J. Appl. Prob. **46,** 1020–1037 (2009)
Printed in England

MODERATE DEVIATIONS FOR WORD COUNTS IN BIOLOGICAL SEQUENCES

SARAH BEHRENS,* *Max Planck Institute for Molecular Genetics*
MATTHIAS LÖWE,** *University of Münster*

Abstract

We derive a moderate deviation principle for word counts (which is extended to counts of multiple patterns) in biological sequences under different models: independent and identically distributed letters, homogeneous Markov chains of order 1 and m, and, in view of the codon structure of DNA sequences, Markov chains with three different transition matrices. This enables us to approximate P-values for the number of word occurrences in DNA and protein sequences in a new manner.

Keywords: Moderate deviations; Markov chain; word counts; motifs; biological sequence analysis

2000 Mathematics Subject Classification: Primary 60F99; 92D20
Secondary 60J10; 60J20; 60G50

1. Introduction

Recent progress in DNA and protein sequencing stressed the necessity to develop statistical methods for the analysis of biological sequences. One probabilistic approach to recognise special features of DNA or protein sequences is to identify words or motifs which occur significantly often or rarely. Reinert *et al.* [19] gave an excellent overview of existing results about statistical properties of words. Much work has been done to examine the distribution of word counts—both exact (see [10], [18], [19], and [20]) and asymptotic results are available. Over the past years, Gaussian (see [17], [19], and [22]), Poisson or compound Poisson (see [19] and [21]), and large deviation based approximations (see [13] and [19]) have been derived yielding approximate P-values to assess the statistical significance of word occurrences. However, from a practical point of view, all approaches have disadvantages: while an exact computation of P-values for long sequences requires a lot of time and memory capacity, the accuracy of the Gaussian approximation decreases as the length of the words increases (rare words) and, although the large deviation approach provides a good approximation for very exceptional words, it cannot manage with words whose expected count is close to the observed one (see [19, Section 6.7.1]).

In this paper we analyse the moderate deviation behaviour of word counts, i.e. we examine the regime between the Gaussian and large deviation regimes. Consequently, we assume that our analysis yields a reasonable approximation for P-values in a moderate deviation regime. As can be seen in Example 1, the moderate deviation based approximation of P-values for moderately appearing words indeed performs better than the Gaussian and large deviation

Received 24 October 2008; revision received 14 August 2009.
* Postal address: Max Planck Institute for Molecular Genetics, Department for Computational Molecular Biology, Ihnestraße 63-73, 14195 Berlin, Germany. Email address: sbehrens@molgen.mpg.de
** Postal address: Fachbereich Mathematik und Informatik, Universität Münster, Einsteinstr. 62, 48149, Münster, Germany. Email address: maloewe@math.uni-muenster.de

based approximations. Thus, our approach should be a good compromise for approximating P-values for word occurrences.

Before stating our main result in Section 2, we now give a survey of the underlying probabilistic models of biological sequences we use and we introduce the concept of moderate deviations. Section 3 is devoted to the proof of the main result and in Section 4 we give some applications in the field of biological sequence analysis.

1.1. Probabilistic models for biological sequences

Usually, a biological sequence, i.e. a DNA or protein sequence, is modelled by a Markov chain (see, e.g. [7]). Let $\mathcal{A}$ be a finite letter alphabet (for DNA sequences, $\mathcal{A}$ contains the four different bases, i.e. $\mathcal{A} = \{\text{A,C,G,T}\}$, and, for protein sequences, the alphabet consists of 20 different amino acids), and let $(X_n)_{n\in\mathbb{N}}$ be a sequence taking values in $\mathcal{A}$.

The simplest model (model M0) assumes that the letters X_i are independent and identically distributed (i.i.d.). While being easy to handle, this model is not very accurate (see, e.g. [2] and [15]). A more general model (model Mm; see [19]) postulates that $(X_n)_{n\in\mathbb{N}}$ is a homogeneous Markov chain of order m. In our paper we will mainly focus on the special case of model M1. To be more precise, we assume that $(X_n)_{n\in\mathbb{N}}$ is a homogeneous, first-order Markov chain with aperiodic and irreducible transition matrix $\Pi = (p(a, b))_{a,b\in\mathcal{A}}$. Note that, under these conditions, the existence of a stationary distribution, i.e. a distribution μ satisfying $\mu\Pi = \mu$, is guaranteed by the ergodic theorem for Markov chains.

When considering coding DNA sequences, a more refined model (model Mm-3; see [19] and [22]) takes the so-called codon structure of these sequences into account: three successive, nonoverlapping bases are translated into one amino acid. In some cases the first two letters of a codon suffice to determine the corresponding amino acid, i.e. the position within the codon may have a different importance. We therefore consider a Markov chain (of order m) with three different transition matrices $\Pi_k = (p_k(a, b))_{a,b\in\mathcal{A}}$, $k = 1, 2, 3$.

Now, given an underlying model Mm or Mm-3, $m \in \mathbb{N}_0$, an alphabet $\mathcal{A}$, and the corresponding random letter sequence $(X_n)_{n\in\mathbb{N}}$, let $w = w_1 \cdots w_\ell$ be a word of length ℓ taking values in $\mathcal{A}$ and let $Y_i(w) := \mathbf{1}_{\{X_i=w_1,\dots,X_{i+l-1}=w_\ell\}}$, $i \in \mathbb{N}$, be the indicator function of the set $\{w \text{ starts at position } i \text{ in } X_1 \cdots X_n\}$. Then the word count of w is defined by

$$N(w) := \sum_{i=1}^{n-\ell+1} Y_i(w).$$

We are interested in the moderate deviations of $N(w) - \mathrm{E}\, N(w)$.

1.2. Moderate deviations

As mentioned above, the moderate deviations are located between the normal and large deviations. Formally, we define a moderate deviation principle as follows.

Definition 1. A sequence $(Z_n)_{n\in\mathbb{N}}$ of real-valued random variables is said to satisfy a moderate deviation principle (MDP) with rate function $I : \mathbb{R} \to [0, \infty]$ if $(Z_n/n^\alpha)_{n\in\mathbb{N}}$ satisfies a large deviation principle (LDP) with rate function I and speed $n^{2\alpha-1}$ for all $\frac{1}{2} < \alpha < 1$, i.e. if

(i) I has closed level sets;

(ii) $\limsup_{n\to\infty}(1/n^{2\alpha-1}) \log \mathrm{P}(Z_n/n^\alpha \in F) \le -\inf_{x\in F} I(x)$ for all closed subsets $F \subset \mathbb{R}$;

(iii) $\liminf_{n\to\infty}(1/n^{2\alpha-1}) \log \mathrm{P}(Z_n/n^\alpha \in O) \ge -\inf_{x\in O} I(x)$ for all open subsets $O \subset \mathbb{R}$.

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

Remark 1. An MDP is often defined on a more general scale b_n satisfying $b_n/n \to 0$ and $b_n^2/n \to \infty$ (where $(b_n)_{n\in\mathbb{N}}$ is a sequence of real positive numbers) rather than on a scale n^α, $\frac{1}{2} < \alpha < 1$. However, for our purposes, the scale n^α is more illustrative.

The aim of this paper is to derive an MDP for $N(w) - \mathrm{E}\, N(w)$, i.e. to study the behaviour of

$$\frac{1}{n^{2\alpha-1}} \log \mathrm{P}\left(\frac{N(w) - \mathrm{E}\, N(w)}{n^\alpha} \in A \right)$$

for open or closed subsets $A \subset \mathbb{R}$.

2. Main result

Let $w = w_1 \cdots w_\ell$ be a word of length ℓ. The set of periods of w is given by

$$\mathcal{P}(w) = \{p \in \{1, \ldots, \ell - 1\} \colon w_i = w_{i+p} \text{ for all } i = 1, \ldots, \ell - p\},$$

e.g. the word 'ATATAT' has periods 2, 4, and 6. Occurrences of periodic words may overlap in a sequence.

The suffix of length q of w is denoted by

$$w_{(q)} = w_{\ell-q+1} \cdots w_\ell,$$

and $ww_{(q)}$ refers to the composite word $w_1 \cdots w_\ell w_{\ell-q+1} \cdots w_\ell$.

Throughout this paper, we will use the notation E_λ and P_λ to emphasise the fact that the initial distribution has been changed to λ. When $\lambda = \delta_a$, we will write E_a and P_a instead, and when referring to the original initial distribution (which is not specified), we will drop the index. The main result of this paper is as follows.

Theorem 1. *Let $(X_n)_{n\in\mathbb{N}}$ be an irreducible, aperiodic, and homogeneous Markov chain with transition matrix $\Pi = (p(a,b))_{a,b\in\mathcal{A}}$, stationary distribution μ, and finite state space $\mathcal{A}$. Then, for all $w \in \mathcal{W} := \{a_1 \cdots a_\ell \in \mathcal{A}^\ell \colon p(a_i, a_{i+1}) > 0 \text{ for all } i \in \{1, \ldots, \ell - 1\}\}$ and any initial distribution, $N(w) - \mathrm{E}_\mu\, N(w)$ satisfies an MDP with rate function $\Lambda_1^*(q) = q^2/2\sigma_1^2(w)$, where*

$$\begin{aligned}
\sigma_1^2(w) &= \mu_1(w) + 2 \sum_{q\in\mathcal{P}(w)} \mu_1(ww_{(q)}) - (2\ell - 1)\mu_1^2(w) \\
&\quad + \frac{2}{\mu(w_1)} \mu_1^2(w) \sum_{k=1}^{\infty} (p^{(k)}(w_\ell, w_1) - \mu(w_1)), \\
\mu_1(w) &= \mu(w_1) p(w_1, w_2) \cdots p(w_{\ell-1}, w_\ell),
\end{aligned}$$

and $p^{(k)}$ denotes the k-step transition probability.

Remarks 2. (i) The limiting variance $\sigma_1^2(w)$ can be rewritten as

$$\begin{aligned}
\sigma_1^2(w) &= \mu_1(w) + 2 \sum_{q\in\mathcal{P}(w)} \mu_1(ww_{(q)}) - (2\ell - 1)\mu_1^2(w) \\
&\quad + 2\mu_1^2(w) \left(\sum_{a\in\mathcal{A}} \mu(a) m_{a,w_1} - m_{w_\ell,w_1} \right),
\end{aligned} \tag{1}$$

where $m_{a,b} = \mathrm{E}_a\, \tau_b = \mathrm{E}_a(\inf\{n \in \mathbb{N} \colon X_n = b\})$ is the mean first passage time from state a to

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

state b, $a, b \in \mathcal{A}$. Firstly, it follows, from [16, Corollary 1] and $m_{w_1,w_1} = 1/\mu(w_1)$, that

$$\frac{1}{\mu(w_1)} \sum_{k=1}^{\infty} (p^{(k)}(w_\ell, w_1) - \mu(w_1)) = \frac{m^{(2)}_{w_1,w_1} + m_{w_1,w_1}}{2m_{w_1,w_1}} - m_{w_\ell,w_1},$$

where $m^{(2)}_{w_1,w_1} = \mathrm{E}_{w_1} \tau^2_{w_1}$. Secondly, we find from [8, Corollary 2.5.5] that

$$m^{(2)}_{w_1,w_1} + m_{w_1,w_1} = 2m_{w_1,w_1} \sum_{a \in \mathcal{A}} \mu(a) m_{a,w_1},$$

which proves (1).

The mean first passage times $m_{a,b}$, $a, b \in \mathcal{A}$, can be computed by solving the following well-known system of linear equations:

$$m_{a,b} = 1 + \sum_{c \in \mathcal{A} \setminus \{b\}} p(a, c) m_{c,b}. \tag{2}$$

(ii) In applications, only words with positive probability of realisation are of interest. So words which are not elements of $\mathcal{W}$ can be neglected.

For the special case of model M0, we obtain the following corollary.

Corollary 1. *Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of i.i.d. random variables with values in a finite alphabet $\mathcal{A}$. Then $N(w) - \mathrm{E}\, N(w)$ satisfies an MDP with rate function $\Lambda^*_0(q) = q^2/2\sigma^2_0(w)$, where*

$$\sigma^2_0(w) = \mu_0(w) + 2 \sum_{q \in \mathcal{P}(w)} \mu_0(w w_{(q)}) - (2\ell - 1)\mu^2_0(w)$$

and $\mu_0(w) = \mathrm{P}(X_1 = w_1) \cdots \mathrm{P}(X_\ell = w_\ell)$.

The result can be easily extended to model Mm.

Theorem 2. *If $(X_n)_{n \in \mathbb{N}}$ is an irreducible, aperiodic, and homogeneous Markov chain of order m with transition matrix $\Pi = (p(a_1 \cdots a_m, b))_{a_1,\dots,a_m,b \in \mathcal{A}}$, stationary distribution μ, and finite state space $\mathcal{A}$, then, for all $w \in \mathcal{W} = \{a_1 \cdots a_\ell \in \mathcal{A}^\ell \colon p(a_{j-m} \cdots a_{j-1}, a_j) > 0$ for all $j \in \{m+1, \dots, \ell\}\}$ and any initial distribution, $N(w) - \mathrm{E}_\mu\, N(w)$ satisfies an MDP with rate function $\Lambda^*_m(q) = q^2/2\sigma^2_m(w)$, where*

$$\begin{aligned}\sigma^2_m(w) &= \mu_m(w) + 2 \sum_{q \in \mathcal{P}(w)} \mu_m(w w_{(q)}) - (2\ell - 1)\mu^2_m(w) \\ &\quad + \frac{2}{\mu(w_1 \cdots w_m)} \mu^2_m(w) \sum_{k=1}^{\infty} (p^{(k)}(w_{\ell-m+1} \cdots w_\ell, w_1 \cdots w_m) - \mu(w_1 \cdots w_m)), \\ \mu_m(w) &= \mu(w_1 \cdots w_m) p(w_1 \cdots w_m, w_{m+1}) \cdots p(w_{\ell-m} \cdots w_{\ell-1}, w_\ell),\end{aligned}$$

and

$$\begin{aligned}&p^{(k)}(w_{\ell-m+1} \cdots w_\ell, w_1 \cdots w_m) \\ &\quad = \mathrm{P}(X_{m+k} = w_{\ell-m+1}, \dots, X_{2m+k-1} = w_\ell \mid X_1 = w_1, \dots, X_m = w_m)\end{aligned}$$

denotes the k-step transition probability from $w_1 \cdots w_m$ to $w_{\ell-m+1} \cdots w_\ell$.

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

In model Mm-3 we consider a Markov chain of order m with finite state space $\mathcal{A}$, with three different strictly positive transition matrices

$$\Pi_k = (p_k(a_1 \cdots a_m, a_{m+1}))_{a_1,\ldots,a_{m+1}\in\mathcal{A}}, \qquad k \in \{1, 2, 3\},$$

where

$$p_k(a_1 \cdots a_m, a_{m+1}) = \mathrm{P}(X_{3j+k} = a_{m+1} \mid X_{3j+k-m} = a_1, \ldots, X_{3j+k-1} = a_m),$$

and with stationary distribution μ defined on $\mathcal{A}^m \times \{1, 2, 3\}$. From the interpretation of the model, it follows that we are interested in the number of occurrences of w starting at position k within the codon, i.e. in the word count

$$N(w, k) := \sum_{i=1}^{n-\ell+1} Y_i(w)\, \mathbf{1}_{\{i \bmod 3 = k\}}, \qquad k \in \{1, 2, 3\}.$$

Additionally, we assume that n and ℓ are both multiples of 3 and that the first letter of the sequence $X_1 \cdots X_n$ is the beginning of a codon. Under these conditions, it can be shown that $N(w, k) - \mathrm{E}_\mu N(w, k)$ satisfies an MDP with rate function $\Lambda_m^*(q) = q^2/2\sigma_m^2(w, k)$. To shorten the analysis, we only state the result for $k = 1$, but analogous results are valid for $k = 2$ and $k = 3$.

Theorem 3. *In model Mm-3 and under the preceding conditions, $N(w, 1) - \mathrm{E}_\mu N(w, 1)$ satisfies an MDP with rate function $\Lambda_m^*(q) = q^2/2\sigma_m^2(w, 1)$, where*

$$\begin{aligned}\sigma_m^2(w, 1) = {} & \frac{1}{3}\mu_m(w, 1) + \frac{2}{3} \sum_{\substack{p \bmod 3 = 3 \\ p\in\mathcal{P}(w)}} \mu_m(w w_{(p)}, 1) - \frac{2\ell - 3}{9}\mu_m^2(w, 1) \\ & + \frac{2}{3}\frac{\mu_m^2(w, 1)}{\mu(w_1 \ldots w_m, 1)} \sum_{\substack{j \bmod 3 = 1 \\ j=1}}^{\infty} (p_1^{(j)}(w_{\ell-m+1} \cdots w_\ell, w_1 \cdots w_m) \\ & \qquad - \mu(w_1 \cdots w_m, 1)),\end{aligned}$$

$$\begin{aligned}& p_1^{(j)}(w_{\ell-m+1} \cdots w_\ell, w_1 \cdots w_m) \\ & \quad = \mathrm{P}(X_{3i+j} = w_1, \ldots, X_{3i+j+m-1} = w_m \mid X_{3i+1-m} = w_{\ell-m+1}, \ldots, X_{3i} = w_\ell),\end{aligned}$$

and

$$\mu_m(w, 1) = \mu(w_1 \cdots w_m, 1)\, p_2(w_1 \cdots w_m, w_{m+1}) \cdots p_3(w_{\ell-m} \cdots w_{\ell-1}, w_\ell).$$

If we want to consider motifs or patterns rather than precise words, i.e. a family $\{w^1, \ldots, w^d\}$ of d words of equal length ℓ, then the following result can be shown in analogy to Theorem 1.

Theorem 4. *Under the conditions of Theorem 1 and for any initial distribution, the vector $(N(w^1) - \mathrm{E}_\mu N(w^1), \ldots, N(w^d) - \mathrm{E}_\mu N(w^d))$, $w^i \in \mathcal{W}$ for all $i \in \{1, \ldots, d\}$, satisfies an MDP with rate function $\Lambda_1^*(q) = \frac{1}{2}\langle q, \Sigma_1^{-1}(w)q\rangle$, where*

$$\Sigma_1(w) = (\Sigma_1(w^i, w^j))_{i,j\in\{1,\ldots,d\}},$$

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

$$\Sigma_1(w^i, w^j) = \mu_1(w^i)\,\mathbf{1}_{\{i=j\}} + \sum_{q\in\mathcal{P}(w^i,w^j)} \mu_1(w^i w^j_{(q)}) + \sum_{q\in\mathcal{P}(w^j,w^i)} \mu_1(w^j w^i_{(q)}) + \frac{\mu_1(w^i)\mu_1(w^j)}{\mu(w^j_1)} \sum_{k=1}^{\infty} (p^{(k)}(w^i_\ell, w^j_1) - \mu(w^j_1)) + \frac{\mu_1(w^i)\mu_1(w^j)}{\mu(w^i_1)} \sum_{k=1}^{\infty} (p^{(k)}(w^j_\ell, w^i_1) - \mu(w^i_1)) - (2\ell-1)\mu_1(w^i)\mu_1(w^j),$$

and $\mathcal{P}(w^i, w^j) = \{p \in \{1, \dots, \ell-1\}\colon w^j_k = w^i_{k+p}$ *for all* $k = 1, \dots, \ell - p\}$ *denotes the set of overlaps between* w^i *and* w^j.

Remarks 3. (i) In analogy to Remarks 2(i), $\Sigma_1(w^i, w^j)$ can be converted to

$$\Sigma_1(w^i, w^j) = \mu_1(w^i)\,\mathbf{1}_{\{i=j\}} + \sum_{q\in\mathcal{P}(w^i,w^j)} \mu_1(w^i w^j_{(q)}) + \sum_{q\in\mathcal{P}(w^j,w^i)} \mu_1(w^j w^i_{(q)}) + \mu_1(w^i)\mu_1(w^j)\Bigg(\sum_{a\in\mathcal{A}} \mu(a)(m_{a,w^j_1} + m_{a,w^i_1}) - m_{w^i_\ell,w^j_1} - m_{w^j_\ell,w^i_1}\Bigg) - (2\ell-1)\mu_1(w^i)\mu_1(w^j), \tag{3}$$

where the mean first passage times $m_{a,b}$, $a, b \in \mathcal{A}$, are defined in Remarks 2(i) and can be computed by solving the system of linear equations given in (2).

(ii) The result can also be extended to the models Mm and Mm-3.

For the special case of model M0 we obtain the following corollary.

Corollary 2. *Let* $(X_n)_{n\in\mathbb{N}}$ *be a sequence of i.i.d. random variables with values in a finite alphabet* $\mathcal{A}$. *Then the vector* $(N(w^1) - \mathrm{E}\,N(w^1), \dots, N(w^d) - \mathrm{E}\,N(w^d))$ *satisfies an MDP with rate function* $\Lambda_0^*(q) = \frac{1}{2}\langle q, \Sigma_0^{-1}(w)q\rangle$, *where* $\Sigma_0(w) = (\Sigma_0(w^i, w^j))_{i,j\in\{1,\dots,d\}}$,

$$\Sigma_0(w^i, w^j) = \mu_0(w^i)\,\mathbf{1}_{\{i=j\}} + \sum_{q\in\mathcal{P}(w^i,w^j)} \mu_0(w^i w^j_{(q)}) + \sum_{q\in\mathcal{P}(w^j,w^i)} \mu_0(w^j w^i_{(q)}) - (2\ell-1)\mu_0(w^i)\mu_0(w^j),$$

and $\mathcal{P}(w^i, w^j)$ *is given in Theorem 4.*

In the following section we will focus on the proof of Theorem 1.

3. Proof of the main result

3.1. Letter occurrences

Before being able to prove Theorem 1, we address the special case of an MDP for counts of single-letter occurrences, i.e. an MDP for $\tilde{N}(w) = \sum_{i=1}^n \mathbf{1}_{\{X_i=w\}}$, where $w \in \mathcal{A}$ is a single letter.

The advantage of this special case is that, in contrast to the general case of counting words of arbitrary length ℓ, there is no extra $(\ell-1)$-dependence of word occurrences added to the dependency structure of the Markov chain.

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

Proposition 1. *Given an irreducible, aperiodic, and homogeneous Markov chain $(X_n)_{n\in\mathbb{N}}$ with finite state space $\mathcal{A}$, transition matrix $\Pi = (p(\alpha,\beta))_{\alpha,\beta\in\mathcal{A}}$, and stationary distribution μ, let*

$$\tilde{N}(w) := \sum_{i=1}^{n} \tilde{Y}_i(w), \quad \text{where} \quad \tilde{Y}_i(w) := \mathbf{1}_{\{X_i=w\}}$$

for all $w \in \mathcal{A}$. Then, for any initial distribution, $\tilde{N}(w) - \mathrm{E}_\mu \tilde{N}(w)$ satisfies an MDP with rate function $\tilde{\Lambda}^*(q) = q^2/2\tilde{\sigma}^2(w)$, where

$$\tilde{\sigma}^2(w) = \lim_{n\to\infty} \frac{1}{n} \mathrm{E}_\mu(\tilde{N}(w) - \mathrm{E}_\mu \tilde{N}(w))^2,$$

and this limit exists.

A result more general than Proposition 1 has been shown by Djellout and Guillin [6], but instead of verifying the conditions stated in their theorem we want to establish a new proof similar to theirs but specific to our situation.

The proof relies on the so-called regeneration method developed by Chung [4, p. 94]. The idea of this method is to decompose the Markov chain into i.i.d. random blocks between visits to a fixed state.

As the underlying Markov chain is finite and irreducible, all states are positive recurrent. Thus, the regeneration method is applicable.

Proof of Proposition 1. Fix an arbitrary $a \in \mathcal{A}$. For all $j, k, n \in \mathbb{N}$, define

$$\begin{aligned}
\tau &:= \tau(1) := \inf\{n \in \mathbb{N} \colon X_n = a\},\\
\tau(k+1) &:= \inf\{n > \tau(k) \colon X_n = a\},\\
\tilde{Z}_j(w) &:= \tilde{Y}_j(w) - \mathrm{E}_\mu \tilde{Y}_j(w),\\
\tilde{\xi}_k(w) &:= \sum_{j=\tau(k)+1}^{\tau(k+1)} \tilde{Z}_j(w),\\
\ell(n) &:= \tau(\tilde{N}(a) \vee 1),\\
e(n) &:= \lfloor n\mu(a) \rfloor.
\end{aligned}$$

Note that both $(\tau(k+1) - \tau(k))_{k\in\mathbb{N}}$ and $(\tilde{\xi}_k(w))_{k\in\mathbb{N}}$ are sequences of i.i.d. random variables with common laws $\mathcal{L}_{P_a}(\tau)$ and $\mathcal{L}_{P_a}(\sum_{k=1}^{\tau} \tilde{Z}_k(w))$, respectively (see [12, Theorem 17.3.1]). The decomposition according to the regeneration method (see [4, p. 94]) is the following:

$$\begin{aligned}
&\tilde{N}(w) - \mathrm{E}_\mu \tilde{N}(w)\\
&\quad= \sum_{k=1}^{n} \tilde{Z}_k(w)\\
&\quad= \sum_{k=1}^{\tau\wedge n} \tilde{Z}_k(w) + \sum_{k=1}^{\tilde{N}(a)-1} \tilde{\xi}_k(w) + \sum_{k=\ell(n)+1}^{n} \tilde{Z}_k(w)\\
&\quad= \sum_{k=1}^{e(n)} \tilde{\xi}_k(w) + \sum_{k=1}^{\tau\wedge n} \tilde{Z}_k(w) + \sum_{k=l(n)+1}^{n} \tilde{Z}_k(w) + \left(\sum_{k=1}^{\tilde{N}(a)-1} \tilde{\xi}_k(w) - \sum_{k=1}^{e(n)} \tilde{\xi}_k(w) \right).
\end{aligned}$$

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

We proceed by analysing each of these four summands and by showing that only the first summand contributes to the moderate deviation behaviour. More precisely, we will prove the following assertions:

(i) $\sum_{k=1}^{e(n)} \tilde{\xi}_k(w)$ satisfies an MDP with rate function $\tilde{\Lambda}^*(q) = q^2/2\tilde{\sigma}^2(w)$, where

$$\tilde{\sigma}^2(w) = \mu(a)\,\mathrm{E}_a\left(\sum_{k=1}^{\tau} \tilde{Z}_k(w)\right)^2;$$

(ii) $$\limsup_{n\to\infty} \frac{1}{n^{2\alpha-1}} \log \mathrm{P}\left(\left|\sum_{k=1}^{\tau\wedge n} \tilde{Z}_k(w)\right| > \varepsilon n^{\alpha}\right) = -\infty;$$

(iii) $$\limsup_{n\to\infty} \frac{1}{n^{2\alpha-1}} \log \mathrm{P}\left(\left|\sum_{k=\ell(n)+1}^{n} \tilde{Z}_k(w)\right| > \varepsilon n^{\alpha}\right) = -\infty;$$

(iv) $$\limsup_{n\to\infty} \frac{1}{n^{2\alpha-1}} \log \mathrm{P}\left(\left|\sum_{k=1}^{\tilde{N}(a)-1} \tilde{\xi}_k(w) - \sum_{k=1}^{e(n)} \tilde{\xi}_k(w)\right| > \varepsilon n^{\alpha}\right) = -\infty;$$

for any initial distribution and all $\varepsilon > 0$. These assertions imply exponential equivalence of $(\tilde{N}(w) - \mathrm{E}_\mu\,\tilde{N}(w))_{n\in\mathbb{N}}$ and $(\sum_{k=1}^{e(n)} \tilde{\xi}_k(w))_{n\in\mathbb{N}}$. Thus, Proposition 1 follows from [5, Theorem 4.2.13], and the fact that

$$\tilde{\sigma}^2(w) = \mu(a)\,\mathrm{E}_a\left(\sum_{k=1}^{\tau} \tilde{Z}_k(w)\right)^2 = \lim_{n\to\infty} \frac{1}{n}\,\mathrm{E}_\mu\left(\tilde{N}(w) - \mathrm{E}_\mu\,\tilde{N}(w)\right)^2. \tag{4}$$

The last equation is valid as $\mathrm{E}_a(\sum_{k=1}^{\tau}|\tilde{Z}_k(w)|)^2 \le \mathrm{E}_a\,\tau^2$ and since $(X_n)_{n\in\mathbb{N}}$ is a finite, irreducible, and aperiodic Markov chain. Namely, this yields the existence of $N \in \mathbb{N}$ with $\Pi^N \gg 0$. Let $\rho := \min_{\alpha,\beta} \Pi^N(\alpha,\beta) > 0$. Then we have

$$\mathrm{P}_a(\tau > kN) \le (1-\rho)^k. \tag{5}$$

Hence, $\mathrm{E}_a\,\tau^2 < \infty$ and (4) follows from [3, pp. 45ff.].

Proof of assertion (i). As mentioned above, $(\tilde{\xi}_k(w))_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables with common law $\mathcal{L}_{\mathrm{P}_a}(\sum_{k=1}^{\tau} \tilde{Z}_k(w))$. Furthermore, $\log \mathrm{E}_a\,\mathrm{e}^{\lambda\tilde{\xi}_1(w)} \le \mathrm{E}_a\,\mathrm{e}^{\lambda\tau} < \infty$ (compare (5)) in some neighbourhood around 0. Thus, a classical MDP for sums of i.i.d. random variables (see, e.g. [5, Theorem 3.7.1]) is applicable, if we substitute $e(n)$ for n. As $e(n)/n = \lfloor n\mu(a)\rfloor/n \to \mu(a)$, we find that $\sum_{k=1}^{e(n)} \tilde{\xi}_k(w)$ satisfies an MDP with rate function $\tilde{\Lambda}^*(q) = q^2/2\tilde{\sigma}^2(w)$, where $\tilde{\sigma}^2(w) = \mu(a)\,\mathrm{E}_a(\sum_{k=1}^{\tau} \tilde{Z}_k(w))^2$.

Proof of assertion (ii). We have $|\sum_{k=1}^{\tau\wedge n} \tilde{Z}_k(w)| \le \tau$. Since $(X_n)_{n\in\mathbb{N}}$ is a homogeneous, irreducible, aperiodic Markov chain with finite state space, we obtain (similarly to (5))

$$\mathrm{P}\left(\left|\sum_{k=1}^{\tau\wedge n} \tilde{Z}_k(w)\right| > \varepsilon n^{\alpha}\right) \le \mathrm{P}(\tau > \varepsilon n^{\alpha}) \le (1-\rho)^{\varepsilon n^{\alpha}/N},$$

where $\rho = \min_{\alpha,\beta} \Pi^N(\alpha,\beta) > 0$. Assertion (ii) follows as an immediate consequence.

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

Proof of assertion (iii). We have

$$\begin{aligned}\Big|\sum_{k=\ell(n)+1}^{n}\tilde{Z}_k(w)\Big| &\le \sum_{k=\ell(n)+1}^{n}|\tilde{Z}_k(w)|\\ &\le \sum_{k=\tau(\tilde{N}(a))+1}^{\tau(\tilde{N}(a)+1)}|\tilde{Z}_k(w)|\\ &\le \max_{1\le k\le n}\sum_{k=\tau(k)+1}^{\tau(k+1)}|\tilde{Z}_k(w)|\\ &\le \max_{1\le k\le n}(\tau(k+1)-\tau(k)).\end{aligned}$$

As $(\tau(k+1)-\tau(k))_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables with common law $\mathcal{L}_{\mathrm{P}_a}(\tau)$, we obtain

$$\begin{aligned}\mathrm{P}\Big(\Big|\sum_{k=\ell(n)+1}^{n}\tilde{Z}_k(w)\Big|>\varepsilon n^{\alpha}\Big) &\le \mathrm{P}\Big(\max_{1\le k\le n}(\tau(k+1)-\tau(k))>\varepsilon n^{\alpha}\Big)\\ &\le \sum_{k=1}^{n}\mathrm{P}(\tau(k+1)-\tau(k)>\varepsilon n^{\alpha})\\ &= n\,\mathrm{P}_a(\tau>\varepsilon n^{\alpha}).\end{aligned}$$

Since $\log n/n^{2\alpha-1}\to 0$ if $n\to\infty$, assertion (iii) follows analogously to (ii).

Proof of assertion (iv). Let $\delta\in(0,\mu(a))$ be fixed but arbitrary. Choose n large enough such that $e(n)\ge\delta n$. Conditioned on $|\tilde{N}(a)-1-e(n)|\le\delta n$ we have

$$\Big|\sum_{k=1}^{\tilde{N}(a)-1}\tilde{\xi}_k(w)-\sum_{k=1}^{e(n)}\tilde{\xi}_k(w)\Big|\le 2\max_{e(n)-\lfloor\delta n\rfloor\le j\le e(n)+\lfloor\delta n\rfloor}\Big|\sum_{k=e(n)-\lfloor\delta n\rfloor}^{j}\tilde{\xi}_k(w)\Big|,$$

and, thus,

$$\begin{aligned}&\mathrm{P}\Big(\Big|\sum_{k=1}^{\tilde{N}(a)-1}\tilde{\xi}_k(w)-\sum_{k=1}^{e(n)}\tilde{\xi}_k(w)\Big|>\varepsilon n^{\alpha}\Big)\\ &\quad\le \mathrm{P}\Big(\max_{1\le j\le 2\lfloor\delta n\rfloor}\Big|\sum_{k=1}^{j}\tilde{\xi}_k(w)\Big|>\frac{\varepsilon}{2}n^{\alpha}\Big)+\mathrm{P}(\tilde{N}(a)-1-e(n)>\delta n)\\ &\qquad+\mathrm{P}(\tilde{N}(a)-1-e(n)<-\delta n)\\ &\quad=: I+II+III.\end{aligned}$$

Let us consider I first. Applying Ottaviani's inequality (see [11, Lemma 6.2]) for the independent random variables $(\tilde{\xi}_k(w))_{k\in\mathbb{N}}$, we obtain

$$\mathrm{P}\Big(\max_{1\le j\le 2\lfloor\delta n\rfloor}\Big|\sum_{k=1}^{j}\tilde{\xi}_k(w)\Big|>\frac{\varepsilon}{2}n^{\alpha}\Big)\le\frac{\mathrm{P}\big(\big|\sum_{k=1}^{2\lfloor\delta n\rfloor}\tilde{\xi}_k(w)\big|>\varepsilon n^{\alpha}/4\big)}{1-\max_{1\le j\le 2\lfloor\delta n\rfloor}\mathrm{P}\big(\big|\sum_{k=j+1}^{2\lfloor\delta n\rfloor}\tilde{\xi}_k(w)\big|>\varepsilon n^{\alpha}/4\big)}.$$

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

From the central limit theorem, it follows that $(1/n^\alpha)\sum_{k=1}^n \tilde{\xi}_k(w) \to 0$ in probability. Thus, for sufficiently large n, we obtain

$$\max_{1\le j\le 2\lfloor \delta n\rfloor} \mathrm{P}\biggl(\biggl|\sum_{k=j+1}^{2\lfloor \delta n\rfloor} \tilde{\xi}_k(w)\biggr| > \frac{\varepsilon}{4}n^\alpha\biggr) \le \frac{1}{2},$$

and, hence,

$$\mathrm{P}\biggl(\max_{1\le j\le 2\lfloor \delta n\rfloor}\biggl|\sum_{k=1}^{j} \tilde{\xi}_k(w)\biggr| > \frac{\varepsilon}{2}n^\alpha\biggr) \le 2\,\mathrm{P}\biggl(\biggl|\sum_{k=1}^{2\lfloor \delta n\rfloor} \tilde{\xi}_k(w)\biggr| > \frac{\varepsilon}{4}n^\alpha\biggr). \tag{6}$$

Similarly to the proof of assertion (i) with $2\lfloor \delta n\rfloor$ instead of $e(n) = \lfloor n\mu(a)\rfloor$, we can show that $\sum_{k=1}^{2\lfloor \delta n\rfloor} \tilde{\xi}_k(w)$ satisfies an MDP with rate function

$$\tilde{\Lambda}^*_\delta(q) = \frac{q^2}{2(2\delta)\,\mathrm{E}_a\,\tilde{\xi}_1^2(w)}.$$

Choosing $F = \{x\colon |x| \ge \varepsilon/4\}$ as an open set and letting $\delta \to 0$, we obtain

$$\lim_{\delta\to 0}\limsup_{n\to\infty} \frac{1}{n^{2\alpha-1}}\log \mathrm{P}\biggl(\biggl|\sum_{k=1}^{2\lfloor \delta n\rfloor} \tilde{\xi}_k(w)\biggr| \ge \frac{\varepsilon}{4}n^\alpha\biggr) = -\infty,$$

and, consequently (see (6)),

$$\lim_{\delta\to 0}\limsup_{n\to\infty} \frac{1}{n^{2\alpha-1}}\log \mathrm{P}\biggl(\max_{1\le j\le 2\lfloor \delta n\rfloor}\biggl|\sum_{k=1}^{j} \tilde{\xi}_k(w)\biggr| > \frac{\varepsilon}{2}n^\alpha\biggr) = -\infty.$$

Let us consider *II*. We have

$$\begin{aligned}\mathrm{P}(\tilde{N}(a) - 1 - e(n) > \delta n) &\le \mathrm{P}(\tau(e(n) + \lfloor \delta n\rfloor) \le n)\\ &\le \mathrm{P}\biggl(\sum_{k=1}^{e(n)+\lfloor \delta n\rfloor} (\tau(k+1) - \tau(k)) \le n\biggr)\\ &= \mathrm{P}\biggl(\frac{1}{e(n)+\lfloor \delta n\rfloor}\sum_{k=1}^{e(n)+\lfloor \delta n\rfloor} (\tau(k+1) - \tau(k)) \le \frac{n}{e(n)+\lfloor \delta n\rfloor}\biggr).\end{aligned}$$

As $(\tau(k+1)-\tau(k))_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables with expectation $\mathrm{E}_a\,\tau = 1/\mu(a)$ and as, for sufficiently large n,

$$\frac{n}{e(n)+\lfloor \delta n\rfloor} = \frac{n}{\lfloor n\mu(a)\rfloor + \lfloor \delta n\rfloor} < \frac{1}{\mu(a)} = \mathrm{E}_a\,\tau,$$

the LDP of Cramér (see [5, Theorem 2.2.3]) is applicable. Since the rate function $I(q) = \sup_{t\in\mathbb{R}}(tq - \log \mathrm{E}_a\,\mathrm{e}^{t\tau})$ governing the large deviations satisfies $I(x) \ge 0$ and $I(x) = 0$ if and only if $x = \mathrm{E}_a\,\tau$, for sufficiently large n, there exists $c > 0$ with

$$\mathrm{P}\biggl(\frac{1}{e(n)+\lfloor \delta n\rfloor}\sum_{k=1}^{e(n)+\lfloor \delta n\rfloor} (\tau(k+1) - \tau(k)) \le \frac{n}{e(n)+\lfloor \delta n\rfloor}\biggr) \le \mathrm{e}^{-nc}.$$

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

Hence,

$$\lim_{n\to\infty} \frac{1}{n^{2\alpha-1}} \log \mathrm{P}(\tilde{N}(a) - 1 - e(n) > \delta n) = -\infty.$$

We approximate *III* analogously. Namely, we find that

$$\lim_{n\to\infty} \frac{1}{n^{2\alpha-1}} \log \mathrm{P}(\tilde{N}(a) - 1 - e(n) < -\delta n) = -\infty.$$

Combining the approximations for *I*, *II*, and *III*, which all hold uniformly in δ, assertion (iv) finally follows. This completes the proof of Proposition 1.

3.2. Word occurrences

In order to generalise the MDP for single-letter counts (see Proposition 1) to an MDP for counts of words of arbitrary length ℓ (see Theorem 1), we consider the Markov chain $(\mathbb{X}_n)_{n\in\mathbb{N}}$, $\mathbb{X}_n = X_n \cdots X_{n+\ell-1}$, with state space $\mathcal{A}^\ell$ and with appropriate transition probabilities (see below). In $\mathcal{A}^\ell$ a word $w = w_1 \cdots w_\ell$ corresponds to a single letter. If we restrict the state space to the set

$$\mathcal{W} = \{a_1 \cdots a_\ell \in \mathcal{A}^\ell \colon p(a_i, a_{i+1}) > 0 \text{ for all } i \in \{1, \dots, \ell - 1\}\}$$

then we can show that $(\mathbb{X}_n)_{n\in\mathbb{N}}$ is an irreducible, aperiodic, and homogeneous Markov chain. This enables us to apply Proposition 1.

Proof of Theorem 1. Consider the Markov chain $(\mathbb{X}_n)_{n\in\mathbb{N}}$, $\mathbb{X}_n = X_n \cdots X_{n+\ell-1}$, with state space $\mathcal{A}^\ell$ and with transition matrix $\tilde{\Pi} := (\tilde{p}(a, b))_{a,b\in\mathcal{A}^\ell}$, where

$$\tilde{p}(a_1 \cdots a_\ell, b_1 \cdots b_\ell) := \begin{cases} p(a_\ell, b_\ell) & \text{if } a_{j+1} = b_j,\ j = 1, \dots, \ell - 1, \\ 0 & \text{otherwise,} \end{cases}$$

for all $a_1, \dots, a_\ell, b_1, \dots, b_\ell \in \mathcal{A}$.

Given the entire state space $\mathcal{A}^\ell$, $\tilde{\Pi}$ is not necessarily irreducible and aperiodic since, for any $M \in \mathbb{N}$, the M-step transition probability $\tilde{p}^{(M)}(a_1 \cdots a_\ell, b_1 \cdots b_\ell)$ cannot be positive if one of the probabilities $p(b_i, b_{i+1})$ is equal to 0. Thus, we restrict the state space to the set $\mathcal{W}$. As $(X_n)_{n\in\mathbb{N}}$ is irreducible and aperodic, there exists $N \in \mathbb{N}$ with $\Pi^n \gg 0$ for all $n \geq N$. Let $n \geq N$ and $a_1 \cdots a_\ell, b_1 \cdots b_\ell \in \mathcal{W}$. Then, by definition of the transition matrix $\tilde{\Pi}$ we have

$$\tilde{p}^{(n+\ell-1)}(a_1 \cdots a_\ell, b_1 \cdots b_\ell) = p^{(n)}(a_\ell, b_1) p(b_1, b_2) p(b_2, b_3) \cdots p(b_{\ell-1}, b_\ell) > 0.$$

Thus, there exists an $M \in \mathbb{N}$ satisfying $\tilde{\Pi}^n_{\mathrm{res}} \gg 0$ for all $n \geq M$, where $\tilde{\Pi}_{\mathrm{res}}$ denotes the restriction of $\tilde{\Pi}$ to $\mathcal{W}$. Hence, $(\mathbb{X}_n)_{n\in\mathbb{N}}$ with state space $\mathcal{W}$ is an irreducible, aperiodic, and homogeneous Markov chain.

Consequently, we can apply Proposition 1. As

$$\{X_i = w_1, \dots, X_{i+\ell-1} = w_\ell\} = \{\mathbb{X}_i = w\}$$

for all $w_1 \cdots w_\ell \in \mathcal{W}$ and, thus, $\tilde{Y}_i(w)$ and $\tilde{N}(w)$ (as functions of $\mathbb{X}_1, \mathbb{X}_2, \dots, \mathbb{X}_n$) correspond to the random variables $Y_i(w)$ and $N(w)$, for any initial distribution and all $w \in \mathcal{W}$, the centred word count $N(w) - \mathrm{E}_\mu N(w)$ satisfies an MDP with rate function $\Lambda_1^*(q) = q^2/2\sigma_1^2(w)$, where

$$\sigma_1^2(w) = \lim_{n\to\infty} \frac{\mathrm{E}_\mu(N(w) - \mathrm{E}_\mu N(w))^2}{n} = \lim_{n\to\infty} \frac{\mathrm{var}_\mu N(w)}{n}.$$

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

Kleffe and Borodovsky [9] computed the second moment of $N(w)$. A nice representation of $\operatorname{var}_\mu N(w)$ is given in [19, Equation (6.4.1)]:

$$\operatorname{var}_\mu N(w) = \mathrm{E}_\mu N(w) + 2 \sum_{q \in \mathcal{P}(w)} \mathrm{E}_\mu N(ww_{(q)}) - (\mathrm{E}_\mu N(w))^2 + \frac{2}{\mu(w_1)} \mu_1^2(w) \sum_{k=1}^{n-2\ell+1} (n - 2\ell + 2 - k) p^{(k)}(w_\ell, w_1).$$

Since $\mathrm{E}_\mu N(w) = (n - l + 1)\mu_1(w)$, this yields

$$\sigma_1^2(w) = \mu_1(w) + 2 \sum_{q \in \mathcal{P}(w)} \mu_1(ww_{(q)}) + \lim_{n\to\infty} \left(\frac{2}{\mu(w_1)} \mu_1^2(w) \frac{1}{n} \sum_{k=1}^{n-2\ell+1} (n - 2\ell + 2 - k) p^{(k)}(w_\ell, w_1) - \frac{(n - \ell + 1)^2}{n} \mu_1^2(w) \right). \tag{7}$$

It can be easily computed that

$$(n - \ell + 1)^2 = 2 \sum_{k=1}^{n-2\ell+1} (n - 2\ell + 2 - k) + n(2\ell - 1) - 3\ell^2 + 4\ell - 1.$$

Thus, we obtain

$$\begin{aligned} &\lim_{n\to\infty} \left(\frac{2}{\mu(w_1)} \mu_1^2(w) \frac{1}{n} \sum_{k=1}^{n-2\ell+1} (n - 2\ell + 2 - k) p^{(k)}(w_\ell, w_1) - \frac{(n - \ell + 1)^2}{n} \mu_1^2(w) \right) \\ &\quad = \lim_{n\to\infty} \left(\frac{2}{\mu(w_1)} \mu_1^2(w) \sum_{k=1}^{n-2\ell+1} \frac{n - 2\ell + 2 - k}{n} (p^{(k)}(w_\ell, w_1) - \mu(w_1)) \right) \\ &\qquad - (2\ell - 1)\mu_1^2(w) \\ &\quad = \frac{2}{\mu(w_1)} \mu_1^2(w) \sum_{k=1}^{\infty} (p^{(k)}(w_\ell, w_1) - \mu(w_1)) - (2\ell - 1)\mu_1^2(w), \end{aligned} \tag{8}$$

where the last equation follows from Kronecker's lemma. Combining (7) and (8) completes the proof of Theorem 1.

Theorem 2 can be obtained analogously, applying Proposition 1 to the first-order Markov chain $(\mathbb{X}_n)_{n\in\mathbb{N}}$, $\mathbb{X}_n = X_n \dots X_{n+\ell-1}$, with state space $\mathcal{A}^\ell$ and with transition matrix $\tilde{\Pi} := (\tilde{p}(a, b))_{a,b\in\mathcal{A}^\ell}$, where this time

$$\tilde{p}(a_1 \cdots a_\ell, b_1 \cdots b_\ell) := \begin{cases} p(a_{\ell-m+1} \cdots a_\ell, b_\ell) & \text{if } a_{j+1} = b_j,\ j = 1, \dots, \ell - 1, \\ 0 & \text{otherwise.} \end{cases}$$

The proof of Theorem 3 is lengthy. However, the only new idea behind it aside from the methods used in the proof of Theorem 1 is to group the sequence $(X_n)_{n\in\mathbb{N}}$ into blocks of

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

length 3, i.e. $Y_1 := X_1X_2X_3$, $Y_2 := X_4X_5X_6$, etc., in order to create a homogeneous Markov chain $(Y_i)_{i\in\mathbb{N}}$. Thus, the proof is omitted. A similar approach can also be found in [22, Chapter 3].

The MDP for the number of occurrences of multiple patterns (see Theorem 4), i.e. the generalisation from dimension 1 (compare Theorem 1) to arbitrary dimension $d \in \mathbb{N}$, can be obtained easily as well.

For more details concerning the proofs of Theorems 2, 3, and 4, see [1].

4. Applications to biological sequence analysis

As mentioned in the introduction, given a probabilistic model and a biological sequence, we are interested in identifying exceptional words (or patterns), i.e. words occurring significantly often or rarely.

Let $N^{\mathrm{obs}}(w)$ denote the observed count of a word w of length ℓ in the given biological sequence. Applying Theorem 1, we obtain the following moderate deviation based approximations for the P-values $\mathrm{P}(N(w) \geq N^{\mathrm{obs}}(w))$ and $\mathrm{P}(N(w) \leq N^{\mathrm{obs}}(w))$, where $f(n) \simeq g(n)$ as $n \to \infty$ is short for $\lim_{n\to\infty}(1/n^{2\alpha-1})(\log(f(n)) - \log(g(n))) = 0$ (logarithmic equivalence).

Corollary 3. *For large n, under the conditions of Theorem 1, we have the following: if $N^{\mathrm{obs}}(w) \geq (n-\ell+1)\mu_1(w)$ then*

$$\mathrm{P}(N(w) \geq N^{\mathrm{obs}}(w)) \simeq \exp\left(-\frac{1}{2\sigma_1^2(w)}\frac{1}{n}(N^{\mathrm{obs}}(w) - (n-\ell+1)\mu_1(w))^2\right)$$

and if $N^{\mathrm{obs}}(w) \leq (n-\ell+1)\mu_1(w)$ then

$$\mathrm{P}(N(w) \leq N^{\mathrm{obs}}(w)) \simeq \exp\left(-\frac{1}{2\sigma_1^2(w)}\frac{1}{n}(N^{\mathrm{obs}}(w) - (n-\ell+1)\mu_1(w))^2\right),$$

where $\sigma_1^2(w)$ and $\mu_1(w)$ are defined in Theorem 1.

Proof. Choosing $F = [q, \infty)$ and $O = (q, \infty)$, $q \geq 0$, in Definition 1 and applying Theorem 1, we obtain

$$\mathrm{P}\left(\frac{N(w) - \mathrm{E}_\mu N(w)}{n^\alpha} \geq q\right) \simeq \exp\left(-n^{2\alpha-1}\frac{q^2}{2\sigma_1^2(w)}\right).$$

As $\mathrm{E}_\mu N(w) = (n-\ell+1)\mu_1(w)$, setting $q := (1/n^\alpha)(N^{\mathrm{obs}}(w) - (n-\ell+1)\mu_1(w))$, which is positive if $N^{\mathrm{obs}}(w) \geq (n-\ell+1)\mu_1(w)$, we obtain the first part of the corollary:

$$\begin{aligned}\mathrm{P}(N(w) \geq N^{\mathrm{obs}}(w)) &= \mathrm{P}\left(\frac{N(w) - \mathrm{E}_\mu N(w)}{n^\alpha} \geq q\right)\\ &= \exp\left(-\frac{1}{2\sigma_1^2(w)}\frac{1}{n}(N^{\mathrm{obs}}(w) - (n-\ell+1)\mu_1(w))^2\right).\end{aligned}$$

The second part follows analogously by choosing $F' = (-\infty, q]$ and $O' = (-\infty, q)$, $q \leq 0$ ($q = (1/n^\alpha)(N^{\mathrm{obs}}(w) - (n-\ell+1)\mu_1(w)) \leq 0$ if $N^{\mathrm{obs}}(w) \leq (n-\ell+1)\mu_1(w)$).

Remark 4. Analogous moderate deviation based approximations for the P-values are also valid in models Mm and Mm-3 (compare Theorems 2 and 3).

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

TABLE 1: Homo sapiens α-globin gene cluster (HBB), mRNA, base range 1–600 (sequence extracted from the National Center for Biotechnology Information; accession number NM_000006).

```
  1   GATCACGCCATTGCACTCCACCCTGGGCGACAGAGCGACGAGACCCCGTATCAAAAAAAA
 61   AAAAAAGAAAGAAAGAAAGAAAAAAGAAAAAAAAAAAGGCCGGGCGCGGTGGCTCACGCCT
121   GTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGAATCACGAGGTCAGGAGTTCGAGACC
181   ATCCTGGCCAACATGGTGAAACCCCGTCTCTACAAAAAAAAAAAAAAAAAATTAGCCGGGC
241   GTGGTGGCGGGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGACAGGAAAATCGCTTGA
301   ACCCGGGAGGCGGAGCTTGCGGTGAGCCGAGATTGCGCCACTGCACTACAGCCTAGGCGA
361   CAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAACACTTGGAAGCCGACAG
421   GAGATCTTTGAGACCTTGGGCGAGGCAGTGACACTAAAGGCAGGAGCGACTACAGAAGAA
481   TAAATTAAACTTCATCAGATTAAAAACTTTACTGCGGCCGGGCGCGGTGGCTCACGCCTG
541   AAATCCCAGCACTTTGGGAGGCCGAGGTGGGCAGATCATGAGATCAGGAGATCTAGACCA
601   ...
```

To illustrate the possibility of applying these approximations, we consider the following example.

Example 1. (*Exceptional words in the human α-globin gene cluster.*) The protein hemoglobin is responsible for the oxygen transport in human red blood cells and consists of four subunits: two α-globin subunits and two β-globin subunits. When considering the DNA sequence (mRNA) of the human α-globin gene cluster (located on chromosome 16), which is 43 058 bases long, we recognise that the word 'CAGG' appears 435 times (the first 600 bases of the α-globin gene cluster can be seen in Table 1). Is this significantly frequent?

To answer this question, we have to determine the model parameters. We have $n = 43\,058$, $w = \text{CAGG}$, $N^{\text{obs}}(w) = 435$, and $\mathcal{P}(w) = \varnothing$. An estimator (MLE) for the transition probabilities $p(a, b)$, $a, b \in \mathcal{A} = \{\text{A,C,G,T}\}$, is given by

$$\hat{p}(a, b) = \frac{N^{\text{obs}}(ab)}{\sum_{c \in \mathcal{A}} N^{\text{obs}}(ac)}$$

(see [19, Section 6.1.2]). Applying this, we obtain the following estimation for the transition matrix Π:

$$\Pi = \begin{pmatrix} 0.29 & 0.22 & 0.32 & 0.17 \\ 0.28 & 0.33 & 0.13 & 0.26 \\ 0.22 & 0.26 & 0.35 & 0.17 \\ 0.16 & 0.26 & 0.35 & 0.23 \end{pmatrix}$$

(in the following we will always state rounded values but continue our calculations with the nonrounded values).

In addition, we need to compute the stationary distribution μ and (referring to (1)) the first passage times $m_{a,b}$, $a, b \in \mathcal{A}$. The first passage times $m_{a,b}$, $a, b \in \mathcal{A}$, can be computed by solving the linear equations

$$m_{a,b} = 1 + \sum_{c \in \mathcal{A} \setminus \{b\}} p(a, c) m_{c,b},$$

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

yielding

$$M = (m_{a,b})_{a,b\in\mathcal{A}} = \begin{pmatrix} 4.19 & 4.11 & 3.60 & 5.17 \\ 4.25 & 3.70 & 4.33 & 4.65 \\ 4.46 & 3.96 & 3.52 & 5.18 \\ 4.79 & 3.94 & 3.51 & 4.83 \end{pmatrix}.$$

Since $\mu(a) = 1/m_{a,a}$ for all $a \in \mathcal{A}$, we also obtain

$$\mu(A) = 0.24, \qquad \mu(C) = 0.27, \qquad \mu(G) = 0.28, \qquad \mu(T) = 0.21.$$

Now we can compute the parameters $\mu_1(\text{CAGG})$ and $\sigma_1^2(\text{CAGG})$ (see (1)):

$$\begin{aligned} \mu_1(\text{CAGG}) &= \mu(\text{C})\,p(\text{C, A})\,p(\text{A, G})\,p(\text{G, G}) = 0.008\,31, \\ \sigma_1^2(\text{CAGG}) &= \mu_1(\text{CAGG}) - 7\mu_1^2(\text{CAGG}) + 2\mu_1^2(\text{CAGG})\left(\sum_{a\in\mathcal{A}} \mu(a)m_{a,\text{C}} - m_{\text{G,C}}\right) \\ &= 0.007\,82. \end{aligned}$$

As $N^{\text{obs}}(w) = 435 \geq 357.73 = (n - l + 1)\mu_1(w)$, applying Corollary 3 yields

$$\begin{aligned} &\text{P}(N(\text{CAGG}) \geq 435) \\ &\quad \simeq \exp\left(-\frac{1}{2\sigma_1^2(\text{CAGG})}\frac{1}{43\,058}(435 - 43\,055\mu_1(\text{CAGG}))^2\right) \\ &\quad = 0.000\,14. \end{aligned}$$

Setting a significance level of 0.02%, we can conclude that, given model M1, the word 'CAGG' occurs significantly often.

In order to compare this moderate deviation approximation based P-value with the exact as well as the Gaussian and large deviation approximation based P-values, we use the programme SPatt (statistic for patterns) which has been developed by Nuel [14] (see also http://stat.genopole.cnrs.fr/spatt/). The programme outputs the following P-values for the word 'CAGG' to occur 435 times or more in a sequence of length 43 058, assuming a Markov chain of order 1 with its parameters estimated from the input sequence, i.e. the sequence of the human α-globin gene cluster:

exact P-value (X-SPatt):	1e-4.61 $\approx$ 0.000 1;
Gaussian approximation based P-value (G-SPatt):	1e-6.29 $\approx$ 0.000 001;
large deviation approximation based P-value (LD-SPatt):	1e-3.61 $\approx$ 0.001.

Thus, we can conclude that our P-value of 0.000 14 approximates the exact P-value very well and that it performs much better than the Gaussian and large deviation based approximations (with regard to this example).

For occurrences of patterns $\{w^1, \ldots, w^d\}$ (see Theorem 4), moderate deviation based approximations can be derived similarly to Corollary 3 by choosing $F = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i \geq q\}$ and $O = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i > q\}$ (or $F' = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i \leq q\}$ and $O' = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i < q\}$; see the proof of Corollary 3).

Corollary 4. *For large n, under the conditions of Theorem 1 and for all* $\frac{1}{2} < \alpha < 1$, *we have*

$$\text{P}\left(\sum_{i=1}^d N(w^i) \geq \sum_{i=1}^d N^{\text{obs}}(w^i)\right) \simeq \exp\left(-\frac{1}{2}n^{2\alpha-1} \inf_{\sum_{i=1}^d x_i \geq q} \langle x, \Sigma_1^{-1}(w)x\rangle\right)$$

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

and

$$\mathrm{P}\Bigg(\sum_{i=1}^{d} N(w^i) \le \sum_{i=1}^{d} N^{\mathrm{obs}}(w^i)\Bigg) \simeq \exp\Bigg(-\frac{1}{2}n^{2\alpha-1} \inf_{\sum_{i=1}^{d} x_i \le q} \langle x, \Sigma_1^{-1}(w)x\rangle\Bigg),$$

where

$$q = \frac{1}{n^{\alpha}}\Bigg(\sum_{i=1}^{d}(N^{\mathrm{obs}}(w^i) - (n-l+1)\mu_1(w^i))\Bigg)$$

and $\Sigma_1(w)$ *is given in Theorem 4.*

Again, we want to illustrate the possibility of applying this approximation in model M1.

Example 2. (*Exceptional motifs in the human* α*-globin gene cluster.*) When considering the DNA sequence of the human α-globin gene cluster a second time, we discover that the word 'CCTG' occurs 487 times (see, among others, the doubly underlined regions in the first 600 bases in Table 1). Thus, for $\{w^1, w^2\} = \{\mathrm{CAGG}, \mathrm{CCTG}\}$, we have $\sum_{i=1}^{2} N^{\mathrm{obs}}(w^i) = 922$. In order to establish whether this count is significantly high, we have to determine the rate function $\Lambda_1^*(q) = \frac{1}{2}\langle q, \Sigma_1^{-1}(w)q\rangle$ (see Corollary 4).

We have $\mathcal{P}(w^1, w^2) = \mathcal{P}(w^2, w^1) = \varnothing$. Using the estimators for $p(a, b)$ and the computed mean first passage times $m_{a,b}$, $a, b \in \mathcal{A}$, from Example 1, we obtain, in analogy to Example 1,

$$\mu_1(w^1) = 0.008\,31, \qquad \mu_1(w^2) = 0.008\,11,$$
$$\Sigma_1(w^1, w^1) = \sigma_1^2(\mathrm{CAGG}) = 0.007\,82, \qquad \Sigma_1(w^2, w^2) = \sigma_1^2(\mathrm{CCTG}) = 0.007\,65,$$

and, referring to (3),

$$\begin{aligned}\Sigma_1(w^1, w^2) &= \Sigma_1(w^2, w^1)\\ &= 2\mu_1(\mathrm{CAGG})\mu_1(\mathrm{CCTG})\Bigg(\sum_{a\in\mathcal{A}} \mu(a)m_{a,\mathrm{C}} - m_{\mathrm{G,C}}\Bigg) - 7\mu_1(\mathrm{CAGG})\mu_1(\mathrm{CCTG})\\ &= -0.000\,48.\end{aligned}$$

Hence,

$$\Sigma_1(w) = \begin{pmatrix} 0.007\,82 & -0.000\,48 \\ -0.000\,48 & 0.007\,65 \end{pmatrix}, \qquad \Sigma_1^{-1}(w) = \begin{pmatrix} 128.36 & 8.00 \\ 8.00 & 131.25 \end{pmatrix},$$

and, thus, we obtain

$$\langle x, \Sigma_1^{-1}(w)x\rangle = 128.36x_1^2 + 16.00x_1x_2 + 131.25x_2^2.$$

In order to apply Corollary 4, we have to compute $\inf_{x_1+x_2\ge q}\langle x, \Sigma_1^{-1}(w)x\rangle$, where

$$q = \frac{1}{n^{\alpha}}(922 - 43\,055(\mu_1(\mathrm{CAGG}) + \mu_1(\mathrm{CCTG}))) = \frac{1}{n^{\alpha}}214.93 > 0.$$

Consequently, we obtain

$$\inf_{x_1+x_2\ge q}\langle x, \Sigma_1^{-1}(w)x\rangle \approx q^2 68.90 = \frac{1}{n^{2\alpha}}3\,182\,727.$$

https://doi.org/10.1239/jap/1261670686 Published online by Cambridge University Press

Thus, applying Corollary 4 we obtain the following approximation:

$$\begin{aligned} \mathrm{P}\Big(\sum_{i=1}^{2} N(w^i) \geq 922\Big) &\simeq \exp\Big(-\frac{1}{2} n^{2\alpha-1} \inf_{\sum_{i=1}^{2} x_i \geq q} \langle x, \Sigma_1^{-1}(w) x\rangle\Big) \\ &\approx \exp\Big(-\frac{1}{2}\frac{1}{n} 3\,182\,727\Big) \\ &= 8.89 \times 10^{-17}. \end{aligned}$$

Hence, the probability of motif {CAGG, CCTG} occurring 922 times or more is vanishingly small and also notably lower than the probability of the single word 'CAGG' occurring 435 times or more (see Example 1).

Acknowledgement

We would like to thank the anonymous referee for many useful comments.

References

[1] Behrens, S. (2008). Moderate und große abweichungen zur statistischen analyse biologischer sequenzen. Doctoral Thesis, Universität Münster.
[2] Blaisdell, B. E. (1985). Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *J. Molec. Evol.* **21,** 278–288.
[3] Chen, X. (1999). Limit theorems for functionals of ergodic Markov chains with general state space. *Mem. Amer. Math. Soc.* **139**.
[4] Chung, K. L. (1967). *Markov Chains With Stationary Transition Probabilities*, 2nd edn. Springer, New York.
[5] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*, 2nd edn. Springer, New York.
[6] Djellout, H. and Guillin, A. (2001). Moderate deviations for Markov chains with atom. *Stoch. Process. Appl.* **95,** 203–217.
[7] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press.
[8] Hunter, J. J. (2008). Variances of first passage times in a Markov chain with applications to mixing times. *Linear Algebra Appl.* **429,** 1135–1162.
[9] Kleffe J. and Borodovsky M. (1992). First and second moment of counts of words in random texts generated by Markov chains. *Comput. Appl. Biosci.* **8,** 433–441.
[10] Kleffe, J. and Langbecker, U. (1990). Exact computation of pattern probabilities in random sequences generated by Markov chains. *Comput. Appl. Biosci.* **6,** 347–353.
[11] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer, Berlin.
[12] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.
[13] Nuel, G. (2001). Grandes déviations et chaînes de Markov pour l'étude des occurrences de mots dans les séquences biologiques. Doctoral Thesis, Université d'Essonne.
[14] Nuel, G. (2006). Numerical solutions for patterns statistics on Markov chains. *Statist. Appl. Genet. Molec. Biol.* **5,** 45 pp.
[15] Nussinov, R. (1981). The universal dinucleotide asymmetry rules in DNA and the amino acid codon choice. *J. Molec. Evol.* **17,** 237–244.
[16] Pitman, J. W. (1974). Uniform rates of convergence for Markov chain transition probabilities. *Z. Wahrscheinlichkeitsth.* **29,** 193–227.
[17] Prum, B., Rodolphe, F. and de Turckheim, È. (1995). Finding words with unexpected frequencies in desoxyribonucleic acid sequences. *J. R. Statist. Soc. B* **57,** 205–220.
[18] Régnier, M. (2000). A unified approach to word occurrence probabilities. *Discrete Appl. Math.* **104,** 259–280.
[19] Reinert, G., Schbath, S. and Waterman, M. S. (2005). Probabilistic and statistical properties of finite words in finite sequences. In *Applied Combinatorics on Words*, eds J. Berstel and D. Perrin, Cambridge University Press.
[20] Robin, S. and Daudin, J. J. (1999). Exact distributions of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36,** 179–193.

[21] Schbath, S. (1995). Compound poisson approximation of word counts in DNA sequences. *ESAIM Prob. Statist.* **1,** 1–16.
[22] Schbath, S. (1995). 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. Doctoral Thesis, Université René Descartes, Paris V.