 will be "translated" into the alphabet $\mathcal{A}$, some overlaps will not appear explicitly in the formulas. In Mm, only the overlaps on m letters or more will be taken into account since the principal periods of $\mathbf{w}^\star$ are the periods of $\mathbf{w}$ that are less or equal to $(h - m)$. The word $\mathbf{w}^\star$ is non-overlapping as soon as $\mathbf{w}$ is not enough self-overlapping.

1.2.4 p-values and scores of exceptionality

The significance of the over-representation of a word $\mathbf{w}$ in a given DNA sequence is measured by the p-value $p(\mathbf{w})$:

$$p(\mathbf{w}) = \mathbb{P}\{N(\mathbf{w}) \geq N_{\text{obs}}(\mathbf{w})\}$$

where $N_{\text{obs}}(\mathbf{w})$ is the observed count of $\mathbf{w}$ in the DNA sequence. If $p(\mathbf{w})$ is close to 0 then the word is exceptionally frequent: there is no chance to observe it so many times in random sequences. On the other hand, the significance of an under-representation is measured by the p-value $p'(\mathbf{w}) = \mathbb{P}\{N(\mathbf{w}) \leq N_{\text{obs}}(\mathbf{w})\}$. If $p'(\mathbf{w})$ is close to 0 then $\mathbf{w}$ is exceptionally rare under the model: there

is no chance that $\mathbf{w}$ occurs so rarely in random sequences. Since the exact distribution of the count $N(\mathbf{w})$ is rarely available in practice, approximate p-values are calculated to detect exceptional words and usually converted into scores of exceptionality.

Approximate p-values A natural way of approximating p-values is to use an approximate distribution of $N(\mathbf{w})$, for instance a Gaussian distribution for highly expected words or a compound Poisson distribution for rarely expected words, as we have seen in Section 1.2.3. Calculating approximate p-values requires just to compute tail of Gaussian or compound Poisson distributions. An efficient algorithm to compute tails of Geometric-Poisson distributions has been proposed by Nuel (2008).
For exceptional words, i.e. words whose count strongly deviates from what is expected, large deviation theory is probably the most accurate way to approximate p-values. This approach has been studied in Nuel (2004). Since it requires sophisticated numerical analysis and longer computation time, this method should be restricted to the most exceptional words (filtered from Gaussian or compound Poisson approximations for instance).

Score of exceptionality In practice, it is often more convenient to manipulate scores from $\mathbb{R}$ than probabilities of the form $p(\mathbf{w}) = \mathbb{P}\{N(\mathbf{w}) \geq N_{\text{obs}}(\mathbf{w})\}$, especially when the ones we are interested in are very close to 0 or very close to 1. For symmetrical reasons we prefer to use the probit transformation rather that the $-\log$ transformation. Therefore, to each probability $p(\mathbf{w})$ we associate the score $u(\mathbf{w})$ such that:

$$\mathbb{P}\{\mathcal{N}(0,1) \geq u(\mathbf{w})\} = p(\mathbf{w}).$$

Therefore, words with a high positive score are exceptionally frequent whereas words with a negative but high absolute value score are exceptionally rare in the observed sequence.

The Gaussian approximation of $N(\mathbf{w})$ has a great practical advantage: it allows to directly calculate the score of exceptionality $u(\mathbf{w})$ without calculating the associated p-value. Indeed, if we put

$$u(\mathbf{w}) = \frac{N(\mathbf{w}) - \widehat{\mathbb{E}}[N(\mathbf{w})]}{\sqrt{\widehat{\sigma}^2(\mathbf{w})}} \tag{1.18}$$

where $\widehat{\mathbb{E}}[N(\mathbf{w})]$ is the estimator of the expected count given by Eq. (1.9), p. 10 and $\widehat{\sigma}^2(\mathbf{w})$ is a plug-in estimator of $(n-h+1)v^2(\mathbf{w})$ (cf. Eq. (1.11), p. 11),

namely

$$\widehat{\sigma}^2(\mathbf{w}) = \widehat{\mathbb{E}}[N(\mathbf{w})] + 2 \sum_{p\in\mathcal{P}(\mathbf{w}),p<h-1} \widehat{\mathbb{E}}[N(\mathbf{w}^p\mathbf{w})] \tag{1.19}$$
$$+\{\widehat{\mathbb{E}}[N(\mathbf{w})]\}^2 \left[\sum_a \frac{[N_{\mathbf{w}}(a+)]^2}{N(a)} - \sum_{a,b} \frac{[N_{\mathbf{w}}(ab)]^2}{N(ab)} + \frac{1-2N_{\mathbf{w}}(w_1+)}{N(w_1)}\right],$$

then we have

$$\mathbb{P}\{N(\mathbf{w}) \geq N_{\text{obs}}(\mathbf{w})\} \simeq \mathbb{P}\{\mathcal{N}(0,1) \geq u(\mathbf{w})\}.$$

1.2.5 Example of DNA motif discovery

Chi motifs in bacterial genomes Chi motifs have been identified in several bacterial genomes and they are not conserved through species. Their identification in a new species is still a challenge. They are involved in the repair of double-strand DNA breaks by homologous recombination. More precisely, they interact specifically with an enzyme that processes along the DNA and degrades it (exonuclease activity): when the enzyme encounters a Chi site, its exonuclease activity is strongly reduced and altered but it still continues to separate the two DNA strands forming then a substrate for homologous pairing and repair of the deleted DNA parts. Since Chi motifs protect the bacterial genome from degradation and stimulate its repair, it seems important that these motifs were as much frequent as possible along the bacterial genome. Biologists expect them to be significantly over-represented.

Moreover, Chi activity is strongly orientation-dependent; The Chi motif is only recognized when the enzyme enters a double-strand DNA molecule from the right side of the motif. In many bacterias for which the Chi motif has been identified, Chi orientation is correlated with the direction of DNA replication, meaning that it occurs preferentially on the leading strand [El Karoui *et al.* (1999), Halpern *et al.* (2007)]. The over-representation of Chi should then be important on the leading strands. Biologists classically measure the asymmetry strand of a motif by calculating its skew. The skew of a motif $\mathbf{w}$ is simply the ratio $N(\mathbf{w})/N(\overline{\mathbf{w}})$ where $\overline{\mathbf{w}}$ is the reverse complementary of the word $\mathbf{w}$; In other words $N(\overline{\mathbf{w}})$ is simply the count of $\mathbf{w}$ in the complementary strand. Therefore, biologists expect Chi to be relatively skewed, i.e. with a skew far from one.

***E. coli* as a learning case** The Chi motif of *E. coli* has been known for long time: it is the 8-letter word `gctggtgg`. If we study the statistical properties of Chi frequency along *E. coli* genome, we can note some significant characteristics. First of all, its 762 occurrences in the complete genome (concatenation of both leading stands, $n = 4.6\,10^6$) are significantly high whatever the model we

choose; In other words, its high frequency cannot be explained by the genome composition. As we can see on Table 1.2, Chi has very high over-representation scores and is always among the 5 most exceptionally frequent 8-letter words. Second, if we restrict the analysis to the *E. coli* backbone[1] ($n = 3.7\,10^6$), Chi becomes the most exceptional 8-letter words in 5 models, especially in the maximal model M6 (see Table 1.2). Analyzing only the backbone seems therefore to reduce the noise produced by the regions which are either highly variable or specific to one or few strains (mobile elements); Indeed, there is a priori no biological reason for Chi to occur in such regions.

	complete genome 762 occurrences				backbone 675 occurrences			
m	$\widehat{\mathbb{E}}_m[N]$	$\widehat{\sigma}^2_m$	u_m	rank	$\widehat{\mathbb{E}}_m[N]$	$\widehat{\sigma}^2_m$	u_m	rank
0	85.9	85.8	72.96	3	73.10	73.02	70.44	3
1	84.9	84.8	73.54	1	71.47	71.32	71.46	1
2	206.8	203.9	38.88	1	186.68	183.82	36.02	1
3	355.5	338.9	22.08	5	315.26	299.68	20.78	1
4	355.3	314.4	22.94	2	309.79	272.90	22.11	2
5	420.9	298.0	19.76	1	376.68	262.42	18.42	1
6	610.1	203.3	10.65	3	539.09	176.02	10.24	1

Table 1.2: Statistics of `gctggtgg` in the complete genome (left) and in the backbone genome (right) of *E. coli* K12 under various models Mm. The rank is obtained while sorting the 65,536 scores by decreasing order.

The choice of the model does not seem to affect the significance of the Chi frequency (it is always exceptional), but this is not a general picture. Note that, when the order of the Markov model increases, the model better fits the sequence composition and less exceptional words are found. This is illustrated by the boxplots of Figure 1.2. Moreover, in a high order model we have a more accurate knowledge about the sequence composition than in a low order model: the significance of a word frequency has then no reason to be the same. This point is illustrated by the plot of Figure 1.2 which compares scores in models M1 and M6. We recognize the Chi motif which is clearly outside the cloud but let us take the case of the word `ggcgctgg`. It occurs 761 times in the *E.coli* backbone, it has a significantly high score of 62.4 in model M1 (it is the second most exceptional word) but has a score of 0.8 in model M6 (rank 17100). It simply means that its high frequency can be explained by the composition of 7-letter words; Indeed it is expected about 749 times in M6.

[1]The backbone of a bacterial genome is composed of the genomic regions conserved in several strains of the bacteria. Here, we used the backbone obtained from the alignment of the three strains K12, O157:H7 and CFT and available at `http://genome.jouy.inra.fr/mosaic/`

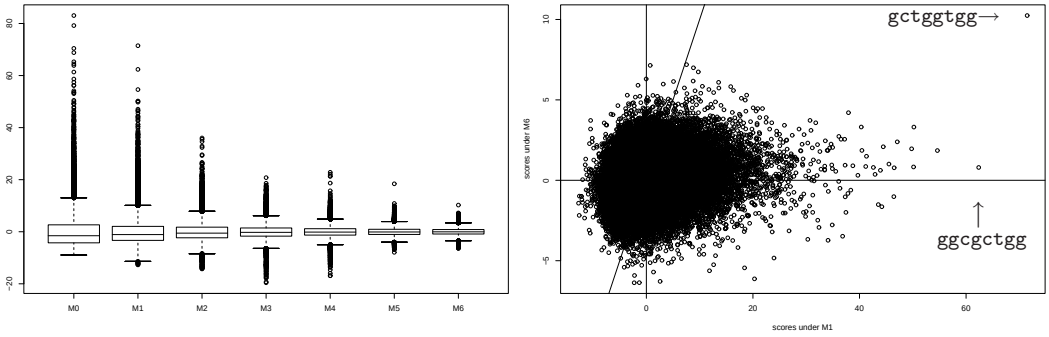


Figure 1.2: Exceptionality scores for the 65,536 8-letter words in the *E.coli* backbone. Left: Boxplots of the scores under models M0 to M6. Right: Scores under models M1 (x-axis) and M6 (y-axis).

The third characteristics of Chi in the *E.coli* backbone is that it is significantly skewed. Its skew is equal to 3.20 and the method described in Section 1.4.1 to assess skew significance gives a score of 6.53 in M6 (p-value of $3.3\,10^{-11}$).

Identification of Chi motif in *S. aureus* We will describe here the strategy used in Halpern *et al.* (2007) to identify the Chi motif in the bacteria *S. aureus*. The first step has been to extract the backbone of the *S. aureus* genome by comparing the genome of six strains of the bacteria. The obtained backbone contains about $2.44\,10^6$ letters.

The second step was to search for motifs which are frequent enough, exceptionally frequent and relatively skewed. They start by analyzing 8-letter words (like for *E. coli*) but none of the most over-represented and skewed motifs were frequent enough to be retained as potential Chi candidates. They thus focused on 7-letter words. Scores of exceptionality were calculated with the Gaussian approximation and in the maximal model, namely model M5. 6 motifs have an exceptionality score greater than 11 (see Table 1.3 or Figure 1.3 for a global view). Two of them have a negative skew score so they were not retained. A biological experiment has then been done to test for *S. aureus* Chi activity of the four candidates: `gaaaatg`, `ggattag`, `gaagcgg` and `gaattag`. The conclusion was that `gaagcgg` is necessary and sufficient to confer Chi activity in *S. aureaus*. This strategy has also been successfully used to predict and validate the Chi motif of three species of the *Streptococcus* genus [Halpern *et al.* (2007)].

w	$N_{\text{obs}}(\mathbf{w})$	$\widehat{\mathbb{E}}_5[N(\mathbf{w})]$	$\widehat{\sigma}^2_5(\mathbf{w})$	$u_5(\mathbf{w})$	Skew	Score
taaaaaa	1542	1214.3	603.4	13.34	1.61	-1.28
gaaaatg	1067	789.9	454.2	13.00	2.48	1.13
taaaatt	1356	1062.6	552.8	12.48	1.04	-1.53
ggattag	266	143.2	97.5	12.43	2.53	1.52
gaagcgg	272	162.4	88.1	11.67	7.56	2.91
gaattag	614	420.7	274.4	11.67	3.89	7.23
gaaaaag	1177	942.1	518.0	10.32	3.52	2.53
taagatt	316	201.3	130.9	10.03	1.07	-2.98
ttaaaag	1059	856.5	431.6	9.75	2.00	3.85
gatttag	657	488.1	305.9	9.66	2.16	4.25

Table 1.3: The 10 most exceptionally frequent 7-letter words under model M5 in the *S. aureaus* complete genome. Columns correspond respectively to the word, its observed count, its estimated expected count, its normalizing factor, its score of over-representation under model M5, its observed skew and its skew score under model M0.

1.3 Words With Exceptional Distribution

The way the occurrences of a given motif **w** are spread along a sequence or among different sequences or sub-sequences may provide functional informations. When the motif (and its functional properties) is known, this gives hints about the function of the regions where it occurs (or where it is avoided). Conversely, new interesting motifs may be discovered by comparing their relative frequencies in different well defined sequences or sub-sequences (e.g. regions of a genome).

1.3.1 Compound Poisson process

For both problems, we need a probabilistic model describing the motif occurrences process to assess the significance of the observed results. In this section, we will focus on the (compound) Poisson process which is simple and provides a surprisingly good approximation of the distribution of the word count [Robin and Schbath (2001)].

In this model, the sequence is viewed as a continuous line. To account for possible overlaps between occurrences, the word is assumed to occur in clumps along the sequence. We assume that the counting process of the clumps $\{C(x)\}_{x\geq 0}$ is an homogeneous Poisson process with intensity λ (in the entire Section 1.3, we will avoided to index the quantities by $(\mathbf{w})$ because there will be no ambiguity). Each clump contains a random number of occurrences, referred to as the clump size. The clump sizes $\{K_1, K_2, \dots\}$ are supposed to be i.i.d. with distribution

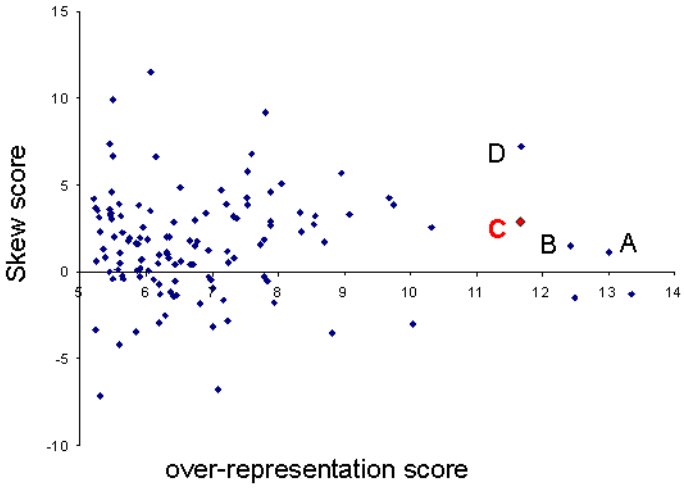


Figure 1.3: Over-representation scores under M5 and skew scores under M0 for the most over-represented 7-letters words (over-representation scores greater than 5) in the complete genome of *S. aureus*. The four best candidates (motifs A to D) are indicated. Motif C (`gaagcgg`) is the functional Chi site of *S. aureus*.

$p(k)$. The counting process $\{N(x)\}_{x\geq 0}$ is hence the compound Poisson process defined as

$$N(x) = \sum_{c=1\ldots C(x)} K_c.$$

In the case of a single fixed word, the clump size has a geometric distribution: $p(k) = (1-a)a^{k-1}$, where a stands for the overlapping probability of the word (see page 14). In the case of more complex motif, $p(k)$ may have a more complicated form [Robin (2002)]. The estimates of parameters λ and a depends on the biological question: empirical estimates will fit the observed word frequency (and clumping), while estimates based on a Markov chain model will account for the sequence composition.

1.3.2 Words significantly unbalanced between two sequences

We first consider the detection of motifs having different frequencies between two sequences $\mathbf{S}_1$ and $\mathbf{S}_2$. Two avoid artifacts and spurious detections, the testing procedure must account for the different lengths and composition of the sequences, and for the fact that the word may have an unexpected frequency in one or both of them.
We only consider here the non-overlapping case (i.e. $a = 0$). In sequence $\mathbf{S}_i$

$(i = 1, 2)$, the count N_i of $\mathbf{w}$ is supposed to have a Poisson distribution

$$N_i \sim \mathcal{P}(\lambda_i), \qquad \lambda_i = k_i \ell_i \mu_i$$

where ℓ_i is the length of $\mathbf{S}_i$, $\mu_i = \mu_i(\mathbf{w})$ is the occurrence probability of $\mathbf{w}$ under a Markov model fitted to the composition of $\mathbf{S}_i$ (see Section 1.2.2) and k_i is the exceptionality coefficient of $\mathbf{w}$ in $\mathbf{S}_i$. This framework is described in Robin *et al.* (2007).

Our purpose is to test if the counts of $\mathbf{w}$ in both sequences deviate from their expected values in the same way; We hence want to test the hypothesis $\mathbf{H}_0 : \{k_1 = k_2\}$ versus $\{k_1 \neq k_2\}$. A test procedure can be derived from the following property: for two independent Poisson variables N_1 and N_2 with respective means λ_1 and λ_2, the conditional distribution of N_1 given the sum $N_1 + N_2$ is binomial $\mathcal{B}(N_1 + N_2, \lambda_1/(\lambda_1 + \lambda_2))$. Hence we have under $\mathbf{H}_0$:

$$N_1 | (N_1 + N_2) \sim \mathcal{B}\left(N_1 + N_2, \ell_1\mu_1/[\ell_1\mu_1 + \ell_2\mu_2]\right).$$

The distribution of the counts of overlapping words is characterized by two parameters (λ and a). For such words, the frequency comparison must be stated in both terms. Assuming that the overlapping probability is the same in the two sequences leads to define the same binomial test procedure as above on the number of clumps (rather than the number of occurrences itself), that is supposed to have a Poisson distribution (see Section 1.2.3).
To illustrate this procedure, we consider the occurrences of the Chi motif $\mathbf{w} =$ `gctggtgg` in the genome of *E. coli*. This genome can be split into a very conserved part (called 'backbone') that is common to various strains of *E. coli* and a remaining part (called 'loops') that is specific to the strain under study: K12. The occurrences of Chi actually never overlap in the whole genome, the number of clumps is the number of occurrences. Chi occurs 691 in the backbone[2] and 66 times in the loops, while the expected numbers of clumps $\ell_i\tilde{\mu}_i$ under model M1 are 73.6 and 11.3, respectively, so $\ell_1\mu_1/(\ell_1\mu_1 + \ell_2\mu_2) = 86.7\%$. It seems therefore more frequent in the backbone than in the loops. To assess the significance of this difference, we calculate the p-value $\Pr\{\mathcal{B}(757, 86.7\%) \geq 691\} = 5.12\,10^{-5}$, which shows that Chi is significantly more frequent in the most conserved region of the genome, which is consistent with its favorable function.

Testing the equality of the two overlapping probabilities ($\mathbf{H}_0 : \{a_1 = a_2\}$) leads to an hyper-geometric test (see Robin *et al.* (2007)).

[2]Contrarily to Section 1.2.5, page 1.2.5, the backbone is here the one obtained from the alignment of two strains: K12 and 0157:H7

1.3.3 Detecting regions significantly enriched or devoid of a word

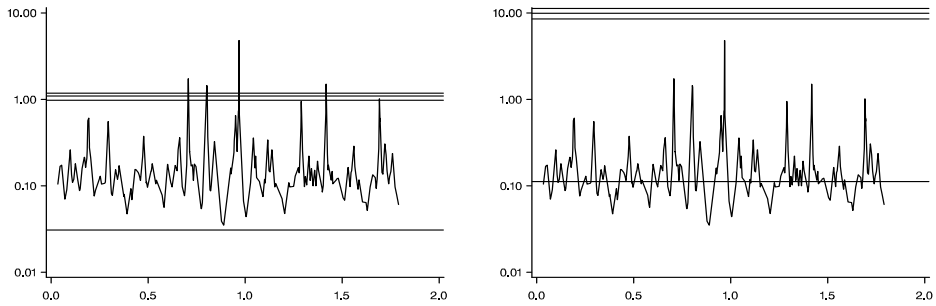
We now want to detect genome regions where the occurrences of a given word **w** are unexpectedly frequent (or rare). The standard strategy in such a situation is to use scan statistics, i.e. distances between successive occurrences. This strategy was first proposed in a genomic context by Karlin and Macken (1991). In this setting, the occurrences are supposed to occur according to an homogeneous Poisson process, which actually corresponds to a non-overlapping word.
Overlapping words can be studied in the compound Poisson model. Since the clump size has a geometric distribution, the distance D between two successive occurrence is either (i) 0 (if the two occurrences belong to the same clump) or (ii) exponential (if they belong to two successive clumps). (i) occurs with probability a and (ii) with probability $(1-a)$. The cdf of D is hence $F(y) = 1-(1-a)e^{-\lambda y}$. The analogous exact distribution is derived in Robin and Daudin (2001) in the Markov chain model. Because the occurrence process is a renewal process, the cdf F_r of the r-scan, i.e. the cumulated distance D^r between the ith occurrence and the $(i+r)$-th is simply the r times self-convolution of F: $F_r = F^{\otimes r}$.

Let $D_1^r, D_2^r, \ldots$ denote the successive r-scans; The richest region in terms of occurrences is characterized by the smallest $D_{\min}^r = \min_i D_i^r$. To check if the observed minimum distance $d_{\min}^r$ is significantly small, we need to evaluate $\Pr\{D_{\min}^r \leq d_{\min}^r\}$. A Poisson approximation strategy is proposed by Dembo and Karlin (1992):

$$\Pr\{D_{\min}^r \leq d_{\min}^r\} \approx 1 - \exp[-(N-r)F_r(d_{\min})].$$

where N is the total number of occurrences; Chen-Stein bounds for this approximation are provided. These results can be applied for both compound Poisson process [Robin (2002)] and Markov chain [Robin and Daudin (2001)] frameworks.
As an illustration, we consider the occurrences of the Chi motif in the genome of *Haemophilus influenzae*, and study their distribution using 3-scans (see page 17 to get the description of the Chi motif). The x-axis of Fig. 1.4 gives the positions in Mbps, the y-axis gives the intensity $3/D^3$ multiplied by 10^3 (in log scale); Peaks correspond to rich regions. We observe several peaks, the highest one being near the center, i.e. near the terminus of repliaction. Chi motifs are expected to be frequent here because this region is crucial in the replication mechanism of the cell. The four horizontal lines give, in ascending order, the theoretical mean intensity, the lower bound of the Chen-Stein approximation, the Chen-Stein threshold and the upper bound. We see that several peaks are significant under the M1 model, but the mean intensity of the occurrence

process is highly underestimated by this model. Using maximum-likelihood estimates, the compound Poisson model fits the observed mean intensity; In this model, even the highest peak turns out to be non-significant any-more.



First order Markov chain Compound Poisson process

Figure 1.4: Significance of the intensity peaks for the occurrences of the Chi site of *H. influenzae*.

1.4 More Sophisticated Patterns

Biological motifs are not always exact and simple words. They often contain some uncertainties (so-called degenerated motifs) like the Chi motif `gntggtgg` of *H. influenzae* (the `n` stands for any of the four DNA letters). In this case, we have to consider the occurrences of a set of words rather than a single word. In the case of transcription factor binding sites, we have to deal with several (exact or not) words that should occur at a constrained distance apart (so-called structured motifs). In Section 1.4.1, we give major extensions required to generalize the results on simple words presented in the previous sections to set of words. Then, we will present some results for structured motifs (Section 1.4.2).

1.4.1 Family of words

Let $\mathcal{W}$ be a set (family) of r words: $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$. To simplify the exposition, we will assume that all of the r words have the same length h. In the general case, one would just make the assumption that no word from the family is part of another word of the family and the results can be easily generalized.

Distribution of the count of a word family (model M1) The number of occurrences of the word family, denoted by $N(\mathcal{W})$, is simply the sum of the counts of each word taken from $\mathcal{W}$:

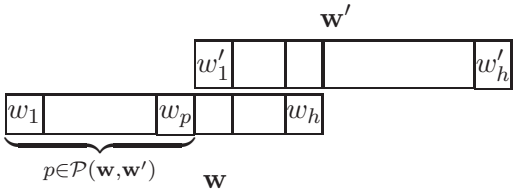
$$N(\mathcal{W}) = \sum_{j=1}^{r} N(\mathbf{w}_j).$$

The expected count $\mathbb{E}[N(\mathcal{W})]$ is then simply the sum of the r expected counts $\mathbb{E}[N(\mathbf{w}_j)]$, $j = 1, \ldots, r$. For the variance, we have $\mathbb{V}[N(\mathcal{W})] = \sum_{j=1}^{r} \mathbb{V}[N(\mathbf{w}_j)] + 2\sum_{j<j'} \mathbb{C}[N(\mathbf{w}_j), N(\mathbf{w}_{j'})]$ so we just need to derive the covariance between two word counts (see below). The Gaussian approximation of $N(\mathcal{W})$ is immediate and it is easy to derive a score of exceptionality for any family of words. For the compound Poisson approximation, it is much more involved. A first strategy could be to approximate separately the clumps of each word, and then to combine the associated Poisson variables [Reinert and Schbath (1998)]. Unfortunately, words from $\mathcal{W}$ can overlap each other and this will lead to a bad approximation for overlapping families. The alternative is to consider clumps of the word family itself, i.e. clumps composed of overlapping occurrences of $\mathcal{W}$ [Roquain and Schbath (2007)]. This leads to a compound Poisson distribution, whose parameters are derived from an overlapping probability matrix $(A(w_j, w_{j'}))_{1\leq j,j'\leq r}$, but which is not a geometric Poisson distribution. Tails of general compound Poisson distribution can be calculated by using the algorithm from Barbour *et al.* (1992a).

Covariance between two word counts in M1 Let two different words $\mathbf{w}$ and $\mathbf{w}'$ of length h. The covariance $\mathbb{C}[N(\mathbf{w}), N(\mathbf{w}')]$ is given by

$$\mathbb{C}[N(\mathbf{w}), N(\mathbf{w}')] = -\mathbb{E}[N(\mathbf{w})]\,\mathbb{E}[N(\mathbf{w}')] + \sum_{i\neq j} \mathbb{E}[Y_i(\mathbf{w})Y_j(\mathbf{w}')].$$

Thanks to symmetry, let us restrict ourself to the calculation of $\mathbb{E}[Y_i(\mathbf{w})Y_{i+d}(\mathbf{w}')]$ for $d > 0$. If $0 < d < h$, an occurrence of $\mathbf{w}'$ at position $i + d$ would overlap an occurrence of $\mathbf{w}$ at position i. We then need to introduce the possible lags between an occurrence of $\mathbf{w}$ and a following overlapping occurrence of $\mathbf{w}'$.



Let $\mathcal{P}(\mathbf{w}, \mathbf{w}')$ be the set of these possible lags, namely

$$p \in \mathcal{P}(\mathbf{w}, \mathbf{w}') \Longleftrightarrow w'_j = w_{j+p}, \quad \forall j \in \{1, \ldots, h - p\}.$$

Overlaps are not necessarily symmetric so $\mathcal{P}(\mathbf{w},\mathbf{w}') \neq \mathcal{P}(\mathbf{w}',\mathbf{w})$. For instance, `atcg` can be overlapped from the right by `cgct` after a lag of 2 ($\mathcal{P}(\texttt{atcg},\texttt{cgct}) = \{2\}$), whereas `cgct` cannot be overlapped from the right by `atcg` ($\mathcal{P}(\texttt{cgct},\texttt{atcg}) = \emptyset$).
If $p \in \mathcal{P}(\mathbf{w},\mathbf{w}')$, let $\mathbf{w}^p\mathbf{w}'$ be the word composed of two overlapping occurrences of $\mathbf{w}$ and $\mathbf{w}'$: $\mathbf{w}^p\mathbf{w}' = w_1 \cdots w_p w'_1 \cdots w'_h$.

By analogy with Equation (1.5), p. 9, one can show that

$$\mathbb{E}[Y_i(\mathbf{w}), Y_{i+d}(\mathbf{w}')] = \begin{cases} 0 & \text{if } 0 \leq d < h, d \notin \mathcal{P}(\mathbf{w},\mathbf{w}'), \\ \mu(\mathbf{w}^d\mathbf{w}') & \text{if } d \in \mathcal{P}(\mathbf{w},\mathbf{w}'), \\ \mu(\mathbf{w})\mu(\mathbf{w}')\frac{\pi^{d-h+1}(w_h,w'_1)}{\mu(w'_1)} & \text{if } d \geq h \end{cases}$$

which finally leads to the following expression for the covariance:

$$\begin{aligned} \mathbb{C}[N(\mathbf{w}), N(\mathbf{w}')] &= -\mathbb{E}[N(\mathbf{w})]\,\mathbb{E}[N(\mathbf{w}')] + \sum_{p \in \mathcal{P}(\mathbf{w},\mathbf{w}')} (n-h-p+1)\mu(\mathbf{w}^p\mathbf{w}') \\ &\quad + \sum_{p \in \mathcal{P}(\mathbf{w}',\mathbf{w})} (n-h-p+1)\mu(\mathbf{w}'^p\mathbf{w}) \\ &\quad + \mu(\mathbf{w})\mu(\mathbf{w}') \sum_{t=1}^{n-2h+1} (n-2h-t+2)\left[\frac{\pi^t(w_h,w'_1)}{\mu(w'_1)} + \frac{\pi^t(w'_h,w_1)}{\mu(w_1)}\right]. \end{aligned}$$

Note that it is also possible to calculate the asymptotic variance of $N(\mathcal{W}) - \sum_j \widehat{\mathbb{E}}[N(\mathbf{w}_j)]$ by using the conditional covariances of $(N(\mathbf{w}_j), N(\mathbf{w}_\ell))$ in the permutation model (see Schbath *et al.* (1995)).

Skew distribution As we have seen in Section 1.2.5, biologists may be interested in the statistical significance of the skew of a word $\mathbf{w}$. The skew is defined like the ratio $N(\mathbf{w})/N(\overline{\mathbf{w}})$ where $\overline{\mathbf{w}}$ is the reverse complementary[3] word of $\mathbf{w}$ (for instance if $\mathbf{w} =$ `gctggtgg` then $\overline{\mathbf{w}} =$ `ccaccagc`). To calculate the significance of the skew one then has to get (or to approximate) the following p-value:

$$\mathbb{P}\left(\frac{N(\mathbf{w})}{N(\overline{\mathbf{w}})} \geq b\right)$$

where b is the observed skew. This requires at least the joint distribution of $(N(\mathbf{w}), N(\overline{\mathbf{w}}))$.
If we assume that $(N(\mathbf{w}), N(\overline{\mathbf{w}}))$ can be approximated by a Gaussian vector with mean $(\widehat{\mathbb{E}}[N(\mathbf{w})], \widehat{\mathbb{E}}[N(\overline{\mathbf{w}})])$ and covariance matrix Σ, the above p-value can be approximated by

$$\mathbb{P}\left(\mathcal{N}(0,1) \geq \frac{b\widehat{\mathbb{E}}[N(\overline{\mathbf{w}})] - \widehat{\mathbb{E}}[N(\mathbf{w})]}{\sqrt{\Sigma_{11} - 2b\Sigma_{12} + b^2\Sigma_{22}}}\right).$$

[3] `a` is the complement of `t` whereas `c` is the complement of `g`

The right term of the above inequality will then be considered like a score to measure the significance of the skew. Typically, Σ_{11} and Σ_{22} are given by Eq. (1.20), p. 17 and Σ_{12} can be obtained similarly thanks to the conditional covariances between counts.
If $N(\mathbf{w})$ and $N(\overline{\mathbf{w}})$ are more likely to be (compound) Poisson distributed, no solution exists for now. If $\mathbf{w}$ and $\overline{\mathbf{w}}$ do not overlap each other, their counts can be approximated by two independent geometric Poisson variables [Reinert and Schbath (1998)] but it does not help to derive an asymptotic distribution for the skew.

Distances between multiple words Because of the possible overlaps between words of the family, the distribution of the inter-site distances between two word family occurrences depends on which word actually occurs first and which word occurs next [Robin (2002)]. Therefore, in the general case, the occurrences of a set of words do not constitute a renewal process and the methodology described in Section 1.3.3 cannot be used to get the r-scan distribution. In the Markov chain framework, the occurrences of a set of words turns out to be a semi-Markov process.

1.4.2 Structured motifs

A structured motif is composed of several words which should occur in a given order and at some distances apart from each other. Let consider the simple case of two fixed words $\mathbf{u}$ and $\mathbf{v}$. We define a structured motif $\mathbf{m}$ like a pattern whose $\mathbf{u}$ is a prefix, $\mathbf{v}$ is a suffix and whose length is $|\mathbf{u}| + d + |\mathbf{v}|$, $d \geq 0$. Moreover we impose that $d_1 \leq d \leq d_2$. Since d_1 can be large (typically 12 to 20 for transcription factor binding sites), it is not reasonable to view a structured motif like a set of words (i.e. a very degenerated word). Dedicated methods should then be provided. The two main questions related to structured motif occurrences are: (i) what is the probability that a random sequence contains at least one occurrence of a given structured motif? (ii) Is this structured motif more over-represented in front of genes than along the whole chromosome? For the first question, an approximate probability has been derived by assuming that the random indicator of occurrence $Y_i(\mathbf{m})$ only depends on $Y_{i-1}(\mathbf{m})$ [Robin *et al.* (2002)]; More recently the generating function of the waiting time for the first occurrence of a structured motif has been proposed [Stefanov *et al.* (2007); See also Stefanov's chapter]. For the second question, one can use the test described in Section 1.3.2 which just requires to compute $\mu(\mathbf{m}) = \mathbb{E}[Y_i(\mathbf{m})]$ the occurrence probability of $\mathbf{m}$; An example of transcription factor binding site discovery method can be found in Touzain *et al.*.

Occurrence probability The probability for $\mathbf{m}$ to occur at a given position in a random sequence $X_1, X_2, \ldots, X_n$ (model M1) is given by:

$$\mu(\mathbf{m}) = \mu(\mathbf{u}) \sum_{d=d_1}^{d_2} \mathbb{P}(D_{\mathbf{u},\mathbf{v}} = d)\mu(\mathbf{v})/\mu(v_1)$$

where $D_{\mathbf{u},\mathbf{v}}$ is the random distance between an occurrence of $\mathbf{u}$ and the next occurrence of $\mathbf{v}$, and v_1 is the first letter of $\mathbf{v}$. The distribution of $D_{\mathbf{u},\mathbf{v}}$ is given in Robin and Daudin (2001) (see also Stefanov's chapter).

1.5 Ongoing Research And Open Problems

Multiple testing problem Multiple testing problems immediately arise in motif detection studies: looking for exceptional 8-letter words leads to perform thousands of tests at the same time. The control of the false discovery rate (FDR, Benjamini and Hochberg (1995)) has received a huge attention in the last few years in the gene expression context, but it is still neglected in most motif statistic studies. The main difficulty comes from the dependency between the counts – and hence between the tests – of all words under study. Under the null (Markov) model, all word counts are correlated, since they are observed on the same sequence. The covariance between any pair of counts is actually known (see Section 1.4.1), but is difficult to account for in multiple testing procedures, partly because of high dimensionality problems.

Sequence classification Many genomes, e.g. bacterial ones, can be characterized in terms of oligo-nucleotides composition; This phenomenon is often refered to as 'genome signature'. Several new genomic approaches aim at classifying sequences with similar origins: comparative genomics aims at finding similarities between complete genomes, typically in an evolutionary perspective; Meta-genome analysis consider sets of hundreds of species living in the same environment (soil, human intestine) and deal with mixtures of sub-sequences coming from these different species.
As seen before, the Mm Markov chain model accounts for the composition of a sequence in $(m+1)$-letter works. Mixture models [McLachlan and Peel (2000)] provide a natural framework to classify objects into unknown groups. Such a model assumes that the sequences actually come from Q groups, each characterized by one transition matrix; Sequence i coming from group number q is a random path with transition matrix $\mathbf{\Pi}_q$. The Expectation-Maximization (E-M) algorithm is the standard way to estimate both group proportions and matrices $\mathbf{\Pi}_q$, which make $(Q-1) + 3Q4^m$ independent parameters. However,

mixture models generally lead to model selection problems, typically to choose the unknown number of groups Q. In the case of sequences, this problem turns out to be very complex because of different sequence lengths: long sequences tend to discriminate very easily from each other, while small sequences have almost no influence on the global model. Combinatorial arguments are needed to evaluate the number of 'efficient' parameters, i.e. the number of transition probabilities for which some information can actually be derived from the data.

High throughput sequencing This new technology is likely to be used in many biological experiments in the next decade, typically in place of micro-arrays. It consists in sequencing a huge number (40 millions) of small DNA fragments (25 nucleotides) in one run. It can be used to count the number of copies of the transcripts of a given gene, to evaluate its expression level, or to explore the meta-genome of a given ecosystem. Dealing with such large datasets is an open problem. Markov models and motif statistics can probably help to organize all these information, be we have to admit that we still do not really know how.

1.6 References

1. Arratia, R., Goldstein, L. and Gordon, L. (1990). Poisson approximation and the Chen-Stein method, *Statistical Science*, **5**, 403–434.

2. Barbour, A. D., Chen, L. H. Y. and Loh, W.-L. (1992a). Compound Poisson approximation for nonnegative random variables via Stein's method, *Annals of Probability*, **20**, 1843–1866.

3. Barbour, A. D., Holst, L. and Janson, S. (1992b). *Poisson approximation*, Oxford-University Press.

4. Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerfull approach to multiple testing, *Journal of the Royal Statistical Society., B*, **57**, 289–300.

5. Cowan, R. (1991). Expected frequencies of DNA patterns using Whittle's formula, *Journal of Applied Probability*, **28**, 886–892.

6. Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes, *Annals of Applied Probability*, **2**, 329–357.

7. El Karoui, M., Biaudet, V., Schbath, S. and Gruss, A. (1999). Characteristics of Chi distribution on several bacterial genomes, *Research in Microbiology*, **150**, 579–587.

8. Erhardsson, T. (1999). Compound Poisson approximation for Markov chains using Stein's method, *Annals of Probability*, **27**, 565–596.

9. Erhardsson, T. (2000). Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains, *Annals of Applied Probability*, **10**, 573–591.

10. 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**, e153.

11. Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate discrete distributions*, Wiley: New-York.

12. Karlin, S. and Macken, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data, *Journal of the American Statistical Association*, **86**, 27–35.

13. Lothaire, M. (2005). *Applied Combinatorics on Words*, volume 105 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press.

14. Lundstrom, R. (1990). *Stochastic models and statistical methods for DNA sequence data*, PhD thesis, University of Utah.

15. McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, Wiley.

16. Nuel, G. (2004). LD-SPatt: Large Deviations Statistics for Patterns on Markov Chains. *Journal of Computational Biology*, **11**, 1023–1033.

17. Nuel, G. (2006). Numerical Solutions for Patterns Statistics on Markov Chains. *Statistical Applications in Genetics and Molecular Biology*, **5**, Article 26.

18. Nuel, G. (2008). Cumulative distribution function of a geometric Poisson distribution, *Journal of Statistical Computation and Simulation*, **78**, 385–394

19. Prum, B., Rodolphe, F. and Turckheim, . (1995). Finding words with unexpected frequencies in DNA sequences, *Journal of the Royal Statistical Society series B*, **57**, 205–220.

20. Reinert, G., Schbath, S. and Waterman, M. (2000). Probabilistic and statistical properties of words, *Journal of Computational Biology*, **7**, 1–46.

21. Reinert, G. and Schbath, S. (1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in markov chains, *Journal of Computational Biology*, **5**, 223–254.

22. Robin, S. (2002). A compound Poisson model for words occurrences in DNA sequences, *Journal of the Royal Statistical Society series C*, **51**, 437–451.

23. Robin, S. and Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters, *Journal of Applied Probability*, **36**, 179–193.

24. Robin, S. and Daudin, J.-J. (2001). Exact distribution of the distances between any occurences of a set of words, *Annals of the Institute of Statistical Mathematics*, **53**, 895–905.

25. Robin, S., Daudin, J.-J., Richard, H., Sagot, M.-F. and Schbath, S. (2002). Occurrence probability of structured motifs in random sequences, *Journal of Computational Biology*, **9**, 761–773.

26. Robin, S., Rodolphe, F. and Schbath, S. (2005). *DNA, Words and Models*, Cambridge University Press, English version of *ADN, mots et modèles*, BELIN 2003.

27. Robin, S., Schbath, S. and Vandewalle, V. (2007). Statistical tests to compare motif count exceptionalities, *BMC Bioinformatics*, **8**, 1–20.

28. Robin, S. and Schbath, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences, *Journal of Computational Biology*, **8**, 349–359.

29. Roquain, E. and Schbath, S. (2007). Improved compound Poisson approximation for the number of occurrences of multiple words in a stationary Markov chain, *Advances in Applied Probability*, **39**, 128–140.

30. Schbath, S. (1995a). Compound Poisson approximation of word counts in DNA sequences, *ESAIM: Probability and Statistics*, **1**, 1–16.

31. Schbath, S. (1995b). *Etude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN*, PhD thesis, Université René Descartes, Paris V.

32. Schbath, S., Prum, B. and de Turckheim, . (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences, *Journal of Computational Biology*, **2**, 417–437.

33. 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.

34. 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**, 1–23.

35. Whittle, P. (1955). Some distribution and moment formulae for the Markov chain, *Journal of the Royal Statistical Society series B*, **17**, 235–242.