CHAPTER 1

Probabilistic and Statistical Properties of Finite Words in Finite Sequences

Gesine Reinert[1]
Sophie Schbath[2]
Michael S. Waterman[3]

Contents

[1]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK.
[2]Unité Mathématique, Informatique & Génome, INRA, 78352 Jouy-en-Josas, France.
[3]Department of Biological Sciences, University of Southern California, Los Angeles CA 90089, USA.

Version June 3, 2003

1.0. Introduction

Statistical and probabilistic properties of words in sequences have been of considerable interest in many fields, such as coding theory and reliability theory, and most recently in the analysis of biological sequences. The latter will serve as the key example in this chapter. We only consider finite words.

Two main aspects of word occurrences in biological sequences are: where do they occur and how many times do they occur? An important problem, for instance, was to determine the statistical significance of a word frequency in a DNA sequence. The naive idea is the following: a word may be significantly rare in a DNA sequence because it disrupts replication or gene expression, (perhaps a negative selection factor), whereas a significantly frequent word may have a fundamental activity with regard to genome stability. Well-known examples of words with exceptional frequencies in DNA sequences are certain biological palindromes corresponding to restriction sites avoided for instance in *E. coli* (Karlin et al. (1992)), and the Cross-over Hotspot Instigator sites in several bacteria (see Biaudet et al. (1998), Chedin et al. (1998), Sourice et al. (1998)). Several papers aim at identifying over- and under-represented words in a particular genome (for instance, Leung et al. (1996), Rocha et al. (1998)).

Statistical methods to study the distribution of the word locations along a sequence and word frequencies have also been an active field of research; the goal of this chapter is to provide an overview of the state of this research.

Because DNA sequences are long, asymptotic distributions were proposed first. Exact distributions exist now, motivated by the analysis of genes and protein sequences. Unfortunately, exact results are not adapted in practice for long sequences because of heavy numerical calculation, but they allow the user to assess the quality of the stochastic approximations when no approximation error can be provided. For example, BLAST is probably the best-known algorithm for DNA matching, and it relies on a Poisson approximation (Altschul et al. (1990)). Approximate p-values can be given; yet the applicability of the Poisson approximation needs to be justified.

Statistical properties of words only make sense with respect to some underlying probability model. DNA sequences are commonly modeled as stationary random sequences. Typical models are homogeneous m-order Markov chains (model Mm) in which the probability of occurrence of a letter at a given position depends only on the m previous letters in the sequence (and not on the position); the independent case is a particular case with $m = 0$. Hidden Markov models (HMMs) reveal however that the composition of a DNA sequence may vary over the sequence (Churchill (1989), Muri (1998), Durbin et al. (1998)) and can be studied with HMMs. However, no statistical properties of words have been yet derived in such heterogeneous models. DNA sequences code for amino acid sequences (proteins) by non-overlapping triplets called *codons*. The three positions of the codons have distinct statistical properties, so that for coding DNA we naturally think of three sequences where the successive letters come from the three codon positions, respectively. The three chains and their transition matrices are denoted as Mm-3. In this chapter, we will focus on the homogeneous models Mm and give existing results for Mm-3.

Because these probabilistic models have to be fitted to the observed biological sequence, we will pay attention to the influence of the model parameter estimation on the statistical results. Some asymptotic results take care of this problem but the exact results require that the true model driving the observed sequence is known.

The choice of the Markov model order depends on the sequence length, because of the data requirements in estimation. One might be able to test hierarchical models using Chi-square tests to assign which order of Markovian dependence is appropriate for the underlying sequence. From a practical point of view, it also depends on the composition of the biological sequence one wants to take into account. Indeed, if the sequence was generated from an m-order Markov chain, then the model Mm provide a good prediction for the $(m + 1)$-letter words.

In this chapter, we are concerned firstly with the occurrences of a single pattern in a sequence. To begin, we discuss the underlying probabilistic models (Section 1.1). The main complication for word occurrences arises from overlaps of words. One might be interested either in overlapping occurrences or in particular non-overlapping ones (Section 1.2). After presenting results for the statistical distribution of word locations along the sequence (Section 1.3), we focus on the distribution of the number of overlapping occurrences (Section 1.4) and the number of renewals (Section 1.5). In Section 1.6, we will study the occurrences of multiple patterns. Section 1.7 gives two applications on how probabilistic and statistical considerations come into play for DNA sequence analysis. Firstly, we look for words with unexpected counts in some DNA sequences. The focus will be on the importance of the order m of the Markov model used and on the interest of using a model of the type Mm-3 (with three transition matrices), when analyzing a coding DNA sequence. We will also take the opportunity to compare exact and asymptotic results on the word count distributions. Secondly, we describe how to analyze so-called SBH chips, a fast and effective method for determining a DNA sequence. These chips provide the ℓ-tuple contents of a DNA sequence, where typically $\ell = 8, 10$ or 12. A non-trivial combinatorial problem arises when determining the probability that a randomly chosen DNA sequence can be uniquely reconstructed from its ℓ-tuple contents. Finally, Section 1.8, meant as an appendix, gives a compilation of more general techniques that are applied in this chapter.

Due to the abundance of literature, the present chapter has no intention of being a complete literature survey (indeed even just a list of references would take up all the space designated to this chapter), but rather to introduce the reader to the major aspects of this field, to provide some techniques and to warn of major pitfalls associated with the analysis of words. For convenience, the main probabilistic tools employed are given in the appendix.

For the same reason we completely omit the algorithmic aspect; an excellent starting point would be Waterman (1995) or Gusfield (1997); for a particular example see also Apostolico et al. (1998), and for a recent exposition see Lonardi (2001).

The material in this chapter can be viewed as an updated version of Reinert et al. (2000). Recent progress on exact distributional results, as well as on compound Poisson approximation and on multivariate normal approximation, has been included.

1.1. Probabilistic models for biological sequences

In this chapter, a biological sequence is either a DNA sequence or a protein sequence, that is, a finite sequence of letters either in the 4-letter DNA alphabet $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ or the 20-letter amino-acid alphabet. To model a biological sequence, we will consider models for random sequences of letters. Even if we observed a finite biological sequence $\underline{S} = s_1 s_2 \cdots s_n$, we consider for convenience in the whole chapter an infinite random sequence $\underline{X} = (X_i)_{i \in \mathbb{Z}}$ on a finite alphabet $\mathcal{A}$, where $\mathbb{Z}$ is the set of integers. We present below two classes of Markov models widely used to analyze biological sequences and how to estimate their parameters according to the observed sequence. Then we give a classical Chi-square test to choose the appropriate order of the Markov model for a given sequence.

However, we will see in Section 1.7.1 that the choice of the model also has to take biological considerations on the sequence composition into account.

1.1.1. Markovian models for random sequences of letters

The simplest model assumes that the letters X_i are independent and take on the value $a \in \mathcal{A}$ with probability $\mu(a) = 1/|\mathcal{A}|$, where $|\mathcal{A}|$ denotes the size of the alphabet. To refine this model, we can simply assume independent letters taking values in $\mathcal{A}$ with probabilities $(\mu(a))_{a\in\mathcal{A}}$ such that $\sum_{a\in\mathcal{A}} \mu(a) = 1$. This is called model M0. Typically for DNA sequences, this model is not very accurate. Therefore, we consider a much more general homogeneous model, the model Mm: an ergodic stationary m-order Markov chain on a finite alphabet $\mathcal{A}$ with transition matrix $\Pi = (\pi(a_1 \cdots a_m, a_{m+1}))_{a_1,\dots,a_{m+1}\in\mathcal{A}}$ such that

$$\pi(a_1 \cdots a_m, a_{m+1}) = \mathbb{P}(X_i = a_{m+1} \mid X_{i-1} = a_m, \dots, X_{i-m} = a_1).$$

In general, a stationary distribution μ of an ergodic stationary Markov chain with transition matrix Π is defined as a solution of $\mu = \mu\Pi$. This implies that the above Markov chain has a unique stationary distribution μ on $\mathcal{A}^m$ defined by

$$\mu(a_1 \cdots a_m) = \mathbb{P}(X_i \cdots X_{i+m-1} = a_1 \cdots a_m), \quad \forall i \in \mathbb{Z}$$

such that the equation

$$\mu(a_1 \cdots a_m) = \sum_{b\in\mathcal{A}} \mu(ba_1 \cdots a_{m-1})\pi(ba_1 \cdots a_{m-1}, a_m)$$

is satisfied for all $(a_1 \cdots a_m) \in \mathcal{A}^m$. The model where the letters $\{X_i\}_{i\in\mathbb{Z}}$ are chosen independently with probabilities $p_1, p_2, \dots, p_{|\mathcal{A}|}$ corresponds to the transition matrix Π with identical rows $(p_1\ p_2\ \cdots\ p_{|\mathcal{A}|})$ and stationary distribution $\mu = (p_1, p_2, \dots, p_{|\mathcal{A}|})$.

A coding DNA sequence is naturally read as successive non-overlapping 3-letter words called codons. These codons are then translated into amino acids via the genetic code to produce a protein sequence. Several different codons can code for the same amino acid, and often the first two letters of a codon suffice to determine the corresponding amino acid. Therefore, letters may have different importance depending on their position with respect to the codon partition. To distinguish the letter probabilities according to their position modulo 3 in the coding DNA sequence, we consider a stationary Markov chain with three distinct transition matrices Π_1, Π_2 and Π_3 such that, for $a_1, \dots, a_{m+1} \in \mathcal{A}$ and $k \in \{1, 2, 3\}$

$$\pi_k(a_1 \cdots a_m, a_{m+1}) = \mathbb{P}(X_{3j+k} = a_{m+1} \mid X_{3j+k-1} = a_m, \dots, X_{3j+k-m} = a_1).$$

This is model Mm-3. The index $k \in \{1, 2, 3\}$ is called *phase* and represents the position of a letter inside a codon. By convention, the phase of a word is the phase of its last letter in the sequence; codons are then 3-letter words in phase

3.
The stationary distribution μ on $\mathcal{A}^m \times \{1,2,3\}$ is given by

$$\mu(a_1 \cdots a_m, k) = \mathbb{P}(X_{3j+k-m+1} \cdots X_{3j+k} = a_1 \cdots a_m), \quad \forall j \in \mathbb{Z}$$

such that the equation

$$\mu(a_1 \cdots a_m, k) = \sum_{b \in \mathcal{A}} \mu(ba_1 \cdots a_{m-1}, k-1)\pi_k(ba_1 \cdots a_{m-1}, a_m)$$

is satisfied for all $(a_1 \cdots a_m, k) \in \mathcal{A}^m \times \{1,2,3\}$.

Some general results for Markov chains will be used in the exposition. For simplicity we concentrate here on the case of a 1-order Markov chain.

The stationary distribution of a Markov chain can be obtained from its transition matrix. For a 1-order Markov chain we diagonalize the transition matrix as follows. Let $(\alpha_t)_{t=1,\ldots,|\mathcal{A}|}$ be the eigenvalues of Π such that $|\alpha_1| \geq |\alpha_2| \geq \cdots \geq |\alpha_{|\mathcal{A}|}|$. The Perron-Frobenius Theorem (see, e.g., Karlin and Taylor (1975)) ensures that $\alpha_1 = 1$ and $|\alpha_2| < 1$; we abbreviate

$$\alpha := \alpha_2. \tag{1.1.1}$$

Then $(1,1,\ldots,1)^T$ is a right eigenvector of Π for the eigenvalue 1 whereas the vector of stationary distribution $(\mu(a), a \in \mathcal{A})$ is a left-eigenvector of Π for the eigenvalue 1. Let $D = \mathrm{Diag}(1, \alpha, \alpha_3, \cdots, \alpha_{|\mathcal{A}|})$. We decompose $\Pi = PDP^{-1}$ such that the first column of P is $(1,1,\ldots,1)^T$; then the first row of P^{-1} is the vector of stationary distribution $(\mu(a), a \in \mathcal{A})$. For all $t \in \{1,\ldots,|\mathcal{A}|\}$, I_t denotes the $|\mathcal{A}| \times |\mathcal{A}|$ matrix such that all its entries are equal to 0 except $I_t(t,t) = 1$, and we define

$$Q_t := PI_tP^{-1}. \tag{1.1.2}$$

We shall use the following decomposition of the h-step transition matrix Π^h

$$\Pi^h = PD^hP^{-1} = \sum_{t=1}^{|\mathcal{A}|} \alpha_t^h Q_t \tag{1.1.3}$$

and that

$$Q_1(a,b) = \mu(b),\ \forall a, b \in \mathcal{A}. \tag{1.1.4}$$

In the exposition, we shall also refer to the *reversed* Markov chain, for a 1-order chain. Its h-step transition probabilities are given by

$$\pi_R^{(h)}(b,a) = \frac{\mu(a)\pi^{(h)}(a,b)}{\mu(b)}.$$

where the $\left(\pi^{(h)}(a,b)\right)$'s are the h-step transition probabilities for the chain itself. Another useful quantity is

$$\rho = 1 - \min\left\{\sum_{b\in\mathcal{A}} \min_{a\in\mathcal{A}} \pi(a,b), \sum_{b\in\mathcal{A}} \min_{a\in\mathcal{A}} \pi_R(a,b)\right\}. \tag{1.1.5}$$

These quantities can easily be generalized to m-order Markov chains, using the following embedding. Let us now assume that the sequence $(X_i)_{i\in\mathbb{Z}}$ is a m-order Markov chain on the alphabet $\mathcal{A}$, with transition probabilities $\pi(a_1\cdots a_m, a_{m+1})$, $a_1,\cdots,a_{m+1} \in \mathcal{A}$. Rewrite the sequence over the alphabet $\mathcal{A}^m$ by defining

$$\mathbb{X}_i = X_i X_{i+1}\cdots X_{i+m-1}, \tag{1.1.6}$$

so that the sequence $(\mathbb{X}_i)_{i\in\mathbb{Z}}$ is a first-order Markov chain on $\mathcal{A}^m$ with transition probabilities, for $\mathbb{A} = a_1\cdots a_m \in \mathcal{A}^m$ and $\mathbb{B} = b_1\cdots b_m \in \mathcal{A}^m$,

$$\Pi(\mathbb{A},\mathbb{B}) = \begin{cases} \pi(a_1\cdots a_m, b_m) & \text{if } a_2\cdots a_m = b_1\cdots b_{m-1} \\ 0 & \text{otherwise.} \end{cases}$$

1.1.2. Estimation of the model parameters

Modeling a biological sequence consists of choosing a probabilistic model (see previous paragraph) and then estimating the model parameters according to the unique realization that is the biological sequence. In the case of model Mm, it means to estimate the transition probabilities $\pi(a_1\cdots a_m, a_{m+1})$; their estimators are classically denoted by $\widehat{\pi}(a_1\cdots a_m, a_{m+1})$.

We now derive the estimators that maximize the likelihood of the M1 model given the observed sequence; we will then give the maximum-likelihood estimators in models Mm and Mm-3.

Assume $X_1\cdots X_n$ is a stationary Markov chain on $\mathcal{A}$ with transition matrix $\Pi = (\pi(a,b))_{a,b\in\mathcal{A}}$ and stationary distribution $(\mu(a))_{a\in\mathcal{A}}$. The likelihood L of the model is

$$L(\pi(a,b), a,b\in\mathcal{A}) = \mu(X_1)\prod_{a,b\in\mathcal{A}} (\pi(a,b))^{N(ab)}$$

where $N(ab)$ denotes the number of occurrences of the 2-letter word ab in the random sequence $X_1\cdots X_n$. To find the transition probabilities that maximize the likelihood, one maximizes the log likelihood

$$\log L(\pi(a,b), a,b\in\mathcal{A}) = \log\mu(X_1) + \sum_{a,b\in\mathcal{A}} N(ab)\log\pi(a,b).$$

One can separately maximize $\sum_{b\in\mathcal{A}} N(ab)\log\pi(a,b)$ for $a\in\mathcal{A}$, keeping in mind that $\sum_{b\in\mathcal{A}}\pi(a,b) = 1$. Let $a\in\mathcal{A}$ and choose $c\in\mathcal{A}$; we have

$$\sum_{b\in\mathcal{A}} N(ab)\log\pi(a,b) = \sum_{b\neq c} N(ab)\log\pi(a,b) + N(ac)\log\left(1 - \sum_{b\neq c}\pi(a,b)\right)$$

and for $b \neq c$

$$\frac{\partial}{\partial \pi(a,b)} \left(\sum_{b \in \mathcal{A}} N(ab) \log \pi(a,b) \right) = \frac{N(ab)}{\pi(a,b)} - \frac{N(ac)}{\pi(a,c)}.$$

All the partial derivatives equal to zero means that

$$\frac{N(ab)}{\pi(a,b)} = \frac{N(ac)}{\pi(a,c)} \quad \forall b \in \mathcal{A};$$

this implies in particular that

$$\frac{N(ab)}{\pi(a,b)} = \frac{\sum_{d \in \mathcal{A}} N(ad)}{\sum_{d \in \mathcal{A}} \pi(a,d)} = \sum_{d \in \mathcal{A}} N(ad) := N(a\bullet) \quad \forall b \in \mathcal{A}.$$

It follows that

$$\widehat{\pi}(a,b) = \frac{N(ab)}{N(a\bullet)} \quad \forall b \in \mathcal{A}.$$

Note that the second partial derivatives of the likelihood function are negative, assuring that we have indeed determined a maximum.

Remark 1.1.1. For notational convenience, the estimators mainly used in the remainder of the chapter will be $\widehat{\pi}(a,b) = N(ab)/N(a)$ since $N(a\bullet) = N(a)$ except for the last letter of the sequence for which the counts differ by 1.

It is important to note that the estimators $\widehat{\pi}(a,b)$ are random variables. Assuming that the biological sequence is a realization of the random sequence, one can calculate a numerical value for the estimator of $\pi(a,b)$; that is

$$\widehat{\pi}^{\text{obs}}(a,b) = \frac{N^{\text{obs}}(ab)}{N^{\text{obs}}(a\bullet)},$$

where $N^{\text{obs}}(\cdot)$ denotes the observed count in the biological sequence. As we will see, some results are obtained assuming that the true parameters $\pi(a,b)$ are known and equal, in practice, to $N^{\text{obs}}(ab)/N^{\text{obs}}(a\bullet)$, and do not take care of the estimation. It is indeed a common practice to substitute the estimator for the corresponding parameter in distributional results, but sometimes it changes the distribution being studied as it is illustrated in Waterman (1995) p. 313.

In the model Mm, the maximum-likelihood estimator of $\pi(a_1 \cdots a_m, a_{m+1})$, $a_1, \ldots, a_{m+1} \in \mathcal{A}$, is

$$\widehat{\pi}(a_1 \cdots a_m, a_{m+1}) = \frac{N(a_1 \cdots a_m a_{m+1})}{N(a_1 \cdots a_m \bullet)},$$

and in model Mm-3, we have $\forall a_1, \ldots, a_{m+1} \in \mathcal{A}$, $\forall k \in \{1,2,3\}$,

$$\widehat{\pi}_k(a_1 \cdots a_m, a_{m+1}) = \frac{N(a_1 \cdots a_m a_{m+1}, k)}{\sum_{b \in \mathcal{A}} N(a_1 \cdots a_m b, k)}.$$

1.1.3. Test for the appropriate order of the Markov model

To test which Markov model would be appropriate for a given sequence of length n, the most straightforward test is a Chi-square test, which can be viewed as a generalized likelihood ratio test. Most well-known is the Chi-square test for independence, see, e.g., Rice (1995). The case of general order Markov chains is reviewed in Billingsley (1961). Suppose we have a sample of size n cross-classified in a table with U rows and V columns. For instance, we could have four rows labeled $\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}$, and four columns labeled $\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}$, and we count how often a letter from the row is followed by a letter from the column in the sequence.

First we test whether we may assume the sequence to consist of independent letters. To this purpose, recall that $N(ab)$ denotes the count in cell (a, b), whereas $N(a\bullet)$ is the ath row count, and let $N(\bullet b)$ is the bth column count. Thus $N(ab)$ counts how often letter a is followed by letter b in the sequence. Let $\pi(a, b)$ be the probability of cell (a, b), let $\pi(a, \bullet)$ be the ath row marginal probability, and let $\pi(\bullet, b)$ be the bth column marginal probability. We test the null hypothesis of independence

$$H_0 : \pi(a, b) = \pi(\bullet, b)$$

against the alternative that the $\pi(a, b)$'s are free. Under H_0, the maximum-likelihood estimate of $\pi(a, b)$ is

$$\hat{\pi}(a, b) = \hat{\pi}(\bullet, b) = \frac{N(\bullet b)}{n-1}.$$

The Pearson chi-square statistic is the sum of the square difference between observed and estimated expected counts, divided by the estimated expected count, where expectations are taken assuming that the null hypothesis is true. Thus, under H_0, for the count $N(ab)$ we expect $(n-1)\mu(a)\pi(\bullet, b)$, estimated by $N(a\bullet)\hat{\pi}(\bullet, b)$, and the chi-square statistic is

$$\chi^2 = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} \frac{(N(ab) - N(a\bullet)N(\bullet b)/(n-1))^2}{N(a\bullet)N(\bullet b)/(n-1)}.$$

Under the null hypothesis, χ^2 follows asymptotically a chi-square distribution with $(|\mathcal{A}| - 1)^2$ degrees of freedom. Thus we would reject the null hypothesis when χ^2 is too large, compared to the corresponding chi-square distribution. As a rule of thumb, this test is applicable when the expected count in each row and column is at least 5. Applying this test to DNA counts, we thus would have to compare χ^2 to a chi-square distribution with $(4-1)^2 = 9$ degrees of freedom. A typical cutoff level would be 5%, or, if one would like to be conservative, 1%. The corresponding critical values are 16.92 for 5 %, and 21.67 for 1 %. Thus, if $\chi^2 > 16.92$, we would reject the null hypothesis of independence at the 5 % level (meaning that, if we repeated this experiment many times, in about 5% of the cases we would reject the null hypothesis when it is true). If $\chi^2 > 21.67$, we could reject the null-hypothesis at the 1 % level (so in only about 1 % of all trials would we reject the null hypothesis when it is true). Otherwise we would not reject the null hypothesis.

If the null hypothesis of independence cannot be rejected at an appropriate level (say, 5 %), then one would fit an independent model. However, if the null hypothesis does get rejected, one would test for a higher-order dependence. The next step would thus be to test for a first-order Markov chain. We desribe here the general case, following Billingsley (1961). Suppose we know that our data come from a Markov chain of order at most m. Let $N(a_1a_2\dots a_{m+1})$ be the count of the vector $(a_1, a_2, \dots, a_{m+1})$ in the sequence $(X_1, \dots, X_n)$, let $N(a_1a_2\dots a_m\bullet)$ be the count of the vector $(a_1, a_2, \dots, a_m)$ in the sequence $(X_1, \dots, X_{n-1})$, let $N(\bullet a_{m-r+1}\dots a_m\bullet)$ be the count of the vector $(a_{m-r+1}, \dots, a_m)$ in the sequence $(X_{r+1}, \dots, X_{n-1})$, $r < m$, and let $N(\bullet a_{m-r+1}\dots a_{m+1})$ be the count of the vector $(a_{m-r+1}, \dots, a_{m+1})$ in the sequence $(X_{r+1}, \dots, X_n)$. Put

$$\hat{\pi}(a_1\dots a_m, a_{m+1}) = \frac{N(\bullet a_{m-r+1}\dots a_{m+1})}{N(\bullet a_{m-r+1}\dots a_m\bullet)}.$$

Then under the null hypothesis of having a Markov chain of order r against the alternative that it is a Markov chain of order higher than r, the test statistic

$$\chi^2 = \sum_{a_1,\dots,a_{m+1}\in\mathcal{A}} \frac{\left(N(a_1a_2\dots a_{m+1}) - N(a_1a_2\dots a_m\bullet)\hat{\pi}(a_1\dots a_m, a_{m+1})\right)^2}{N(a_1a_2\dots a_m\bullet)\hat{\pi}(a_1\dots a_m, a_{m+1})}$$

is asymptotically chi-square distributed with $(|\mathcal{A}|^{m+1} - |\mathcal{A}|^m) - (|\mathcal{A}|^{r+1} - |\mathcal{A}|^r)$ degrees of freedom.

Although this test can be carried out for arbitrary orders, caution is advised: for higher order, a longer sequence of observations is required; see for example Guthrie and Youssef (1970).

1.2. Overlapping and non-overlapping occurrences

Statistical inference is often based on independence assumptions. Even if the sequence letters are independent and identically distributed, the different random indicators of word occurrences are not independent due to overlaps. For example, if $w =$ `atat` occurs at position i in the sequence, then another occurrence of w is much more likely to occur at position $i+2$ than if w did not occur at position i, and an occurrence of w at position $i+1$ is not possible. Many of the arguments needed for a probabilistic and statistical analysis of word occurrences deal with disentangling this overlapping structure.

Let $w = w_1 \cdots w_\ell$ be a word of length ℓ on a finite alphabet $\mathcal{A}$. Two occurrences of w may overlap in a sequence if and only if w is periodic, meaning that there exists a period $p \in \{1, \dots, \ell-1\}$ such that $w_i = w_{i+p}$, $i = 1, \dots, \ell-p$. A word may have several periods: for instance `gtgtgtg` admits three periods, 2, 4, 6, and aacaa has the periods 3 and 4. The set $\mathcal{P}(w)$ of the periods of w is defined by

$$\mathcal{P}(w) := \{p \in \{1, \ldots, \ell - 1\} : w_i = w_{i+p}, \forall i = 1, \ldots, \ell - p\}.$$

A word w is not periodic if and only if $\mathcal{P}(w)$ is empty. As we will see later, not all periods of a word will have the same importance; we distinguish the multiples of the minimal period $p_0(w)$ of w from the so-called *principal* periods of w, namely the periods that are not strictly multiples of the minimal period. We denote by $\mathcal{P}'(w)$ the set of the *principal* periods of w. For instance, $\mathcal{P}'(\texttt{gtgtgtg}) = \{2\}$ and $\mathcal{P}'(\texttt{aacaa}) = \{3, 4\}$.

Occurrences of periodic words tend to overlap in a sequence. There are 4 occurrences of aacaa in the sequence tg<u>aacaaacaacaa</u>tag<u>aacaa</u>aa, starting respectively at positions 3, 7, 10 and 18. The first 3 occurrences overlap and form a clump. A *clump* of w in a sequence is a maximal set of overlapping occurrences of w in the sequence. By definition two clumps of w in a sequence cannot overlap. A clump composed of exactly k overlapping occurrences of w is called a k-clump of w. There are 2 clumps of aacaa in the previous sequence, the first one is a 3-clump starting at position 3 and the second one is a 1-clump starting at position 18. Let $\mathcal{C}_k(w)$ be the set of the concatenated words composed of exactly k overlapping occurrences of w. For example, $\mathcal{C}_1(\texttt{aacaa}) = \{\texttt{aacaa}\}$ and $\mathcal{C}_2(\texttt{aacaa}) = \{\texttt{aacaacaa}, \texttt{aacaaacaa}\}$.

For a word $w = w_1 \cdots w_\ell$ we use the following prefix and suffix notation:

$$\begin{aligned} w^{(p)} &= w_1 \cdots w_p \quad \text{denotes the prefix of } w \text{ of length } p \\ w_{(q)} &= w_{\ell-q+1} \cdots w_\ell \quad \text{denotes the suffix of } w \text{ of length } q, \end{aligned} \tag{1.2.1}$$

and $w^{(p)}w = w_1 \cdots w_p w_1 \cdots w_\ell$ is the concatenated word obtained by two overlapping occurrences starting p positions apart. Sometimes we shall also use the notation $w^{(p)}.w$ for $w^{(p)}w$ for clarification. If $p \in \mathcal{P}(w)$ then $w^{(p)}$ is called a *root* of w; if $p \in \mathcal{P}'(w)$, $w^{(p)}$ is called a *principal root* of w.

Related to the set of periods is the autocorrelation polynomial $\mathcal{Q}(z)$ associated with w (see, e.g. Guibas and Odlyzko (1980), Li (1980), Biggins and Cannings (1987)) defined by

$$\mathcal{Q}(z) = 1 + \sum_{p \in \mathcal{P}(w)} \frac{\mu(w)}{\mu(w^{(\ell-p)})} z^p. \tag{1.2.2}$$

Renewals are another type of non-overlapping occurrences of interest that require scanning the sequence from one end to the other: the first occurrence of w in the sequence is a renewal and a given occurrence of w is a renewal if and only if it does not overlap a previous renewal. Renewals of w do not overlap in a sequence. In the above example, there are 3 renewals of aacaa starting at position 3, 10 and 18.

Depending on the problem, one could be interested in studying the overlapping occurrences of w in a sequence, or in restricting attention to non-overlapping occurrences: the beginnings of clumps, the beginnings of k-clumps or the renewals. We now introduce notation related to occurrences of a word $w = w_1 \cdots w_\ell$, of a clump of w, of a k-clump of w, of a renewal of w in a sequence, and to the corresponding counts.

Occurrence and number of overlapping occurrences An occurrence of w starts at position i in the sequence $\underline{X} = (X_i)_{i\in\mathbb{Z}}$ if and only if $X_i \cdots X_{i+\ell-1} = w_1 \cdots w_\ell$. Let $Y_i(w)$ be the associated random indicator

$$Y_i(w) := \mathbb{I}\{w \text{ starts at position } i \text{ in } \underline{X}\}. \tag{1.2.3}$$

For convenience in some sections, $Y_i(w)$ will be the random indicator that an occurrence of w ends at position i in $\underline{X}$; it will be made precise in that case.

In the stationary m-order Markovian model, the expectation of $Y_i(w)$, that is, the probability that an occurrence of w occurs at a given position in the sequence, is denoted by $\mu_m(w)$ and is given by

$$\mu_m(w) = \mu(w_1 \cdots w_m)\pi(w_1 \cdots w_m, w_{m+1}) \cdots \pi(w_{\ell-m} \cdots w_{\ell-1}, w_\ell). \tag{1.2.4}$$

When there is no ambiguity, the index m referring to the order of the model will be omitted.

The number of overlapping occurrences of w in the sequence $(X_i)_{i=1,\ldots,n}$, simply called *count* of w in this chapter, is defined by $N(w) = N_n(w) = \sum_{i=1}^{n-\ell+1} Y_i(w)$ (or $N(w) = \sum_{i=\ell}^{n} Y_i(w)$ if $Y_i(w)$ is associated with an occurrence of w ending at position i).

Clump and declumped counts A clump of w starts at position i in the infinite sequence $\underline{X}$ if and only if there is an occurrence of w starting at position i that does not overlap a previous occurrence of w. It follows that

$$\begin{aligned}\widetilde{Y}_i(w) &:= \mathbb{I}\{\text{a clump of } w \text{ starts at position } i \text{ in } \underline{X}\} \\ &= Y_i(w)(1 - Y_{i-1}(w)) \cdots (1 - Y_{i-\ell+1}(w)).\end{aligned} \tag{1.2.5}$$

Often $\widetilde{Y}_i(w)$ is zero, depending on the overlapping structure of w. Using the principal periods, it turns that

$$\widetilde{Y}_i(w) = Y_i(w) - \sum_{p\in\mathcal{P}'(w)} Y_{i-p}(w^{(p)}w) \tag{1.2.6}$$

with the notation from (1.2.1). Equation (1.2.6) is obtained from the two following steps: (i) note that an occurrence of w starting at position i overlaps a previous occurrence of w if and only if it is directly preceded by an occurrence of a principal root of w, meaning that a principal root $w^{(p)}$, $p \in \mathcal{P}'(w)$, occurs at position $i - p$, (ii) note that the events $E_p = \{Y_{i-p}(w^{(p)}) = 1\}$, $p \in \mathcal{P}'(w)$, are disjoint. To prove (ii), we assume that two different principal roots $w^{(p)}$ and $w^{(q)}$ occur simultaneously at position $i - p$ and $i - q$. If so, the minimal root $w^{(p_0)}$ of w could be decomposed into $w^{(p_0)} = xy = yx$ where x and y are two nonempty words. Proposition 1.3.2 from Lothaire (1983) says that two words commute if and only if they are powers of the same word. Thus, we would obtain the contradiction that the minimal root is not minimal (see Schbath (1995a) for more details).

It follows from Equation (1.2.6) that the probability $\widetilde{\mu}(w)$ that a clump of w starts at a given position in $\underline{X}$ is given by

$$\begin{aligned}\widetilde{\mu}(w) &= \mu(w) - \sum_{p\in\mathcal{P}'(w)} \mu(w^{(p)}w) \\ &= (1 - A(w))\mu(w) \qquad (1.2.7)\end{aligned}$$

where $A(w)$ is the probability for an occurrence of w to be overlapped from the left by a previous occurrence of w:

$$A(w) = \sum_{p\in\mathcal{P}'(w)} \frac{\mu(w^{(p)}w)}{\mu(w)}. \qquad (1.2.8)$$

The number $\widetilde{N}(w)$ of clumps of w in the finite sequence $X_1 \cdots X_n$ (or the declumped count) may be different from the sum $\widetilde{N}_{\mathrm{inf}}(w) = \sum_{i=1}^{n-\ell+1} \widetilde{Y}_i(w)$ because of a possible clump of w that would start in $\underline{X}$ before position 1 and would stop after position $\ell - 1$. The difference $\widetilde{N}(w) - \widetilde{N}_{\mathrm{inf}}(w)$ is either equal to 0 or equal to 1. In fact, it can be shown that $\mathbb{P}(\widetilde{N}(w) \neq \widetilde{N}_{\mathrm{inf}}(w)) \leq (\ell - 1)(\mu(w) - \widetilde{\mu}(w))$ (see Reinert and Schbath (1998)).

k-clump and number of k-clumps A *k-clump* of w starts at position i in $\underline{X}$ if and only if there is an occurrence of a concatenated word $c \in \mathcal{C}_k(w)$ starting at position i that does not overlap any other occurrence of w in the sequence $\underline{X}$. As we proceeded for a clump occurrence, an occurrence of $c \in \mathcal{C}_k(w)$ is a k-clump of w in $\underline{X}$ if and only if it is not directly preceded by any principal root $w^{(p)}$ of w and it is not directly followed by any suffix $w_{(q)} = w_{\ell-q+1} \cdots w_\ell$ with $q \in \mathcal{P}'(w)$. Some straightforward calculation yields the expression (Schbath (1995a)):

$$\begin{aligned}\widetilde{Y}_{i,k}(w) &:= \mathbb{1}\{\text{a } k\text{-clump of } w \text{ starts at position } i \text{ in } \underline{X}\} \qquad (1.2.9) \\ &= \sum_{c\in\mathcal{C}_k(w)} \left(Y_i(c) - \sum_{p\in\mathcal{P}'(w)} Y_{i-p}(w^{(p)}c) - \sum_{q\in\mathcal{P}'(w)} Y_i(cw_{(q)}) \right. \\ &\qquad \left. + \sum_{p,q\in\mathcal{P}'(w)} Y_{i-p}(w^{(p)}cw_{(q)}) \right),\end{aligned}$$

with the notation (1.2.1). It follows that the probability for a k-clump to start at a given position is given by

$$\widetilde{\mu}_k(w) = \sum_{c\in\mathcal{C}_k(w)} \mu(c) - 2 \sum_{c'\in\mathcal{C}_{k+1}(w)} \mu(c') + \sum_{c''\in\mathcal{C}_{k+2}(w)} \mu(c'').$$

This formula can be simplified. Note that $\mathcal{C}_{k+1}(w) = \{w^{(p)}c, c \in \mathcal{C}_k(w), p \in \mathcal{P}'(w)\}$ and $\mu(w^{(p)}c) = \mu(c)\frac{\mu(w^{(p)}c)}{\mu(c)} = \mu(c)\frac{\mu(w^{(p)}w)}{\mu(w)}$. By using the overlap probability $A(w)$ given in (1.2.8), we have that

$$\sum_{c' \in \mathcal{C}_{k+1}(w)} \mu(c') = A(w) \sum_{c \in \mathcal{C}_k(w)} \mu(c)$$

and it follows that

$$\begin{aligned}
\widetilde{\mu}_k(w) &= (1 - A(w))^2 \sum_{c \in \mathcal{C}_k(w)} \mu(c) \\
&= (1 - A(w))^2 A(w) \sum_{c \in \mathcal{C}_{k-1}(w)} \mu(c) \\
&\vdots \\
&= (1 - A(w))^2 A(w)^{k-1} \mu(w). \qquad (1.2.10)
\end{aligned}$$

As for the declumped count, the number of k-clumps of w in the finite sequence may be different from the sum $\widetilde{N}_{\mathrm{inf}}^{(k)}(w) = \sum_{i=1}^{n-\ell+1} \widetilde{Y}_{i,k}(w)$ because of possible end effects. The probability that these counts are not equal can be explicitly bounded, see (1.4.10), (1.4.11) below. Moreover, possible end effects may lead to a difference between the count $N(w)$ and $\sum_{k>0} k\widetilde{N}_{\mathrm{inf}}^{(k)}(w)$ but this can also be controlled (Reinert and Schbath (1998)).

Renewal and renewal count A renewal of w starts at position i in $X_1 \cdots X_n$ if and only if there is an occurrence of w starting at position i that either is the first one or does not overlap a previous renewal of w. Let $\mathbb{I}_i(w)$ be the associated random indicator:

$$\begin{aligned}
\mathbb{I}_i(w) &= \mathbb{1}\{\text{a renewal of } w \text{ starts at position } i \text{ in } X_1 \cdots X_n\} \\
&= Y_i(w) \prod_{j=i-\ell+1}^{i-1} (1 - \mathbb{I}_j(w)) \qquad (1.2.11)
\end{aligned}$$

with the convention that $\mathbb{I}_j(w) = 0$ if $j < 1$. Thus, for $i \leq \ell$, a renewal occurrence of w at position i is exactly a clump occurrence of w at i in the finite sequence. The renewal count makes extensive use of the linear ordering in the sequence: it is defined by $R(w) = R_n(w) = \sum_{i=1}^{n-\ell+1} \mathbb{I}_i(w)$.

1.3. Word locations along a sequence

The first analysis of the restriction sites in *E. coli* has been carried out by Churchill et al. (1990) while analyzing the distance between those sites. Here we are concerned with the length of the gaps between word occurrences. First we describe how to obtain the exact distribution of the distance between successive occurrences of a word, and then we give asymptotic results.

A related problem concerns the position of the first occurrence of a word; it is treated in Rudander (1996) and more recently in Stefanov (2003).

1.3.1. Exact distribution of the distance between word occurrences

Let $w = w_1 \cdots w_\ell$ be a word of length ℓ on a finite alphabet $\mathcal{A}$. We assume that $X_1 \cdots X_n$ is a stationary first-order Markov chain on $\mathcal{A}$ with transition matrix $\Pi = (\pi(a,b))_{a,b\in\mathcal{A}}$ and stationary distribution $(\mu(a))_{a\in\mathcal{A}}$. Here we are interested in the statistical distribution of the distance D between two successive occurrences of w and more precisely in the probabilities

$$\begin{aligned} f(d) &= \mathbb{P}(D = d) \\ &= \mathbb{P}(w \text{ occurs at } i + d \text{ and there is no occurrence of } w \\ &\qquad \text{between } i+1 \text{ and } i+d-1 \mid w \text{ occurs at } i), \quad d \geq 1. \end{aligned}$$

In this section, we say that a word w occurs at position i if an occurrence of w ends at position i; it happens with probability $\mu(w)$ given in (1.2.4).

The probability $f(d)$ can be obtained via a recursive formula (Robin and Daudin (1999)) as it was first proposed for independent and uniformly distributed letters by Blom and Thorburn (1982). It is clear that, if $1 \leq d \leq \ell - 1$ and $d \notin \mathcal{P}(w)$, then $f(d) = 0$. If $d \in \mathcal{P}(w)$ or if $d \geq \ell$ then we decompose the event

$$E = \{w \text{ occurs at } i + d\}$$

into the disjoint events

$$\begin{aligned} E_1 = \{&w \text{ occurs at } i+d \text{ and there is no occurrence of } w \text{ between } i+1 \\ &\text{and } i+d-1\} \end{aligned}$$

and

$$\begin{aligned} E_2 = \{&w \text{ occurs at } i+d \text{ and there are some occurrences of } w \text{ between } i+1 \\ &\text{and } i+d-1\}. \end{aligned}$$

Thus $\{E_1 \mid w \text{ at } i\}$ has probability $f(d)$. Moreover E_2 is itself decomposed as $E_2 = \cup_{h=1}^{d-1} E_2(h)$, where

$$\begin{aligned} E_2(h) = \{&\text{there is no occurrence of } w \text{ between } i+1 \text{ and } i+h-1, \\ &w \text{ occurs at } i+h \text{ and } i+d\} \end{aligned}$$

are again disjoint events.

If $1 \leq d \leq \ell - 1$ and $d \in \mathcal{P}(w)$, then $\mathbb{P}(E \mid w \text{ at } i) = \mu(w)/\mu(w^{(\ell-d)})$. Moreover, if there are occurrences at positions $i+h$ and $i+d$, for some $h < d$, then the occurrences necessarily overlap, and this is only possible for $d - h \in \mathcal{P}(w)$; in this case, $\mathbb{P}(E_2(h) \mid w \text{ at } i) = f(h)\mu(w)/\mu(w^{(\ell-d+h)})$. Thus, we have

$$\frac{\mu(w)}{\mu(w^{(\ell-d)})} = f(d) + \sum_{\substack{1\leq h\leq d-1 \\ d-h\in\mathcal{P}(w)}} f(h)\frac{\mu(w)}{\mu(w^{(\ell-d+h)})}.$$

If $d \geq \ell$, then $\mathbb{P}(E \mid w \text{ at } i) = \Pi^{d-\ell+1}(w_\ell, w_1)\mu(w)/\mu(w_1)$. If there is an occurrence at positions $i + h$ and $i + d$, for some $h < d$, then we distinguish two cases depending on the possible overlap between the occurrences at $i+h$ and $i+d$: if $d-\ell+1 \le h \le d-1$, they overlap and we use previous calculation; if $1 \le h \le d-\ell$, they do not overlap and $\mathbb{P}(E_2(h) \mid w \text{ at } i) = f(h)\Pi^{d-\ell-h+1}(w_\ell, w_1)\mu(w)/\mu(w_1)$. Thus, from

$$\mathbb{P}(E \mid w \text{ at } i) = \mathbb{P}(E_1 \mid w \text{ at } i) + \sum_{h=1}^{d-1} \mathbb{P}(E_2(h) \mid w \text{ at } i)$$

we get

$$\Pi^{d-\ell+1}(w_\ell, w_1)\frac{\mu(w)}{\mu(w_1)} = f(d) + \sum_{1\le h\le d-\ell} f(h)\Pi^{d-\ell-h+1}(w_\ell, w_1)\frac{\mu(w)}{\mu(w_1)} + \sum_{\substack{d-\ell+1\le h\le d-1\\ d-h\in\mathcal{P}(w)}} f(h)\frac{\mu(w)}{\mu(w^{(\ell-d+h)})}.$$

This is the proof of the next theorem, see Robin and Daudin (1999).

THEOREM 1.3.1. *The distribution $f(d) = \mathbb{P}(D = d)$ of the distance D between two successive occurrences of a word w in a Markov chain is given by the following recursive formula:*

If $1 \le d \le \ell-1$ and $d \notin \mathcal{P}(w)$, then $f(d) = 0$.
If $1 \le d \le \ell-1$ and $d \in \mathcal{P}(w)$,

$$f(d) = \frac{\mu(w)}{\mu(w^{(\ell-d)})} - \sum_{\substack{1\le h\le d-1\\ d-h\in\mathcal{P}(w)}} f(h)\frac{\mu(w)}{\mu(w^{(\ell-d+h)})}.$$

If $d \ge \ell$,

$$f(d) = \Pi^{d-\ell+1}(w_\ell, w_1)\frac{\mu(w)}{\mu(w_1)} - \sum_{1\le h\le d-\ell} f(h)\Pi^{d-\ell-h+1}(w_\ell, w_1)\frac{\mu(w)}{\mu(w_1)} - \sum_{\substack{d-\ell+1\le h\le d-1\\ d-h\in\mathcal{P}(w)}} f(h)\frac{\mu(w)}{\mu(w^{(\ell-d+h)})}.$$

Since D is the distance between two successive occurrences of w, note that, even if $d \in \mathcal{P}(w)$, $f(d)$ can be null. For instance, by taking $w = \mathtt{aaa}$, we have $\mathcal{P}(\mathtt{aaa}) = \{1, 2\}$, and $f(1) = \mu(\mathtt{aaa})/\mu(\mathtt{aa}) = \pi(\mathtt{a}, \mathtt{a})$, $f(2) = \pi^2(\mathtt{a}, \mathtt{a}) - f(1)\pi(\mathtt{a}, \mathtt{a}) = 0$.

Note that the recurrence formula on $f(d)$ is not a "finite" recurrence since calculating $f(d)$ requires the calculation of $f(d-1), \ldots, f(1)$, involving substantial numerical calculations for large d. One can approach this computation problem by using the generating function defined by $\Phi_D(t) := \mathbb{E}(t^D) = \sum_{d\ge1} f(d)t^d$. The key argument is that the $\Phi_D(t)$ expression is a rational function (Theorem 1.3.2 and Remark 1.3.3; see Robin and Daudin (1999)) of the form $P(t)/Q(t)$ so the coefficient $f(d)$ of t^d can be expressed by a recurrence formula whose order is the degree of the polynomial $Q(t)$ (see Section 1.8.4).

THEOREM 1.3.2. *The generating function of D is*

$$\Phi_D(t) = 1 - \mu^{-1}(w)\left(\sum_{\substack{u=0\\ u\in\mathcal{P}(W)\cup\{0\}}}^{\ell-1} \frac{t^u}{\mu(w^{(\ell-u)})} + \frac{1}{\mu(w_1)}\sum_{u\geq 1}\Pi^u(w_\ell, w_1)t^{\ell+u-1}\right)^{-1}.$$

REMARK 1.3.3. If the transition matrix Π is diagonalizable, there exists δ_i, $\beta_i \in \mathbb{C}$, $i = 2\cdots|\mathcal{A}|$, such that

$$\frac{1}{\mu(w_1)}\sum_{u\geq 1}\Pi^u(w_\ell, w_1)t^{\ell+u-1} = \frac{t^\ell}{1-t}\left(1 + \frac{1-t}{\mu(w_1)}\sum_{i=2}^{|\mathcal{A}|}\frac{\delta_i}{1-t\beta_i}\right)$$

implying that the above expression is a rational function with a pole at $t = 1$. Recently, Stefanov (2003) obtained another expression for the generating function that avoid the calculation of the "infinite" sum of the Π^u's.

REMARK 1.3.4. Since $\Phi_D(t) = \sum_{d\geq 1} f(d)t^d$, we have the general following properties:

$$\begin{aligned}\mathbb{E}(D) &= \Phi'_D(1) = \mu^{-1}(w)\\ \mathrm{Var}(D) &= \Phi''_D(1) + \Phi'_D(1)(1-\Phi'_D(1)).\end{aligned}$$

Successive derivatives of $\Phi_D(t)$ are obtained using the decomposition stated in the previous remark.

Proof

The proof of Theorem 1.3.2 is not complicated since one just has to develop the sum $\sum_{d\geq 0} f(d)t^d$ with $f(d)$ given by Theorem 1.3.1, but it is very technical. We thus only give the main lines of the calculation. By replacing $f(d)$ given by Theorem 1.3.1 in $\sum_{d\geq 0} f(d)t^d$, we obtain a sum of five term

$$\Phi_D(t) = K_1 - K_2 + K_3 - K_4 - K_5$$

with

$$\begin{aligned}K_1 &= \sum_{\substack{d=1\\ d\in\mathcal{P}(W)}}^{\ell-1} \frac{\mu(w)}{\mu(w^{(\ell-d)})}t^d\\ K_2 &= \sum_{\substack{d=1\\ d\in\mathcal{P}(W)}}^{\ell-1}\sum_{\substack{h=1\\ d-h\in\mathcal{P}(W)}}^{d-1} f(h)\frac{\mu(w)}{\mu(w^{(\ell-d+h)})}t^d\\ &= \sum_{h=1}^{\ell-2} f(h)\sum_{\substack{u=1\\ u\in\mathcal{P}(W)}} \frac{\mu(w)}{\mu(w^{(\ell-u)})}t^{h+u}\end{aligned}$$

$$
\begin{aligned}
K_3 &= \sum_{d\geq \ell} \Pi^{d-\ell+1}(w_\ell, w_1)\frac{\mu(w)}{\mu(w_1)} t^d \\
&= \frac{\mu(w)}{\mu(w_1)} t^{\ell-1} \sum_{u\geq 1} \Pi^u(w_\ell, w_1) t^u \\
K_4 &= \sum_{d\geq \ell} \sum_{h=1}^{d-\ell} f(h)\Pi^{d-\ell-h+1}(w_\ell, w_1)\frac{\mu(w)}{\mu(w_1)} t^d \\
&= \frac{\mu(w)}{\mu(w_1)} t^{\ell-1} \sum_{h\geq 1} f(h) t^h \sum_{z\geq h} \Pi^{z-h+1}(w_\ell, w_1) t^{z-h+1} \\
&= \frac{\mu(w)}{\mu(w_1)} t^{\ell-1} \Phi_D(t) \sum_{u\geq 1} \Pi^u(w_\ell, w_1) t^u
\end{aligned}
$$

and

$$
\begin{aligned}
K_5 &= \sum_{d\geq \ell} \sum_{\substack{h=d-\ell+1 \\ d-h\in\mathcal{P}(W)}}^{d-1} f(h)\frac{\mu(w)}{\mu(w^{(\ell-d+h)})} t^d \\
&= \sum_{h=1}^{\ell-1} f(h) \sum_{\substack{z=1 \\ z+\ell-h-1\in\mathcal{P}(W)}}^{h} \frac{\mu(w)}{\mu(w^{(h-z+1)})} t^{z+\ell-1} \\
&\quad + \sum_{h\geq \ell} f(h) t^h \sum_{\substack{z=h-\ell+2 \\ z+\ell-h-1\in\mathcal{P}(W)}}^{h} \frac{\mu(w)}{\mu(w^{(h-z+1)})} t^{z-h+\ell-1} \\
&= \sum_{h=1}^{\ell-1} f(h) \sum_{\substack{u=\ell-h \\ u\in\mathcal{P}(W)}}^{\ell-1} \frac{\mu(w)}{\mu(w^{(\ell-u)})} t^{h+u} + \sum_{h\geq \ell} f(h) t^h \sum_{\substack{u=1 \\ u\in\mathcal{P}(W)}}^{\ell-1} \frac{\mu(w)}{\mu(w^{(\ell-u)})} t^u.
\end{aligned}
$$

Grouping $K_1 - K_2 - K_5$ and $K_3 - K_4$ leads to

$$
\Phi_D(t) = (1-\Phi_D(t)) \left(\sum_{\substack{u=1 \\ u\in\mathcal{P}(W)}}^{\ell-1} \frac{\mu(w)}{\mu(w^{(\ell-u)})} t^u + \frac{\mu(w)}{\mu(w_1)} t^{\ell-1} \sum_{u\geq 1} \Pi^u(w_\ell, w_1) t^u \right),
$$

hence

$$
\Phi_D(t) = 1 - \left(1 + \sum_{\substack{u=1 \\ u\in\mathcal{P}(W)}}^{\ell-1} \frac{\mu(w)}{\mu(w^{(\ell-u)})} t^u + \frac{\mu(w)}{\mu(w_1)} t^{\ell-1} \sum_{u\geq 1} \Pi^u(w_\ell, w_1) t^u \right)^{-1}
$$

Using $\mu(w)/\mu(w^{(\ell)}) = 1$ establishes the theorem.

■

Similar results are derived in Robin and Daudin (2001) and Stefanov (2003) for the probability distribution of the distance between any word in a given set. They will not be presented in Section 1.6 but are useful for instance for the purpose of calculating the occurrence probability of a structured motif (Robin et al. (2002)). Indeed, a structured motif is of the form $v(d_1 : d_2)w$ that is a word v separated from a word w by a distance between d_1 and d_2; v and w can be approximate patterns. Efficient algorithms exist to find such structured motifs. These are of particular interest since they are candidate promotors for transcription (Marsan and Sagot (2000)).

The distance D between two successive occurrences of w can be seen as the distance between the j-th and $(j+1)$-th occurrence of w in the sequence, since we use a homogeneous model. It may be useful to study the distance $D^{(r)}$ between the j-th and $(j+r)$-th occurrence of w, the so-called r-scan by Karlin and colleagues (e.g. Dembo and Karlin (1992)). The distance $D^{(r)}$ is the sum of r independent and identically distributed random variables with same distribution as D. So we have

$$\Phi_{D^{(r)}}(t) = \big(\Phi_D(t)\big)^r.$$

We obtain the exact distribution of $D^{(r)}$ from the Taylor expansion of $\Phi_{D^{(r)}}(t)$: the probability $\mathbb{P}(D^{(r)} = d)$ is the coefficient of t^d in the series.

1.3.2. Asymptotic distribution of r-scans

In the preceding paragraph, we presented how to obtain the exact distribution of an r-scan $D^{(r)}$, the distance between a word occurrence and the $(r-1)$-th next one, in a stationary Markov chain of first order. Often one is interested in the occurrence of any element of a subset of words; such a subset is called a *motif*. When analyzing a biological sequence, assume we observe $(h+1)$ occurrences of a given motif, so that we observe h distances $D_1, \ldots, D_h$ between occurrences of the motif. Thus we observe $(h-r+1)$ so-called r-scans $D_i^{(r)} = \sum_{j=i}^{i+r-1} D_j$. To detect poor and rich regions with this motif, one is interested in studying the significance of the smallest and the largest r-scans, or more generally the kth smallest r-scan, denoted by m_k, and the kth largest r-scan, denoted by M_k. In this section, we present a Poisson approximation for the statistical distribution of the extreme value m_k obtained by Dembo and Karlin (1992) using the Chen-Stein method. A similar result is available for M_k by following an identical setup, so it will not be explained in detail here.

We begin by defining the Bernoulli variables that will be used in the Chen-Stein method (see Section 1.8.2):

$$W_i^-(d) := \mathbb{1}\{D_i^{(r)} \leq d\},\ d \geq 0.$$

Denote by

$$W^-(d) = \sum_{i=1}^{h-r+1} W_i^-(d)$$

the number of r-scans less or equal to d. Note the duality principle

$$\{W^-(d) < k\} = \{m_k > d\},\ d \geq 0.$$

We now use Theorem 1.8.2 to get a Poisson approximation for the distribution of $W^-(d)$. To apply this theorem, we first need to choose a neighborhood of dependence for each indicator variable; ideally the indicator variables with indices not from the neighborhood of dependence are independent of that indicator variable. Secondly there are three quantities to bound, called b_1, b_2, and b_3, given in (1.8.1), (1.8.2), and (1.8.3). Piecing this together gives a bound on the total variation distance between the distributions. Here we proceed as follows.

For $i \in \{1, \ldots, h-r+1\}$, we choose the neighborhood $B_i = \{j \mid |i-j| < r\}$, so that $D_i^{(r)}$ is independent of $D_j^{(r)}$ if $j \notin B_i$ (recall the distances $D_1, \ldots, D_h$ are independent). Let Z_{λ^-} be the Poisson variable with expectation λ^-, where

$$\begin{aligned}
\lambda^- &= \mathbb{E}W^-(d) \\
&= (h-r+1)\mathbb{E}W_i^-(d) \\
&= (h-r+1)\mathbb{P}(D^{(r)} \leq d).
\end{aligned}$$

Theorem 1.8.2 gives that

$$\begin{aligned}
d_{\mathrm{TV}}\left(\mathcal{L}(W^-(d)), \mathcal{L}(Z_{\lambda^-})\right) \leq \frac{1-\mathrm{e}^{-\lambda^-}}{\lambda^-} \Bigg(& \sum_{i=1}^{h-r+1} \sum_{j \in B_i} \mathbb{E}W_i^-(d)\mathbb{E}W_j^-(d) \\
& + \sum_{i=1}^{h-r+1} \sum_{j \in B_i \setminus \{i\}} \mathbb{E}(W_i^-(d)W_j^-(d)) \Bigg).
\end{aligned}$$

Indeed the neighborhood B_i has been chosen so that $W_i^-(d)$ is independent of $W_j^-(d)$, $\forall j \notin B_i$, leading to $b_3 = 0$. For $j > i$, we have

$$\begin{aligned}
\mathbb{E}(W_i^-(d)W_j^-(d)) &= \mathbb{P}(D_i^{(r)} \leq d, D_j^{(r)} \leq d) \\
&= \mathbb{P}(D_j^{(r)} \leq d \mid D_i^{(r)} \leq d)\mathbb{P}(D_i^{(r)} \leq d) \\
&= \mathbb{P}(D_{j-i+1}^{(r)} \leq d \mid D_1^{(r)} \leq d)\mathbb{P}(D^{(r)} \leq d).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{h-r+1} \sum_{j \in B_i \setminus \{i\}} & \mathbb{E}(W_i^-(d)W_j^-(d)) \\
& \leq 2(h-r+1)\mathbb{P}(D^{(r)} \leq d) \sum_{s=2}^{r} \mathbb{P}(D_s^{(r)} \leq d \mid D_1^{(r)} \leq d) \\
& \leq 2\lambda^- \sum_{s=2}^{r} \mathbb{P}(D_s^{(r)} \leq d \mid D_1^{(r)} \leq d).
\end{aligned}$$

It can be shown that

$$\mathbb{P}(D_s^{(r)} \leq d \mid D_1^{(r)} \leq d) \leq \mathbb{P}\left(\sum_{i=r+1}^{s+r-1} D_i \leq d\right) = \mathbb{P}(D^{(s-1)} \leq d).$$

We finally get

$$d_{\mathrm{TV}}\left(\mathcal{L}(W^-(d)), \mathcal{L}(Z_{\lambda^-})\right) \leq \left((2r-1)\mathbb{P}(D^{(r)} \leq d) + 2\sum_{s=1}^{r-1} \mathbb{P}(D^{(s)} \leq d)\right) \times (1 - \mathrm{e}^{-\lambda^-}).$$

From the duality principle,

$$|\mathbb{P}(m_k > d) - \mathbb{P}(Z_{\lambda^-} < k)| \leq \left((2r-1)\mathbb{P}(D^{(r)} \leq d) + 2\sum_{s=1}^{r-1} \mathbb{P}(D^{(s)} \leq d)\right) \times (1 - \mathrm{e}^{-\lambda^-}).$$

This approximation is very useful for the comparison between the expected distribution of the r-scans and the one observed in the biological sequence. It has been first applied in Karlin and Macken (1991) to the *E. coli* genome by approximating the r-scan distribution given in Section 1.3.1 by a sum of $r-1$ independent exponential random variables. Robin and Daudin (2001) refined this approximation using the exact distribution of the r-scans. Related work is presented by Robin (2002) but using a compound Poisson model for the word occurrences rather than a Markov model for the sequence of letters. This new approach has the advantage of taking the eventual unexpected frequency of the word of interest into account (next section) when analyzing its location along a sequence. See Glaz et al. (2001) for more material and applications of scan statistics.

1.4. Word count distribution

Let again $w = w_1 \cdots w_\ell$ be a word of length ℓ on a finite alphabet $\mathcal{A}$ and $\underline{X} = (X_i)_{i\in\mathbb{Z}}$ be a random sequence on $\mathcal{A}$. This section is devoted to the statistical distribution of the count $N(w)$ of w in the sequence $X_1 \cdots X_n$. First we state how to compute the exact distribution in the model M1, using recursion techniques. For long sequences, however, asymptotic results are obtainable, and, in general, easier to handle. Here the appropriate asymptotic regime depends crucially on the length ℓ of the target word relative to the sequence length n. For very short words, the law of large numbers can be applied to approximate the word count by the expected word count. This being a very crude estimate, one can easily improve on it by employing the Central Limit Theorem, stating that the word count distribution is asymptotically normal. This approximation will be satisfactory when the words are not too long. For rare words, as a rule of thumb words of length $\ell \asymp \log n$, a compound Poisson approximation will give better results. For the latter, the error made in the approximation can be bounded in terms of the sequence length, the word length, and word probabilities, so that it is possible to assess when a compound Poisson approximation will be a good choice. Moreover, the error bound can be incorporated to give conservative confidence intervals, as will be explained below.

1.4.1. Exact distribution

If $\underline{X}$ is a stationary first-order Markov chain, the exact distribution of the count $N(w)$ can be easily obtained using the distribution of the successive positions $(T_j)_{j\geq 1}$ of the j-th occurrence of w in $X_1 \cdots X_n$, using the duality principle

$$\{N(w) \geq j\} = \{T_j \leq n\}.$$

The exact distribution of T_j can be obtained as in Section 1.3.1, by deriving the Taylor expansion of the generating function $\Phi_{T_j}(t)$ of T_j. If $j = 1$, the generating function $\Phi_{T_1}(t)$ can be obtained as $\Phi_D(t)$ (see Theorem 1.3.2). We just state the result of Robin and Daudin (1999):

$$\Phi_{T_1}(t) = \frac{t^\ell}{1-t}\left(\sum_{\substack{u=0\\ u\in\mathcal{P}(W)\cup\{0\}}}^{\ell-1} \frac{t^u}{\mu(w^{(\ell-u)})} + \frac{1}{\mu(w_1)}\sum_{u\geq 1}\Pi^u(w_\ell, w_1)t^{\ell+u-1}\right)^{-1}.$$

As $T_j - T_1$ is a sum of $j-1$ independent and identically distributed random variables with the same distribution as D, we have $\Phi_{T_j}(t) = \Phi_{T_1}(t)\big(\Phi_D(t)\big)^{j-1}$. Now $\mathbb{P}(T_j = a) = g_j(a)$ is equal to the coefficient of t^a in the Taylor expansion of $\Phi_{T_j}(t)$. Using the duality principle, we obtain

$$\mathbb{P}(N(w) = j) = \sum_{a=\ell}^{n} g_j(a) - g_{j+1}(a).$$

This generalizes the exact result that Gentleman and Mullin (1989) obtained for the case that the sequence is composed of i.i.d. letters, where each letter occurs with equal probability. In this case, Gentleman (1994) also gives an algorithm for calculating the word frequency distribution. Moreover, in the Markov case the exact distribution of the count can also be obtained by other techniques: Kleffe and Langbecker (1990) as well as Nicodème et al. (1999) used an automaton built on the pattern structure matrix, whereas Régnier (2000) used a language decomposition approach to obtain the generating function of the count.

Language decomposition approach Here we give a flavor of the approach used in Régnier (2000). A set of basic languages, called *initial*, *minimal*, and *tail languages*, is defined, and the counting problem is rewritten as a problem of text decomposition over these basic languages. The term *language* in our context can just be interpreted as a set of finite strings of letters.

Let $\mathcal{R}$, the initial language, be the language of words which end with w and contain no other occurrence of w. Let the minimal language $\mathcal{M} \subset \mathcal{R}$ be the language of words v which end with w and such that the concatenated word wv[4] contains exactly two occurrences of w. Finally let $\mathcal{U}$, the tail language, be the language of words v such that wv contains exactly one occurrence of w. The language $\mathcal{T}_j$, composed of the words v that contain exactly j occurrences of w, $j = 1, 2, \ldots$, can thus be written as

$$\mathcal{T}_j = \mathcal{R}.\mathcal{M}^{j-1}.\mathcal{U}.$$

Here, for two languages $\mathcal{L}_1$ and $\mathcal{L}_2$,

$$\mathcal{L}_1.\mathcal{L}_2 = \{vw : v \in \mathcal{L}_1, w \in \mathcal{L}_2\}.$$

Similarly the language $\mathcal{T}$, composed of the words v that contain at least one occurrence of w can be written as

$$\mathcal{T} = \mathcal{R}.\mathcal{M}^*.\mathcal{U},$$

with

$$\mathcal{M}^* = \sum_{j \geq 0} \mathcal{M}^j.$$

Any language equation translates into equations for corresponding generating functions. For a language $\mathcal{L}$, the generating function of $\mathcal{L}$ under a probability model that assigns probability $\mu(v)$ to $v \in \mathcal{L}$ is defined as

$$L(z) = \sum_{v \in \mathcal{L}} \mu(v) z^{|v|},$$

where $|v|$ denotes the number of letters in v. Régnier (2000) shows that the generating function

$$T(z, u) := \sum_{n \geq 0} z^n \sum_{j \geq 0} \mathbb{P}(N(w) = j) u^j$$

satisfies

$$T(z, u) = \frac{uR(z)U(z)}{1 - uM(z)},$$

where R, U and M are the generating functions of the languages $\mathcal{R}$, $\mathcal{U}$, and $\mathcal{M}$, respectively. Expressions for $R(z)$, $U(z)$ and $M(z)$ are obtained in terms of the autocorrelation polynomial $\mathcal{Q}(z)$ given in (1.2.2). Refined expressions for $T(z, u)$ can also be found in Régnier and Denise (2003).
The probabilities $\mathbb{P}(N(w) = j)$ are the coefficients of $z^n u^j$ in the Taylor expansion of $T(z, u)$. These are obtained through symbolic calculation.

[4]For simplicity, we avoid the classical dot operator for concatenation of words.

1.4.2. The weak law of large numbers

As a crude first approximation, the weak law of large numbers states that the observed counts will indeed converge towards the expected counts. Indeed we may use Chebyshev's inequality to bound the expected deviation of the observed counts from the expected number of occurrences. This approximation is valid only for relatively short words, and in this case a normal approximation gives more information. Such an approximation will be derived in the following subsection.

1.4.3. Asymptotic distribution: the Gaussian regime

We assume that $\underline{X} = (X_i)_{i\in\mathbb{Z}}$ is a stationary m-order Markov chain on $\mathcal{A}$, $0 \leq m \leq \ell - 2$, with transition probabilities $\pi(a_1 \cdots a_m, a_{m+1})$ and stationary distribution $\mu(a_1 \cdots a_m)$, $a_1, \ldots, a_{m+1} \in \mathcal{A}$. For convenience in this particular subsection, we consider $N(w) = \sum_{i=\ell}^{n} Y_i(w)$ and

$$Y_i = Y_i(w) = \mathbb{1}\{w \textit{ ends} \text{ at position } i \text{ in } \underline{X}\}.$$

If the model is known, the asymptotic normality of $(N(w) - \mathbb{E}N(w))/\sqrt{n}$ directly follows from a Central Limit Theorem for Markov chains and the variance of $N(w)$ is given by Kleffe and Borodovsky (1992). When $m = 1$, the expectation and variance of $N(w)$ are

$$\begin{aligned} \mathbb{E}N(w) &= (n-\ell+1)\mu_1(w) \\ \mathrm{Var}N(w) &= \mathbb{E}N(w) + 2\sum_{p\in\mathcal{P}(w)} \mathbb{E}N(w^{(p)}w) - \mathbb{E}^2N(w) \\ &\quad + \frac{2}{\mu(w_1)}\mu_1^2(w)\sum_{d=1}^{n-2\ell+1}(n-2\ell+2-d)\Pi^d(w_\ell, w_1) \end{aligned} \tag{1.4.1}$$

where $\mu_1(w)$ is given in Eq. (1.2.4).

In the problem of finding exceptional words in biological sequences, the model is unknown and its parameters are estimated from the observed sequence. The expected mean of $N(w)$ is not available and is approximated by an estimator $\widehat{N}_m(w)$. In this paragraph, we derive both the asymptotic normality of $(N(w) - \widehat{N}_m(w))/\sqrt{n}$ and the asymptotic variance. This is not a trivial problem since the estimation changes the variance expression fundamentally.

The expected mean of $N(w)$ is given by $\mathbb{E}N(w) = (n - \ell + 1)\mu(w)$ where $\mu(w) = \mu_m(w)$ is the probability that an occurrence of w ends at a given position in the sequence (see Eq. (1.2.4)). Estimating each parameter by its maximum likelihood estimator (with the simplification from Remark 1.1.1) gives an estimator $\widehat{N}_m(w)$ of $\mathbb{E}N(w)$:

$$\widehat{N}_m(w) = \frac{N(w_1 \cdots w_{m+1}) \cdots N(w_{\ell-m} \cdots w_\ell)}{N(w_2 \cdots w_{m+1}) \cdots N(w_{\ell-m} \cdots w_{\ell-1})}. \tag{1.4.2}$$

Maximal model Let us first consider the maximal model ($m = \ell-2$), which is mainly used to find exceptional words (Brendel et al. (1986), Leung et al. (1996), Rocha et al. (1998)). To shorten the formulas, we introduce the notation

$$\begin{aligned} w^- &:= w_1 \cdots w_{\ell-1} && \text{first } \ell-1 \text{ letters of } w \\ {}^-w &:= \ \ w_2 \cdots w_\ell && \text{last } \ell-1 \text{ letters of } w \\ {}^-w^- &:= w_2 \cdots w_{\ell-1} && \ell-2 \text{ central letters of } w. \end{aligned}$$

Under the maximal model, the estimator of $N(w)$ is

$$\widehat{N}_{\ell-2}(w) = \frac{N(w_1 \cdots w_{\ell-1}) N(w_2 \cdots w_\ell)}{N(w_2 \cdots w_{\ell-1})} = \frac{N(w^-)N({}^-w)}{N({}^-w^-)};$$

moreover, the asymptotic normality of $(N(w) - \widehat{N}_{\ell-2}(w))/\sqrt{n}$ and the asymptotic variance can be obtained in an elegant way using martingale techniques. (For an introduction to martingales, see, e.g., Chung (1974)). Indeed, $\widehat{N}_{\ell-2}(w)$ is a natural estimator of $N(w^-)\pi({}^-w^-, w_\ell)$, and $N(w) - N(w^-)\pi({}^-w^-, w_\ell)$ is approximately a martingale as it is shown below.

We introduce the martingale $M_n = \sum_{i=\ell}^n (Y_i - \mathbb{E}(Y_i \mid \mathcal{F}_{i-1}))$ with $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$; it is easy to verify that $\mathbb{E}(M_n \mid \mathcal{F}_{n-1}) = M_{n-1}$. Moreover, we have

$$\begin{aligned} \mathbb{E}(Y_i \mid \mathcal{F}_{i-1}) &= \mathbb{P}(w^- \text{ ends at } i-1 \text{ and } w_\ell \text{ occurs at } i \mid \mathcal{F}_{i-1}) \\ &= \mathbb{I}\{w^- \text{ ends at } i-1\}\pi({}^-w^-, w_\ell), \end{aligned}$$

and

$$\sum_{i=\ell}^n \mathbb{E}(Y_i \mid \mathcal{F}_{i-1}) = \big(N(w^-) - \mathbb{I}\{w^- \text{ ends at } n\}\big)\pi({}^-w^-, w_\ell).$$

Therefore,

$$\begin{aligned} \frac{1}{\sqrt{n}} M_n = {} & \frac{1}{\sqrt{n}} \big(N(w) - N(w^-)\pi({}^-w^-, w_\ell)\big) \\ & - \frac{1}{\sqrt{n}} \mathbb{I}\{w^- \text{ ends at } n\}\pi({}^-w^-, w_\ell). \end{aligned} \tag{1.4.3}$$

Note that $n^{-1/2}\mathbb{I}\{w^- \text{ ends at } n\}\pi({}^-w^-, w_\ell)$ tends to zero as $n \to \infty$. The next proposition establishes the asymptotic normality of $M_n/\sqrt{n}$.

PROPOSITION 1.4.1. *Let $V = \mu(w^-)\pi({}^-w^-, w_\ell)(1 - \pi({}^-w^-, w_\ell))$. We have*

$$\frac{1}{\sqrt{n}} M_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, V) \text{ as } n \to \infty.$$

Proof

This is an application of Theorem 1.8.7 for the one-dimensional random variable $\xi_{n,i} = n^{-1/2}(Y_i - \mathbb{E}(Y_i \mid \mathcal{F}_{i-1}))$. Three conditions have to be satisfied. Condition (i) holds from $\mathbb{E}(\xi_{n,i} \mid \mathcal{F}_{i-1}) = 0$. We then have to check that

$\sum_{i=\ell}^{n} \mathrm{Var}(\xi_{n,i} \mid \mathcal{F}_{i-1})$ converges to V as $n \to \infty$. Since Y_i is a 0-1 random variable, we have

$$\begin{aligned}\mathrm{Var}(Y_i \mid \mathcal{F}_{i-1}) &= \mathbb{E}(Y_i \mid \mathcal{F}_{i-1}) - \big(\mathbb{E}(Y_i \mid \mathcal{F}_{i-1})\big)^2 \\ &= \mathbb{I}\{w^- \text{ ends at } i-1\}\pi({}^-w^-, w_\ell)\big(1 - \pi({}^-w^-, w_\ell)\big).\end{aligned}$$

We thus obtain

$$\begin{aligned}\sum_{i=\ell}^{n} \mathrm{Var}(\xi_{n,i} \mid \mathcal{F}_{i-1}) &= \frac{1}{n}\sum_{i=\ell}^{n} \mathrm{Var}(Y_i \mid \mathcal{F}_{i-1}) \\ &= \frac{1}{n} N(w^-)\pi({}^-w^-, w_\ell)(1 - \pi({}^-w^-, w_\ell)) \\ &\quad -\frac{1}{n}\mathbb{I}\{w^- \text{ ends at } i-1\}\pi({}^-w^-, w_\ell)(1 - \pi({}^-w^-, w_\ell)) \\ &\longrightarrow V \text{ as } n \to \infty;\end{aligned}$$

the convergence follows from the Law of Large Numbers: $N(w^-)/n \to \mu(w^-)$. Finally, $|\xi_{n,i}| \leq \frac{2}{\sqrt{n}}$, so that $\forall \varepsilon > 0$, $\forall n > 4/\varepsilon^2$, $\mathbb{P}(|\xi_{n,i}| > \varepsilon) = 0$, establishing condition (iii). Using Theorem 1.8.7 proves the proposition. ∎

Proposition 1.4.1 and Equation (1.4.3) also yield that

$$\frac{1}{\sqrt{n}}\big(N(w) - N(w^-)\pi({}^-w^-, w_\ell)\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V) \text{ as } n \to \infty.$$

We want to prove such convergence for

$$T_n = \frac{1}{\sqrt{n}}\big(N(w) - N(w^-)\widehat{\pi}({}^-w^-, w_\ell)\big),$$

where

$$\widehat{\pi}({}^-w^-, w_\ell) = \frac{N({}^-w)}{N({}^-w^-)}.$$

To this purpose, we decompose T_n as follows:

$$\begin{aligned}T_n &= \frac{1}{\sqrt{n}}\big(N(w) - N(w^-)\pi({}^-w^-, w_\ell)\big) \\ &\quad -\frac{1}{\sqrt{n}}N(w^-)\big(\widehat{\pi}({}^-w^-, w_\ell) - \pi({}^-w^-, w_\ell)\big) \\ &= \frac{1}{\sqrt{n}}\big(N(w) - N(w^-)\pi({}^-w^-, w_\ell)\big) \\ &\quad -\frac{1}{\sqrt{n}}\frac{N(w^-)}{N({}^-w^-)}\big(N({}^-w) - N({}^-w^-)\pi({}^-w^-, w_\ell)\big) \\ &= \frac{1}{\sqrt{n}}M_n - \frac{1}{\sqrt{n}}\frac{N(w^-)}{N({}^-w^-)}M'_n + \mathrm{o}(1), \qquad (1.4.4)\end{aligned}$$

where M'_n is the martingale $M'_n = \sum_{i=\ell}^{n} (Y_i(^-w) - \mathbb{E}(Y_i(^-w) \mid \mathcal{F}_{i-1}))$. Now, using Theorem 1.8.7 gives

$$\frac{1}{\sqrt{n}} \begin{pmatrix} M_n \\ M'_n \end{pmatrix} \longrightarrow \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} V & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \right) \tag{1.4.5}$$

with

$$V_{21} = V_{12} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=\ell}^{n} \mathbb{E}\Big(\big(Y_i - \mathbb{E}(Y_i \mid \mathcal{F}_{i-1})\big)\big(Y_i(^-w) - \mathbb{E}(Y_i(^-w) \mid \mathcal{F}_{i-1})\big)\Big)$$

and

$$V_{22} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=\ell}^{n} \mathrm{Var}(Y_i(^-w) \mid \mathcal{F}_{i-1}).$$

With the same technique as for the derivation of V, as $Y_i Y_i(^-w) = Y_i$, we get $V_{21} = V_{12} = V$ and $V_{22} = \mu(^-w^-)\pi(^-w^-, w_\ell)(1 - \pi(^-w^-, w_\ell))$. Note that the Law of Large Numbers guarantees that, almost surely,

$$\frac{N(w^-)}{N(^-w^-)} \to \frac{\mu(w^-)}{\mu(^-w^-)} \text{ as } n \to \infty. \tag{1.4.6}$$

From (1.4.4)–(1.4.6), we are now able to deduce that T_n converges in distribution to $\mathcal{N}(0, \sigma^2_{\ell-2}(w))$ with

$$\begin{aligned}
\sigma^2_{\ell-2}(w) &= V_{11} - 2\frac{\mu(w^-)}{\mu(^-w^-)} V_{12} + \left(\frac{\mu(w^-)}{\mu(^-w^-)}\right)^2 V_{22} \\
&= \mu(w^-)\left(1 - \frac{\mu(w^-)}{\mu(^-w^-)}\right) \pi(^-w^-, w_\ell)(1 - \pi(^-w^-, w_\ell)) \\
&= \frac{\mu(w)}{\mu(^-w^-)} \big(\mu(^-w^-) - \mu(w^-)\big)(1 - \pi(^-w^-, w_\ell)) \\
&= \frac{\mu(w)}{\mu(^-w^-)} \big(\mu(^-w^-) - \mu(w^-) - \mu(^-w) + \mu(w)\big) \\
&= \frac{\mu(w)}{\mu(^-w^-)^2} \big(\mu(^-w^-) - \mu(^-w)\big)\big(\mu(^-w^-) - \mu(w^-)\big).
\end{aligned}$$

We have just proved the following theorem.

THEOREM 1.4.2. *As $n \to \infty$, we have*

$$\frac{1}{\sqrt{n}} \left(N(w) - \widehat{N}_{\ell-2}(w) \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{\ell-2}(w))$$

with

$$\sigma^2_{\ell-2}(w) = \frac{\mu(w)}{\mu(^-w^-)^2} (\mu(^-w^-) - \mu(^-w))(\mu(^-w^-) - \mu(w^-))$$

and

$$\frac{N(w)-\widehat{N}_{\ell-2}(w)}{\sqrt{n\widehat{\sigma}^2_{\ell-2}(w)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

where $n\widehat{\sigma}^2_{\ell-2}(w)$ *is the plug-in estimator of* $n\sigma^2_{\ell-2}(w)$*:*

$$n\widehat{\sigma}^2_{\ell-2}(w) = \frac{\widehat{N}_{\ell-2}(w)}{N(^-w^-)^2}\big(N(^-w^-)-N(^-w)\big)\big(N(^-w^-)-N(w^-)\big).$$

Non-maximal model In the non-maximal models ($m < \ell-2$), it is straightforward to extend the previous martingale approach to prove the asymptotic normality of $(N(w)-\widehat{N}_m(w))/\sqrt{n}$ and to derive the asymptotic variance. Indeed, for each value of $\ell-m$, the difference $N(w)-\widehat{N}_m(w)$ has to be decomposed as a linear combination of martingales, exactly as it was done for T_n. For instance, if $w = abcde$ and $m = 1$, one has to write

$$\begin{aligned}
N(abcde)-\widehat{N}_1(abcde) &= N(abcde) - \frac{N(ab)N(bc)N(cd)N(de)}{N(b)N(c)N(d)}\\
&= N(abcde) - N(abcd)\frac{N(de)}{N(d)}\\
&\quad + \frac{N(de)}{N(d)}\left(N(abcd)-N(abc)\frac{N(cd)}{N(c)}\right)\\
&\quad + \frac{N(de)N(cd)}{N(d)N(c)}\left(N(abc)-N(ab)\frac{N(bc)}{N(b)}\right).
\end{aligned}$$

Another approach uses the δ-method as proposed in Lundstrom (1990). The idea is to consider $N(w)-\widehat{N}_m(w)$ as $f(\underline{N})$, where $\underline{N}$ is the count vector

$$\begin{aligned}
\underline{N} = (&N(w), N(w_1\cdots w_{m+1}),\dots,N(w_{\ell-m}\cdots w_\ell),\\
&N(w_2\cdots w_{m+1}),\dots,N(w_{\ell-m}\cdots w_{\ell-1}))
\end{aligned}$$

(see Eq. (1.4.2)). There exists a covariance matrix Σ such that

$$\frac{1}{\sqrt{n}}(\underline{N}-\mathbb{E}\underline{N}) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\Sigma)$$

(see Waterman (1995) for an exposition). The next step is to use the δ-method (Theorem 1.8.5) to transfer this convergence to $f(\underline{N})$:

$$\frac{1}{\sqrt{n}}(f(\underline{N})-f(\mathbb{E}\underline{N})) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\nabla\Sigma\nabla^t),$$

where $\nabla = \left(\frac{\partial f(x_1,\dots,x_{2(\ell-m)})}{\partial x_j}\,|_{\mathbb{E}\underline{N}}\right)_{j=1,\dots,2(\ell-m)}$ is the partial derivative vector of f. Since $f(\mathbb{E}\underline{N}) = 0$, we finally obtain

$$\frac{1}{\sqrt{n}}\left(N(w)-\widehat{N}_m(w)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\nabla\Sigma\nabla^t).$$

However, this method does not easily provide an explicit formula for the asymptotic variance since the function f and its derivative depends on $\ell - m$.

The conditional approach of Prum et al. (1995) provides an alternative to the problem. Initially presented for the model M1, this method has been generalized to the general model Mm, $0 \le m \le \ell - 2$ (Schbath (1995b)). The principle is to work conditionally on the sufficient statistic $\mathcal{S}_m$ of the model Mm, namely the collection of counts $\{N(a_1 \cdots a_{m+1}),\ a_1, \ldots, a_{m+1} \in \mathcal{A}\}$ and the first m letters of the sequence. Using the technique developed by Cowan (1991), one can derive both the conditional expectation $\mathbb{E}(N(w) \mid \mathcal{S}_m)$ and the conditional variance of $N(w)$. The key arguments are first that the conditional expectation is asymptotically equivalent to $\widehat{N}_m(w)$, leading to the asymptotic normality of $(N(w) - \mathbb{E}(N(w) \mid \mathcal{S}_m))/\sqrt{n}$, and second, that $n^{-1}\mathrm{Var}(N(w) \mid \mathcal{S}_m)$ has the limiting value $\sigma_m^2(w)$ with

$$\sigma_m^2(w) = \mu(w) + 2 \sum_{p \in \mathcal{P}(w),\, p \le \ell - m - 1} \mu(w^{(p)}w) + \mu(w)^2 \left(\sum_{a_1, \ldots, a_m} \frac{n(a_1 \cdots a_m \bullet)^2}{\mu(a_1 \cdots a_m)} - \sum_{a_1, \ldots, a_{m+1}} \frac{n(a_1 \cdots a_{m+1})^2}{\mu(a_1 \cdots a_{m+1})} + \frac{1 - 2n(w_1 \cdots w_m \bullet)}{\mu(w_1 \cdots w_m)} \right), \qquad (1.4.7)$$

where $n(\cdot)$ denotes the number of occurrences inside w, and $n(a_1 \cdots a_m \bullet)$ stands for $\sum_{b \in \mathcal{A}} n(a_1 \cdots a_m b)$. Since the conditional moment of order 4 of $N(w)/\sqrt{n}$ is bounded, it follows that

$$\frac{1}{\sqrt{n}} \left(N(w) - \widehat{N}_m(w) \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_m^2(w)).$$

The overlapping structure of w clearly appears in the limiting variance. It is an exercise to verify that the limiting variances given by Theorem 1.4.2 and Equation (1.4.7) with $m = \ell - 2$ are identical.

Taking the phase into account Both martingale and conditional approaches can be extended to the Mm-3 model (see Section 1.1 for definition and notation). When one wants to distinguish the occurrences of w in a coding DNA sequence according to a particular phase $k \in \{1, 2, 3\}$ (k represents the position of the word with respect to the codons), one is interested in the count $N(w, k)$ of w in phase k in $X_1 \cdots X_n$; recall that the word phase is the phase of its last letter. Here we state the result in the maximal model; see Schbath (1995b) for the general case.

THEOREM 1.4.3. *Assume $\underline{X} = (X_i)_{i \in \mathbb{Z}}$ is a stationary $(\ell - 2)$-order Markov chain on $\mathcal{A}$ with transition probabilities $\pi_k(a_1 \cdots a_{\ell-2}, b)$ and stationary distribution $\mu(a_1 \cdots a_{\ell-2}, k)$, $a_1, \ldots, a_{\ell-2}, b \in \mathcal{A}$, $k \in \{1, 2, 3\}$. As $n \to \infty$, we have*

$$\frac{1}{\sqrt{n}} \left(N(w, k) - \frac{N(w^-, k-1) N({}^-w, k)}{N({}^-w^-, k-1)} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_{\ell-2}^2(w, k))$$

with

$$\sigma^2_{\ell-2}(w,k) = \frac{\mu(w,k)}{\mu(^-w^-,k-1)^2}\Big(\mu(^-w^-,k-1) - \mu(^-w,k)\Big) \times \Big(\mu(^-w^-,k-1) - \mu(w^-,k-1)\Big)$$

and

$$\begin{aligned} \mu(w^-,k-1) &= \mu(w_1\cdots w_{\ell-2},k-2)\pi_{k-1}(w_1\cdots w_{\ell-2},w_{\ell-1}) \\ \mu(^-w,k) &= \mu(^-w^-,k-1)\pi_k(^-w^-,w_\ell) \\ \mu(w,k) &= \mu(w^-,k-1)\pi_k(^-w^-,w_\ell). \end{aligned}$$

Error bound for the approximation Using Stein's method for normal approximations, namely Theorem 1.8.1. the following result was obtained by Huang (2002). It provides a bound on the distance to the normal distribution; however, it does not take the estimation of parameters into account.

Recall $v^2 = \mathrm{Var}N(w)$ from (1.4.1), and α given in (1.1.1). Using Theorem 1.8.1, Huang (2002) derived the following result.

THEOREM 1.4.4. *Assume $\underline{X} = (X_i)_{i\in\mathbb{Z}}$ is a stationary 1-order Markov chain. Let w be a word of length ℓ and $Z \sim \mathcal{N}((n-\ell+1)\mu(w), v^2)$. There are constants c and C_1, C_2, C_3 such that*

$$|\mathbb{P}(N(w)\le x) - \mathbb{P}(Z\le x)| \le c \min_{\ell\le s\le \frac{n}{2}} B_s,$$

where

$$\begin{aligned} B_s = {} & 2(4s-3)v^{-1} + 2n(2s-1)(4s-3)v^{-3}(|\log v^{-1}| + \log n) \\ & +C_1 n v^{-1}\mu(w)|\alpha|^{s-\ell+1} \\ & +C_2(|\log v^{-1}| + \log n)(2s-1)|\alpha|^{s-\ell+1} \\ & +C_3(|\log v^{-1}| + \log n)(n-2s+1)n\mu^2(w)v^{-2}|\alpha|^{s-\ell+1}. \end{aligned}$$

For a discussion on the constant c, which is not easy to describe, see Huang (2002) and references therein. The multivariate generalization will be presented in Theorem 1.6.1, where the explicit forms of the constants C_1, C_2, and C_3 will be given. The result in the independent case was first presented in Reinert et al. (2000).

1.4.4. Asymptotic distribution: the Poisson regime

In the previous section, we showed that the count $N(w)$ of a word w in a random sequence of length n can be approximated by a Gaussian distribution for large n. This Gaussian approximation is in fact not good when the expected count $(n-\ell+1)\mu(w)$ is very small, meaning that w is a rare word. Poisson approximations are appropriate for counts of rare events. As an illustration, it is well-known that a sum of independent Bernoulli variables can be either approximated by a Gaussian distribution or a Poisson distribution, depending on the asymptotic behavior of the expected value.

When the sequence letters are independent, Poisson and compound Poisson approximations for $N(w)$ have been widely studied in the literature (Chryssaphinou and Papastavridis (1988a), Chryssaphinou and Papastavridis (1988b), Arratia et al. (1990), Godbole (1991), Hirano and Aki (1993), Godbole and Schaffner (1993), Fu (1993)). Markovian models under different conditions have then been considered (Rajarshi (1974), Godbole (1991), Godbole and Schaffner (1993), Hirano and Aki (1993), Geske et al. (1995), Schbath (1995a), Erhardsson (1997)), but few works concern general periodic words and provide explicit parameters of the limiting distribution. As we will see, a Poisson distribution is not satisfactory for periodic words because of possible overlaps; a compound Poisson distribution is proposed. Two classes of tools can be used: generating functions, which do not provide any approximation error, and the Chen-Stein method, which gives a bound for the total variation distance between the two distributions (see Section 1.8.2 for details). In this section, we chose to present the Chen-Stein approach under a first-order Markovian model with known parameters; generalizations to higher order and to estimated parameters are presented at the end of the section. No assumption is made on the overlapping structure of the word w. Our two basic references are Arratia et al. (1990) and Schbath (1995a).

We assume that $\underline{X} = (X_i)_{i\in\mathbb{Z}}$ is a stationary first-order Markov chain on $\mathcal{A}$, with transition probabilities $\pi(a,b)$ and stationary distribution $\mu(a)$, $a,b\in\mathcal{A}$. Let $w = w_1\cdots w_\ell$ be a word of length ℓ on $\mathcal{A}$. Here, $Y_i = Y_i(w) = \mathbb{I}\{w \text{ starts at position } i \text{ in } \underline{X}\}$ and $\mu(w) = \mathbb{E}Y_i(w)$. Moreover, we make the *rare word assumption* $n\mu(w) = \mathrm{O}(1)$. Note that $n\mu(w) = \mathrm{O}(1)$ also means $\ell = \mathrm{O}(\log n)$.

Applying Theorem 1.8.2 to the Bernoulli variables Y_i, we obtain a bound $b_1 + b_2 + b_3$ for the total variation distance between the distribution of $N(w)$ and the Poisson distribution with mean $(n-\ell+1)\mu(w)$ that does not converge to 0 under the rare word assumption. The problem comes from the b_2 term and the possible overlaps of periodic words. Indeed, let w be a periodic word; its set of periods $\mathcal{P}(w)$ is not empty. Take $B_i = \{i-2\ell+1,\ldots,i+2\ell-1\}$ for the neighborhood of $i\in I = \{1,\ldots,n-\ell+1\}$; then b_1 and b_3 tend to 0 as $n\to+\infty$. We obtain

$$b_2 := \sum_{i\in I}\sum_{j\in B_i\setminus\{i\}} \mathbb{E}Y_iY_j = 2(n-\ell+1)\sum_{p\in\mathcal{P}(w)}\mu(w^{(p)}w) + \mathrm{O}(n\ell\mu^2(w));$$

this quantity can be of order $\mathrm{O}(1)$ if $\mathcal{P}(w)$ contains small periods p. The Poisson approximation is however valid for the count of non-periodic words because the set of periods is empty. For periodic words, the crucial argument is to consider clumps, as by definition they cannot overlap. We first prove that the declumped count $\widetilde{N}(w)$ can be approximated by a Poisson distribution with mean $(n-\ell+1)\widetilde{\mu}(w)$ (see Eq. (1.2.7)) by applying Theorem 1.8.2 to the Bernoulli variables $\widetilde{Y}_i(w)$ defined in (1.2.5). For simplicity, the variables $\widetilde{Y}_i(w)$ are denoted by $\widetilde{Y}_i$. In the next section we prove a compound Poisson approximation for $N(w)$.

Poisson approximation for the declumped count Our aim is to approximate the vector $\underline{\widetilde{Y}} = (\widetilde{Y}_i(w))_{i\in I}$ of Bernoulli variables by a vector $\underline{Z} = (Z_i)_{i\in I}$ with independent Poisson coordinates of mean $\mathbb{E}Z_i = \mathbb{E}\widetilde{Y}_i(w) = \widetilde{\mu}(w)$, where $\widetilde{\mu}(\cdot)$ is defined in (1.2.7). To apply Theorem 1.8.2, we choose the following neighborhood of $i \in I$:

$$B_i := \{j \in I : |j-i| \le 3\ell - 3\}.$$

The neighborhood is such that, for j not in B_i, there are no letters X_h common to $\widetilde{Y}_i$ and $\widetilde{Y}_j$, and moreover, the X_h's defining $\widetilde{Y}_i$ and those defining $\widetilde{Y}_j$ are separated by at least ℓ positions. It is important to consider a lag converging to infinity with n since it leads to the exponential decay of the b_3 term given by Theorem 1.8.2 as we will see below. Deriving a bound for the total variation distance between $\underline{\widetilde{Y}}$ and $\underline{Z}$ consists of bounding the quantities b_1, b_2 and b_3 given in (1.8.1), (1.8.2) and (1.8.3). Bounding b_1 presents no difficulty:

$$b_1 := \sum_{i\in I}\sum_{j\in B_i} \mathbb{E}\widetilde{Y}_i\mathbb{E}\widetilde{Y}_j \le (n-\ell+1)(6\ell-5)\widetilde{\mu}^2(w) = \mathrm{O}\left(\frac{\log n}{n}\right).$$

Since clumps of w do not overlap in the sequence, $\widetilde{Y}_i\widetilde{Y}_j = 0$ for $|j-i| < \ell$. Therefore, we get

$$b_2 := \sum_{i\in I}\sum_{j\in B_i\setminus\{i\}} \mathbb{E}\widetilde{Y}_i\widetilde{Y}_j \le 2\sum_{i\in I}\sum_{j=i+\ell}^{i+3\ell-3} \mathbb{E}\widetilde{Y}_i\widetilde{Y}_j$$

using the symmetry of B_i. Now we have

$$\mathbb{E}\widetilde{Y}_i\widetilde{Y}_j \le \mathbb{E}\widetilde{Y}_i Y_j = \widetilde{\mu}(w)\Pi^{j-i-\ell+1}(w_\ell, w_1)\frac{\mu(w)}{\mu(w_1)}$$

and

$$b_2 \le \frac{2}{\mu(w_1)}(n-\ell+1)\widetilde{\mu}(w)\mu(w)\sum_{s=1}^{2\ell-2}\Pi^s(w_\ell, w_1) = \mathrm{O}\left(\frac{\log n}{n}\right).$$

Bounding b_3 is a little more involved but we give all the steps because the same technique is used for the compound Poisson approximation of the count and will not be described in detail there. By definition we have

$$b_3 := \sum_{i\in I}\mathbb{E}|\mathbb{E}(\widetilde{Y}_i - \mathbb{E}\widetilde{Y}_i \mid \sigma(\widetilde{Y}_j, j\notin B_i))|.$$

Since $\sigma(\widetilde{Y}_j, j \notin B_i) \subset \sigma(X_1, \ldots, X_{i-2\ell+1}, X_{i+2\ell-1}, \ldots, X_n)$, properties of conditional expectation and the Markov property give

$$\begin{aligned} b_3 &\le \sum_{i\in I} \mathbb{E}|\mathbb{E}(\widetilde{Y}_i - \mathbb{E}\widetilde{Y}_i \mid X_{i-2\ell+1}, X_{i+2\ell-1})| \\ &\le \sum_{i\in I}\sum_{x,y\in\mathcal{A}} |\mathbb{E}(\widetilde{Y}_i - \mathbb{E}\widetilde{Y}_i \mid X_{i-2\ell+1} = x, X_{i+2\ell-1} = y)| \\ &\qquad\qquad \times \mathbb{P}(X_{i-2\ell+1} = x, X_{i+2\ell-1} = y). \end{aligned}$$

To evaluate the right-hand term, we introduce the set of possible words of length $\ell - 1$ preceding a clump of w:

$$\mathcal{G}(w) = \{g = g_1 \cdots g_{\ell-1} : \text{ for all } p \in \mathcal{P}(w), g_{\ell-p} \cdots g_{\ell-1} \ne w^{(p)}\}. \quad (1.4.8)$$

Thus a clump of w starts at position i in $(X_i)_{i\in\mathbb{Z}}$ if and only if one of the words gw, $g \in \mathcal{G}(w)$, starts at position $i - \ell + 1$. Therefore, we can write

$$\widetilde{Y}_i(w) = \sum_{g\in\mathcal{G}(w)} Y_{i-\ell+1}(gw). \quad (1.4.9)$$

This gives

$$\begin{aligned} &b_3 \\ &\le \sum_{i\in I}\sum_{x,y\in\mathcal{A}}\sum_{g\in\mathcal{G}(w)} |\mathbb{E}(Y_{i-\ell+1}(gw) - \mathbb{E}Y_{i-\ell+1}(gw) \mid X_{i-2\ell+1} = x, X_{i+2\ell-1} = y)| \\ &\qquad\qquad \times \mathbb{P}(X_{i-2\ell+1} = x, X_{i+2\ell-1} = y) \\ &= \sum_{i\in I}\sum_{x,y\in\mathcal{A}}\sum_{g\in\mathcal{G}(w)} |\mathbb{P}(X_{i-2\ell+1} = x, Y_{i-\ell+1}(gw) = 1, X_{i+2\ell-1} = y) \\ &\qquad\qquad - \mu(gw)\mathbb{P}(X_{i-2\ell+1} = x, X_{i+2\ell-1} = y)| \\ &= \sum_{i\in I}\sum_{x,y\in\mathcal{A}}\sum_{g\in\mathcal{G}(w)} \left|\mu(x)\Pi^\ell(x,g_1)\frac{\mu(gw)}{\mu(g_1)}\Pi^\ell(w_\ell,y) - \mu(gw)\mu(x)\Pi^{4\ell-2}(x,y)\right|. \end{aligned}$$

We now use the diagonalization (1.1.3) and (1.1.4), with α given in (1.1.1), yielding

$$\begin{aligned} b_3 \le (n-\ell+1)|\alpha|^\ell \sum_{g\in\mathcal{G}(w)} \mu(gw) \sum_{x,y\in\mathcal{A}} \mu(x) &\left| \frac{1}{\mu(g_1)} \sum_{(t,t')} \frac{\alpha_t^\ell \alpha_{t'}^\ell}{\alpha^\ell} Q_t(x,g_1) Q_{t'}(w_\ell,y) \right. \\ &\qquad \left. - \sum_{t=1}^{|\mathcal{A}|} \frac{\alpha_t^{4\ell-2}}{\alpha^\ell} Q_t(x,y) \right| \\ = (n-\ell+1)|\alpha|^\ell \sum_{g\in\mathcal{G}(w)} \mu(gw) \sum_{x,y\in\mathcal{A}} \mu(x) &\left| \frac{1}{\mu(g_1)} \sum_{(t,t')\ne(1,1)} \frac{\alpha_t^\ell \alpha_{t'}^\ell}{\alpha^\ell} Q_t(x,g_1) Q_{t'}(w_\ell,y) \right. \end{aligned}$$

$$
\left. - \sum_{t=2}^{|\mathcal{A}|} \frac{\alpha_t^{4\ell-2}}{\alpha^\ell} Q_t(x,y) \right|
$$

$$
\leq (n-\ell+1)|\alpha|^\ell \sum_{g\in\mathcal{G}(w)} \mu(gw)\gamma(\ell, w_\ell),
$$

where

$$
\gamma(\ell, a) = \max_{b\in\mathcal{A}} \sum_{x,y\in\mathcal{A}} \mu(x) \left| \frac{1}{\mu(b)} \sum_{(t,t')\neq(1,1)} \frac{\alpha_t^\ell \alpha_{t'}^\ell}{\alpha^\ell} Q_t(x,b) Q_{t'}(a,y) - \sum_{t=2}^{|\mathcal{A}|} \frac{\alpha_t^{4\ell-2}}{\alpha^\ell} Q_t(x,y) \right|.
$$

Note that $\gamma(\ell, w_\ell) = \mathrm{O}(1)$. From (1.4.8) we have $\sum_{g\in\mathcal{G}(w)} \mu(gw) = \widetilde{\mu}(w)$ and

$$
b_3 \leq (n-\ell+1)\widetilde{\mu}(w)\gamma(\ell, w_\ell)|\alpha|^\ell = \mathrm{O}(|\alpha|^\ell).
$$

We have proved the next theorem.

THEOREM 1.4.5. *Let $\underline{Z} = (Z_i)_{i\in I}$ be independent Poisson variables with expectation $\mathbb{E}Z_i = \mathbb{E}\widetilde{Y}_i(w) = \widetilde{\mu}(w)$. We have*

$$
\begin{aligned}
d_{TV}\left(\mathcal{L}(\underline{\widetilde{Y}}), \mathcal{L}(\underline{Z})\right) \leq (n-\ell+1)\widetilde{\mu}(w) \Bigg\{ & (6\ell-5)\widetilde{\mu}(w) + \gamma(\ell, w_\ell)|\alpha|^\ell \\
& + \frac{2}{\mu(w_1)}\mu(w) \sum_{s=1}^{2\ell-2} \Pi^s(w_\ell, w_1) \Bigg\}.
\end{aligned}
$$

The declumped count $\widetilde{N}(w)$ can be approximated by $\widetilde{N}_{\mathrm{inf}}(w) := \sum_{i\in I} \widetilde{Y}_i(w)$ since

$$
\begin{aligned}
d_{\mathrm{TV}}\left(\mathcal{L}(\widetilde{N}(w)), \mathcal{L}(\widetilde{N}_{\mathrm{inf}}(w))\right) &\leq \mathbb{P}(\widetilde{N}(w) \neq \widetilde{N}_{\mathrm{inf}}(w)) \\
&\leq (\ell-1)(\mu(w) - \widetilde{\mu}(w)). \qquad (1.4.10)
\end{aligned}
$$

Using the triangle inequality leads to the following corollary.

COROLLARY 1.4.6. *Let Z be a Poisson variable with expectation $\mathbb{E}Z = (n-\ell+1)\widetilde{\mu}(w)$. We have*

$$
\begin{aligned}
d_{TV}\left(\mathcal{L}(\widetilde{N}(w)), \mathcal{L}(Z)\right) \leq (n-\ell+1)\widetilde{\mu}(w) \Bigg\{ & (6\ell-5)\widetilde{\mu}(w) + \gamma(\ell, w_\ell)|\alpha|^\ell \\
& + \frac{2}{\mu(w_1)}\mu(w) \sum_{s=1}^{2\ell-2} \Pi^s(w_\ell, w_1) \Bigg\} \\
& + (\ell-1)(\mu(w) - \widetilde{\mu}(w)).
\end{aligned}
$$

Estimation of the parameters When the transition probabilities are unknown and can only be estimated from the observed sequence, we need to evaluate the total variation distance between the word count distribution and the distribution of $\sum_{k\geq 1} kZ'_k$, the Z'_k's being independent Poisson variables with expectation $(n-\ell+1)\widehat{\widetilde{\mu}}_k(w)$, where $\widehat{\widetilde{\mu}}_k(w)$ is the observed value of the plug-in maximum likelihood estimator of $\widetilde{\mu}_k(w)$. Similarly, we want to know the total variation distance between the declumped count, $\widetilde{N}(w)$, and the Poisson variable with expectation $(n-\ell+1)\widehat{\widetilde{\mu}}(w)$. For this we use the triangle inequality and the fact that the total variation distance between two Poisson variables with expectation λ and λ' is less than $|\lambda-\lambda'|$:

$$d_{\mathrm{TV}}\left(\mathcal{L}(\widetilde{N}(w)), \mathcal{P}\mathrm{o}((n-\ell+1)\widehat{\widetilde{\mu}}(w))\right) \leq d_{\mathrm{TV}}\left(\mathcal{L}(\widetilde{N}(w)), \mathcal{P}\mathrm{o}((n-\ell+1)\widetilde{\mu}(w))\right) + (n-\ell+1)|\widehat{\widetilde{\mu}}(w)-\widetilde{\mu}(w)|.$$

Using the Law of Iterated Logarithm for Markov chains (Senoussi (1990)) and Equation (1.2.4), one can show that

$$\widehat{\mu}(w) = \mu(w)\left(1+\mathrm{O}\left(\frac{\ell\sqrt{\log\log n}}{\sqrt{n}}\right)\right) \qquad \text{almost surely (a.s.)}$$

(see Schbath (1995b)). Under the rare word condition $n\mu(w)=\mathrm{O}(1)$, we get

$$n\widehat{\mu}(w) - n\mu(w) = \mathrm{O}\left(\frac{\ell\sqrt{\log\log n}}{\sqrt{n}}\right) \qquad \text{a.s.}$$

Now, using Equation (1.2.7), we obtain

$$n\widehat{\widetilde{\mu}}(w) - n\widetilde{\mu}(w) = \mathrm{O}\left(\frac{\ell^2\sqrt{\log\log n}}{\sqrt{n}}\right) \qquad \text{a.s.}$$

This quantity converges to zero as $n\to\infty$, because the rare word condition implies that $\ell=\mathrm{O}(\log n)$. Thus,

$$\begin{aligned} & d_{\mathrm{TV}}\left(\mathcal{L}(\widetilde{N}(w)), \mathcal{P}\mathrm{o}\big((n-\ell+1)\widehat{\widetilde{\mu}}(w)\big)\right) \\ & \leq d_{\mathrm{TV}}\left(\mathcal{L}(\widetilde{N}(w)), \mathcal{P}\mathrm{o}\big((n-\ell+1)\widetilde{\mu}(w)\big)\right) + \mathrm{O}\left(\frac{\ell^2\sqrt{\log\log n}}{\sqrt{n}}\right). \end{aligned}$$

The approximation follows from Corollary 1.4.8.

We do not have an explicit bound for this additional error term. However, for long sequences the error term due to the maximum-likelihood estimation will be small compared to the bound on the Poisson approximation error.

1.4.5. Asymptotic distribution: the Compound Poisson regime

Here we present two approaches for a compound Poisson approximation for the count. Firstly, such an approximation can be derived using a Poisson process approximation for the Bernoulli variables $\widetilde{Y}_{i,k}(w)$ defined in (1.2.9) and by using that $N(w)$ is asymptotically equivalent to $\sum_{i\in I}\sum_{k\geq 1} k\widetilde{Y}_{i,k}(w)$ in probability. For simplicity, the variables $\widetilde{Y}_{i,k}(w)$ are denoted by $\widetilde{Y}_{i,k}$. Secondly, a direct approximation for $N(w)$ can be obtained using Stein's method for compound Poisson approximation. The second method yields better bounds on the approximation, whereas the first method is easier to generalize to multivariate results, as will be shown in Section 1.6.

Compound Poisson approximation via Poisson process To approximate the distribution of the count $N(w)$, we first use that $N(w)$ is asymptotically equivalent to $N_{\text{inf}}(w) := \sum_{i=1}^{n-\ell+1}\sum_{k\geq 1} k\widetilde{Y}_{i,k}$ in probability (Reinert and Schbath (1998)):

$$\begin{aligned} d_{\text{TV}}\left(\mathcal{L}(N(w)), \mathcal{L}(N_{\text{inf}}(w))\right) &\leq \mathbb{P}(N(w) \neq N_{\text{inf}}(w)) \\ &\leq 2(\ell-1)(\mu(w) - \widetilde{\mu}(w)). \end{aligned} \tag{1.4.11}$$

Our goal is now to approximate the vector $(\widetilde{Y}_{i,k})_{(i,k)\in I}$, $I = \{1,\ldots,n-\ell+1\} \times \{1,2,\ldots\}$, of Bernoulli variables by a vector $(Z_{i,k})_{(i,k)\in I}$ with independent Poisson coordinates of expectation $\mathbb{E}Z_{i,k} = \mathbb{E}\widetilde{Y}_{i,k} = \widetilde{\mu}_k(w)$ where $\widetilde{\mu}_k(\cdot)$ is given in Equation (1.2.10). The neighborhood $B_{i,k}$ of (i,k) is such that, for (j,k') not in $B_{i,k}$, the letters X_h's defining $\widetilde{Y}_{i,k}$ and those defining $\widetilde{Y}_{j,k}$ are separated by at least ℓ positions. Since $\widetilde{Y}_{i,k}$ can be described by at most $X_{i-\ell+1},\ldots,X_{i+(k+1)(\ell-1)}$, we consider

$$B_{i,k} := \{(j,k') \in I : -(k'+3)(\ell-1) \leq j-i \leq (k+3)(\ell-1)\}.$$

We bound successively the quantities given in (1.8.1), (1.8.2) and (1.8.3). By definition

$$\begin{aligned} b_1 &:= \sum_{(i,k)\in I}\ \sum_{(j,k')\in B_{i,k}} \mathbb{E}\widetilde{Y}_{i,k}\mathbb{E}\widetilde{Y}_{j,k'} \\ &\leq \sum_{i=1}^{n-\ell+1}\sum_{k\geq 1}\sum_{k'\geq 1}\sum_{j=i-(k'+3)(\ell-1)}^{i+(k+3)(\ell-1)} \widetilde{\mu}_k(w)\widetilde{\mu}_{k'}(w) \\ &\leq (n-\ell+1)\sum_{k\geq 1}\sum_{k'\geq 1}\big((k+k'+6)(\ell-1)+1\big)\widetilde{\mu}_k(w)\widetilde{\mu}_{k'}(w). \end{aligned}$$

From (1.2.7) and (1.2.10), we use that

$$\sum_{k\geq 1}\widetilde{\mu}_k(w) = \widetilde{\mu}(w)\,, \tag{1.4.12}$$

$$\sum_{k\geq 1}k\widetilde{\mu}_k(w) = \mu(w)\,, \tag{1.4.13}$$

to obtain

$$b_1 \leq (n-\ell+1)\left(2(\ell-1)\widetilde{\mu}(w)\mu(w) + (6\ell-5)\widetilde{\mu}(w)^2\right).$$

The b_2 term involves products such as $\widetilde{Y}_{i,k}\widetilde{Y}_{j,k'}$ with $(j,k') \in B_{i,k}$. Since a k-clump of w at position i cannot overlap a k'-clump of w, many of these products are zero. To identify them, we need to describe in more detail the compound words $c \in \mathcal{C}_k(w)$ and $c' \in \mathcal{C}_{k'}(w)$ that may occur at positions i and j. For this purpose, we introduce the set of words of length $\ell - 1$ that can follow a clump of w:

$$\mathcal{D}(w) = \{d = d_1 \cdots d_{\ell-1} : \forall p \in \mathcal{P}(w), d_1 \cdots d_p \neq w_{\ell-p+1} \cdots w_\ell\}.$$

Therefore, we can write

$$\widetilde{Y}_{i,k}(w) = \sum_{g\in\mathcal{G}(w), c\in\mathcal{C}_k(w), d\in\mathcal{D}(w)} Y_{i-\ell+1}(gCd). \tag{1.4.14}$$

For convenience, we write $\sum_{gcd}$ for the sum over $g \in \mathcal{G}(w)$, $c \in \mathcal{C}_k(w)$, $d \in \mathcal{D}(w)$, and, similarly, $\sum_{g'c'd'}$ for the sum over $g' \in \mathcal{G}(w)$, $c' \in \mathcal{C}_{k'}(w)$ and $d' \in \mathcal{D}(w)$. This gives

$$\begin{aligned} b_2 &:= \sum_{(i,k)\in I} \sum_{(j,k')\in I\setminus\{(i,k)\}} \mathbb{E}\widetilde{Y}_{i,k}\widetilde{Y}_{j,k'} \\ &= \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \sum_{k'\geq 1} \sum_{gcd} \sum_{g'c'd'} \sum_{j=i-(k'+3)(\ell-1)}^{i+(k+3)(\ell-1)} \mathbb{E}Y_{i-\ell+1}(gcd)Y_{j-\ell+1}(g'c'd'). \end{aligned}$$

For $i - |c'| < j < i + |c|$, we have that $Y_{i-\ell+1}(gcd)Y_{j-\ell+1}(g'c'd') = 0$ because clumps do not overlap. We distinguish two cases:

(1) $g'c'd'$ at position $j - \ell + 1$ overlaps gcd at position $i - \ell + 1$ (this is only possible over at most $2(\ell - 1)$ letters); that is, for

$$j \in \{i - |c'| - 2\ell + 3, \ldots, i - |c'|\} \cup \{i + |c|, \ldots, i + |c| + 2\ell - 3\}\,;$$

let b_{21} denote the associated term.

(2) $g'c'd'$ at position $j - \ell + 1$ does not overlap gcd at position $i - \ell + 1$; that is, for

$$j \in \{i-(k'+3)(\ell-1), \ldots, i-|c'|-2\ell+2\} \cup \{i+|c|+2\ell-2, \ldots, i+(k+3)(\ell-1)\}\,;$$

let b_{22} denote the associated term.

By symmetry, we have

$$b_{21} \leq 2 \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \sum_{k'\geq 1} \sum_{gcd} \sum_{g'c'd'} \sum_{j=i+|c|}^{i+|c|+2\ell-3} \mathbb{E}Y_{i-\ell+1}(gCd)Y_{j-\ell+1}(g'C'd')\,.$$

Summing over k', g', c' and d' gives

$$b_{21} \leq 2 \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \sum_{gcd} \sum_{j=i+|c|}^{i+|c|+2\ell-3} \mathbb{E}Y_{i-\ell+1}(gcd)\widetilde{Y}_j(w)\,;$$

now, summing over d and using that $\widetilde{Y}_j(w) \leq Y_j(w)$ leads to

$$b_{21} \leq 2 \sum_{i=1}^{n-\ell+1} \sum_{k\geq 1} \sum_{gc} \sum_{j=i+|c|}^{i+|c|+2\ell-3} \mathbb{E} Y_{i-\ell+1}(gc) Y_j(w) \,.$$

An occurrence of gc at position $i-\ell+1$ does not overlap an occurrence of w at position $j \geq i+|c|$; thus it follows that

$$\mathbb{E} Y_{i-\ell+1}(gc) Y_j(w) = \mu(gc) \Pi^{j-i-|c|+1}(w_\ell, w_1) \frac{\mu(w)}{\mu(w_1)} \,,$$

and

$$b_{21} \leq 2(n-\ell+1) \frac{\mu(w)}{\mu(w_1)} \sum_{s=1}^{2\ell-2} \Pi^s(w_\ell, w_1) \sum_{k\geq 1} \sum_{gc} \mu(gc).$$

Finally, note that

$$\sum_{k\geq 1} \sum_{gc} \mu(gc) = \sum_{k\geq 1} \sum_{k^*\geq k} \widetilde{\mu}_{k^*}(w) = \sum_{k^*\geq 1} k^* \widetilde{\mu}_{k^*}(w) = \mu(w) \,,$$

which leads to

$$b_{21} \leq 2(n-\ell+1) \frac{\mu^2(w)}{\mu(w_1)} \sum_{s=1}^{2\ell-2} \Pi^s(w_\ell, w_1) = \mathrm{O}\left(\frac{\log n}{n}\right) .$$

The b_{22} term is easier to bound and we get

$$b_{22} \leq 2(n-\ell+1) \frac{\widetilde{\mu}(w)}{\mu_{\min}} \left((\ell-2)\mu(w) + \widetilde{\mu}(w)\right) = \mathrm{O}\left(\frac{\log n}{n}\right) ,$$

where $\mu_{\min}$ is the smallest value of $\{\mu(a), a \in \mathcal{A}\}$.

Combining these bounds, we have

$$b_2 \leq 2(n-\ell+1) \frac{\mu^2(w)}{\mu(w_1)} \sum_{s=1}^{2\ell-2} \Pi^s(w_\ell, w_1) + 2(n-\ell+1) \frac{\widetilde{\mu}(w)}{\mu_{\min}} \left((\ell-2)\mu(w) + \widetilde{\mu}(w)\right) .$$

Bounding b_3 consists of following the different steps previously detailed for the declumped count and using the decomposition (1.4.14) instead of (1.4.9). Since there is no interest in repeating this technical part, we just give the bound of b_3 and state the theorem:

$$b_3 \leq (n-\ell+1) \widetilde{\mu}(w) \gamma_2(\ell) |\alpha|^\ell$$

with

$$\gamma_2(\ell) = \sum_{x,y\in\mathcal{A}} \mu(x) \max_{a,b\in\mathcal{A}} \left(\frac{1}{\mu(b)} \sum_{(t,t')\neq(1,1)} \left| \frac{\alpha_t^\ell \alpha_{t'}^\ell}{\alpha^\ell} Q_t(x,b) Q_{t'}(a,y) \right| \right.$$
$$\left. + \sum_{t=2}^{|\mathcal{A}|} \left| \frac{\alpha_t^{5\ell-3}}{\alpha^\ell} Q_t(x,y) \right| \right) .$$

THEOREM 1.4.7. *Let $(Z_{i,k})_{(i,k)\in I}$ be independent Poisson variables with expectation $\mathbb{E}Z_{i,k} = \mathbb{E}\widetilde{Y}_{i,k}(w) = \widetilde{\mu}_k(w)$. With the previous notation, we have*

$$\begin{aligned}
&d_{TV}\Big(\mathcal{L}\big((\widetilde{Y}_{i,k}(w))_{(i,k)\in I}\big), \mathcal{L}\big((Z_{i,k})_{(i,k)\in I}\big)\Big)\\
&\le (n-\ell+1)\widetilde{\mu}(w)\Big(2(\ell-1)\mu(w) + (6\ell-5)\widetilde{\mu}(w) + \gamma_2(\ell)|\alpha|^\ell\Big)\\
&\quad +2(n-\ell+1)\left\{\frac{\mu^2(w)}{\mu(w_1)}\sum_{s=1}^{2\ell-2}\Pi^s(w_\ell,w_1) + \frac{\widetilde{\mu}(w)}{\mu_{min}}\big((\ell-2)\mu(w)+\widetilde{\mu}(w)\big)\right\}.
\end{aligned}$$

From the total variation distance properties, we have

$$d_{\mathrm{TV}}\Big(\mathcal{L}\big(\sum_{(i,k)\in I} k\widetilde{Y}_{i,k}\big), \mathcal{L}\big(\sum_{(i,k)\in I} kZ_{i,k}\big)\Big) \le d_{\mathrm{TV}}\Big(\mathcal{L}\big((\widetilde{Y}_{i,k}(w))_{(i,k)\in I}\big), \mathcal{L}\big((Z_{i,k})_{(i,k)\in I}\big)\Big).$$

Since the $Z_{i,k}$'s are independent Poisson variables, $\sum_{(i,k)\in I} kZ_{i,k}$ has the same distribution as $\sum_{k\ge 1} kZ_k$, where the Z_k's are independent Poisson variables with expectation $(n-\ell+1)\widetilde{\mu}_k(w)$. Note that the latter has a compound Poisson distribution with parameters $\big((n-\ell+1)\widetilde{\mu}(w), (\widetilde{\mu}_k(w)/\widetilde{\mu}(w))_k\big)$. Because of the expressions of $\widetilde{\mu}(w)$ and $\widetilde{\mu}_k(w)$ given by (1.2.7) and (1.2.10), this compound Poisson distribution reduces to a Polýa-Aeppli distribution. Using the triangle inequality leads to the following corollary.

COROLLARY 1.4.8. *Let $(Z_k)_{k\ge 1}$ be independent Poisson variables with expectation $\mathbb{E}Z_k = (n-\ell+1)\widetilde{\mu}_k(w)$. Let*

$$CP = CP\Big((n-\ell+1)\mu(w)(1-A(w)), \big((1-A(w))A^{k-1}(w)\big)_{k\ge 1}\Big) \quad (1.4.15)$$

denote the compound Poisson distribution of $\sum_{k\ge 1} kZ_k$. With the previous notation, we have
$d_{TV}(\mathcal{L}(N(w)), CP) \le$

$$\begin{aligned}
&(n-\ell+1)\widetilde{\mu}(w)\Big(2(\ell-1)\mu(w) + (6\ell-5)\widetilde{\mu}(w) + \gamma_2(\ell)|\alpha|^\ell\Big)\\
&+2(n-\ell+1)\left\{\frac{\mu^2(w)}{\mu(w_1)}\sum_{s=1}^{2\ell-2}\Pi^s(w_\ell,w_1) + \frac{\widetilde{\mu}(w)}{\mu_{min}}\big((\ell-2)\mu(w)+\widetilde{\mu}(w)\big)\right\}\\
&+2(\ell-1)(\mu(w)-\widetilde{\mu}(w))\\
&= O\left(\frac{\log n}{n}\right).
\end{aligned}$$

The special case of runs of 1 in a random sequence of letters in the binary alphabet $\{0,1\}$ has been extensively studied: Erdös and Rényi (1970) gave the asymptotic behavior of the longest run in a sequence of Bernoulli trials, and of the length of the longest segment that contains a proportion of 1 greater than a predescribed level α. Their result was refined by Guibas and Odlyzko (1980), Deheuvels et al. (1986), and Gordon et al. (1986). The compound Poisson approximation for counts of runs in case where the sequence letters are independent has been considered by Eichelsbacher and Roos (1999), also employing the Chen-Stein method using results by Barbour and Utev (1998) (the limiting distribution is the same as the one given above, reduced to this special case). Barbour and Xia (1999) obtained a more accurate limiting approximation for the case of runs of length 2; this approximation is based on a perturbation of a Poisson distribution.

Such a bound on the total variation distance between, for instance, the word count distribution and the associated compound Poisson distribution has the great advantage of providing confidence intervals (see Section 1.8.2). Indeed, using notation from Corollary 1.4.8, for all $t \in \mathbb{R}$, we have

$$\left| \mathbb{P}(N(w) \geq t) - \mathbb{P}\left(\sum_{k \geq 1} k Z_k \geq t \right) \right| \leq d_{\mathrm{TV}} \left(\mathcal{L}(N(w)), \mathcal{L}\left(\sum_{k \geq 1} k Z_k \right) \right).$$

Direct compound Poisson approximation Empirically, often a compound Poisson approximation also gives good results when the underlying words are not so rare, indicating that the theoretical bounds are not sharp. Using the direct compound Poisson approximation Theorem 1.8.4, Barbour et al. (2001) obtained improved bounds for $N(w)$. They chose as neighborhoods in Theorem 1.8.4

$$B(i,k) = \{(j,k') : -(k'-2)(\ell-1) - r + 1 \leq j - i \leq (k+2)(\ell-1) + r - 1\},$$

where $r \geq 1$ can be chosen. In Theorem 1.4.5 we had $r = \ell$. Recall (1.2.10), ρ from (1.1.5), Γ from (1.5.5), and CP from (1.4.15). Barbour et al. (2001) obtain the following result.

THEOREM 1.4.9. *If $A(w) \leq \frac{1}{5}$, then*

$$d_{TV}(\mathcal{L}(N(w)), CP) \leq \frac{1 - A(w)}{1 - 5A(w)} \left(\Delta_1 + \sqrt{(n - \ell + 1)\mu(w)\Delta_0} \right) + 2(\ell - 1)\mu(w),$$

where

$$\begin{aligned}
\Delta_0 &= 2\rho^r \left(2 + 3\rho^{3(\ell-1)+r} + 2\rho^r \right), \\
\Delta_1 &= 2\mu(w) \left\{ 3(\ell - 1) + r + 2(\ell - 1)\frac{A(w)}{1 - A(w)} \right. \\
&\quad \left. + \Gamma \left(\frac{2A(w)(\ell - 1 - p_0)}{(1 - A(w))^3} + \frac{2(\ell - 1) + r - 1}{(1 - A(w))^2} \right) \right\},
\end{aligned}$$

and p_0 is the shortest period of w. The value r can be chosen to minimize the estimates.

Barbour et al. (2001) give a more general form of the result, and also an estimate for the Kolmogorov distance. Using the approach by Erhardsson (1999), they also derive a slightly less explicit but asymptotically better bound in terms of stopping times for a Markov chain.

Indeed, in Erhardsson (1997), Erhardsson (1999) and Erhardsson (2000), a different approach based on the direct compound Poisson approximation Theorem 1.8.3 is developed. The idea is to express counts of events as numbers of visits of a certain Markov chain to a rare set, and to use regeneration cycles for suitable couplings. It results in bounds that are formulated in terms of stopping times of Markov chains. Results of this type are less explicit, but they have asymptotic order $O(n^{-1})$ under the typical regime $n\mu(w) = O(1)$, see also Barbour et al. (2001) and Gusto (2000), whereas the bounds in Theorem 1.4.9 and in Corollary 1.4.8 (which is from Schbath (1995a)) are of order $O(n^{-1}\log n)$ under the same regime.

Numerical experiments in Barbour et al. (2001) display that the bound in Theorem 1.4.9 and the bound from the Erhardsson (1997)-approach perform better than the bound in Corollary 1.4.8 for the word `acgacg` in the bacteriophage *Lambda* ($n = 48,502$) under three different transition matrices.

In contrast, Gusto (2000) compared the result from Erhardsson (1999) to the one in Schbath (1995a) and did not observe any marked improvement for all words of length 8 in the bacteriophage *Lambda*.

This may illustrate that, whereas the compound Poisson approximation via a Poisson process approximation works well in the case of rare words, it does not yield the best bounds in the case of not so rare words.

Estimation of the parameters When estimating the parameters, as in Subsection 1.4.4, the total variation distance between the two compound Poisson distributions is bounded by

$$d_{\mathrm{TV}}\left(\mathcal{L}\left(\sum_{k\geq 1} kZ_k\right), \mathcal{L}\left(\sum_{k\geq 1} kZ'_k\right)\right) \leq \sum_{k\geq 1} |n\widehat{\widetilde{\mu}}_k(w) - n\widetilde{\mu}_k(w)|.$$

Using Equation (1.2.10), this quantity tends to zero as $n \to \infty$ when $n\mu(w) = \mathrm{O}(1)$ (see Schbath (1995b)).

Again, for long sequences the error term due to the maximum-likelihood estimation will be small compared to the bound on the compound Poisson approximation error.

Generalization to Mm Let us now assume that the sequence $(X_i)_{i\in\mathbb{Z}}$ is a m-order Markov chain on the alphabet $\mathcal{A}$, with transition probabilities $\pi(a_1 \cdots a_m,$

$a_{m+1})$, $a_1, \cdots, a_{m+1} \in \mathcal{A}$. The basic idea is to rewrite the sequence over the alphabet $\mathcal{A}^m$ using the embedding (1.1.6),

$$\mathbb{X}_i = X_i X_{i+1} \cdots X_{i+m-1},$$

so that the sequence $(\mathbb{X}_i)_{i\in\mathbb{Z}}$ is a first-order Markov chain on $\mathcal{A}^m$ with transition probabilities ($\mathbb{A} = a_1 \cdots a_m \in \mathcal{A}^m$, $\mathbb{B} = b_1 \cdots b_m \in \mathcal{A}^m$)

$$\Pi(\mathbb{A}, \mathbb{B}) = \begin{cases} \pi(a_1 \cdots a_m, b_m) & \text{if } a_2 \cdots a_m = b_1 \cdots b_{m-1} \\ 0 & \text{otherwise.} \end{cases}$$

Denote by $\mathbb{W} = \mathbb{W}_1 \cdots \mathbb{W}_{\ell-m+1}$ the word $w = w_1 \dots w_\ell$ written using the alphabet $\mathcal{A}^m$, so that $\mathbb{W}_j = w_j \dots w_{j+m-1}$. The results presented below are valid for the number $N(\mathbb{W})$ of overlapping occurrences and the number $\widetilde{N}(\mathbb{W})$ of clumps of $\mathbb{W}$ in $\mathbb{X}_1 \cdots \mathbb{X}_{n-m+1}$. Since an occurrence of w at position i in $X_1 \cdots X_n$ corresponds to an occurrence of $\mathbb{W}$ at position $i-m+1$ in $\mathbb{X}_1 \cdots \mathbb{X}_{n-m+1}$, we simply have $N(w) = N(\mathbb{W})$. In contrast, clumps of $\mathbb{W}$ in $\mathbb{X}_1 \cdots \mathbb{X}_{n-m+1}$ are different from clumps of w in $X_1 \cdots X_n$ because $\mathbb{W}$ is less periodic than w, leading to $\widetilde{N}(\mathbb{W}) \neq \widetilde{N}(w)$. Let us take a simple example: $w = \texttt{ata}$ and $m = 2$. Put $\mathbb{A} = \texttt{at} \in \mathcal{A}^2$ and $\mathbb{B} = \texttt{ta} \in \mathcal{A}^2$; we then have $\mathbb{W} = \mathbb{AB}$. The sequence `tatatatat` contains a unique clump of `at` whereas the associated sequence $\mathbb{BABABABA}$ contains 3 clumps of $\mathbb{AB}$. Indeed, $\mathbb{AB}$ has no period and `ata` has one period. In fact, the periods of $\mathbb{W}$ are those periods of w that are strictly less than $\ell - m + 1$. Therefore, the Poisson approximation for the declumped count in a m-order Markov chain does not follow immediately from the case $m = 1$; a rigorous proof would require applying the Chen-Stein theorem with an adapted neighborhood and to bound the new quantities b_1, b_2 and b_3 in Mm, but this has not yet been carried out.

Since $N(w) = N(\mathbb{W})$, Corollary 1.4.8 ensures that $N(w)$ can be approximated by a sum $\sum_{k\geq 1} kZ_k$, where Z_k is a Poisson variable whose expectation is $(n - \ell + 1)$ times the probability that a k-clump of $\mathbb{W}$ starts at a given position in $\mathbb{X}_1 \cdots \mathbb{X}_{n-m+1}$. From Equation (1.2.10), we obtain

$$\mathbb{E}Z_k = (n - \ell + 1)(1 - A'(w))^2 A'(w)^{k-1} \mu(w)$$

with

$$A'(w) = \sum_{p \in \mathcal{P}'(w) \cup \{1, \dots, \ell - m\}} \frac{\mu(w^{(p)} w)}{\mu(w)}.$$

An important consequence is that, in Mm, the compound Poisson approximation for words that cannot overlap on more than $m - 1$ letters becomes a single Poisson approximation.

1.4.6. Large deviation approximations

For long sequences, the probability that a given word occurs more than a certain number of times can be approximated using a Gaussian or a compound Poisson distribution (Sections 1.4.3 and 1.4.5). The aim of this section is to show that large deviation techniques can also be used to approximate the probability that a given word frequency deviates from its expected value by more than a certain amount. Let $w = w_1 \cdots w_\ell$ be a word of length ℓ; recall that $\mu(w)$ denotes the probability that w occurs at a given position in $X_1 \cdots X_n$. We aim to provide good approximations for $\mathbb{P}(\frac{1}{n-\ell+1}N(w) \geq \mu(w)+b)$ and $\mathbb{P}(\frac{1}{n-\ell+1}N(w) \leq \mu(w)-b)$ with $0 < b < 1$.

We assume that $X_1 \cdots X_n$ is a stationary first-order Markov chain on a finite alphabet $\mathcal{A}$ with transition probabilities $\pi(a,b) > 0$, $a, b \in \mathcal{A}$. (Generalization to Mm follows the same setup as in Section 1.4.5, using (1.1.6).) To use Theorem 1.8.6 for $\frac{1}{n-\ell+1}N(w)$, we need to consider the irreducible Markov chain $\mathbb{X}_1$, $\ldots$, $\mathbb{X}_{n-\ell+1}$ on $\mathcal{A}^\ell$ where $\mathbb{X}_i = X_i \cdots X_{i+\ell-1}$, with transition matrix $\mathbb{\Pi} = (\Pi(u,v))_{u,v\in\mathcal{A}^\ell}$ such that

$$\Pi(u_1 \cdots u_\ell, v_1 \cdots v_\ell) = \begin{cases} \pi(u_\ell, v_\ell) & \text{if } u_{j+1} = v_j,\ j = 1 \cdots \ell-1, \\ 0 & \text{otherwise.} \end{cases}$$

The count $N(w)$ can then be written as

$$\begin{aligned} N(w) &= \sum_{i=1}^{n-\ell+1} \mathbb{I}\{X_i \cdots X_{i+\ell-1} = w_1 \cdots w_\ell\} \\ &= \sum_{i=1}^{n-\ell+1} \mathbb{I}\{\mathbb{X}_i = w\} \end{aligned}$$

Let I be the function

$$I(x) = \sup_{\theta\in\mathbb{R}} (\theta x - \log \lambda(\theta)),$$

$x \in \mathbb{R}$, where $\lambda(\theta)$ is the largest eigenvalue of the matrix $\mathbb{\Pi}_\theta = (\Pi_\theta(u,v))_{u,v\in\mathcal{A}^\ell}$ defined by

$$\Pi_\theta(u,v) = \begin{cases} \mathrm{e}^\theta \Pi(u,v) & \text{if } v = w, \\ \Pi(u,v) & \text{otherwise.} \end{cases}$$

Let $0 < b < 1$; applying Theorem 1.8.6 with the function $f(u) = \mathbb{I}\{u = w\}$ to the closed subset $[\mu(w)+b, +\infty]$ and the open subset $(\mu(w)+b, +\infty)$, we obtain

$$\lim_{n\to+\infty} \frac{1}{n-\ell+1} \log \mathbb{P}\left(\frac{1}{n-\ell+1}N(w) \geq \mu(w)+b\right) = -I(\mu(w)+b)\,;$$

indeed, the rate function I is convex and minimal at $\mathbb{E}f(\mathbb{X}_i) = \mu(w)$. Similarly we have

$$\lim_{n\to+\infty} \frac{1}{n-\ell+1} \log \mathbb{P}\left(\frac{1}{n-\ell+1}N(w) \leq \mu(w)-b\right) = -I(\mu(w)-b).$$

Denoting the observed count of w in the biological sequence by $N^{\text{obs}}(w)$, as a consequence we have for large n:

if $N^{\text{obs}}(w) > (n-\ell+1)\mu(w)$ and $b := \frac{N^{\text{obs}}(w)}{n-\ell+1} - \mu(w)$, then

$$\mathbb{P}(N(w) \geq N^{\text{obs}}(w)) \simeq \exp\left(-(n-\ell+1)I\left(\frac{N^{\text{obs}}(w)}{n-\ell+1}\right)\right),$$

if $N^{\text{obs}}(w) < (n-\ell+1)\mu(w)$ and $b := \mu(w) - \frac{N^{\text{obs}}(w)}{n-\ell+1}$, then

$$\mathbb{P}(N(w) \leq N^{\text{obs}}(w)) \simeq \exp\left(-(n-\ell+1)I\left(\frac{N^{\text{obs}}(w)}{n-\ell+1}\right)\right).$$

Note that this approximation has been obtained assuming the transition probabilities $\pi(a,b)$, $a,b \in \mathcal{A}$ are known. Moreover, since $\lambda(\theta)$ is an eigenvalue of a $|\mathcal{A}|^\ell \times |\mathcal{A}|^\ell$ matrix, the word length ℓ is a limiting factor for the numerical calculation, even if $|\mathcal{A}| = 4$.

Nuel (2001) obtained a more practical approximation using a large deviation principle for the empirical distribution of the ℓ-letter words. This empirical distribution is defined as the random measure $L_{n,\ell}$ on $\mathcal{A}^\ell$ such that, for $w \in \mathcal{A}^\ell$,

$$L_{n,\ell}(w) = \frac{1}{n-\ell+1}\sum_{i=1}^{n-\ell+1} Y_i(w),$$

so that $L_{n,\ell}(w) = N(w)$. The large deviation rate function is easier to implement than the above rate function I. However, its definition and its mathematical treatment are rather involved.

1.5. Renewal count distribution

As a particular case of non-overlapping occurrence counts, in this section we count renewals of a word $w = w_1 w_2 \dots w_\ell$ in a random sequence $X_1 \cdots X_n$ as defined in Section 1.2. We then consider the renewal count $R_n(w) = \sum_{i=1}^{n-\ell+1} \mathbb{I}_i(w)$, where $\mathbb{I}_i(w)$ is the random indicator that a renewal of w starts at position i in $X_1 \cdots X_n$ (see (1.2.11)). For a classical introduction see Chapter 13 in Feller (1950).

Exact results for the distribution of R_n can be found in Régnier (2000); in particular, a combinatorial approach and language decompositions are used to derive the moment generating function of the renewal count. Because those tools are very different from the ones used in this chapter, we only present asymptotic results. First we derive the expected renewal count.

Expected renewal count If the random indicators $\mathbb{I}_i(w)$ had the same expectation, say $\mu_R(w)$, then $\mathbb{E}R_n(w) = (n-\ell+1)\mu_R(w)$. This is the commonly used expectation (see Breen et al. (1985) or Tanushev (1996) for instance), but it ignores the end effect. For $i > \ell$, the $\mathbb{I}_i(w)$'s are effectively identically distributed by stationarity of the Markov process, but it is not the case for $1 \leq i \leq \ell$.

We start with the calculation of $\mu_R(w)$. Recall that $\mathcal{P}(w)$ is the set of periods of w and that $w^{(p)} = w_1 w_2 \cdots w_p$ denotes the word composed of the first p letters of w. When the Markov process is in stationarity, we have from renewal theory that

$$\mu_R(w) = \frac{\mu(w)}{\mathcal{Q}(1)} \tag{1.5.1}$$

with $\mathcal{Q}$ given in (1.2.2). To understand this formula, note that we can decompose the event {there is an occurrence of w starting at position i}, $i > \ell$, as the disjoint union of {there is a renewal of w starting at position i} and {there is a renewal of w starting at position j directly followed by the letters $w_{\ell-i+j+1} \cdots w_\ell$ and $j - i$ is a period of w}, for $j \in \{i - \ell + 1, \ldots, i - 1\}$. This can be written as follows

$$\begin{aligned} Y_i(w) &= \sum_{j=i-\ell+1}^{i} \mathbb{I}_j(w) Y_{j+\ell}(w_{\ell-i+j+1} \cdots w_\ell) \mathbb{I}\{i - j \in \mathcal{P}(w) \cup \{0\}\} \\ &= \sum_{p \in \mathcal{P}(w) \cup \{0\}} \mathbb{I}_{i-p}(w) Y_{i+\ell-p}(w_{\ell-p+1} \cdots w_\ell). \end{aligned}$$

Taking expectations on both sides thus gives

$$\mu(w) = \sum_{p \in \mathcal{P}(w) \cup \{0\}} \mu_R(w) \mu(w_{\ell-p} \cdots w_\ell) \frac{1}{\mu(w_{\ell-p})}.$$

Hence

$$\mu_R(w) = \frac{\mu(w)}{1 + \sum_{p \in \mathcal{P}(w)} \pi(w_{\ell-p}, w_{\ell-p+1}) \cdots \pi(w_{\ell-1}, w_\ell)},$$

which gives the result (1.5.1).

As previously noted, the first variables $\mathbb{I}_1(w), \ldots, \mathbb{I}_\ell(w)$ are not identically distributed because of boundary effects. For the asymptotic results we are interested in in this section, this end effect may be ignored.

1.5.1. Gaussian approximation

When the letters $X_1, \ldots, X_n$ are independent and identically distributed, the asymptotic distribution of the renewal count was studied by Breen et al. (1985); Tanushev (1996) proved a Central Limit Theorem in the Markovian case. For the i.i.d. case see also Tanushev and Arratia (1997). The main technique being generating functions, no bound on the rate of convergence is obtained. Here we present the result from Tanushev (1996).

Note that, once the asymptotic variance is established, the normal approximation follows from the Markov Renewal Central Limit Theorem. Calculating the asymptotic variance is a little more involved than calculating the mean, relying much on the autocorrelation polynomial. To this purpose, similarly to

Tanushev (1996) we define $\mathbf{1}$ as the $|\mathcal{A}| \times |\mathcal{A}|$ matrix where all the entries equal 1. With Π denoting the Markovian transition matrix, put

$$Z = \sum_{k=1}^{\infty} (\Pi - \underline{\mu}\mathbf{1})^k. \tag{1.5.2}$$

Put

$$\sigma^2 = \mu_R^2(w) \left((1 - 2\ell) + 2\frac{\mathcal{Q}'(1)}{\mathcal{Q}(1)} + \frac{2Z(w_\ell, w_1)}{\mu(w_1)} \right).$$

Then Tanushev (1996) proved the following theorem, using generating functions. (It is a special case of his theorem, which in fact proves a multivariate approximation.)

THEOREM 1.5.1. *We have that, as $n \to \infty$,*

$$\frac{R_n(w) - n\mu_R(w)}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2).$$

The theorem is much easier to prove in the i.i.d. case, which is carried out in Waterman (1995). Note that we do not have a corresponding result when mean and standard deviation are estimated.

1.5.2. Poisson approximation

Similarly as with the declumped count, we can also derive a Poisson approximation for the renewal count under the rare word condition $n\mu(w) = \mathrm{O}(1)$. Indeed this is very simple. Recall (1.2.3)

$$Y_i(w) := 1\!\!1\{w \text{ starts at position } i \text{ in } \underline{X}\}.$$

We can write, for $i > \ell$,

$$\begin{aligned}
\mathbb{I}_i(w) &= Y_i(w) \prod_{j=i-\ell+1}^{i-1} (1 - \mathbb{I}_j(w)) \\
&= Y_i(w) \prod_{j=i-\ell+1}^{i-1} (1 - Y_j(w)) \\
&\quad + Y_i(w) \left(\prod_{j=i-\ell+1}^{i-1} (1 - \mathbb{I}_j(w)) - \prod_{j=i-\ell+1}^{i-1} (1 - Y_j(w)) \right) \\
&= \widetilde{Y}_i(w) + Y_i(w) \left(\prod_{j=i-\ell+1}^{i-1} (1 - \mathbb{I}_j(w)) - \prod_{j=i-\ell+1}^{i-1} (1 - Y_j(w)) \right) \qquad (1.5.3)
\end{aligned}$$

whereas $\mathbb{I}_i(w) = Y_i(w) \prod_{j=1}^{i-1} (1 - Y_j(w))$ if $1 \le i \le \ell$. Note that a renewal occurrence in the first ℓ positions is a clump occurrence observed in the finite sequence, and conversely. Thus we have

$$\begin{aligned} R_n(w) &= \sum_{i=1}^{n-\ell+1} \mathbb{I}_i(w) \\ &= \widetilde{N}(w) + \sum_{i=\ell+1}^{n-\ell+1} Y_i(w) \left(\prod_{j=i-\ell+1}^{i-1} (1-\mathbb{I}_j(w)) - \prod_{j=i-\ell+1}^{i-1} (1-Y_j(w)) \right). \end{aligned}$$

We have already derived a Poisson approximation for the number of clumps $\widetilde{N}(w)$ (see Section 1.4.5). Let us consider the difference

$$R_n(w) - \widetilde{N}(w) = \sum_{i=\ell+1}^{n-\ell+1} Y_i(w) \left(\prod_{j=i-\ell+1}^{i-1} (1-\mathbb{I}_j(w)) - \prod_{j=i-\ell+1}^{i-1} (1-Y_j(w)) \right).$$

For a summand to be nonzero, firstly we need that $Y_i(w) = 1$. Note that a renewal always implies an occurrence, so that

$$\prod_{j=i-\ell+1}^{i-1} (1-\mathbb{I}_j(w)) \geq \prod_{j=i-\ell+1}^{i-1} (1-Y_j(w)).$$

The product being always 0 or 1, the two products are different if and only if $\prod_{j=i-\ell+1}^{i-1}(1-\mathbb{I}_j(w)) = 1$ and $\prod_{j=i-\ell+1}^{i-1}(1-Y_j(w)) = 0$. This implies that there is no renewal between the positions $i-\ell+1$ and $i-1$, but that there must be an occurrence not only at position i but also at some position j between $i-\ell+1$ and $i-1$. This occurrence again cannot be a renewal, so that it must be part of a larger clump; repeating this argument we see that the occurrence at i must be part of a clump that started before position $i-\ell+1$. This implies that there had to be an occurrence of w somewhere between $i-2\ell+2$ and $i-\ell$, and this occurrence is in the same clump as the occurrence at i. Thus

$$\begin{aligned} \mathbb{P}(R_n(w) \neq \widetilde{N}(w)) &\leq \sum_{i=\ell+1}^{n-\ell+1} \sum_{j=i-2\ell+2}^{i-\ell} \mathbb{E} Y_i(w) Y_j(w) \\ &\leq (n-2\ell+1)(\ell-1)\mu(w)^2 \frac{1}{\mu(w_1)}. \qquad (1.5.4) \end{aligned}$$

This quantity will be small under the asymptotic framework $n\mu(w) = \mathrm{O}(1)$. Thus we may use the Poisson bound for the number of clumps derived above, and just add an error term of order $\log n/n$. Indeed this has been the idea behind the proof of Geske et al. (1995), although Geske et al. (1995) prove the result only for words having at most one principal period. Related results have been obtained by Chryssaphinou and Papastavridis (1988b).

Chryssaphinou et al. (2001) proceed differently and obtain another type of bound. Put

$$\Gamma = \Gamma(w) = \sup_{t \geq 1} \frac{\pi^{(t)}(w_\ell, w_1)}{\mu(w_1)}. \qquad (1.5.5)$$

Recall ρ given in (1.1.5), and $\mathbb{E}R_n(w)$ is given in (1.5.1). Chryssaphinou et al. (2001) prove the following theorem.

THEOREM 1.5.2. *We have that*

$$d_{TV}\left(\mathcal{L}(R_n(w)), \mathcal{P}o(\mathbb{E}R_n(w))\right) \leq \left(1 - e^{-\mathbb{E}\widetilde{N}(w)}\right) D_1 + \min\left\{1, \sqrt{\frac{2}{e\mathbb{E}\widetilde{N}(w)}}\right\} D_2 + D_3,$$

where

$$\begin{aligned} D_1 &= (2\ell - 5)(\widetilde{\mu}(w) + \Gamma\mu(w)) - \Gamma(2\ell - 1)\mu(w), \\ D_2 &= 2\mathbb{E}R_n(w)\rho^\ell\left(2 + 2\rho^\ell + \rho^{3\ell-2}\right), \\ D_3 &= \left(1 + \min\left\{1, (\mathbb{E}R_n(w))^{-\frac{1}{2}}\right\}\right)(\mathbb{E}R_n(w) - \mathbb{E}\widetilde{N}(w)). \end{aligned}$$

It is also of interest to consider the case that $n \to \infty$, for a sequence of words $w^{(n)}$ of length $\ell^{(n)}$, where $\ell^{(n)}$ may grow with n. Chryssaphinou et al. (2001) show that under the conditions

(i) $\lim_{n\to\infty} \mathbb{E}R_n(w^{(n)}) = \lambda < \infty$

(ii) $\lim_{n\to\infty} \frac{\ell^{(n)}}{n} = 0$,

the bound in Theorem 1.5.2 is of order $O\left(\frac{\ell^{(n)}}{n}\right)$, which converges to zero for $n \to \infty$. Thus $R_n(w^{(n)})$ converges in distribution to a Poisson variable with mean λ.

1.6. Occurrences and counts of multiple patterns

In biological sequence analysis often the distribution of the joint occurrences of multiple patterns rather than that of single words is of relevance, for example when characterizing protein families via short motifs, or when assessing the statistical significance of the count of degenerated words such as `a(c` or `g)g(a` or `t)`, describing the family of words $\{$`acga`, `agga`, `acgt`, `aggt`$\}$.

The multivariate generating function of the counts of multiple words can be found in Régnier (2000) and can be derived from Robin and Daudin (2001). The methods used are extensions of the ones presented in Subsection 1.4.1. Since the exact distribution of the counts of multiple words is not easily calculated in practice, we will focus in this section on the asymptotic point of view.

Indeed, asymptotic results, similar to the above approximations, are available for the distribution of joint occurrences and joint counts of multiple patterns and we will present them in this section. As we will see, the main new feature one has to consider are the possible overlaps between different words from the target family.

Consider the family of q words $\{w^1, \ldots, w^q\}$, where $w^r = w_1^r w_2^r \cdots w_{\ell_r}^r$. For two words $w^1 = w_1^1 w_2^1 \cdots w_{\ell_1}^1$ and $w^2 = w_1^2 w_2^2 \cdots w_{\ell_2}^2$ on $\mathcal{A}$, we describe the possible overlaps between w^1 and w^2 by defining

$$\mathcal{P}(w^1, w^2) := \{p \in \{1, \ldots, \ell_1 - 1\} : w_i^2 = w_{i+p}^1, \forall i = 1, \ldots, (\ell_1 - p) \wedge \ell_2\}.$$

Thus $\mathcal{P}(w^1, w^2) \neq \emptyset$ means that an occurrence of w^2 can overlap an occurrence of w^1 from the right, and $\mathcal{P}(w^2, w^1) \neq \emptyset$ means that w^2 can overlap w^1 from the left. Note the lack of symmetry; for example, if $w^1 =$ `aaagaagaa` and $w^2 =$ `aagaatca`, we have $\mathcal{P}(w^1, w^2) = \{4, 7, 8\}$ and $\mathcal{P}(w^2, w^1) = \{7\}$. To avoid trivialities, we make the following assumption.

(A1) $\forall r \neq r'$, w^r is not a substring of $w^{r'}$.

Thus $\{w^1, \ldots, w^q\}$ is a *reduced* set of words. Again we model the sequence $\{X_i\}_{i \in \mathbb{Z}}$ as a stationary ergodic Markov chain.

We introduce the notation

$$\begin{aligned} \ell &= \max_{1 \leq r \leq q} \ell_r \\ \ell_{\min} &= \min_{1 \leq r \leq q} \ell_r. \end{aligned} \tag{1.6.1}$$

1.6.1. Gaussian approximation for the joint distribution of multiple word counts

We assume the general model Mm, $m \leq \ell_{\min} - 2$. We will show the asymptotic normality of the vector $n^{-1/2}(N(w^r) - \widehat{N}_m(w^r))_{r=1,\ldots,q}$:

$$\frac{1}{\sqrt{n}}\big(N(w^r) - \widehat{N}_m(w^r)\big)_{r=1,\ldots,q} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_m).$$

To prove this result, we use a multivariate martingale central limit theorem. The estimated count $\widehat{N}_m(w^r)$ is given by (1.4.2). The novelty consists here of deriving the asymptotic covariance matrix $\Sigma_m = (\Sigma_m(w^r, w^{r'}))_{r,r'=1,\ldots,q}$.

Suppose all the words w^r have the same length ℓ and $m = \ell - 2$ (the maximal model) then the martingale technique (see Section 1.4.3) leads to

$$\begin{aligned} \Sigma_{\ell-2}(w^r, w^{r'}) = \mu(w^r)\mu(w^{r'}) \Bigg(& \frac{\mathbb{I}\{w^r = w^{r'}\}}{\mu(w^r)} - \frac{\mathbb{I}\{(w^r)^- = (w^{r'})^-\}}{\mu((w^r)^-)} \\ & - \frac{\mathbb{I}\{{}^-(w^r) = {}^-(w^{r'})\}}{\mu({}^-(w^r))} + \frac{\mathbb{I}\{{}^-(w^r)^- = {}^-(w^{r'})^-\}}{\mu({}^-(w^r)^-)} \Bigg). \end{aligned}$$

Note that when $r = r'$, this formula reduces to the asymptotic variance $\sigma_{\ell-2}^2(w^r)$ of Section 1.4.3.

More generally, for $r \neq r'$, the conditional approach (see Section 1.4.3) leads to

$$\Sigma_m(w^r, w^{r'}) = \sum_{\substack{p \in \mathcal{P}(w^r, w^{r'}) \\ p \leq \ell_r - m - 1}} \mu\Big((w^r)^{(p)} w^{r'}\Big) + \sum_{\substack{p \in \mathcal{P}(w^{r'}, w^r) \\ p \leq \ell_{r'} - m - 1}} \mu\Big((w^{r'})^{(p)} w^r\Big)$$

$$
\begin{aligned}
&+\mu(w^r)\mu(w^{r'})\left(\sum_{a_1,\ldots,a_m}\frac{n(a_1\cdots a_m\bullet)n'(a_1\cdots a_m\bullet)}{\mu(a_1\cdots a_m)}\right.\\
&-\sum_{a_1,\ldots,a_{m+1}}\frac{n(a_1\cdots a_{m+1})n'(a_1\cdots a_{m+1})}{\mu(a_1\cdots a_{m+1})}-\frac{n(w_1^{r'}\cdots w_m^{r'}\bullet)}{\mu(w_1^{r'}\cdots w_m^{r'})}\\
&\left.+\frac{\mathbb{I}\{w_1^r\cdots w_m^r=w_1^{r'}\cdots w_m^{r'}\}-n'(w_1^r\cdots w_m^r\bullet)}{\mu(w_1^r\cdots w_m^r)}\right),
\end{aligned}
$$

where $n(\cdot)$ denotes the number of occurrences inside w^r and $n'(\cdot)$ denotes the number of occurrences inside $w^{r'}$. (When $r = r'$, the formula reduces to Equation (1.4.7).)

Note that, if one wants to study the total number of occurrences of a word family $\{w^r, r = 1, \ldots, q\}$, we have

$$
\frac{1}{\sqrt{n}}\left(\sum_{r=1}^{q}N(w^r)-\sum_{r=1}^{q}\widehat{N}_m(w^r)\right)\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,\sum_{r,r'}\Sigma_m(w^r,w^{r'})\right).
$$

Error bound for the normal approximation Similarly to Theorem 1.4.4, Huang (2002) gives a bound on the approximation when the parameters do not have to be estimated. Let $\mathbf{w} = \{w^1, \ldots, w^q\}$ be the word set and

$$
\mathbf{N}(\mathbf{w}) = (N(w^1), \ldots, N(w^q))
$$

be the vector of word counts. Denote its covariance matrix by

$$
\mathcal{L}_n = \mathcal{L}_n(\mathbf{w}) = \mathrm{Cov}(\mathbf{N}(\mathbf{w})) = \left(\mathrm{Cov}(N(w^i), N(w^j)\right)_{i,j=1,\ldots,q}.
$$

A calculation similar to (1.4.1) shows that, for two different words u and v of length ℓ_u and ℓ_v such that u is not a substring of v and v is not a substring of u:

$$
\begin{aligned}
&\mathrm{Cov}(N(u),N(v)) =\\
&\quad\sum_{p\in\mathcal{P}(u,v)}\mathbb{E}N(u^{(p)}v)+\sum_{p\in\mathcal{P}(v,u)}\mathbb{E}N(v^{(p)}u)-\mathbb{E}N(u)\mathbb{E}N(v)\\
&\quad+\mu_1(u)\mu_1(v)\sum_{d=1}^{n-\ell_u-\ell_v+1}(n-\ell_u-\ell_v+2-d)\left[\frac{\Pi^d(u_{\ell_u},v_1)}{\mu_1(v_1)}+\frac{\Pi^d(v_{\ell_v},u_1)}{\mu_1(u_1)}\right].
\end{aligned}
$$

In a different form this was calculated also in Lundstrom (1990). In particular, $\mathcal{L}_n$ is invertible (see Huang (2002)).

Some more notation is needed. Let $\mathcal{H}$ denote the collection of convex sets in $\mathbb{R}^q$, and let

$$
\beta = \beta(\mathbf{w}) = \max_{1\le r\le q}\mu(w^r).
$$

Recall, from the transition matrix diagonalization, α given in (1.1.1) and Q_t given in (1.1.2). Using Theorem 1.8.1, Huang (2002) derived the following result.

THEOREM 1.6.1. *Assume the Markov model M1. Let* $\mathbf{Z} \sim \mathcal{N}(\mathbb{E}(\mathbf{N}(w), \mathcal{L}_n)$. *There are constants* c *and* C_1, C_2, C_3 *such that, for any set* $\mathbf{w}$ *of* q *words with maximal length* ℓ,

$$\sup_{A \in \mathcal{H}} |\mathbb{P}(\mathbf{N}(\mathbf{w}) \in A) - \mathbb{P}(\mathbf{Z} \in A)| \leq c \min_{\ell \leq s \leq \frac{n}{2}} B_s,$$

where

$$\begin{aligned}
B_s &= 2q^{\frac{3}{2}}(4s-3)\left|\mathcal{L}_n^{-1}\right|^{\frac{1}{2}} \\
&+2q^{\frac{1}{2}}n(2s-1)(4s-3)\left(q^2\sqrt{\left|\mathcal{L}_n^{-1}\right|}\right)^3\left(|\log(q^2\sqrt{\left|\mathcal{L}_n^{-1}\right|})|+\log n\right) \\
&+C_1 n\left|\mathcal{L}_n^{-1}\right|^{\frac{1}{2}}\beta|\alpha|^{s-\ell+1} \\
&+C_2\left(|\log(q^2\sqrt{\left|\mathcal{L}_n^{-1}\right|})|+\log n\right)(2s-1)|\alpha|^{s-\ell+1} \\
&+C_3\left(|\log(q^2\sqrt{\left|\mathcal{L}_n^{-1}\right|})|+\log n\right)(n-2s+1)nq^4\beta^2\left|\mathcal{L}_n^{-1}\right||\alpha|^{s-\ell+1}.
\end{aligned}$$

Here,

$$\begin{aligned}
C_1 &= \max\left\{\sum_{a,b\in\mathcal{A}}\mu(a)C_{1,1}(a,b), \sum_{a\in\mathcal{A}}C_{1,2}(a), \sum_{a\in\mathcal{A}}C_{1,3}(a)\right\} \\
C_2 &= n\left|\mathcal{L}_n^{-1}\right|q^4\beta(2\beta+1)C_1 \\
C_3 &= \max_{a,b\in\mathcal{A}}\left\{\sum_{t\geq 2}\frac{|Q_t(a,b)|}{\mu(a)}\right\}
\end{aligned}$$

and

$$\begin{aligned}
C_{1,1}(a,b) &= \max_{x,y\in\mathcal{A}}\left\{\sum_{t\geq 2 \text{ or } t'\geq 2}\frac{|Q_t(a,x)Q_{t'}(b,y)|}{\mu(x)}\right\} + \sum_{t\geq 2}|Q_t(a,b)| \\
C_{1,2}(a) &= \max_{b\in\mathcal{A}}\left\{\sum_{t\geq 2}|Q_t(b,a)|\right\} \\
C_{1,3}(a) &= \max_{b\in\mathcal{A}}\left\{\sum_{t\geq 2}\frac{|Q_t(a,b)|}{\mu(b)}\right\}.
\end{aligned}$$

The constant c is not easy to describe, see Huang (2002) and references therein. Note that convergence on the class of convex sets is not as strong as convergence in total variation. As in Rinott and Rotar (1996), Huang (2002) considers more general classes of test functions as well, but not as general as to cover total variation. Indeed, approximating discrete counts by a continuous multivariate variable might not be expected to be very good in total variation distance.

1.6.2. Poisson and compound Poisson approximations for the joint distribution of declumped counts and multiple word counts

We assume the model M1 since generalization to Mm follows the single pattern case. To give a bound on the error for a Poisson process approximation for overlapping counts, Reinert and Schbath (1998) define the following quantities for all r and r' in $\{1,\dots,q\}$, and for all $a \in \mathcal{A}$:

$$
\begin{aligned}
\Omega_r &= \sum_{s=1}^{3\ell-\ell_r-2} \Pi^s, \\
\Omega_{r,r'} &= \sum_{s=1}^{\ell_r+\ell_{r'}-2} \Pi^s, \\
M(w^r, w^{r'}) &= \begin{cases} \displaystyle\sum_{p\in\mathcal{P}(w^r,w^{r'})} \frac{1}{\mu((w^{r'})^{(\ell_r-p)})} & \text{if } r \neq r', \\ 0 & \text{if } r = r', \end{cases} \\
T_1(w^r, w^{r'}) &= (2n-\ell_r-\ell_{r'}+2)\mu(w^r)\widetilde{\mu}(w^{r'}) \left(\frac{\Omega_{r'}(w^{r'}_{\ell_{r'}}, w^r_1)}{\mu(w^r_1)} + M(w^{r'}, w^r) \right), \\
T_2(w^r, w^{r'}) &= (n-\ell_r+1)\Big((\ell-1)(\widetilde{\mu}(w^r)\mu(w^{r'}) + \mu(w^r)\widetilde{\mu}(w^{r'})) \\
&\quad + (6\ell-5)\widetilde{\mu}(w^r)\widetilde{\mu}(w^{r'}) \Big), \\
T_3(w^r, w^{r'}) &= (n-\ell_r+1)\mu(w^r)\mu(w^{r'}) \left(\frac{\Omega_{r,r'}(w^r_{\ell_r}, w^{r'}_1)}{\mu(w^{r'}_1)} + \frac{\Omega_{r,r'}(w^{r'}_{\ell_{r'}}, w^r_1)}{\mu(w^r_1)} \right) \\
&\quad + \frac{(n-\ell_r+1)(6\ell-3\ell_r-3\ell_{r'}+2)}{\mu_{\min}} \widetilde{\mu}(w^r)\widetilde{\mu}(w^{r'}) \qquad (1.6.2)\\
&\quad + \frac{(n-\ell_r+1)(\ell-2)}{\mu_{\min}} \Big(\mu(w^r)\widetilde{\mu}(w^{r'}) + \mu(w^{r'})\widetilde{\mu}(w^r) \Big) \\
&\quad + (n-\ell_r+1)\mu(w^r)\mu(w^{r'}) \Big(M(w^r, w^{r'}) + M(w^{r'}, w^r) \Big), \\
\gamma_1(\ell_r, \ell, a) &= \sum_{x,y\in\mathcal{A}} \mu(x) \max_{b\in\mathcal{A}} \Bigg| \frac{1}{\mu(b)} \sum_{(t,t')\neq(1,1)} \frac{\alpha_t^{2\ell-\ell_r}\alpha_{t'}^{2\ell-\ell_r}}{\alpha^\ell} Q_t(x,b)Q_{t'}(a,y) \\
&\quad - \sum_{t=2}^{|\mathcal{A}|} \frac{\alpha_t^{4\ell-2}}{\alpha^\ell} Q_t(x,y) \Bigg|, \\
\gamma_2(\ell_r, \ell) &= \sum_{x,y\in\mathcal{A}} \mu(x) \max_{a,b\in\mathcal{A}} \left(\frac{1}{\mu(b)} \sum_{(t,t')\neq(1,1)} \left| \frac{\alpha_t^{2\ell-\ell_r}\alpha_{t'}^{2\ell-\ell_r}}{\alpha^\ell} Q_t(x,b)Q_{t'}(a,y) \right| \right.
\end{aligned}
$$

$$+\sum_{t=2}^{|\mathcal{A}|}\left|\frac{\alpha_t^{5\ell-3}}{\alpha^\ell}Q_t(x,y)\right|\Bigg).$$

Here we choose as index set $I = \{1, 2, \ldots, q(n+1) - \sum_{s=1}^q \ell_s\}$; it can be written as the disjoint union $I = \bigcup_{r=1}^q I_r$ with

$$I_r = \left\{(r-1)(n+1) - \sum_{s=1}^{r-1}\ell_s + 1, \ldots, r(n+1) - \sum_{s=1}^{r}\ell_s\right\}. \qquad (1.6.3)$$

We define $[i]$ by

$$[i] := i - (r-1)(n+1) + \sum_{s=1}^{r-1}\ell_s \quad \text{with } r = r(i) \text{ such that } i \in I_r. \quad (1.6.4)$$

Joint distribution of declumped counts To apply Theorem 1.8.2, the Bernoulli process $\underline{\widetilde{Y}} = (\widetilde{\mathbb{Y}}_i)_{i\in I}$ and the Poisson process $\underline{Z} = (Z_i)_{i\in I}$ are given by

$$\begin{aligned}\widetilde{\mathbb{Y}}_i &= \widetilde{Y}_{[i]}(w^r),\\ Z_i &\sim \mathrm{Po}(\widetilde{\mu}(w^r)),\end{aligned} \qquad (1.6.5)$$

where r is such that $i \in I_r$. For $i \in I$, we choose the neighborhood $B_i := \{j \in I : |[j]-[i]| \le 3\ell - 3\}$.

Then Reinert and Schbath (1998) prove the following results. Recall the notation (1.6.2), (1.6.5), and (1.6.1).

THEOREM 1.6.2. *Under assumption (A1) we have*

$$\begin{aligned}&d_{TV}\left(\mathcal{L}(\underline{\widetilde{Y}}), \mathcal{L}(\underline{Z})\right)\\ &\le (n-\ell_{min}+1)(6\ell-5)\left(\sum_{r=1}^q \widetilde{\mu}(w^r)\right)^2 + \sum_{1\le r,r'\le q} T_1(w^r, w^{r'})\\ &\quad + |\alpha|^\ell \sum_{r=1}^q \gamma_1(\ell_r, \ell, w^r_{\ell_r})(n-\ell_r+1)\widetilde{\mu}(w^r).\end{aligned}$$

COROLLARY 1.6.3. *Let $(Z_r)_{r=1,\ldots,m}$ be independent Poisson variables with $\mathbb{E}Z_r = (n-\ell_r+1)\widetilde{\mu}(w^r)$. With the previous notation and under assumption (A1), we have*

$$\begin{aligned}&d_{TV}\left(\mathcal{L}\big((\widetilde{N}(w^r))_{r=1,\ldots,q}\big), \mathcal{L}\big((Z_r)_{r=1,\ldots,q}\big)\right)\\ &\le (n-\ell_{min}+1)(6\ell-5)\left(\sum_{r=1}^q \widetilde{\mu}(w^r)\right)^2 + \sum_{1\le r,r'\le q} T_1(w^r, w^{r'})\\ &\quad + |\alpha|^\ell \sum_{r=1}^q \gamma_1(\ell_r, \ell, w^r_{\ell_r})(n-\ell_r+1)\widetilde{\mu}(w^r) + \sum_{r=1}^q (\ell_r - 1)\big(\mu(w^r) - \widetilde{\mu}(w^r)\big).\end{aligned}$$

The proof is a direct application of Theorem 1.8.2, similar as in Section 1.4.

Distribution of multiple word counts In a similar way a compound Poisson approximation for the numbers of occurrences can be obtained. Choose as index set

$$I = \left\{ 1, 2, \ldots, q(n+1) - \sum_{s=1}^{q} \ell_s \right\} \times \{1, 2, \ldots\}.$$

To apply Theorem 1.8.2, the Bernoulli process $\underline{\widetilde{\mathbb{Y}}} = (\widetilde{\mathbb{Y}}_{i,k})_{(i,k)\in I}$ and the Poisson process $\underline{\mathbb{Z}} = (Z_{i,k})_{(i,k)\in I}$ are now defined as

$$\begin{aligned} \widetilde{\mathbb{Y}}_{i,k} &= \widetilde{Y}_{[i],k}(w^r), \\ Z_{i,k} &\sim \mathrm{Po}(\widetilde{\mu}_k(w^r)), \end{aligned}$$

where $r = r(i)$ is such that $i \in I_r$; I_r and $[i]$ are given by (1.6.3) and (1.6.4). For $(i,k) \in I$, the neighborhood is still $B_{i,k} := \{(j,k') \in I : -(k'+3)(\ell-1) \leq [j] - [i] \leq (k+3)(\ell-1)\}$.
We make the following weak assumption on the overlap structure.

(A2) $\forall r \neq r'$, w^r is not a substring of any composed word in $\mathcal{C}_2(w^{r'})$.

THEOREM 1.6.4. *Under assumptions (A1), (A2) and with the notation (1.6.2), we have*

$$\begin{aligned} d_{TV}\left(\mathcal{L}(\underline{\widetilde{\mathbb{Y}}}), \mathcal{L}(\underline{\mathbb{Z}})\right) \leq &\sum_{1\leq r,r'\leq q} T_2(w^r, w^{r'}) + \sum_{1\leq r,r'\leq q} T_3(w^r, w^{r'}) \\ &+ |\alpha|^\ell \sum_{r=1}^{q} \gamma_2(\ell_r, \ell)(n - \ell_r + 1)\widetilde{\mu}(w^r)\,. \end{aligned}$$

The following corollary is easily obtained.

COROLLARY 1.6.5. *Let $(Z_k)_{k\geq 1}$ be independent Poisson variables with expectation $\mathbb{E}Z_k = \sum_{r=1}^{q}(n - \ell_r + 1)\widetilde{\mu}_k(w^r)$; CP denotes the (compound Poisson) distribution of $\sum_{k\geq 1} kZ_k$. Under assumptions (A1), (A2) and with the notation (1.6.2), we have*

$$\begin{aligned} d_{TV}\left(\mathcal{L}\left(\sum_{r=1}^{q} N(w^r)\right), \mathrm{CP}\right) \leq &\sum_{1\leq r,r'\leq q} T_2(w^r, w^{r'}) + \sum_{1\leq r,r'\leq q} T_3(w^r, w^{r'}) \\ &+ |\alpha|^\ell \sum_{r=1}^{q} \gamma_2(\ell_r, \ell)(n - \ell_r + 1)\widetilde{\mu}(w^r) \\ &+ 2\sum_{r=1}^{q} (\ell_r - 1)\big(\widetilde{\mu}(w^r) - \mu(w^r)\big)\,. \end{aligned}$$

Again, empirically the compound Poisson approximation may perform better than the bound suggests, in the case of not so rare words.

Recently, Chen and Xia (2002) obtained a much improved bound for the independent model, in the Wasserstein metric (which is weaker than the total variation metric), for the Poisson approximation of counts of palindromes, assuming the four-letter alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ and that $p_{\mathtt{a}} = p_{\mathtt{t}}, p_{\mathtt{c}} = p_{\mathtt{g}}$.

Expected count of mixed clumps For the family $\mathbf{w} = (w^1, \dots, w^q)$ of words it is also interesting to consider the number of *mixed* clumps of occurrences. Let

$$Y_i^c(\mathbf{w}) = \sum_{r=1}^{q} Y_i(w^r) \prod_{j=i-\ell_r+1}^{i-1} \left\{ 1 - \sum_{r'=1}^{q} Y_j(w^{r'}) \right\},$$

that is, $Y_i^c(\mathbf{w}) = 1$ if there is an occurrence of a word from the family $\mathbf{w}$ at i, and if there is no previous occurrence of any word in $\mathbf{w}$ that overlaps position i. Thus the mixed clumps can be composed of any words from $\mathbf{w}$, whereas for $\widetilde{Y}_i$ the clumps are composed of the same word. Note that, for $q = 1$, $Y_i^c(\mathbf{w}) = \widetilde{Y}_i(w^1)$. Let

$$N^c(\mathbf{w}) = \sum_{i=1}^{n-\ell_{\min}+1} Y_i^c(\mathbf{w}).$$

be the number of mixed clumps in the sequence. To calculate $\mathbb{E}N^c(\mathbf{w})$, introduce the quantities

$$e_{r,s} = \sum_{p \in \mathcal{P}(w^r, w^s)} \frac{\mu\left(w^s_{(\ell_s - \ell_r + p + 1)}\right)}{\mu(w^r_{\ell_r})},$$

where the summands are the probabilities of observing the last $(\ell_s - \ell_r + p)$ letters of w^s successively given that the last letter of w^r has just occured. Chryssaphinou et al. (2001) show that

$$\mathbb{E}N^c = (n - \ell + 1) \sum_{r=1}^{q} y_r, \tag{1.6.6}$$

where $(y_1, \dots, y_q)$ is the solution of the $q \times q$ linear system of equations

$$\sum_{r=1}^{q} y_r e_{r,s} = \mu(w^s), \quad s = 1, \dots, q.$$

1.6.3. Competing renewal counts

Related results to the above for renewal counts are available. As in Tanushev (1996) we consider non-overlapping occurrences in competition with each other.

For example, in the sequence `cgtatattaaaaatattaga`, the set of words `tat`, `tta` and `aa` has renewal occurrences of `tat` at position 3 and 14, of `tta` at position 7, and of `aa` at positions 10 and 12. The occurrences of `tat` at position 5, of `tta` at position 16, and of `aa` at positions 9 and 11 are not counted because they overlap with some already counted words.

Let

$$\mathbb{I}_i^c(w^r) = \mathbb{1}\{\text{a competing renewal of } w^r \text{ starts at position } i \text{ in } X_1 \cdots X_n\},$$

and let

$$R_n^c(w^r) = \sum_{i=1}^{n-\ell_r+1} \mathbb{I}_i^c(w^r)$$

be the number of competing renewals of w^r in the sequence $X_1 X_2 \cdots X_n$.

The mean $\mu_R^c(w^r) = \mathbb{E}R_n^c(w^r)$ has been derived in Tanushev (1996). Some more notation is needed for this. For a matrix A denote its transposed matrix by A^T, and, if A is a square matrix, $Diag(A)$ represents the vector of the diagonal elements of A. Define the probabilities of ending a word for $1 \leq j \leq \ell_r - 1$ as

$$\begin{aligned}\mathbb{P}_r(j) &= \mathbb{P}(\text{ collect final } j \text{ letters of } w^r | \text{ start with correct } \ell_r - j \text{ initial} \\ &\qquad\qquad \text{letters of } w^r) \\ &= \frac{\mu(w^r)}{\mu((w^r)^{(\ell_r - j)})}.\end{aligned}$$

Then, in analogy to (1.2.2), the correlation polynomials are defined as

$$\mathcal{Q}_{r,r'}(z) = 1 + \sum_{p \in \mathcal{P}(w^r, w^{r'})} z^p \mathbb{P}_{r'}(p).$$

Define the $q \times q$ matrix

$$\Delta(z) = (\mathcal{Q}_{r,r'}(z))_{r,r'=1,\ldots,q}$$

and

$$\begin{aligned}\Lambda(z) &= (\Delta^{-1})(z)^T \\ \Lambda &= \Lambda(1).\end{aligned}$$

Moreover put $K_r = \mu(w_1^r)\mathbb{P}_r(\ell_r - 1)$ and define the vector

$$K = (K_1, \ldots, K_q)^T.$$

Then the means $\mu_R^c(w^r), r = 1, \ldots, q$, are given by

$$(\mu_R^c(w^1), \ldots, \mu_R^c(w^q))^T = \Lambda K.$$

The mean of the total number of competing renewals, $\sum_{r=1}^q R_n^c(w^r)$, has recently been presented in a slightly simpler form by Chryssaphinou et al. (2001).

Gaussian approximation for the joint distribution of competing renewal counts A multivariate normal approximation has been obtained by Tanushev (1996). The main problem here is to specify the covariance structure. To state the result, quite a bit of notation is needed. Define

$$\widetilde{K}_r(z) = z^{\ell_r - 1}\mathbb{P}_r(\ell_r - 1)$$

and the vector

$$\widetilde{K}(z) = (\widetilde{K}_1(z), \dots, \widetilde{K}_q(z))^T.$$

Denote by

$$Diag(\widetilde{K}(z))$$

the $q \times q$ diagonal matrix with the components of $\widetilde{K}(z)$ as diagonal elements. Put

$$\begin{aligned} \widetilde{K} &= \widetilde{K}(1) \\ H(z) &= \frac{d}{dz}\Lambda(z) \\ H &= H(1). \end{aligned}$$

Define the vector

$$L = (\ell_1 K_1, \dots, \ell_q K_q),$$

and the matrix

$$\widetilde{Z} = Z_{[\psi]},$$

where Z is defined in (1.5.2), and for a matrix A the matrix $A_{[\psi]}$ is the $q \times q$ matrix whose (r, r') entry is the element of A at the row corresponding to the last letter $w^r_{\ell_r}$ of the word w^r, and at the columns corresponding to the first letter $w^{r'}_1$ of $w^{r'}$. Define the variance-covariance matrix

$$\begin{aligned} C = {} & \frac{1}{2}\big(\Lambda K(\Lambda K - 2HK - 2\Lambda L)^T + (\Lambda K - 2HK - 2\Lambda L)(\Lambda K)^T\big) \\ & + Diag(\Lambda K)\widetilde{Z} Diag(\widetilde{K})\Lambda^T + \Lambda Diag(\widetilde{K})\widetilde{Z}^T Diag(\Lambda K) + Diag(\Lambda K). \end{aligned}$$

Now we have all the ingredients to state the normal approximation obtained by Tanushev (1996).

Theorem 1.6.6. *Under Assumption (A1) we have*

$$\left(\frac{R^c_n(w^r) - n\mu^c_R(w^r)}{\sqrt{n}}\right)_{r=1,\dots,q} \xrightarrow{\mathcal{D}} \mathcal{N}(0, C).$$

In the case of a single pattern, this theorem reduces to Theorem 1.5.1.

Poisson approximation for the renewal count distribution For a Poisson approximation, the problem can be reduced to declumped counts, as in the case of a single word. For a Poisson process approximation (and, following from that, a Poisson approximation for the counts), we want to assess $\mathbb{P}(\mathbb{I}_i^c(w^r) \neq \widetilde{Y}_i(w^r))$. First consider $\mathbb{P}(\mathbb{I}_i^c(w^r) = 1, \widetilde{Y}_i(w^r) = 0)$. Note that, from (1.5.3), for $i > \ell_r$, to have $\mathbb{I}_i^c(w^r) = 1, \widetilde{Y}_i(w^r) = 0$, there must be an occurrence of w^r at position i, and this occurrence cannot be the start of a clump of w^r, so that there must be an overlapping occurrence of w^r at some position $j = i - \ell_r + 1, \ldots, i-1$. Moreover, this occurrence cannot be a competing renewal, so there must be another word $w^{r'}$ overlapping this occurrence. Hence we may bound

$$\begin{aligned}
&\mathbb{P}(\mathbb{I}_i^c(w^r) = 1, \widetilde{Y}_i(w^r) = 0) \\
&\leq \mu^2(w^r) \sum_{p \in \mathcal{P}(w^r)} \frac{1}{\mu((w^r)^{(\ell_r - p)})} \sum_{r'=1}^{q} \mu(w^{r'}) M(w^{r'}, w^r),
\end{aligned}$$

with M given in (1.6.2). For $i \leq \ell_r$ the above bound is still valid (the probability is even smaller since there is not always enough space for these clumps to occur). Secondly, consider $\mathbb{P}(\mathbb{I}_i^c(w^r) = 0, \widetilde{Y}_i(w^r) = 1)$. For $\mathbb{I}_i^c(w^r) = 0, \widetilde{Y}_i(w^r) = 1$ to occur, there must be an occurrence of w^r at position i, overlapped by an occurrence of a different word $w^{r'}$, so that we may bound

$$\mathbb{P}(\mathbb{I}_i^c(w^r) = 0, \widetilde{Y}_i(w^r) = 1) \leq \mu(w^r) \sum_{r'=1}^{q} \mu(w^{r'}) M(w^{r'}, w^r).$$

Again, for $i \leq \ell_r$ the above bound remains valid. Thus we have

$$\begin{aligned}
\mathbb{P}(\underline{\mathbb{I}^c(w^r)} \neq \underline{\widetilde{Y}(w^r)}) \leq{}& (n - \ell_r + 1)\mu(w^r) \sum_{r'=1}^{q} \mu(w^{r'}) M(w^{r'}, w^r) \\
&\left(1 + \mu(w^r) \sum_{p \in \mathcal{P}(w^r)} \frac{1}{\mu((w^r)^{(\ell_r - p)})}\right).
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{P}(\underline{\mathbb{I}}^c \neq \underline{\widetilde{Y}}) \leq{}& \sum_{r=1}^{q} (n - \ell_r + 1)\mu(w^r) \sum_{r'=1}^{q} \mu(w^{r'}) M(w^{r'}, w^r) \\
&\left(1 + \mu(w^r) \sum_{p \in \mathcal{P}(w^r)} \frac{1}{\mu((w^r)^{(\ell_r - p)})}\right).
\end{aligned}$$

Thus we obtain as a corollary of Theorem 1.6.2

Corollary 1.6.7. *Under assumption (A1) and with the notation (1.6.2) and*

(1.6.5), we have

$$d_{TV}(\mathcal{L}(\underline{\mathbb{I}}^c), \mathcal{L}(\underline{Z})) \le (n - \ell_{min} + 1)(6\ell - 5)\left(\sum_{r=1}^{q} \widetilde{\mu}(w^r)\right)^2 + \sum_{1\le r,r'\le q} T_1(w^r, w^{r'})$$
$$+ |\alpha|^\ell \sum_{r=1}^{q} \gamma_1(\ell_r, \ell, w^r_{\ell_r})(n - \ell_r + 1)\widetilde{\mu}(w^r)$$
$$+ \sum_{r=1}^{q} (n - \ell_r + 1)\mu(w^r) \sum_{r'=1}^{q} \mu(w^{r'})M(w^{r'}, w^r)$$
$$\left(1 + \mu(w^r) \sum_{p\in\mathcal{P}(w^r)} \frac{1}{\mu((w^r)^{(\ell_r - p)})}\right).$$

Note that the order of the approximation is the same as in Theorem 1.6.2; the additional error terms are comparable to T_1 and T_2, respectively. A Poisson approximation for the competing renewal counts follows immediately.

Poisson approximation for competing renewal counts Alternatively to the above approach, in Chryssaphinou et al. (2001), a Poisson approximation similar to Theorem 1.5.2 for the number of competing renewals is given. Recall $\mathbb{E}N^c(\mathbf{w})$ from (1.6.6), and Γ from (1.5.5).

THEOREM 1.6.8. *We have that*

$$d_{TV}\left(\mathcal{L}(\sum_{r=1}^{q} R_n^c(w^r)), \mathcal{P}o\left(\sum_{r=1}^{q} \mathbb{E}R_n^c(w^r)\right)\right)$$
$$\le \left(1 - e^{-\mathbb{E}N^c(\mathbf{w})}\right) D_1 + \min\left\{1, \sqrt{\frac{2}{e\mathbb{E}N^c(\mathbf{w})}}\right\} D_2 + D_3,$$

where

$$D_1 = (2\ell - 5)\left(\mathbb{E}Y_i^c(\mathbf{w}) + \Gamma \sum_{r=1}^{q} \mu(w^r)\right) - \Gamma(2\ell_{min} - 1)\sum_{r=1}^{q} \mu(w^r),$$
$$D_2 = 2\mathbb{E}N^c(\mathbf{w})\rho^\ell\left(2 + 2\rho^\ell + \rho^{3\ell-2}\right),$$
$$D_3 = \left(1 + \min\left\{1, (\mathbb{E}N^c(\mathbf{w}))^{-\frac{1}{2}}\right\}\right)\left(\sum_{r=1}^{q} \mathbb{E}R_n^c(w^r) - \mathbb{E}N^c(\mathbf{w})\right).$$

It is again interesting to consider the case that $n \to \infty$, for a sequence of words $\mathbf{w}^{(n)} = (w^{1,n}, \ldots, w^{q,n})$ of maximal length $\ell^{(n)}$, where $\ell^{(n)}$ may grow with n. Chryssaphinou et al. (2001) show that under the conditions

(i) $\lim_{n\to\infty} \sum_{r=1}^{q} \mathbb{E}R_n^c(w^{r,n}) = \lambda < \infty$

(ii) $\lim_{n\to\infty} \frac{\ell^{(n)}}{n} = 0$,

the bound in Theorem 1.5.2 is of order $O\left(\frac{\ell^{(n)}}{n}\right)$, so that $\sum_{r=1}^{q} R_n^c(w^{r,n})$ converges in distribution to a Poisson variable with mean λ.

1.7. Some applications to DNA sequences

1.7.1. Detecting exceptional words in DNA sequences

We call *exceptional* word a word w that appears in an observed sequence with a significantly high or low frequency. This significance is measured under a given probabilistic model by the p-value $\mathbb{P}(N(w) \geq N_{\mathrm{obs}}(w))$ using the distribution of the count $N(w)$. Depending on the sequence length and on the expected count of the word it is often not realistic to use the exact distribution of the count since it is time consuming to calculate In this section, we will first give some elements of comparison between the p-values obtained using the exact distribution (Section 1.4.1) and the ones obtained using the Gaussian approximation (Section 1.4.3) or the compound Poisson approximation (Section 1.4.5) or using the large deviation technics (Section 1.4.6). For convenience, we will manipulate scores from $\mathbb{R}$ of the form $\phi^{-1}(\mathrm{p}-value)$ rather than the p-values, where ϕ is the cumulative distribution function of the standard Gaussian distribution (probit normalization). Exceptionally frequent words would then have high positive scores whereas exceptionally rare words would have high negative scores.

Quality of the approximate p-values For each word of length 3, 6 and 9 of the complete genome of the *Lambda* phage ($\ell = 48\,502$), Robin and Schbath (2001) compared the exact scores under the Bernoulli model M0 with the approximate ones using either the Gaussian approximation or the compound Poisson distribution (the parameters are assumed to be known). The results are presented on Figure 1.1 together with the approximate scores obtained with the large deviation approach of Nuel (2001): the x-axis of each plot is for the exact scores of 3-words (first row), 6-words (second row) and 9-words (last row). The y-axis is for the scores approximated with the Gaussian approximation (first column), the compound Poisson distribution (second column) and the large deviation approach (last column). Due to numerical errors the exact score of 5 words of length 3 have not been calculated successfully. We observe that the accuracy of the Gaussian approximation decreases as the length of the words increases (rare words). The compound Poisson approximation is surprisingly satisfactory even for short (frequent) words. This agrees with the evolution of the total variation distance between the exact distribution of the count and both approximate distributions as shown in Robin and Schbath (2001); when the expected count of the word is close to 100 or greater then the accuracy of the Gaussian approximation is very good. The large deviation approach seems also to provide a good approximation for the exceptional words. However, it cannot manage with words having an estimated expected count too close to the observed one. The p-value is then set to 1/2 in this case and the flatness of the

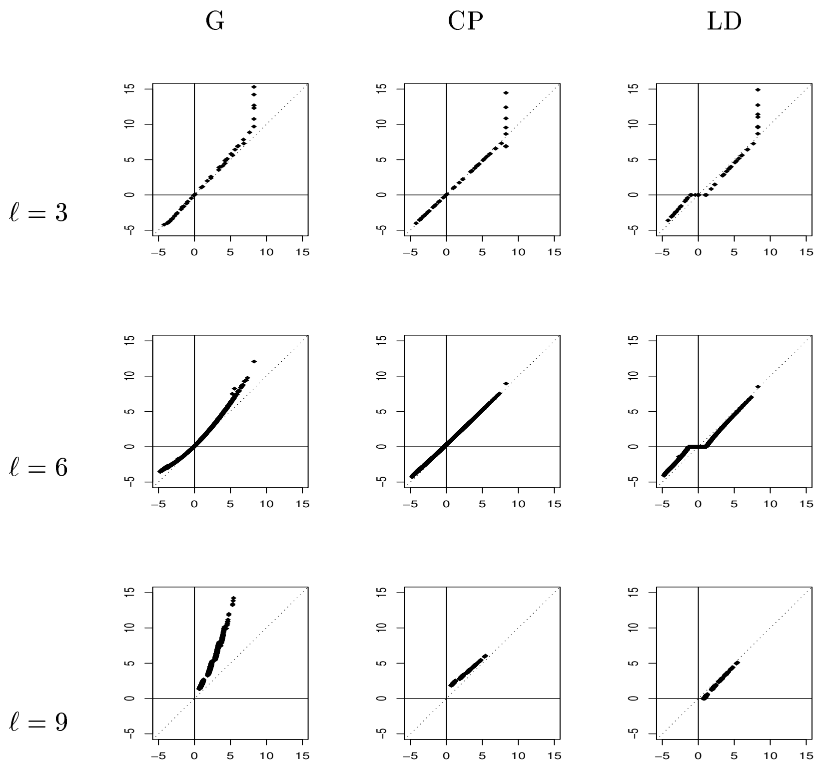

Figure 1.1. Normalized p-values of the counts of all the words of length 3, 6 and 9 in the genome of the *Lambda* phage ($\ell = 48\,502$). Comparison of the Gaussian, the compound Poisson and the large deviation approximations (y-axis) with the exact scores (x-axis).

curves is an artifact. An important feature is that every method to calculate or to approximate the p-values seems to classify the words in the same way; the score ranks are almost the same. Moreover, in this example, the three methods agree on the fact that there are no exceptionally rare words of length 9.

Influence of the model Whatever the word count distribution used to calculate the normalized p-value, the choice of the model, in particular the order m of the Markov chain, is important to interpret the exceptionality of a given word. Using the model Mm means taking into account the 1- to $(m+1)$-letter word composition of the sequence. Therefore, the greater the order m of the model, the closer the random sequences will be to the observed sequence, and fewer unexpected words will be found. As an example, Figure 1.2 shows the discrepancy of the scores for the 8-letter words in the complete genome of *E. coli* ($\ell = 4\,638\,858$) under models M0 to M6. For each model, the box contains half of the 65536 scores, the horizontal line is drawn through the box at the median of the data, the upper and lower ends of the box are at upper and lower quartiles (25% and 75%) and vertical lines go up and down from the box to the extremes of the data, except for the outliers, which are plotted by themselves. Here the outliers are the scores that are separated from the box by at least 3 times the inter-quartile range (height of the box). In models M7 and higher, all the 8-letter words have a null score since their counts are included in these models: they are expected as they occur. M6 is then the maximal model for words of length 8.

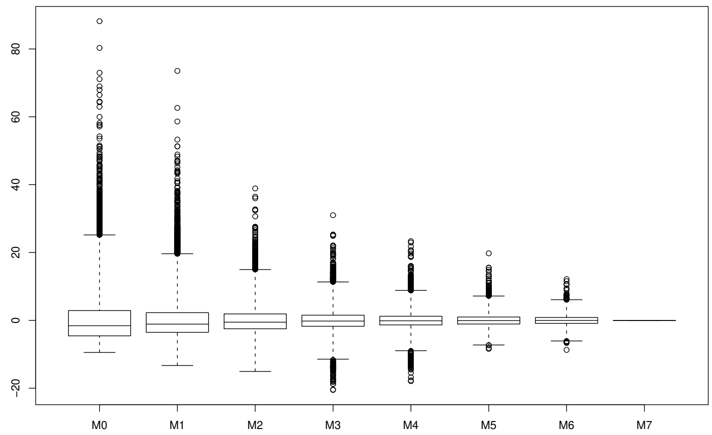

Figure 1.2. Boxplots of the 8-letter word scores in the complete genome of *E. coli* under models M0 to M7.

To analyze the frequency of a ℓ-letter word, the maximal model is of order $m = \ell - 2$; in this model the exceptionality of a word of length ℓ cannot be explained by an unexpected sub-word, since all the sub-word frequencies are included into the maximal model. On the contrary, in small models such contamination by exceptional sub-words may occur. As an illustration let us consider the following example: Figure 1.3 compares the scores (using the Gaussian approximation) of all the 256 4-letter words in the complete *Lamdba* genome under the models M1 (x-axis) and M2 (y-axis). The most over-represented 4-word under M1 is `ccgg`, and it remains significantly over-represented under M2 while taking into account the counts of `ccg` and `cgg`. However, many words lose their exceptionality when the order of the model increases. For example, `gctg` loses its exceptionality as soon as one takes into account the fact that `ctg` occurs 1169 times and is thus a significantly frequent 3-letter word (see Table 1.2). The model M2 says that the 406 occurrences of `gctg` are expected according to the 3-letter word composition of the sequence: `gctg` is expected 394

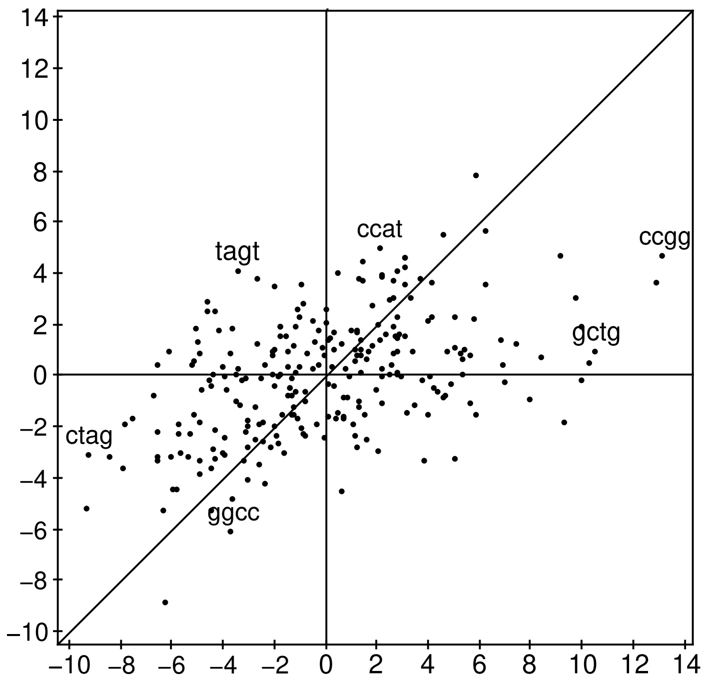

Figure 1.3. Scores of the 4-letter words in the *Lambda* genome under M1 (x-axis) and M2 (y-axis).

times under M2 (see Table 1.1). Its exceptionality under M1 (expected only 255 times) is an artifact due to the important over-representation of its sub-word `ctg`. The number of times that we see `gctg` is not surprising given the number of occurrences of `ctg`. This is what we call a contamination. Another such example is `ctag`: it is exceptionally rare under M1 but not under M2. On the other hand, some exceptionality may be hidden in small models and be revealed in higher models, leading to very interesting interpretations. As an example, `ccat` is not exceptional under M1 and becomes one of the most over-represented word under M2. If we look at its two sub-words of length 3, `cca` and `cat` are slightly under-represented (see Table 1.2). Given their low frequency, `ccat` is expected only 191 times under M2, which is is significantly less than the 218 observed occurrences. So, `cca` and `cat` are slightly avoided in the sequence but they are preferentially overlapping in the sequence. This is more pronounced for `tagt` which is composed of the most avoided 3-word `tag` and is declared under-represented under M1 (contamination in fact), but it seems that there is an important constraint for these occurrences of `tag` to be followed by a `t`.

		Model M1				Model M2			
w	$N(w)$	$\widehat{N}_1(w)$	$\sigma_1(w)$	score	rank	$\widehat{N}_2(w)$	$\sigma_2(w)$	score	rank
`ctag`	14	101.8	9.5	-9.21	2	28.7	4.7	-3.10	27
`tagt`	71	104.0	9.6	-3.42	57	47.3	5.8	4.07	246
`ccat`	218	191.1	12.6	2.12	180	168.6	10.0	4.94	253
`gctg`	406	255.2	14.3	10.52	254	394.6	11.9	0.96	170
`ccgg`	328	169.7	12.0	13.16	256	273.5	11.6	4.68	252

Table 1.1. Statistics of some 4-letter words in the *Lambda* genome under the models M1 and M2. The rank of the scores are obtained while sorting the 256 scores by increasing order.

w	$N(w)$	$\widehat{N}_1(w)$	$\sigma_1(w)$	score	rank
`tag`	217	481.2	17.6	-15.04	1
`cat`	803	869.4	21.6	-3.07	18
`cca`	675	706.5	19.9	-1.58	25
`agt`	595	590.2	19.1	0.25	34
`gct`	856	806.6	20.7	2.39	46
`cgg`	963	772.1	21.0	9.10	60
`ccg`	884	684.3	19.7	10.15	61
`ctg`	1169	802.4	20.8	17.63	63

Table 1.2. Statistics of some 3-letter words in the *Lambda* genome under model M1. The rank of the scores are obtained while sorting the 64 scores by increasing order.

Utility of models Mm_3 Coding DNA sequences are composed of successive trinucleotides called codons. Each base in the sequence is associated to a phase k in $\{1,2,3\}$ depending on its position in the associated codon. In the general model Mm_3, the transition probabilities of a letter depend on its phase and word occurrences can be analyzed separately for each phase or for all phases together (see p. 29); note that $N(w) = \sum_k N(w,k)$. Recall that the phase of an occurrence is by convention in this chapter the phase of its last letter. It is well-known to biologists that there exists a bias in the codon usage: codons that code for the same amino acid are not used uniformly. The following analysis illustrates the importance of taking the 3-letter word composition on each phase into account, in particular the codon composition (3-words on phase 3). Let us consider 36 genes of *E. coli* ($\ell = 44\,856$) and analyze the trinucleotide frequency. Figure 1.4 shows that the majority of the trinucleotides have the same behavior under M1 or M1_3; however, some trinucleotides are less exceptional when one takes the phase into account. If we now calculate the scores of the trinucleotides on phase 1, on phase 2 and on phase 3 under M1_3, we see that the main exceptional trinucleotides are the ones on phase 3: the codons. Figure 1.5 presents the discrepancy of theses scores: codons are much more exceptional than the trinucleotides on phase 1 and 2.

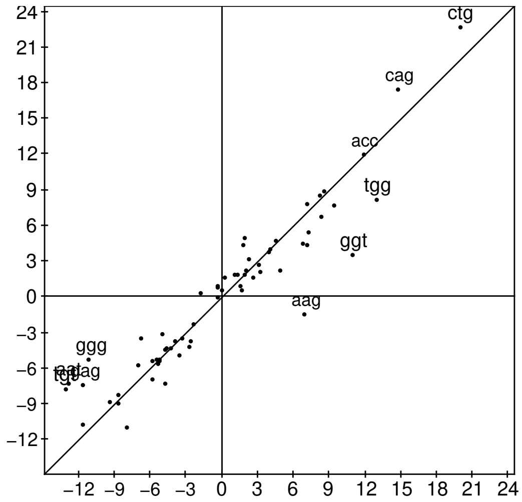

Figure 1.4. Scores of the 3-letter words in 36 genes of *E. coli* under the models M1 (x-axis) and M1_3 (y-axis).

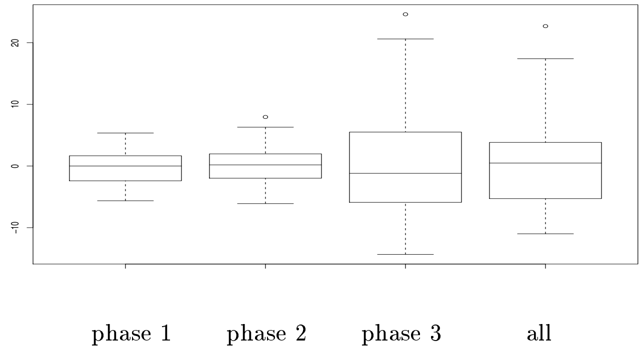

Figure 1.5. Boxplots of the scores of the 3-letter words for each phase and for all phases in 36 genes of *E. coli* under the model M1_3.

Figure 1.6 compares the scores of the 4-words on phase 1 under M1_3 (the codon composition is not taken into account) and M2_3 (the codon composition is taken into account). Note that a 4-word on phase 1 starts with a codon. The 3 most over-represented codons are `ctg`, `cag` and `tgg`. This over-representation is responsible of the exceptionality of `ctgg`, `tggt`, `tggc` and `cagc`. The over-representation of `cagg` seems to be a strong constraint since it is still exceptional given the high frequency of `cag`. When analyzing coding sequences, to be sure to find exceptional words that are not contaminated by the codon usage, the minimal model to use is the model M2_3.

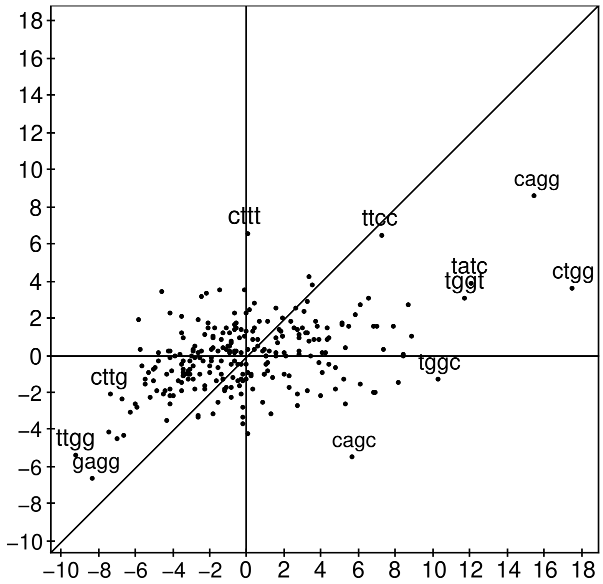

Figure 1.6. Scores of the 4-letter words on phase 1 in 36 genes of *E. coli* under the models M1_3 (x-axis) and M2_3 (y-axis).

1.7.2. Sequencing by hybridization

As a slightly more involved example of how statistics and probability on words are applied in DNA sequence analysis, we describe a problem related to sequencing by hybridization; see Arratia et al. (1996) for more details. Sequencing by hybridization is an approach to determine a DNA sequence from the unordered list of all ℓ-tuples contained in this sequence; typical numbers for ℓ are $\ell = 8, 10, 12$. It is based on the fact that DNA nucleotides bind or hybridize with each other: `a` and `t` hybridize, and `c` and `g` hybridize. DNA strands have a polarity (5', 3'), and hybridizing sequences must be of opposite polarity. To avoid introducing notation to show polarity, we present complementary strands written in reverse direction. For example, the sequence `tgtgtgagtg` hybridizes with `acacactcac`. In a sequencing chip, all 4^ℓ possible oligonucleotides ("probes") of length ℓ are attached to the surface of a substrate, each fragment at a distinct location.

To use an SBH chip, the single-stranded target DNA is amplified, labeled by a fluorescent, and exposed to the sequencing chip. The probes on the chip will hybridize to a copy of the single-stranded target DNA if the substring complementary to the probe exists in the target. These probes are then detected with a spectroscopic detector. For example, if $\ell = 4$, the sequence `tgtgtgagtg` will hybridize to the probes `acac`, `actc`, `caca`, `cact`, `ctca` and `tcac`.

As chips can be washed and used again, and due to automatization, this method is not only fast but also inexpensive. There are still technical difficulties in producing an error-free chip; moreover the SBH image may be difficult to read. We remark that the microarray industry grew out of attempts to make SBH technology practical. However, even if these sources of errors are eliminated, a major drawback of the SBH procedure is that more than one sequence may produce the same SBH data. For example, if $\ell = 4$, the sequence `acactcacac` will hybridize to the same probes as the sequence `acacactcac`.

To control this error resulting from non-unique recoverability, we are interested in an estimate for the probability that a sequence is uniquely recoverable. This probability will depend on the probe length ℓ, on the length n of the target sequence, and on the frequencies of the different nucleotides, `a`, `c`, `g` and `t`, in the sequence. Furthermore we need to bound the error made in estimating the probability of unique recoverability in order to make assertions about the reliability of the chip.

As a simplification, we assume that we not only know the set of all ℓ-tuples in the sequence but also their multiplicity (but not the order in which they occur). This multiset is called the ℓ-spectrum of the sequence. In the sequel, unique recoverability is understood to mean unique recoverability of a sequence from its ℓ-spectrum.

Pevzner (1989) characterizes unique recoverability from the ℓ-spectrum using the de Bruijn-graph (see van Lint and Wilson (1992)) whose vertices are the $(\ell-1)$-tuples in the sequence. Two vertices v and w are joined by a directed edge from v to w if the ℓ-spectrum contains an ℓ-tuple for which the first $(\ell-1)$ nucleotides coincide with v and the last $(\ell-1)$ nucleotides coincide with w. Pevzner (1989) showed that a sequence is uniquely recoverable from its ℓ-spectrum if and only if there is a unique (Eulerian) path connecting all the vertices. Ukkonen (1992) conjectured and Pevzner (1995) proved that there are exactly three structures that prevent unique recoverability:

1. Rotation. The sequence starts and ends with the same $(\ell-1)$-tuple. In this case, the de Bruijn-graph is a cycle, and any vertex could be chosen as the starting point.

2. Transposition with a three-way repeat. If an $(\ell-1)$-tuple occurs three times in the sequence, then the de Bruijn-graph has two loops at this vertex, and the order in which these loops are passed is not fixed.

3. Transposition with two interleaved pairs of repeats. There are two "interleaved" pairs of $(\ell-1)$-tuple repeats, i.e. in the de Bruijn-graph there are two vertices x and y connected by a path of the form $\ldots c \ldots y \ldots x \ldots y \ldots$, where we described a path connecting all the vertices by listing the vertices in the order they are used in the path. This implies that there are two ways of going from x to y in the graph.

Example 1.7.1. The sequence `acacactcac` possesses as 4-spectrum the multiset {`acac`, `acac`, `caca`, `cact`, `actc`, `ctca`, `tcac`}. The competing sequence `acactcacac` has the same 4-spectrum. The de Bruijn-graph for the sequence `acacactcac` has as vertices `aca`, `cac`, `act`, `ctc` and `tca`. There are two directed edges from `aca` to `cac`, and one directed edge each from `cac` to `aca`, from `cac` to `act`, from `act` to `ctc`, from `ctc` to `tca`, and from `tca` to `cac`. The competing sequence `acactcacac` has the same de Bruijn-graph. For the sequence `acacactcac`, a path connecting all vertices is

$$\mathtt{aca}, \mathtt{cac}, \mathtt{aca}, (\mathtt{cac}, \mathtt{act}, \mathtt{ctc}, \mathtt{tca}), \mathtt{cac}.$$

The alternate path

$$\mathtt{aca}, (\mathtt{cac}, \mathtt{act}, \mathtt{ctc}, \mathtt{tca}), \mathtt{cac}, \mathtt{aca}, \mathtt{cac}$$

also connecting all the vertices, corresponds to the sequence, `acactcacac`, with the same 4-spectrum.

Thus unique recoverability can be described in terms of possibly overlapping repeats of $(\ell-1)$-tuples within a single sequence. We use the model M0. For a sequence to be uniquely recoverable, the event of an $(\ell-1)$-tuple repeat should be rare. This implies that we consider the occurrence of $(\ell-1)$-tuples under a Poisson regime. (Note that we are interested in the configuration in which the repeats occur; hence we need a Poisson process approximation for the process of repeats rather than a Poisson approximation for the number of repeats.) If repeats are rare, then three-way repeats are negligible, and so is the probability that a sequence starts and ends with the same $(\ell-1)$-tuple. After bounding these probabilities, we thus restrict our attention to interleaved pairs of repeats. Under the Poisson regime, if there are k pairs of repeats, then the occurrences of these repeats are discrete uniform. Additional randomization makes the position of the repeats continuously uniform, so that all orderings of these pairs will be approximately equally likely. This allows the application of a combinatorial argument using Catalan numbers to obtain that the number of interleaved pairs of repeats, if k repeats are present, is approximately $2^k/(k+1)!$. If λ is the expected number of repeats of ℓ-tuples in a single sequence, we hence get, for the probability P_ℓ that $X_1 X_2 \ldots X_n$ is uniquely recoverable from its ℓ-spectrum,

$$P_\ell \approx e^{-\lambda} \sum_{k \geq 0} \frac{(2\lambda)^k}{k!(k+1)!}.$$

The Chen-Stein method for Poisson approximation provides explicit bounds for the error terms in this approximation, as follows.

In the sequence $X_1 \dots X_n$ of independent identically distributed letters, let $p = \sum_{a\in\mathcal{A}} \mu^2(a)$ be the probability that two random letters match. We write t for $\ell - 1$, as we are interested in $(\ell-1)$-repeats. Again we have to declump: Define $Y_{i,i} = 0$ for all i, and

$$Y_{i,j} = \begin{cases} \mathbb{1}\{X_1 \cdots X_t = X_{j+1} \cdots X_{j+t}\} & \text{if } i = 0 \\ (1 - \mathbb{1}\{X_i = X_j\})\mathbb{1}\{X_{i+1} \cdots X_{i+t} = X_{j+1} \cdots X_{j+t}\} & \text{otherwise.} \end{cases}$$

Thus $Y_{i,j} = 1$ if and only if there is a leftmost repeat starting after i and j. Put $I = \{(i,j), 1 \le i,j \le n - \ell + 1\}$. A careful analysis (see Arratia et al. (1996)) yields that the process $\underline{Y} = (Y_\alpha)_{\alpha\in I}$ is sufficient to decide whether a sequence is uniquely recoverable from its ℓ-spectrum (although $\underline{Y}$ contains strictly less information than the process of indicators of occurrences).

For a Poisson process approximation, we first identify the expected number λ of leftmost repeats. If $\alpha = (i,j)$ does not have self-overlap, that is, if $j - i > t$, then

$$\mathbb{E}(Y_\alpha) = \begin{cases} p^t & \text{if } i = 0 \\ (1-p)p^t & \text{otherwise.} \end{cases}$$

Hence the expected number λ^* of repeats without self-overlap is

$$\lambda^* = \binom{n-2t}{2}(1-p)p^t + (n-2t)p^t.$$

If α does have self-overlap, then, in order to have a leftmost repeat at α, for indices in the overlapping set, two matches are required, and for indices in the non-overlapping set, one match is required. Let $d = j - i$; then $E(Y_\alpha)$ depends on the decomposition of $t + d$ into a quotient q of d and a remainder r (such that $t + d = qd + r$): if p_q is the probability that q random letters match, then

$$\mathbb{E}(Y_\alpha) = \begin{cases} p_{q+1}^r p_q^{d-r} & \text{if } i = 0 \\ (p_q - p_{q+1})^r p_q^{d-r} & \text{otherwise.} \end{cases}$$

If λ^* is bounded away from 0 and infinity, which corresponds to having $t = 2log_{1/p(}n) + c$ for some constant c, then it can be seen that

$$\lambda \approx \frac{n^2}{2}(1-p)p^t.$$

Under the regime that λ is bounded away from 0 and infinity, here is a short version of the general result in Arratia et al. (1996). Let $\mu_{\max} = \max_a \mu(a)$ be the probability of the most likely letter.

THEOREM 1.7.2. *Let $\underline{Z} \equiv (Z_\alpha)_{\alpha\in I}$ be a process with independent Poisson distributed coordinates Y_α, with $\mathbb{E}Z_\alpha = \mathbb{E}Y_\alpha, \alpha \in I$. Then*

$$d_{TV}(\underline{Y}, \underline{Z}) \le b(n,t),$$

where the error term $b(n,t)$ is such that

$$b(n,t) \sim \begin{cases} 16\lambda^2 \frac{t}{n} & \text{in the uniform case} \\ n\mu_{\max}^t & \text{in the nonuniform case.} \end{cases}$$

This bound has been improved by Shamir and Tsur (2001).

In Arratia et al. (1996), a more general result is derived for general alphabets, and explicit bounds are obtained. These bounds can be used to approximate the probability of unique recoverability. Arratia et al. (2000) have obtained results on the number of possible reconstructions for a given sequence (when the reconstruction is not unique).

1.8. Some probabilistic and statistical tools

1.8.1. Stein's method for normal approximation

Stein's method for the normal approximation was first published by Stein (1972). Rinott and Rotar (1996) applied it to obtain multivariate normal approximations with a bound on the error in the distance of suprema over convex sets.

Let $\mathcal{H}$ denote the class of convex sets in $\mathbb{R}^d$. Let $\mathbf{Y}_j, j = 1, \ldots, n$ be random vectors taking values in $\mathbb{R}^d$, and let $\mathbf{W} = \sum_{j=1}^n \mathbf{Y}_j$ be the vector of sums. Assume there is a constant B such that $|\mathbf{Y}_j| := \sum_{i=1}^d |Y_{(j,i)}| \leq B$. Let $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$ have the $d-$dimensional standard multivariate normal distribution.

THEOREM 1.8.1. *Let $\mathcal{S}_i$ and $\mathcal{N}_i$ be subsets of $\{1, \ldots, n\}$, such that $i \in \mathcal{S}_i \subset \mathcal{N}_i, i = 1, \ldots, n$. Assume that there exist constants $D_1 \leq D_2$ such that*

$$\max\{|\mathcal{S}_i|, i = 1, \ldots, n\} \leq D_1$$

and

$$\max\{|\mathcal{N}_i|, i = 1, \ldots, n\} \leq D_2,$$

where for sets $|\cdot|$ denotes cardinality. Then, for $d = 1$ there exists a universal constant c such that

$$\begin{aligned} &\sup_{x \in \mathbb{R}} |\mathbb{P}(\mathbf{W} \leq x) - \mathbb{P}(\mathbf{Z} \leq x)| \\ &\leq c\{2D_2 B + n(2 + \sqrt{\mathbb{E}W^2}) D_1 D_2 B^3 + \chi_1 + \chi_2 + \chi_3\}. \end{aligned}$$

For $d \geq 1$ there exists a constant c depending only on the dimension d such that

$$\begin{aligned} \sup_{A \in \mathcal{H}} |\mathbb{P}(\mathbf{W} \in A) - \mathbb{P}(\mathbf{Z} \in A)| \leq c\{&2\sqrt{d} D_2 B + 2\sqrt{d} n D_1 D_2 B^3 (|\log B| + \log n) \\ &+ \chi_1 + (|\log B| + \log n)(\chi_2 + \chi_3)\}, \end{aligned}$$

where

$$\chi_1 = \sum_{j=1}^n \mathbb{E} \left| \mathbb{E}(\mathbf{Y}_j \mid \sum_{k \notin \mathcal{S}_j} \mathbf{Y}_k) \right|$$

$$\chi_2 = \sum_{j=1}^{n} \mathbb{E}\left| \mathbb{E}(\mathbf{Y}_j(\sum_{k \in \mathcal{S}_j} \mathbf{Y}_k)^T) - \mathbb{E}(\mathbf{Y}_j(\sum_{k \in \mathcal{S}_j} \mathbf{Y}_k)^T | \sum_{l \notin \mathcal{N}_j} \mathbf{Y}_l) \right|$$

$$\chi_3 = \left| I - \sum_{j=1}^{n} \mathbb{E}(\mathbf{Y}_j(\sum_{k \in \mathcal{S}_j} \mathbf{Y}_k)^T) \right|.$$

(Indeed, Rinott and Rotar (1996) derive the result for more general classes of test functions.) Note that there are no explicit assumptions on the mean vector and the variance-covariance matrix; however, for a good approximation it would be desirable to have the mean vector close to zero, and the variance-covariance matrix close to the identity.

1.8.2. The Chen-Stein method for Poisson approximation

The Chen-Stein method is a powerful tool for deriving Poisson approximations and compound Poisson approximations in terms of bounds on the total variation distance. For any two random processes $\underline{Y}$ and $\underline{Z}$ with values in the same space E, the total variation distance between their probability distributions is defined by

$$\begin{aligned} d_{\mathrm{TV}}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z})) &= \sup_{B \subset E, \text{measurable}} |\mathbb{P}(\underline{Y} \in B) - \mathbb{P}(\underline{Z} \in B)| \\ &= \sup_{h: E \to [0,1], \text{measurable}} |\mathbb{E}h(\underline{Y}) - \mathbb{E}h(\underline{Z})|. \end{aligned}$$

First published by Chen (1975) as the Poisson analog to Stein's method for normal approximations (Stein (1972)), it has found widespread application; word counts being just one of them. A friendly exposition is found in Arratia et al. (1989) and a description with many examples can be found in Arratia et al. (1990) and Barbour et al. (1992a). The key theorem for word counts in stationary Markov chains is Theorem 1 in Arratia et al. (1990) with an improved bound by Barbour et al. (1992a) (Theorem 1.A and Theorem 10.A), giving the following theorem.

THEOREM 1.8.2. *Let I be an index set. For each $\alpha \in I$, let Y_α be a Bernoulli random variable with $p_\alpha = \mathbb{P}(Y_\alpha = 1) > 0$. Suppose that, for each $\alpha \in I$, we have chosen $B_\alpha \subset I$ with $\alpha \in B_\alpha$. Let Z_α, $\alpha \in I$, be independent Poisson variables with mean p_α. The total variation distance between the dependent Bernoulli process $\underline{Y} = (Y_\alpha, \alpha \in I)$ and the Poisson process $\underline{Z} = (Z_\alpha, \alpha \in I)$ satisfies*

$$d_{TV}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z})) \le b_1 + b_2 + b_3,$$

where

$$b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta \tag{1.8.1}$$

$$b_2 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha, \beta \neq \alpha} \mathbb{E}(Y_\alpha Y_\beta) \tag{1.8.2}$$

$$b_3 = \sum_{\alpha \in I} \mathbb{E}|\mathbb{E}\{Y_\alpha - p_\alpha | \sigma(Y_\beta, \beta \notin B_\alpha)\}| \,. \tag{1.8.3}$$

Moreover, if $W = \sum_{\alpha \in I} Y_\alpha$ *and* $\lambda = \sum_{\alpha \in I} p_\alpha < \infty$, *then*

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq \frac{1 - e^{-\lambda}}{\lambda}(b_1 + b_2) + \min\left(1, \sqrt{\frac{2}{\lambda e}}\right) b_3.$$

Note that $b_3 = 0$ if Y_α is independent of $\sigma(Y_\beta, \beta \notin B_\alpha\}$. We think of B_α as a neighborhood of strong dependence of Y_α.

One consequence of this theorem is that for any indicator of an event, i.e. for any measurable functional h from E to $[0,1]$, there is an error bound of the form $|\mathbb{E}h(\underline{Y}) - \mathbb{E}h(\underline{Z})| \leq d_{\mathrm{TV}}(\mathcal{L}(\underline{Y}), \mathcal{L}(\underline{Z}))$. Thus, if $T(\underline{Y})$ is a test statistic then, for all $t \in \mathbb{R}$,

$$|\mathbb{P}\left(T(\underline{Y}) \geq t\right) - \mathbb{P}\left(T(\underline{Z}) \geq t\right)| \leq b_1 + b_2 + b_3 \,,$$

which can be used to construct confidence intervals and to find p-values for tests based on this statistic.

Note that this method can also be used to prove compound Poisson approximations. For multivariate compound Poisson approximations it is very convenient. For univariate compound Poisson approximations, better bounds are at hand, as will be illustrated in the next subsection.

1.8.3. Stein's method for direct compound Poisson approximation

Much of this subsection is based on the recent overview of Barbour and Chryssaphinou (2001). A drawback of the point process approach to compound Poisson approximation is that the bounds are not very accurate. Instead it is possible to set up a related method for obtaining a compound Poisson approximation directly, in the univariate case. This approach started with Barbour et al. (1992b); see also Roos (1993), Barbour and Utev (1998). Recently much attention has been given to this problem, and the reader is referred to the references in Barbour and Chryssaphinou (2001).

Denote by $CP(\lambda, \underline{\nu})$ the *compound Poisson distribution* with parameters λ and $\underline{\nu}$, that is, the distribution of the random variable $\sum_{k \geq 1} kM_k$, where $(M_k)_{k \geq 1}$ are independent, and $M_k \sim \mathcal{P}o(\lambda \nu_k), k = 1, 2, \ldots$.

The particular case where $\lambda = n\phi(1-p)$ and $\nu_k = p^{k-1}(1-p)$ for some $\phi > 0$ and $0 < p < 1$, is called the *Polya-Aeppli* distribution.

Again, let I be an index set, and let

$$W = \sum_{\alpha \in I} V_\alpha,$$

where $(V_\alpha)_{\alpha \in I}$ are nonnegative integer-valued random variables. Similarly to the Poisson case, for each $\alpha \in I$ decompose the index set into disjoint sets as

$$I = \alpha \cup \mathcal{S}_\alpha \cup \mathcal{W}_\alpha \cup \mathcal{U}_\alpha.$$

Here, $\mathcal{S}_\alpha$ would correspond to the set of indices with strong influence on α, $\mathcal{W}_\alpha$ would correspond to the set of indices with weak influence on α, and $\mathcal{U}_\alpha$ collects the remaining indices. Put

$$\begin{aligned} S_\alpha &= \sum_{\beta \in \mathcal{S}_\alpha} V_\alpha \\ W_\alpha &= \sum_{\beta \in \mathcal{W}_\alpha} V_\alpha \\ U_\alpha &= \sum_{\beta \in \mathcal{U}_\alpha} V_\alpha. \end{aligned}$$

Then, for $\alpha \in I$,

$$W = V_\alpha + S_\alpha + W_\alpha + U_\alpha.$$

Define the *canonical* parameters $(\lambda, \underline{\nu})$ of the corresponding compound Poisson distribution by

$$\begin{aligned} \lambda\nu_k &= \frac{1}{k} \sum_{\alpha \in I} \mathbb{E}\{V_\alpha \mathbb{1}(V_\alpha + S_\alpha = k)\}, \quad , k \geq 1 \\ \lambda &= \sum_{k \geq 1} k\nu_k. \end{aligned} \tag{1.8.4}$$

Put

$$q_{jk}^{(\alpha)} = \frac{\mathbb{P}(V_\alpha = j, S_\alpha = k)}{m_{i,1}}, \quad \geq 1, k \geq 0,$$

and

$$\begin{aligned} m_{1,k} &= \mathbb{E}V_\alpha \\ m_1 &= \mathbb{E}W = \sum_{\alpha \in I} m_{1,\alpha}. \end{aligned}$$

Similarly to the Poisson case, we shall need the quantities

$$\begin{aligned} \delta_1 &= \sum_{\alpha \in I} m_{1,\alpha} \sum_{j \geq 1} \sum_{k \geq 1} q_{jk}^{(\alpha)} \mathbb{E} \left| \frac{\mathbb{P}(V_\alpha = j, S_\alpha = k | W_\alpha)}{\mathbb{P}(V_\alpha = j, S_\alpha = k)} - 1 \right| \\ \delta_2 &= 2 \sum_{\alpha \in I} \mathbb{E}\{ V_\alpha d_{\mathrm{TV}}(\mathcal{L}(W_\alpha | V_\alpha, S_\alpha); \mathcal{L}(W_\alpha)\} \\ \delta_3 &= \sum_{\alpha \in I} \{\mathbb{E}(V_\alpha U_\alpha) + \mathbb{E}V_\alpha \mathbb{E}(V_\alpha + S_\alpha + U_\alpha)\}. \end{aligned}$$

Then, roughly, δ_3 corresponds to $b_1 + b_2$ in the Poisson case, whereas δ_1 and δ_2 correspond to b_3 in the Poisson case. (Indeed, in Barbour and Chryssaphinou (2001) there is an additional, alternative quantity given in terms of the Wasserstein distance between two distributions.)

Barbour and Chryssaphinou (2001) give the following theorem, which in their paper is also phrased in the Kolmogorov distance, and slightly more general.

THEOREM 1.8.3. *There exist constants* $H_0 = H_0(\lambda, \underline{\nu})$, $H_1 = H_1(\lambda, \underline{\nu})$, *independent of* W, *such that, with* $(\lambda, \underline{\nu})$ *given in (1.8.4),*

$$d_{TV}(\mathcal{L}(W), CP(\lambda, \underline{\nu})) \leq H_0 \min(\delta_1, \delta_2) + H_1 \delta_3,$$

and

$$H_0, H_1 \leq \min(1, (\lambda \nu_1)^{-1}) e^{\lambda}.$$

If in addition

$$k\nu_k \geq (k+1)\nu_{k+1}, \quad k \geq 1, \tag{1.8.5}$$

then, with $\gamma = \lambda(\nu_1 - 2\nu_2)$,

$$\begin{aligned} H_0 &\leq \min\left\{1, \frac{1}{\sqrt{\gamma}}\left(2 - \frac{1}{\sqrt{\gamma}}\right)\right\} \\ H_1 &\leq \min\left\{1, \frac{1}{\sqrt{\gamma}}\left(\frac{1}{4\gamma} + \log^+(2\gamma)\right)\right\}. \end{aligned}$$

Instead of Condition (1.8.5), improved bounds on H_0 and H_1 are also available under the condition that $m^{-1}(m_2 - m_1) < 1/2$, where $m_2 = \sum_{k \geq 1} k^2 \nu_k$, see Barbour and Chryssaphinou (2001).

An important special case is the *declumped* situation, that is, W can be written as

$$W = \sum_{\alpha \in I} \sum_{k \geq 1} k \mathbb{1}_{\alpha k},$$

where

$$\mathbb{1}_{\alpha k} = \mathbb{1}(\alpha \text{ is the index of the representative of a clump of size } k).$$

For $\alpha \in I, k \in \mathbb{N}$, let $B(\alpha, k) \subset I \times \mathbb{N}$ contain $\{\alpha\} \times \mathbb{N}$; this set can be viewed intuitively as the neighborhood of strong dependence of (α, k).

The canonical parameters are now

$$\begin{aligned} \lambda &= \sum_{\alpha \in I} \sum_{k \geq 1} \mathbb{E} \mathbb{1}_{\alpha k} \\ \nu_k &= \lambda^{-1} \sum_{\alpha \in I} \mathbb{E} \mathbb{1}_{\alpha k}. \end{aligned} \tag{1.8.6}$$

For example, if $\mathbb{I}_{\alpha k} = \widetilde{Y}_{i,k}$, then $W = N(w)$, and the canonical parameters are $(n-\ell+1)\widetilde{\mu}_k, k \geq 1$, and $\tilde{\lambda} = \mathbb{E}\widetilde{N}(w)$, so that the approximating distribution is as before, $\mathcal{L}(\sum_{k\geq 1} kZ_k)$ with Z_k's independent Poisson variables with parameters $(n-\ell+1)\widetilde{\mu}_k$. Thus it is the same distribution as in Corollary 1.4.8.

Similarly as in the Poisson case, we shall need the quantities

$$\begin{aligned}
b_1^* &= \sum_{(\alpha,k)\in I\times\mathbb{N}} \sum_{(\beta,k')\in B(\alpha,k)} k'k\mathbb{E}\mathbb{I}_{\alpha k}\mathbb{E}\mathbb{I}_{\beta k'} \\
b_2^* &= \sum_{(\alpha,k)\in I\times\mathbb{N}} \sum_{(\beta,k')\in B(\alpha,k)\setminus\{(\alpha,k)\}} k'k\mathbb{E}\mathbb{I}_{\alpha k}\ \mathbb{I}_{\beta k'} \\
b_3^* &= \sum_{(\alpha,k)\in I\times\mathbb{N}} k\mathbb{E}|\mathbb{E}\{\mathbb{I}_{\alpha k} - \mathbb{E}\mathbb{I}_{\alpha k}|\sigma(\mathbb{I}_{\beta k'}, (\beta,k') \notin B(\alpha,k)\}|\,.
\end{aligned}$$

Then the following theorem can be found in Barbour and Chryssaphinou (2001).

THEOREM 1.8.4. *With the parameters as in (1.8.6), we have that*

$$d_{TV}(\mathcal{L}(W), CP(\lambda,\underline{\nu})) \leq b_3^* H_0 + (b_1^* + b_2^*)H_1.$$

Barbour and Chryssaphinou (2001) also give refined versions of this approach as well as results in Kolmogorov distance. Barbour and Mansson (2002) give related results in Wasserstein distance.

1.8.4. Moment-generating function

Here is a short outline of moment-generating functions; see, e.g., Rice (1995). The *moment-generating function* M of a random variable X is defined as

$$\Phi_X(t) = \mathbb{E}(e^{tX}).$$

So, if X has a discrete distribution p, we have that

$$\Phi_X(t) = \sum_x e^{tx} p(x).$$

If the moment-generating function exists for all t in an open interval containing zero, it uniquely determines the probability distribution.

In particular, under regularity conditions the moments of a random variable can be obtained via the moment-generating function using differentiation. Namely, if $\Phi_X(t)$ is finite, we have

$$\Phi_X'(t) = \frac{d}{dt}\mathbb{E}(e^{tX}) = \mathbb{E}(Xe^{tX}).$$

Thus

$$\Phi_X'(0) = \mathbb{E}X$$

if both sides of the equation exist. Similarly, differentiating r times we obtain

$$\Phi_X^{(r)}(0) = \mathbb{E}(X^r).$$

A special case is when the moment-generating function $\Phi_X(t)$ is rational, that is, when $\Phi_X(t)$ can be written as

$$\Phi_X(t) = \frac{p_0 + p_1 t + \ldots + p_r t^r}{q_0 + q_1 t + \ldots + q_s t^s} = \sum_d f(d) t^d,$$

for some r, s and coefficients $p_1, \ldots, p_r, q_1, \ldots, q_s$. By normalization we may assume $q_0 = 1$. Then

$$p_0 + p_1 t + \ldots + p_r t^r = \sum_d f(d) t^d (1 + q_1 t + \ldots + q_s t^s).$$

Identification of the coefficients of t^i on both sides yields

$$p_i = \sum_{d=0}^{i} f(d) q_{i-d} \text{ for } i \le r$$
$$0 = \sum_{d=0}^{i} f(d) q_{i-d} \text{ for } i > r.$$

This gives a recurrence formula for the coefficients $f(d)$; we have

$$f(0) = p_0$$
$$f(d) = p_d - \sum_{i=1}^{\min(d,s)} f(d-i) q_i, \quad d \ge 1$$

where $p_d = 0$ for $d > r$.

1.8.5. The δ-method

In general, the δ-method, or *propagation of error*, is a linear approximation (Taylor expansion) of a nonlinear function of random variables. Here we are particularly interested in the validity of a normal approximation for functions of random vectors (see, e.g., Rice (1995)). The following theorem can be found on p.313 in Waterman (1995).

THEOREM 1.8.5. *Let $\underline{X}_n = (X_{n1}, X_{n2}, \ldots, X_{nk})$ be a sequence of random vectors satisfying*

$$b_n(\underline{X}_n - \underline{\mu}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$$

with $b_n \to \infty$. The vector valued function $\underline{g}(\underline{x}) = (g_1(\underline{x}), \ldots, g_\ell(\underline{x}))$ has real valued $g_i(\underline{x})$ with non-zero differential

$$\frac{\partial g_i}{\partial g_{\underline{x}}} = \left(\frac{\partial g_i}{\partial g_{x_1}}, \ldots, \frac{\partial g_i}{\partial g_{x_k}} \right).$$

Define $\mathbf{D} = (d_{i,j})$ *where* $d_{i,j} = \frac{\partial g_i}{\partial g_{x_j}}(\underline{\mu})$. *Then*

$$b_n(g(\underline{X}_n) - g(\underline{\mu})) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{D\Sigma D^T}).$$

1.8.6. A large deviation Principle

Assume $X_1 \cdots X_n$ is an irreducible Markov chain on a finite alphabet $\mathcal{A}$ with transition probabilities $\pi(a,b)$, $a, b \in \mathcal{A}$. The following theorem for Markov chains can be found on p.78 in Bucklew (1990).

THEOREM 1.8.6 (Miller). *Let* f *be a function mapping* $\mathcal{A}$ *into* $\mathbb{R}$. *Then,* $n^{-1}\sum_{i=1}^n f(X_i)$ *obeys a large deviation principle with rate function* I *defined below: for every closed subset* $F \subset \mathbb{R}$ *and every open subset* $O \subset \mathbb{R}$,

$$\limsup_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) \in F\right) \le -\inf_{x\in F} I(x),$$
$$\liminf_{n\to+\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) \in O\right) \ge -\inf_{x\in O} I(x).$$

The rate function I *is positive, convex, uniquely equal to zero at* $x = \mathbb{E}f(X_1)$ *and given by*

$$I(x) = \sup_\theta(\theta x - \log\lambda(\theta))$$

where $\lambda(\theta)$ *is the largest eigenvalue of the matrix* $\left(e^{\theta f(b)}\pi(a,b)\right)_{a,b\in\mathcal{A}}$.

1.8.7. A CLT for martingales

The following theorem can be found in Dacunha-Castelle and Duflo (1983) p.80.

THEOREM 1.8.7. *Let* $(\xi_{n,i})_{i=1,\ldots,n}$ *be a triangular array of* d*-dimensional random vectors such that* $\mathbb{E}||\xi_{n,i}||_2^2 < \infty$, *and* V *be a positive* $d \times d$ *matrix. Put* $\mathcal{F}_{n,i} = \sigma(\xi_{n,1}, \ldots, \xi_{n,i})$; $\mathbb{E}(\xi_{n,i} \mid \mathcal{F}_{n,i-1})$ *denotes the conditional expectation vector of* $\xi_{n,i}$ *and* $Cov(\xi_{n,i} \mid \mathcal{F}_{n,i-1})$ *denotes the conditional covariance matrix of* $\xi_{n,i}$. *If as* $n \to \infty$

(i) $\displaystyle\sum_{i=1}^n \mathbb{E}(\xi_{n,i} \mid \mathcal{F}_{n,i-1}) \xrightarrow{\mathbb{P}} 0$,

(ii) $\displaystyle\sum_{i=1}^n Cov(\xi_{n,i} \mid \mathcal{F}_{n,i-1}) \to V$,

(iii) $\displaystyle\forall\varepsilon > 0, \sum_{i=1}^n \mathbb{P}(|\xi_{n,i}| > \varepsilon \mid \mathcal{F}_{n,i-1}) \xrightarrow{\mathbb{P}} 0$,

then $\sum_{i=1}^n \xi_{n,i} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V)$.

Acknowledgments. We would like to thank Simon Tavaré for many helpful comments. Thanks goes also to Xueying Xie for pointing out inconsistencies in a previous version concerning testing for the order of a Markov chain.

This work was supported in part by Sandia National Laboratories, operated by Lockheed Martin for the U.S. Department of Energy under contract No. DE-AC04-94AL85000, and by the Mathematics, Information, and Computational Science Program of the Office of Science of the U.S. Department of Energy. Gesine Reinert was supported in part by EPSRC grant aGR/R52183/01. Michael Waterman was partially supported by Celera Genomics.

An earlier version of this paper appeared as *Probabilistic and statistical properties of words: an overview* in the *Journal of Computational Biology*, Vol. 7 (2000), pp. 1–46. We thank Mary Ann Liebert, Inc. Publishers for permission to include that material here.

Bibliography

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool, *J. Mol. Biol.*, *215*, 403–410.

Apostolico, A., Bock, M., and Xuyan, X. (1998). Annotated statistical indices for sequence analysis, In *Proceedings of Compression and Complexity of SEQUENCES 97*, pp. 215–229. IEEE Computer Society Press.

Arratia, R., Bollobas, B., Coppersmith, D., and Sorkin, G. (2000). Euler circuits and DNA sequencing by hybridization, *Discrete Applied Mathematics*, *104*, 63–69.

Arratia, R., Goldstein, L., and Gordon, L. (1989). Two moments suffice for Poisson approximations : the Chen-Stein method, *Ann. Prob.*, *17*, 9–25.

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

Arratia, R., Martin, D., Reinert, G., and Waterman, M. S. (1996). Poisson approximation for long repeats in a random sequence with application to sequencing by hybridization, *J. Comp. Biol.*, *3*, 425–463.

Barbour, A. and Xia, A. (1999). Poisson Perturbations, *ESAIM: Probability and Statistics*, *3*, 131–150.

Barbour, A. D., Chryssaphinou, O., and Vaggelatou, E. (2001). Applications of compound Poisson approximation, In Charalambides, C. A., Koutras, M. V., and Balakrishnan, N. (Eds.), *Probability and Statistical Models with Applications*. Chapman and Hall, Boca Raton. 41–62.

Barbour, A. D., Holst, L., and Janson, S. (1992a). *Poisson Approximation*. Oxford University Press.

Barbour, A., Chen, L., and Loh, W.-L. (1992b). Compound Poisson approximation for nonnegative random variables via Stein's method, *Ann. Prob.*, *20*, 1843–1866.

Barbour, A. and Chryssaphinou, O. (2001). Compound Poisson approximation: a user's guide, *Ann. Appl. Prob.*, *11*, 964–1002.

Barbour, A. and Mansson, M. (2002). Compound Poisson process approximation, *Ann. Prob.*, *30*, 1492–1537.

Barbour, A. and Utev, S. (1998). Solving the Stein equation in compound Poisson approximation, *Adv. Appl. Prob.*, *30*, 449–475.

Biaudet, V., El Karoui, M., and Gruss, A. (1998). Codon usage can explain GT-rich islands surrounding Chi sites on the *Escherichia coli* genome, *Mol. Microbiol.*, *29*, 666–669.

Biggins, J. and Cannings, C. (1987). Markov renewal processes, counters and repeated sequences in Markov chains, *Ann. Appl. Prob.*, *19*, 521–545.

Billingsley, P. (1961). Statistical Methods in Markov Chains, *Annals of Mathematical Statistics*, *2*, 12–40.

Blom, G. and Thorburn, D. (1982). How many random digits are required until given sequences are obtained?, *J. Appl. Prob.*, *19*, 518–531.

Breen, S., Waterman, M. S., and Zhang, N. (1985). Renewal theory for several patterns, *J. Appl. Prob.*, *22*, 228–234.

Brendel, V., Beckmann, J. S., and Trifonov, E. N. (1986). Linguistics of nucleotide sequences : Morphology and comparison of vocabularies, *J. Biomol. Struct. Dynamics*, *4*, 11–21.

Bucklew, J. A. (1990). *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley.

Chedin, F., Noirot, P., Biaudet, V., and Ehrlich, S. (1998). A five-nucleotide sequence protects DNA from exonucleolytic degradation by AddAB, the RecBCD analogue of *Bacillus subtilis*, *Mol. Microbiol.*, *31*, 1369–1377.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials, *Ann. Prob.*, *3*, 534–545.

Chen, L. H. Y. and Xia, A. (2002). Stein's method, Palm theory and Poisson process approximation, Preprint.

Chryssaphinou, O. and Papastavridis, S. (1988a). A limit theorem for the number of non-overlapping occurrences of a pattern in a sequence of independent trials, *J. Appl. Prob.*, *25*, 428–431.

Chryssaphinou, O. and Papastavridis, S. (1988b). A limit theorem on the number of overlapping appearances of a pattern in a sequence of independent trials, *Prob. Theory Rel. Fields*, *79*, 129–143.

Chryssaphinou, O., Papastavridis, S., and Vaggelatou, E. (2001). Poisson approximation for the non-overlapping appearances of several words in Markov chains, *Combinatorics, Probability and Computing*, *10*, 293–308.

Chung, K. L. (1974). *A Course in Probability Theory.* Academic Press, 2^{nd} ed.

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences, *Bull. Math. Biol*, *51*, 79–94.

Churchill, G. A., Daniels, D. L., and Waterman, M. S. (1990). The distribution of restriction enzyme sites in *Escheria coli*, *Nucl. Acids Res.*, *18*, 589–597.

Cowan, R. (1991). Expected frequencies of DNA patterns using Whittle's formula, *J. Appl. Prob.*, *28*, 886–892.

Dacunha-Castelle, D. and Duflo, M. (1983). *Probabilités et Statistiques 2.Problèmes à Temps Mobile.* Masson.

Deheuvels, P., Devroye, L., and Lynch, J. (1986). Exact convergence rate in the limit theorems of Erdös-Rényi and Shepp, *Ann. Prob.*, *14*, 209–223.

Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes, *Ann. Appl. Prob.*, *2*, 329–357.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis.* Cambridge University Press.

Eichelsbacher, P. and Roos, M. (1999). Compound Poisson approximation for dissociated random variables via Stein's method, *Combinatorics, Probability and Computing*, *8*, 335–346.

Erdös, P. and Rényi, A. (1970). On a new law of large numbers, *J. Analyse Math.*, *23*, 103–111.

Erhardsson, T. (1997). *Compound Poisson approximation for Markov chains.* Ph.D. thesis, Royal Institute of Technology, Stockholm.

Erhardsson, T. (1999). Compound Poisson approximation for Markov chains using Stein's method, *Ann. Prob.*, *27*, 565–596.

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.

Feller, W. (1950). *An Introduction to Probability Theory and Its Applications*, Vol. 1. Wiley, 3rd edition.

Fu, J. C. (1993). Poisson convergence in reliability of a large linearly connected system as related to coin tossing, *Statistica Sinica*, *3*, 261–275.

Gentleman, J. (1994). The distribution of the frequency of subsequences in alphabetic sequences, as exemplified by deoxyribonucleic acid, *Appl. Statist.*, *43*, 404–414.

Gentleman, J. and Mullin, R. (1989). The distribution of the frequency of occurrence of nucleotide subsequences, based on their overlap capability, *Biometrics*, *45*, 35–52.

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.

Glaz, J., Naus, J., and Wallenstein, S. (2001). *Scan statistics.* Springer.

Godbole, A. P. (1991). Poisson approximations for runs and patterns of rare events, *Adv. Appl. Prob.*, *23*, 851–865.

Godbole, A. P. and Schaffner, A. A. (1993). Improved Poisson approximations for word patterns, *Adv. Appl. Prob.*, *25*, 334–347.

Gordon, L., Schilling, M., and Waterman, M. (1986). An extreme value theory for long head runs, *Prob. Theor. Rel. Fields*, *72*, 279–287.

Guibas, L. and Odlyzko, A. (1980). Long repetitive patterns in random sequences, *Z. Wahrscheinlichkeitstheorie*, *53*, 241–262.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press.

Gusto, G. (2000). Approximation par une loi de Poisson composée de la loi du comptage d'un mot rare dans une chaîne de Markov, Master's thesis, DEA Modélisation Stochastique et Statistique, Université Paris XI, Orsay.

Guthrie, D. and Youssef, M. (1970). Empirical evaluation of some chi-square tests for the order of a Markov chain, *J. Amer. Statist. Soc.*, *65*, 631–634.

Hirano, K. and Aki, S. (1993). On number of occurrences of success runs of specified length in a two-state Markov chain, *Statistica Sinica*, *3*, 313–320.

Huang, H. (2002). Error bounds on multivariate normal approximations for word count statistics, *Adv. Appl. Prob.*, *34*, 559–587.

Karlin, S., Burge, C., and Campbell, A. M. (1992). Statistical analyses of counts and distributions of restriction sites in DNA sequences, *Nucl. Acids Res.*, *20*, 1363–1370.

Karlin, S. and Macken, C. (1991). Assessment of inhomogeneities in an *E. coli* physical map, *Nucl. Acids Res.*, *19*, 4241–4246.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes.* Academic Press, 2^{nd} ed.

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.

Kleffe, J. and Langbecker, U. (1990). Exact computation of pattern probabilities in random sequences generated by Markov chains, *Comp. Applic. Biosci.*, *6*, 347–353.

Leung, M. Y., Marsh, G. M., and Speed, T. P. (1996). Over and underrepresentation of short DNA words in Herpesvirus genomes, *J. Comp. Biol.*, *3*, 345–360.

Li, S.-Y. (1980). A martingale approach to the study of occurrence of sequence patterns in repeated experiments, *Ann. Prob.*, *8*, 1171–1176.

Lonardi, S. (2001). *Global detectors of unusual words design, implementation and applications to pattern discovery in biosequences.* Ph.D. thesis, Purdue University, West Lafayette, Indiana.

Lothaire, M. (1983). *Combinatorics on Words.* Addison-Wesley.

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

Marsan, L. and Sagot, M.-F. (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification, *J. Comp. Biol.*, *7*, 345–362.

Muri, F. (1998). Modelling Bacterial Genomes using Hidden Markov Models,. IN *Compstat'98 Proceedings in Computational Statistics* (eds Payne, R. and Green, P.J.), 89–100. Physica-Verlag, Heidelberg.

Nicodème, P., Salvy, B., and Flajolet, P. (1999). Motif statistics, In *European Symposium on Algorithms - ESA99*, pp. 194–211. Lecture Notes in Computer Science **1643**.

Nuel, G. (2001). *Grandes déviations et chaînes de Markov pour l'étude des mots exceptionnels dans les séquences biologiques.* Ph.D. thesis, Université d'Evry Val d'Essonne.

Pevzner, P. A. (1989). l-tuple DNA Sequencing: Computer Analysis, *J. Biomol. Struct. Dynamics*, *7*, 63–73.

Pevzner, P. A. (1995). DNA physical mapping and alternating Eulerian cycles in colored graphs, *Algorithmica*, *13*, 77–105.

Prum, B., Rodolphe, F., and Turckheim, É. (1995). Finding words with unexpected frequencies in DNA sequences, *J. R. Statist. Soc. B*, *57*, 205–220.

Rajarshi, M. B. (1974). Success runs in a two-state Markov chain, *J. Appl. Prob.*, *11*, 190–192.

Régnier, M. (2000). A unified approach to word occurrence probabilities, *Discrete Applied Mathematics*, *104*, 259–280.

Régnier, M. and Denise, A. (2003). Rare events on random strings, Preprint.

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.

Reinert, G., Schbath, S., and Waterman, M. (2000). Probabilistic and statistical properties of words: an overview, *J. Comp. Biol.*, *7*, 1–46.

Rice, J. (1995). *Mathematical Statistics and Data Analysis.* Duxbury Press.

Rinott, Y. and Rotar, V. (1996). On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted U-statistics, *J. Multivariate Analysis*, *56*, 333–350.

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*(4), 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.

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.

Rocha, E., Viari, A., and Danchin, A. (1998). Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons, *Nucl. Acids Res.*, *26*, 2971–2980.

Roos, M. (1993). *Stein-Chen method for compound Poisson approximation.* Ph.D. thesis, University of Zurich.

Rudander, J. (1996). *On the first occurrence of a given pattern in a semi-Markov process.* Ph.D. thesis, Uppsala, Sweden.

Schbath, S. (1995a). Compound Poisson approximation of word counts in DNA sequences, *ESAIM: Probability and Statistics*, *1*, 1–16. (`http://www.emath.fr/ps/`).

Schbath, S. (1995b). *Étude 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.* Ph.D. thesis, Université René Descartes, Paris V.

Senoussi, R. (1990). Statistique asymptotique presque-sûre de modèles statistiques convexes, *Ann. Inst. Henri Poincaré*, *26*, 19–44.

Shamir, R. and Tsur, D. (2001). Large scale sequencing by hybridization, In *Proc. 5th Annual Int. Conf. on Computational Molecular Biology – RECOMB*, pp. 269–277. ACM Press.

Sourice, S., Biaudet, V., El Karoui, M., Ehrlich, S., and Gruss, A. (1998). Identification of the Chi site of *Haemophilus influenzae* as several sequences related to the *Escherichia coli* Chi site, *Mol. Microbiol.*, *27*, 1021–1029.

Stefanov, V. (2003). The intersite distances between pattern occurrences in strings generated by general discrete - and continuous- time models: an algorithmic approach, To appear.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, In *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, Vol. 2, pp. 583–602. Univ. California Press, Berkeley.

Tanushev, M. S. (1996). Central limit theorem for several patterns in a Markov chain sequence of letters., Preprint.

Tanushev, M. S. and Arratia, R. (1997). Central limit theorem for renewal theory for several patterns, *J. Comp. Biol.*, *4*, 35–44.

Ukkonen, E. (1992). Approximate string-matching with q-grams and maximal matches, *Theoret. Comp. Science*, *92*, 191–211.

van Lint, J. H. and Wilson, R. M. (1992). *A Course in Combinatorics.* Cambridge University Press.

Waterman, M. (1995). *Introduction to Computational Biology.* Chapman & Hall.